Name: Nichols, Cory

**MSDS 6370 Sampling Statistics Final Exam (part II)**

**Spring 2017**

This exam is due at midnight CT on Monday, April 17, 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. (8 pts) Select the best answer to each question below:

(i) What is the purpose of using a post stratification adjustment?

A. To reduce the variance of the estimator due to under coverage.

B. To reduce the variance of the estimator due to the sample design’s unequal selection probabilities.

C. To reduce the bias of the estimator due to under coverage.

D. To reduce the bias of the estimator due to the sample design’s unequal selection probabilities.

**C – post stratification adjusts for under-coverage and non-response typically to weight up or down to a target, resulting in less bias in our estimator and OFTEN less variance too.**

**Sources: http://www.stata.com/manuals13/svypoststratification.pdf**

htt<p://www.atlas.illinois.edu/support/stats/resources/spss/create-post-stratification-wei>ghts-for-survery-analysis.pdf

(ii) A SRS of 100 patients from a particular doctor’s practice is chosen in order to estimate the total unpaid charges for all patients. All bills from the past 12 months for each sampled patient are selected, and amount of the unpaid charges on each bill is recorded. The doctor billed a total of 900 patients during the year.

This sample (all bills from sampled patients) can be thought of as which of the following? \_\_ (enter a, b, or c)

(a) a simple random sample of bills

(b) a cluster sample of bills

(c) a stratified sample of bills.

**B – patients are PSUs, containing our unit of interest, charges or billings, this is a cluster sample of bills using patients as the cluster**

2. (15 pts) A **simple random sample** of 100 of the 1000 housing units in a small community is sampled. Their water meters are monitored during the restricted watering portion of one day in order to estimate the total water usage in the community for that day, which fell in the drought season. The sample mean and sample variance are found to be.

1. Construct a 95% confidence interval for the **total gallons** of water used during the restricted watering portion of that day for the whole community.

This is a finite population with 100 observations and a lot of variability s = 353.6

Total Estimate: 130.0 \* 1000 = 130,000 gallons of water estimated

Standard Error: sqrt(125000/100 \* (1-100/1000)) = 33.54 FPC adjusted standard error

Mean Analysis:

Mean Half width: 1.96 \* 33.54 = 65.74

Mean Interval

130 65.74 = [64.26, 195.74]

Total Analysis:

Total Standard Deviation: 1000 \* 33.54 = 33540

Total Half Width: 1.96 \* 33540 = 65738.4

**Total 95% Confidence Interval**

130000 65738.4 = [64261.6, 195738.4]

1. Suppose that the city decided to save money next year in data collection by selecting the sample of 100 **by using city blocks as a frame**. Suppose there are 100 blocks in the community, **each with an average of 10 housing units**. Their plan was to randomly **select 10 blocks and sample all the houses on those blocks**, instead of the plan in part (a). Would the margin of error for this design be likely to be larger or smaller than what you calculated in part (a)? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. Carefully explain your reasoning.

ANSWER:

This is a cluster sampling design, where city blocks are the cluster with households being units in each cluster. The margin of error would be larger than part A because correlation between observations would likely be higher than a simple random or stratified sample, resulting in higher variance associated with estimates. Because of such correlation, new information would be less available and our sample would also be less representative.

1. You are the statistical consultant for the community. They ask you to **determine how many blocks they would need to sample in order to achieve the SAME margin of error as they did in the analysis in (a), but using the city blocks as sampling units**. You tell them you would need to know the value of the intra-cluster correlation, . They don’t have any data on that; however, water usage is highly correlated with size of the lot for the housing unit, for which = 0.4. You decide to use that for planning purposes. **How many blocks must they sample?**

**ANSWER:**

We need to estimate the intra-cluster correlation, so we’ll use a proxy variable, water usage:

Intra-cluster Correlation: 1 + p(N-1) = 1 + 0.4(10-1) = 4.6

We sampled 100 people in A. Therefore, we must sample 100\*4.6 = 460 housing units, or 460/10 = 46 blocks to achieve roughly the same margin of error as we did in analysis A when taking into account the average of 10 household units per block.

3. (12 pts) The SMU student directory contains 94 pages of student listings. Each page has 115 lines of text and 4 columns of names. The number of students listed on each page can vary, since different students have a different number of lines.

**Examples of sample listings in SMU directory**



1 page, 4 columns and 115 lines

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

115 lines

Consider the following sampling plans for **selecting a sample of SMU students listed in the directory**. ***For each, note whether it is a probability design or not, and explain why***.

For each of the sample designs you identified as probability samples, specify how you would define the features in PROC SURVEYMEANS so that a valid estimator of the mean and its standard error can be computed. In particular, for each so designated design, specify (i) the weight for each sampled student; (ii) Do you include the STRATUM statement in PROC SURVEYMEANS? If you include STRATUM statement then what is the STRATUM variable? (iii) Do you include CLUSTER statement in PROC SURVEYMEANS?

If you include CLUSTER statement then what is the CLUSTER variable?

1. You use a random number table to select a simple random sample (SRS) of 10 of the 94 pages. Then you select all the students on the selected pages into your sample. Is this a probability sample? Circle **yes** or no. Why?

ANSWER:

This is a probability sample since we can calculate the probabilities of selection for each unit. In this case, the probability is 10/94 or 10.6%. The design is a cluster design with weights of 94/10 or 9.4. We would not include the STRATUM statement here, since we are not stratifying. We would include the cluster statement and our variable would be the page number, the pages are not rolled up so we would need to include a CLUSTER statement. If they were rolled up, we would be able to run SURVEYMEANS without a CLUSTER option, similar to how we run a SRS in SAS.

Sample syntax with placeholders:

**PROC** **SURVEYMEANS** data = students clmean sum clsum mean total = num\_pages;

weight samp\_wt;

var variable\_to\_estimate;

cluster page;

**RUN**;

1. For every one of the 94 pages in the directory, you select a SRS of 2 of the 4 columns. Then you select at random 1 of the 115 lines from each of those 2 columns. Then you select the student to whom that line pertains. Circle **yes** or no. Why?

ANSWER:

This is a probability sample since we can calculate the probabilities of selection for each unit. For this example, a complex design is used with stratification based on the page and a two-stage design where columns represent the clusters and lines associated with each student are the SSUs.

For each stratum, the probability of selection is:

(2/4) \* ni/115 where ni are the number of lines associated with student i.

The weights for each student would be unequal, depending on the number of lines they represent. Weights would be:

wi = 4/2 \* 115/ni

We’d need to identify the strata because each page could contain a different number of students based on the number of lines. Further, units are not aggregated for each page, so strata need to be identified. We’d also need to identify the cluster as the column because this is a two-stage design, where columns represent our clusters and we sub-sample from within the columns selected.

**PROC** **SURVEYMEANS** data = students clmean sum clsum mean total = str\_sizes;

var variable\_to\_estimate;

(class if categorical)

stratum page;

cluster col;

weight base\_wt;

**RUN**;

Weights would need to be calculated for each student directly before running SAS code.

1. You select a SRS of 40 of the 94 pages. Then you select the first and last students listed on each selected page. Circle yes or **no.** Why?

ANSWER:

This is not a probability sample! The first and last students have probabilities of selection, however, everyone in between does not. Therefore, this is not a probability sample.

4. (10 pts)

A population of 16 balls (ordered as shown below) is made up of 7 white balls and 9 red balls as shown below:

R R W R W R W W R R R W W R R W

A systematic sample of size 3 is selected from the population and the proportion of red balls in the population is estimated from the sample using as the estimator

.

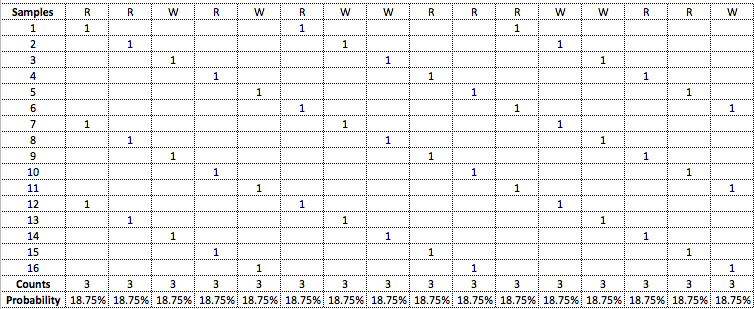
Find the sampling distribution of .

Our interval to be used for systematic sampling is not an integer (16/3) = 5.33. In order to avoid selection bias and ensure each observation has a chance to be selected, a roundabout method is used. We consider a random start between 1 and N instead of between 1 and k (our interval) to avoid biasing our sample or excluding observations. For instance, if we use 16/3 = 5.33 = 5, the last observation would never be selected. If we use 6 for our interval and a roundabout method, observation 1 would be selected twice. Therefore, randomization from 1 to N instead of 1 to k is appropriate to ensure all observations have probability of selection and selections are not biased. We round our interval of 5.33 down to 5 and use the roundabout method to ensure all observations are selected and have an equal probability of selection.

SOURCE: <https://stats.stackexchange.com/questions/73741/when-doing-systematic-sampling-what-should-be-done-if-the-sampling-interval-i>

|  |  |
| --- | --- |
| **Distribution** | **Histogram** |
| R, R, R  R, W, W  W, W, W  R, R, R  W, R, R  R, R, W  R, W, W  R, W, W  W, R, R  R, R, R  W, R, W  W, R, R  R, W, W  R, W, W  R, R, R  W, W, R | ../../../../../Screen%20Shot%202017-04-17%20at%208.44.25%20AM.png |

Round-about method results:



5. (15 pts) For this exercise, you will examine the HRS data (download from Asynchronous Lecture 11.4.1).(Same data set you used for lab11).

Consider the question of whether or not diabetes (DIABETES) is independent of race/ethnicity (RACECAT). Using the Rao-Scott Chi-square test, determine if diabetes and race/ethnicity are independent or not. Include your SAS code and output, as well as 6 steps of hypothesis.

**ANSWER:**

**Step 1: State null and alternative hypotheses**

: Diabetes is independent of race/ethnicity

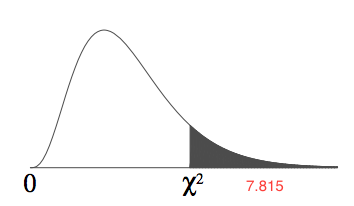
: Diabetes is *not* independent of race/ethnicity

**Step 2: Choose significance level**

For this exercise, our significance level is 0.05, meaning our acceptable probability of making a type I error is 5%.

**Step 3: Determine the critical values**

For a 5% level of significance and three degrees of freedom (4-1)(2-1), the critical value for a chi-square test from a chi-square table is **7.815.**



**Step 4: Determine the test statistic**

To determine the appropriate test statistic for assessing independence of race and diabetes, we must ensure we consider the sampling design. In this case, a complex design is used with stratification and clusters included in each stratum. The Rao-Scott chi-square test statistic is calculated using the pearson test statistic adjusted for the design.

When considering the sampling design, the resulting Rao-Scott chi-square test statistic is **211.04.**

**Step 5: Compare test statistic and critical values and level of significance**

Our test statistic of 211.04 falls firmly into the rejection region of our chi-squared distribution. This is confirmed with a p-value of < 0.0001.

**Step 6: Write conclusion**

There is evidence to show (p < 0.0001), for a significance level of 5%, that race and diabetes status are not independent of one another.

**SAS CODE:**

**data** hrs;

set "\\Client\C$\Users\patrickcorynichols\Desktop\hrs\_analysis.sas7bdat" ;

**run**;

**proc** **datasets**;

contents data=\_all\_;

**run**;

**proc** **tabulate** data=hrs;

class DIABETES racecat;

table DIABETES,racecat;

**run**;

**proc** **surveyfreq** data=hrs;

weight KWGTR;

strata stratum;

cluster SECU;

table DIABETES \*racecat / chisq;

**run**;