**MSDS 6370 Final Exam**

**Spring 2018**

**Submitted by Hari Narayan Sanadhya**

The Attached excel sheet contains all the formulas behind the manual calculations shown in this examination solutions for reference.

****

**You are required to work alone in completing the examination. You may not consult with anyone else in completing your answers. If you do not adhere to these requirements, you will be in violation of SMU’s honor code for which there is consequences.**

**Q1(10 pts):** From a face to face survey last year you learned that 2 out of 6 farmers in Green County raised chickens. This year you decide to take a sample of 3 farmers with a phone survey to determine if the percentage of farmers raising chickens changed.

1. Calculate the probability distribution of the sample proportion of farmers raising chickens assuming there is no change from last year.

**Solution: -**

|  |  |
| --- | --- |
| Farmer | Raised Chicken |
| 1 | y |
| 2 | y |
| 3 | n |
| 4 | n |
| 5 | n |
| 6 | n |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Farmer Chosen | 2,3 | 2,4 | 2,5 | 2,6 | 3,4 | 3,5 | 3,6 | 4,5 | 4,6 | 5,6 |
| 1 | 1,2,3 | 1,2,4 | 1,2,5 | 1,2,6 | 1,3,4 | 1,3,5 | 1,3,6 | 1,4,5 | 1,4,6 | 1,5,6 |
| 2 | x | x | x | x | 2,3,4 | 2,3,5 | 2,3,6 | 2,4,5 | 2,4,6 | 2,5,6 |
| 3 | x | x | x | x | x | x | x | 3,4,5 | 3,4,6 | 3,5,6 |
| 4 | x | x | x | x | x | x | x | x | x | 4,5,6 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Probability | | | | | | | | | | |
| Farmer Chosen | **2,3** | **2,4** | **2,5** | **2,6** | **3,4** | **3,5** | **3,6** | **4,5** | **4,6** | **5,6** |
| 1 | .667 | 0.667 | 0.667 | 0.667 | 0.33 | 0.33 | 0.33 | 0.33 | 0.33 | 0.33 |
| 2 | x | x | x | x | 0.33 | 0.33 | 0.33 | 0.33 | 0.33 | 0.33 |
| 3 | x | x | x | x | x | x | x | 0 | 0 | 0 |
| 4 | x | x | x | x | x | x | x | x | x | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| Value | Freq | Prob | Prob\*(value-mean)2 |
| 0 | 4 | 0.2 | 0.022045 |
| 0.33 | 12 | 0.6 | 2.4E-06 |
| 0.67 | 4 | 0.2 | 0.022849 |
|  |  | **Mean** | 0.332 |
|  |  | **Variance** | 0.044896 |
|  |  | **Std Error** | 0.211887 |

2. What is the probability that this year your sample will indicate no change from last year in the proportion of farmers raising chickens? Explain how you arrived at your answer.

**Solution: -**

In the first part of the question, we already determined that there are 20 possible samples of 3 farmers that can be picked from a set of 6 farmers .

For the proportion of farmers raising chickens to not change, the possible cases are: -

1. All 6 farmers are raising chicken: In case a, all these 20 samples have a value of 1 (everyone in sample is raising chicken). Every sample in this case will show that the proportion of farmers raising chicken from last year has changed. (Since probability 1 was last seen in the last year)
2. 5 out of 6 farmers are raising chicken: of these 20 sample, if farmer 1 have a value of 0 (i.e. he is not raising chicken), then the probability table would be

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Probability | | | | | | | | | | |
| Farmer Chosen | **2,3** | **2,4** | **2,5** | **2,6** | **3,4** | **3,5** | **3,6** | **4,5** | **4,6** | **5,6** |
| 1 | .667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 |
| 2 | x | x | x | x | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | x | x | x | x | x | x | x | 1 | 1 | 1 |
| 4 | x | x | x | x | X | x | x | x | x | 1 |

Which means that of the 10, 10 samples will show that the proportion has changed (probability of 1 was not seen in last year but was visible this year in 10 cases out of 20)

1. 4 out of 6 farmers are raising chicken – In this case the Contingency table showing probability would look like

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Probability | | | | | | | | | | |
| Farmer Chosen | **2,3** | **2,4** | **2,5** | **2,6** | **3,4** | **3,5** | **3,6** | **4,5** | **4,6** | **5,6** |
| 1 | .33 | .33 | .33 | .33 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 |
| 2 | x | x | x | x | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 | 0.667 |
| 3 | x | x | x | x | x | x | x | 1 | 1 | 1 |
| 4 | x | x | x | x | X | x | x | x | x | 1 |

4 outcomes out of 20 give a probability of 1

1. 3 out of 6 farmers are raising chicken

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Probability | | | | | | | | | | |
| Farmer Chosen | **2,3** | **2,4** | **2,5** | **2,6** | **3,4** | **3,5** | **3,6** | **4,5** | **4,6** | **5,6** |
| 1 | 0 | .33 | .33 | .33 | .33 | .33 | .33 | 0.667 | 0.667 | 0.667 |
| 2 | x | x | x | x | .33 | .33 | .33 | 0.667 | 0.667 | 0.667 |
| 3 | x | x | x | x | x | x | x | .33 | .33 | .33 |
| 4 | x | x | x | x | X | x | x | x | x | 1 |

1. 1 out of 6 farmers is raising chicken

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Probability | | | | | | | | | | |
| Farmer Chosen | **2,3** | **2,4** | **2,5** | **2,6** | **3,4** | **3,5** | **3,6** | **4,5** | **4,6** | **5,6** |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.33 |
| 2 | x | x | x | x | 0 | 0 | 0 | 0 | 0 | 0.33 |
| 3 | x | x | x | x | x | x | x | 0 | 0 | .33 |
| 4 | x | x | x | x | X | x | x | x | X | .33 |

1. None of the 6 farmers are raising chicken

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Probability | | | | | | | | | | |
| Farmer Chosen | **2,3** | **2,4** | **2,5** | **2,6** | **3,4** | **3,5** | **3,6** | **4,5** | **4,6** | **5,6** |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | x | x | x | x | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | x | x | x | x | x | x | x | 0 | 0 | 0 |
| 4 | x | x | x | x | X | x | x | x | X | 0 |

So of the total 120 possible samples, in 35 cases, the surveyor may get a probability of 1 which is the case where he can say with 100% confidence that the proportion of farmers raising chicken has changed this year as compared to last year.

Probability = 35/120 = 0.2917

**Q2(15 pts):** The MotorsHere Rental Car Company has 100 rental sites. Each site has ni cars for i=(1,…,100). Each site has 60 parking spaces numbered 1-60 with all the spaces occupied. The parking spaces are separated into 3 rows of twenty. Small cars take one parking space, medium cars take two parking spaces, and large cars take three parking spaces. Each car has a unique registration number.

You are considering the sampling plans below:

**Plan A:** Randomly select 10 sites and then randomly select 20 vehicles at each site using the vehicle registration number.

**Plan B:** Randomly select 10 sites and then select all the vehicles in row 2 at the site.

**Plan C:** From every site randomly select two rows and then randomly select a parking spot in each row and take the car in the parking spot.

For each sampling plan specify the following and explain your answers:

**Solution: -**

Based of the information given in the problem statement:

Where ni is the number of cars at the ith rental site such that:

Where rij = number of cars in row j of the ith rental site such that

Where sij = number of small cars in row j of the ith rental site, mij = number of medium cars in row j of the ith rental site, lij = number of large cars in row j of the ith rental site

Other information given in question: -

For any given value of i and j,

For any values of i,

So based on this information, we can say that the minimum possible value of ni is 21 (when each of the 3 rows at the ith site has 6 large cars and one medium car) and the maximum possible value of ni is 60 (all cars at the ith site are small cars)

1. Is the sampling plan a probability design or not? Explain your answer.

**Solution: -**

**Plan A:** This is a probability design since the sampling unit is the number of cars for the site which is already known (ni). Probability of selection of a rental site is (1/100) and from that site, the probability of selection of a car is 1(/ni). This makes the probability of selection of the car as (1/(100\*ni))

**Plan B:** This is a non-probability design as the number of cars in the rows of the site is not known. Probability of selection of a rental site is (1/100) and since row 2 is the row selected from the selected random site, the probability of selection of row is 1. The total probability of selection of all the cars in row 2 is 1 but the probability of individual cars selection is not known beforehand.

**Plan C:** This is a probability sample since the units sampled is parking spot from the selected row of the selected site. Probability of selection of a site is 1/100, the probability of selection of a row within the selected site is (1/3) and the probability of selection of a parking spot as 1/20. Overall the probability of selection of the parking spot is (1/(100\*3\*20)).

1. What would be the weight for each type of car?

**Solution: -**

**Plan A:** Weight would be 1 for small cars, 2 for the medium cars and 3 for the large cars

**Plan B:** Weight would be 1 for small cars, 2 for the medium cars and 3 for the large cars

**Plan C:** Weight would be 1 for all the cars since the parking spot is getting selected, so the large car can be picked 3 times, a medium car can be picked 2 times and a small car can be picked once. This make the sampling style to already apply the weights by default while picking the sample.

1. Does the design contain strata and if so, what is the strata variable?

**Solution: -**

**Plan A:** Yes, it’s a stratified design with site as the strata variable.

**Plan B:** Yes, it’s a stratified design with site as the strata variable.

**Plan C:** Yes, it’s a stratified design with site as the strata variable.

1. Does the design include clusters and if so, what is the cluster variable?

**Solution: -**

**Plan A:** Not a clustering design

**Plan B:** Clustering design with rows within the site as the cluster.

**Plan C:** Clustering design with rows within the site as the cluster.

**Q3(20 pts):** True and False

1. National Surveys use sample size within PSUs, strata, and segments so that households have an Equal Probability of Selection. (TRUE OR FALSE)

**FALSE** – the main purpose of National Surveys use sample size within PSUs, strata, and segments is to make the data collection cheaper. For equal probability of selection, national surveys use composite weights which is the product of weight from the design, a weight from non-response adjustment, and a weight from a problem with the frame that is under coverage of the frame

2. An initial stratification of data is necessary before using the cum *f* method of stratification. (TRUE OR FALSE)

**FALSE** – initial knowledge of the number of strata in the dataset is required but not actual stratification. The number of strata can be identified by using previous research or using educated guess or by using professional judgement.

3. Replication methods for variance estimation are preferred with large samples.

(TRUE OR FALSE)

**FALSE** – Replication methods of variance estimation are preferred with smaller samples wherein its very costlier to get large samples of data.

4. In the Jackknife method of variance estimation, replicate weights are determined by the sample size. (TRUE OR FALSE)

**FALSE** – Replicate weights in Jackknife method are the actual weights leaving one PSU at a time.

5. Weighting data increases the bias of an estimator, but lowers its variance. (TRUE OR FALSE)

**FALSE** – Weighting the data correctly helps in producing an unbiased estimate of mean. An unweighted design may produce correct estimate values but the standard error in that case may not be correct.

6. The best number of certainty strata in a data set is determined by dividing the sum of the observations by the number of strata to obtain a value k and using the observations with the highest values that sum to k. (TRUE OR FALSE)

**FALSE** - The statistical reason for selecting a certainty stratum is to capture and isolate the largest values containing records so that their extremely large values do not influence sampling variability. This allows more stable estimates and an overall sample size reduction while meeting the desired confidence and precision goals. To gain higher comfort level or better estimates, its likely that people choose large number of records in the certainty stratum. The method mentioned to compute the cutoff value for certainty stratum as sum/# of strata may be one of the ways to select the certainty stratum but since its dependent on number of strata, I don’t agree with this way being the best. I would prefer professional judgment or by possibly evaluating results of methods to form strata cutoffs

7. Ratio estimators are a good idea when there is large correlation between the variables used in the ratio estimation. (TRUE OR FALSE)

**TRUE** – Ratio estimators are preferred when where is the correlation between x and y variables. If the correlation value is large then its highly likely that this criterion will be satisfied since most of the times the CV (corresponding coefficient of variation in the population is not known)

8. Estimators from cut-off samples are similar to ratio estimators when the relationship between x that is available from the whole population and the variable of interest y is well understood and stable. (TRUE OR FALSE)

**FALSE** – Cut-off samples since are non-probability samples, they when applied can lead to results far away or close to the actual values depending on whether it was safe to apply it or not. Cutoff samples are safe to apply and yield good results when the relationship between x that is available from the whole population and the variable of interest y is well understood and stable. But there are other conditions as well like if the largest units in the population does not cover a large fraction of the total for the variable of interest, then there may be bias which could reduce the estimation quality of the cut-off sample.

9. Categorical variables are dependent when the cell probabilities in a contingency table equal the product of row and column probabilities. (TRUE OR FALSE)

**FALSE** – Categorical variables are independent when the cell probabilities in a contingency table equal the product of row and column probabilities i.e. πij=πi+πj

10. When you have a sampling design with clusters, you always have strata defined for the population. (TRUE OR FALSE)

**FALSE** – When you have sampling design with clustering, which is a probability sample in which each sampling unit is a collection, or cluster, of elements, we don’t need to have strata defined until or unless complex sampling design is involved wherein the data is divided into primary sampling units (strata or cluster) and then each PSU is again divided into SSU's (secondary sampling units).

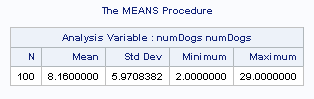
**Q4(20 pts):** A county in Texas has 100 animal shelters and wants to estimate the mean number of dogs at each shelter every month. Last month the County Clerk recorded the number of dogs at each shelter for all the shelters. The data collected by the Clerk is in finalDatSpring2018.xlsx, sheet dogDat. The County Chairperson thinks it is costly to survey all the shelters each month and comes to you for advice on taking a sample of 30 shelters. In your consultation with the County Chairperson you say a sample is a good idea and that the county may want to consider stratification in the sample design.

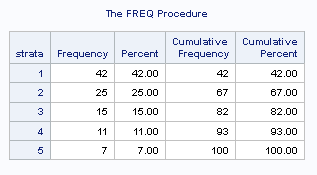
1. (5 pts) Stratify the data into five strata using the number of dogs at each shelter using the Equal Sum of Values in Strata method of stratification.

**Solution: -**

Steps

* Sort data in descending order
* Calculate the length of the strata interval which is the sum of the data element divided by the number of strata
* Calculate the cumulative sum of the data element
* Assign the elements to strata based on the strata interval and cumulative distribution.





**SAS CODE**

libname xl XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/Final Exam/finalDatSpring2018.xlsx';

libname xl list ;

data dogData;

set xl.dogdat;

run;

/\* verify data import \*/

proc print data=dogData(obs=2); run;

/\* Get population mean \*/

proc means data=dogdata;

var numdogs;

run;

/\* Step 1 - sort the data by the data column \*/

proc sort data=dogData out=dogdata;

by numdogs;

run;

/\* Step 2 - compute the strata interval \*/

proc sql noprint;

select sum(numdogs)/5 into :strata\_interval from dogdata;

run;

%put &strata\_interval;

/\* Step 3 - Compute the commulative frequency \*/

/\* Step 4 - Divide data into stratum \*/

data dogdata;

set dogdata;

retain Cummulative\_Dist 0;

Cummulative\_Dist = Cummulative\_Dist + numdogs;

strata = 5;

if Cummulative\_Dist <= &strata\_interval then strata=1;

else if Cummulative\_Dist <= &strata\_interval\*2 then strata=2;

else if Cummulative\_Dist <= &strata\_interval\*3 then strata=3;

else if Cummulative\_Dist <= &strata\_interval\*4 then strata=4;

run;

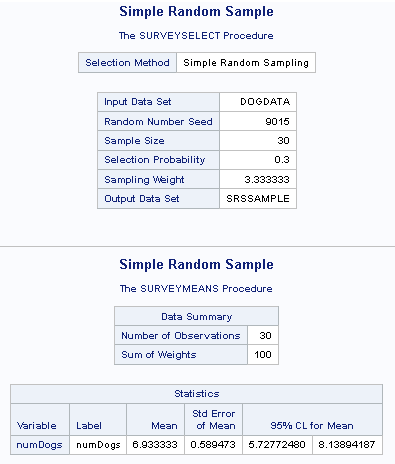
/\* View the frequency of the observation in each strata \*/

proc freq data=dogdata(keep=strata);

run;

1. (5 pts) With a sample of size 30, use SAS to calculate the mean and standard error of the number of dogs per shelter without using stratification. Use a random number seed of 9015. Show your SAS code.

**Solution: -**



**SAS Code**

proc surveyselect data=dogdata sampsize=30 out=SRSSAMPLE method=SRS seed=9015;

title "Simple Random Sample";

run;

data SRSSAMPLE;

set SRSSAMPLE;

samplingWeight = 100/30;

run;

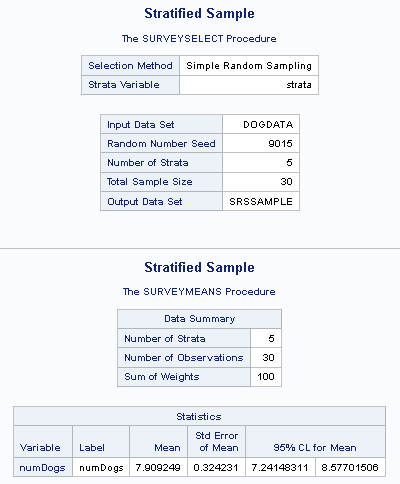
proc surveymeans data=SRSSAMPLE total=100 mean stderr clm;

var numdogs;

weight samplingWeight;

run;

1. (5 pts) With a sample of size 30, calculate the mean and standard error of the number of dogs per shelter using the stratification you computed in question 1. Use a proportional sampling approach in selecting your sample. Use a random number seed of 9015. Show your SAS code.



**SAS Code**

proc sql;

create table stratumsize as

select strata, count(strata) as \_total\_ from dogdata group by strata;

run;

proc sql;

select strata, (\_total\_ \* 30/100) as approx\_samples\_from\_strat from stratumsize;

run;

/\* Used output from above statement to determine the value

of sampsize parameter used in surveyselect method \*/

proc surveyselect data=dogdata sampsize=(13, 7, 5, 3, 2) out=SRSSAMPLE method=srs seed=9015;

strata strata;

title "Stratified Sample";

run;

proc surveymeans data=SRSSAMPLE total=stratumsize mean stderr clm;

strata strata;

var numdogs;

weight samplingWeight;

run;

1. (5 pts) What would you advise the County Chairperson about stratifying the data for analysis?

**Solution: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean | Std Err | 95% CL for Mean | | |
| Population | 8.16 |  |  |  |
| Simple Random Sample | 6.93 | 0.589473 | 5.72772480 | 8.13894187 |
| Stratified Sampling | 7.91 | 0.324231 | 7.24148311 | 8.57701506 |

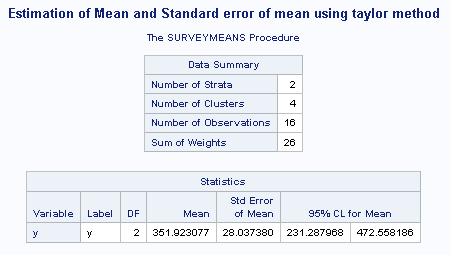
Based on the results obtained, my suggestion to the County's Chairperson would be to go with the stratification of data for the analysis. The reasons are:

1. Mean estimate from Stratified Sampling are closer to population mean then compared with the estimate obtained from simple random sampling.
2. Standard error obtained from stratified design is smaller then what is obtained from simple random design which would mean that the variation of estimates obtained from stratified design would be smaller proving results to be more accurate.
3. The true mean lies in the 95% confidence interval of the Mean obtained from stratified sampling proving it to be statistical significant as well. Whereas for Simple Random Design, the true mean doesn’t lie in the 95% confidence interval of the mean.

**Q5(15 pts.):** The Hobo Rail Freight company has a fleet of trains. Each train consists of a number of box cars. The data scientist hired to analyze the data has stratified the box cars into two strata based on the volume of the interior of the car. To estimate the mean weight of each box car, the company randomly samples 2 trains and then randomly chooses 8 box cars from each stratum. The data collected is in finalDatSpring2018.xlsx, sheet railDat.

1. (4 pts) Using SAS, find the mean and standard error of the estimate with the Taylor Method. Show your SAS code and results.

**Solution: -**



**SAS Code**

libname xl XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/Final Exam/finalDatSpring2018.xlsx';

libname xl list ;

data railData;

set xl.raildat;

if y<>.;

\*Remove the blank records, the given excel has some additional records in the end which does not contain any data;

run;

proc print data=railData; run;

PROC SURVEYMEANS DATA = railData DF mean stderr clm varmethod=taylor;

STRATUM strata;

CLUSTER psu;

WEIGHT samplingWeight;

VAR y;

title "Estimation of Mean and Standard error of mean using taylor method";

run;

1. (4 pts) Manually compute the standard error of the estimate you calculated in question 1 using the jackknife method.

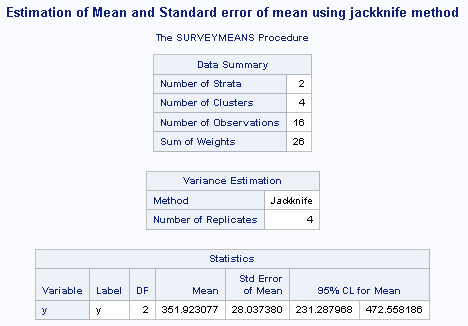
**Solution: -**

Estimate of standard error of mean from Jackknife method = 28.03738



1. (4 pts) Using SAS, find the mean and standard error of the estimate with the Jackknife Method. Show your SAS code and results.

**Solution: -**



**SAS Code**

libname xl XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/Final Exam/finalDatSpring2018.xlsx';

libname xl list ;

data railData;

set xl.raildat;

if y<>.;

\*Remove the blank records, the given excel has some additional records in the end which does not contain any data;

run;

PROC SURVEYMEANS DATA = railData DF mean stderr clm varmethod=jackknife;

STRATUM strata;

CLUSTER psu;

WEIGHT samplingWeight;

VAR y;

title "Estimation of Mean and Standard error of mean using jackknife method";

run;

1. (3 pts) Compare the two methods and comment on which method to estimate the standard error is preferred for this problem.

**Solution: -**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Std Err | 95% CL of Mean | |
| Taylor Method | 351.923077 | 28.037380 | 231.287968 | 472.558186 |
| JackKnife Method | 351.923077 | 28.037380 | 231.287968 | 472.558186 |

Since both the Jackknife and Taylor method gave the exact same answer, the preferred method to estimate standard error is Taylor method as Jackknife is computationally more intensive.

**Q6(10 pts):** To control its health insurance expense a company requires its employees to join its on-site fitness center or some other center off-site. The company knows how many times a month on average each employee enrolled at the on-site facility used the facility. It also knows the salary of each of its employees and believes this amount is related to fitness center usage. The company has 110 employees. The data the company has collected is in finalDatSpring2018.xlsx, sheet fitnessDat.

1. (2 pts) Manually, calculate the mean number of times an employee uses the fitness center based only on the data for people enrolled in the on-site center.

**Solution: -**

Mean = sum/count = (Sum of time usage of employees enrolled in onsite center) / (Number of employees enrolled in onsite center) = 240/24 = 10

Based of the data of people enrolled in on-site fitness facility, an employee uses 10 times on average the fitness center in a month.

1. (2 pts) Assuming the data come from a random sample, estimate the standard error of the mean calculated in question 1.

**Solution: -**



Estimate of Standard error of mean = 5.364942562

1. (2 pts) Using ratio estimation, manually calculate the mean of the average number of times an employee uses a fitness center.

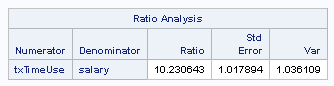
**Solution: -**



1. (2 pts) Using SAS, calculate the variance of the ratio estimator you calculated in question 3. Show your SAS code and results.

**Solution: -**

Variance using ratio estimator: 1.036109



**SAS code**

libname xl XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/Final Exam/finalDatSpring2018.xlsx';

libname xl list ;

data fitnessData;

set xl.fitnessdat;

run;

proc print data=fitnessData(obs=5);

run;

data fitnessSample;

set fitnessData;

if timesUse=. then delete;

txTimeUse = timesUse \* 8633000/110;

\*8633000 is the value of Tx, since its mean calculation, divide tx by 110;

SamplingWeight = 1; \* weight = populationSize/SampleSize;

run;

proc print data=fitnesssample; run;

PROC SURVEYMEANS DATA = fitnesssample sum mean var TOTAL=110;

VAR txTimeUse salary;

WEIGHT SamplingWeight;

TITLE "Simple random sample with ratio estimation";

RATIO txTimeUse/salary;

run;

1. (2 pts) Would you advise the company to use a ratio estimator to find an estimate of the mean number of times employees uses the fitness center?

**Solution: -**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Std Err | Variance |
| Simple Random Sample | 10 | 5.365 | 28.7826 |
| Ratio Estimate | 10.523 | 1.018 | 1.036 |

Mean of sample x (salary) values = 78481.81818

Mean of (estimated) population y (timesUse) values = 10

Standard deviation of known x values = 17841.57

Standard deviation of y values = 5.365

So CVy = 5.365/10 = 0.5365 and CVx = 17841.57/78481.81818 = 0.22733

The ratio estimates are preferred when . Here in this case, we have already been provided that the salary amount is related to fitness center usage which means that (correlation between x and y) is high and should definitely be greater than 0.212 satisfying the condition making the ratio estimator a good method for the given data.

So yes I will advise the company to use ratio estimator to find an estimate of the mean number of times employees uses the fitness center.

**Q7(10 pts):** A pharmaceutical company wants to evaluate if the there is a difference in blood pressure for patients taking three different doses of its medicine. It collects a sample of patients with all the same initial blood pressure and measures each patient’s blood pressure after taking the medicine for two months. The data collected is shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Blood Pressure\Dose** | **10mg** | **20mg** | **30mg** |
| **>130** | 36 | 80 | 20 |
| **<=130** | 40 | 124 | 100 |

1. (5 pts) Manually calculate the Chi-square statistic and evaluate whether blood pressure is independent of dose.

**Solution: -**

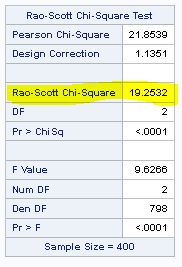


1. (5 pts) Use the sample information in finalDatSpring2018.xlsx.xlsx, sheet doseDat to calculate the Rao Chi-square with SAS and evaluate whether blood pressure is independent of dose. Comment on why the Rao Chi-square gives different estimates from what you calculated in question 1. Show your SAS code and results.

**Solution: -**

Rao-Scott Chi-Squared is adjusted chi-square statistic. In order to calculate it, the unweighted estimates of proportion are replaced with the weighted estimates of proportion while calculating the expected frequencies or the expected counts in the cells. This method was designed by Rao and Scott (1984) which derives first order correction to Pearson Chi-Square statistic that takes into account the survey design. It deflates the chi-square statistic by a generalized degree of freedom (GDEF) in the output and compares it to the usual reference distribution, which is a chi-squared distribution with (r – 1)\*(c -1) degrees of freedom where r and c are the rows and columns of the contingency table. The reason for Rao-Scott Chi-Squared estimate is different then chi-squared estimate is due to the design correction. The Pearson chi-squared divided by the design correction to obtain the Rao-Scott statistic value.





**SAS Code**

libname xl XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/Final Exam/finalDatSpring2018.xlsx';

libname xl list ;

data dosageData;

set xl.dosedat;

run;

proc print data=dosagedata(obs=5);

run;

/\* Run the Chi-square test \*/

proc freq data = dosageData;

tables dose \* bp/exact chisq crosslist all expected plots=all;

run;

/\* Rao-Scott Chi Sqaure test \*/

proc surveyfreq data = dosageData;

weight weight;

table dose \* bp/ chisq col expected;

run;