University of Toronto Mississauga

STA304H5F - Fall 2012
Instructor: Ramya Thinniyam

Term Test #2 - November 15th, 2012
Version 1

Family Name (print): *(the name in large print on your T-card)*	SOLUTIONS - V1
Given Names (print): *(the names in small print on your T-card)*	
Signature:	
Student Number:	
Tutorial (circle one):	Fridays 12-1pm Fridays 2-3pm

Aids Allowed: Non-programmable Calculator (without a text keyboard)
Aids Provided: Formula sheet

INSTRUCTIONS:
-There are 4 questions – answer all questions.
-There are 8 pages total. Make sure you have all pages before starting the test.
-For all true/false and fill in the blank questions, circle or put your final answers in blanks as instructed. Only final answers will be marked.
-For all other questions, show your work to earn full marks and then circle the final answer. Correct answers with no justifications will not receive any marks.
-You may use formulas/results from formula sheet without proof unless you are asked to specifically prove that formula.
-Simplify answers and round to **4 decimal places** where appropriate.
-Recall: **SRS**=Simple Random Sample without replacement
-STRS= Stratified Random Sample

BEST WISHES! ☺

Question	**1.** (/5)	**2.** (/5)	**3.** (/20)	**4.** (/10)	**TOTAL:(/40)**
Marks					

[5 marks - 1 each]

1. TRUE/FALSE: *If the statement is true under all conditions, circle T ; otherwise circle F.*

(a) A two-stage cluster sample is a self-weighting sample. T $\boxed{F}$

(b) ICC will be high when there is homogeneity within clusters. $\boxed{T}$ F

(c) A STRS usually gives estimators with smaller variance than a SRS of the same size. (when strata are large) $\boxed{T}$ F

(d) Ratio estimators are unbiased. T $\boxed{F}$

(e) STRS with proportional allocation is a self-weighting sample. $\boxed{T}$ F

[5 marks]

2. SHORT ANSWER: *Give a short answer (2/3 sentences) to the following questions:*

[3 marks]

a) STRS vs. Cluster Sampling. Explain the difference between Stratified and One-Stage Cluster Sampling in terms of how the groups (strata/clusters) should be chosen in order to increase precision. Justify.

2/3 sentences explaining the following:

In STRS, we sample from each strata so the groups should be chosen so that elements within strata are homogeneous and between strata are heterogeneous.

A One-Stage Cluster sample is the opposite since we sample all elements from some clusters, we don't want to repeat information if there is homogeneity within clusters. So, elements within clusters should be heterogeneous and between clusters should be homogeneous (ie. each cluster is a mini representation of the population).

[2 marks]

b) Ratio Estimation. We have two variables, *x* (auxiliary variable) and *y* (response / variable of interest) and assume we know the population total and mean for *x*. We use the ratio estimator $\hat{B} = \frac{\bar{y}}{\bar{x}}$. One would think that $\tilde{B} = \frac{\bar{y}}{\bar{x}_U}$ is a natural estimator of the population ratio since $\bar{x}_U$ is known. Explain why we use $\hat{B}$ rather than $\tilde{B}$ as the ratio estimator. (no proof required)

We use $\hat{B}$ as it increases precision: since *x* and *y* are correlated, the sampling distribution of $\hat{B}$ has a smaller variance than that of $\tilde{B}$.

[20 marks- 1 each blank : except c) is 1 mark total for all blanks]

3. <u>FILL IN THE BLANKS</u>: *You may do rough work on the back of the pages or in empty space, but only answers filled in the blanks will be marked.*

<u>A.</u> There are 10 Introductory Spanish courses offered at a community college. 5 of these classes are randomly selected. Each student in the sampled classes is given a vocabulary test and their scores are recorded. We wish to estimate the mean vocabulary test score for all students in this community college. (A student cannot be registered in more than one of these courses at a time). Below is the data description and some 'R' output:

'class' = Class number
'score'= Score on vocabulary test (out of 100)
'trip' = 1 if plan a trip to a Spanish-speaking country, 0 otherwise

```
> spanish <- read.csv("spanish.csv")
> attach(spanish)

> cl <- cluster(spanish,c="class",size=5,method="srswor")

> mysample <- getdata(spanish,cl)
> mysample
```

	score	trip	class	ID	Prob
111	30	0	20	111	0.5
103	71	1	20	103	0.5
104	90	1	20	104	0.5
.					
.					
142	41	0	23	142	0.5
150	81	1	23	150	0.5
.					
.					
6	45	0	34	6	0.5
16	59	0	34	16	0.5
4	62	0	34	4	0.5
.					
.					
192	69	0	39	192	0.5
193	69	0	39	193	0.5
.					
85	64	0	69	85	0.5
86	48	0	69	86	0.5

```
> attach(mysample)

> a <- tapply(score,class,length)

> a

20 23 34 39 69

21 14 22 21 22

> sum(a)

[1] 100

> b <- tapply(score,class,mean)

> b

      20          23          34          39          69

66.95238 67.42857 57.59091 83.19048 62.13636

> c = sum(a*b) / sum(a)

> c

[1] 67.31

> sum( a^2 *(b-c)^2 )

[1] 169948.7

>  tapply(score,class,var)

      20          23          34          39          69

234.9476 267.4945 113.3961 159.5619 172.5996

> f <- tapply(score,class,sum)

> f

  20    23   34    39   69

1406  944 1267 1747 1367

>sum(f)

[1] 6731

> var(f)

[1] 83171.7
```

(a) This is an example of a **____One-Stage Cluster____** sample.

[name the sampling method- choose from SRS, STRS, One-Stage Cluster, Two-Stage Cluster, Systematic.]

(b) The number of students in the above sample is**___100____**.

(c) The following classes were selected to be given vocabulary tests: **20, 23, 34, 39, 69.**
[write the class numbers]

(d) The estimate for the population mean is **___67.31___** with a standard error of **___3.2591___**.

(e) The sampling weight for the *j*th student from the *i*th class is **_____2______**.

(f) You are now informed that the community college has a total of 196 students taking an introductory Spanish course. The estimate for the population mean is **__68.6837__** with a standard error of **____4.6530___**.

B. In total, 63 students plan to visit a Spanish-speaking country, whereas 133 don't plan to. Now, we take a stratified sample instead, grouping by the variable 'trip'. Below is 'R' output:

```
> tapply(spanish$score,spanish$trip,length)

 0    1

133  63

> tapply(spanish$score,spanish$trip,mean)

      0         1

61.88722 77.15873

> tapply(spanish$score,spanish$trip,var)

      0         1

274.7523 169.6841

> length(spanish$score)

[1] 196

> mean(spanish$score)

[1] 66.79592
```

```
> str<-strata(spanish,c("trip"),size=c(25,25),method=c("srswor"))

> strs.sample <- getdata(spanish,str)
> strs.sample

    class score trip   ID   Prob        Stratum

5     34    42    0     5     0.1879699     1

12    34    70    0    12     0.1879699     1

42    60    66    0    42     0.1879699     1

51    60    60    0    51     0.1879699     1

.

.

.

153   23    94    1   153   0.3968254     2

177   39    93    1   177   0.3968254     2

179   39    94    1   179   0.3968254     2

180   39    91    1    180  0.3968254     2

> attach(strs.sample)

> tapply(score,trip,length)

 0  1

25 25

> tapply(score,trip,mean)

   0      1

64.68   77.04

> tapply(score,trip,var)

      0          1

268.4767   202.7067
```

(g) The number of students in the above stratified sample is____**50**____.

(h) Every student who does not want to visit a Spanish-speaking country has an inclusion probability of _______**0.1880**________.

(i) A 95% CI for the mean college score is [___**64.4856**___ , ___**72.8202**____] .

(j) The estimate for the total college score has expected value _____**13092.0003**_____ and true variance _____**174110.6309**____.

(k) The estimated mean score for students who plan to visit a Spanish-speaking country is ______**77.04**_______ with estimated variance ______**4.8907**_______ .

(l) Suppose we want to take another STRS, this time with 100 students. Using proportional allocation, we would sample ___**32**____ students who are planning to visit a Spanish-speaking country and ___**68**___ students who are not planning to do so.

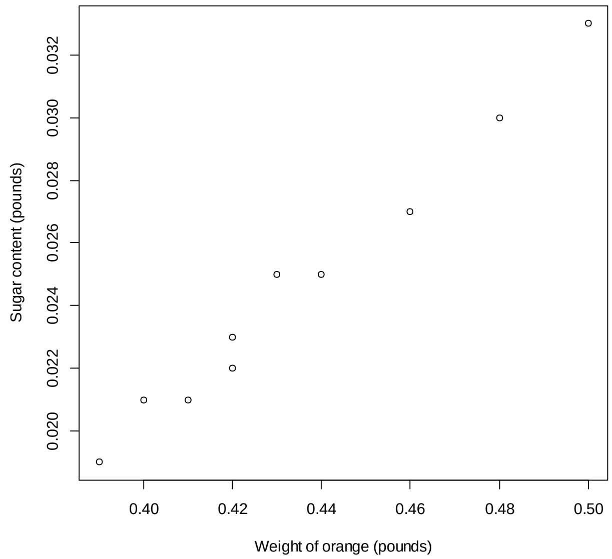
(m) Suppose we want to take yet another STRS of 100 students and assume the standard deviation of the *trip*=0 group is twice as much as the *trip*=1 group. Using Neyman allocation, we would sample ___**19**____ students who are planning to visit a Spanish-speaking country and ___**81**___ students who are not planning to do so.

[10 marks]

4. A truckload of oranges has just arrived and you wish to make inference about the sugar content of the oranges. The total number of oranges is unknown (and you do not want to count them). You obtain the total weight of all the oranges, as 1800 pounds, by first weighing the truck loaded then unloaded. Then you take a random sample of n=10 oranges - summary of data is given below: y_i x_i ${e_i}^2$

	Sugar content (pounds)	Weight of orange (pounds)	$(\textbf{Sugar content} - \hat{\boldsymbol{B}} * \textbf{weight})^2$
Total:	0.246	4.35	0.000053

$\bar{y} = 0.0246$ $\bar{x} = 0.435$ $s_e^2 = \frac{0.000053}{9}$



<u>Using proper notation, show your work and then circle the final answer</u> *for the following:*

[3 marks]
a) What type of estimation would be reasonable to use in this situation? Justify.

Since N is unknown, use ratio estimation which seems reasonable because weight and sugar content are positively correlated (from scatterplot).

[2 marks]
b) Estimate N, the total number of oranges in the truckload.

$N \approx \frac{t_x}{\bar{x}} = \frac{1800}{0.435} = 4137.9310$ so approximately $\boxed{\mathbf{4138\ oranges}}$.

[5 marks]
c) You are now informed that that the truckload contained 3000 oranges.
Find a 95% CI for B, the mean sugar content per pound of an orange.

$N = 3000$

$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{0.0246}{0.435} = 0.0566 \qquad SE(\hat{B}) = \sqrt{\left(1 - \frac{10}{3000}\right)\frac{0.000053}{9(10)(0.435)^2}}$

95% CI is: $\boxed{[\mathbf{0.0531}, \mathbf{0.0601}]}$