NAME (PRINT): _____
Last/Surname First /Given Name

STUDENT #: _____ **SIGNATURE:** _____

UNIVERSITY OF TORONTO
STA304/1003 H1F - PRACTICE EXAM
Surveys, Sampling and Observational Data
Dr. Ramya Thinniyam
Duration - 3 hours
Aids: Non-Programmable Calculators;
Formula Sheet (use the posted Exam Formula sheet for practice)

INSTRUCTIONS:

-There are 7 questions – answer all questions.
-For all true/false and fill in the blank questions - circle or put your final answers in blanks as instructed. You may do rough work on scrap paper but 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.
-For questions with 'R' output given, you may copy the appropriate numbers without calculating yourself. Do your own calculations when specified to do from scratch.
-Final answers should be in reduced form. Round your answers to 4 decimal places where appropriate.

BEST WISHES! ☺

Question	1	2	3	4	5	6	7	TOTAL
Value	10	10	20	20	10	15	15	**100**
Mark								

[10 marks - 1each]

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

(a) In cases that require Cluster sampling, we can instead use STRS to increase precision. T F

(b) A Systematic Sample produces more precise estimators than an SRS of the same size. T F

(c) A small R_a^2 is desirable for a Cluster sample. T F

(d) The sampled population is a subset of the target population. T F

(e) Non-response bias is a type of non-sampling error. T F

(f) Voluntary surveys tend to have selection bias. T F

(g) A probability sample ensures a representative sample. T F

(h) In STRS, Neyman and Optimal allocation are the same when the variances in each stratum are the same. T F

(i) Ratio estimates have smaller variances than SRS estimates. T F

(j) A company wants to assess their costumers opinions on their product, 'X'. They ask this on the questionnaire: *"We have recently upgraded X's features to become a first class tool. What are your thoughts on X ?"* This is an example of a leading question. T F

[10 marks - 2 each]

2. <u>NAME THE APPROPRIATE METHOD</u>: *Below is a list of different sampling/estimation/ design methods.*

I. Simple Random Sample without replacement (SRS)
II. Simple Random Sample With Replacement (SRSWR)
III. Stratified Random Sample (STRS)
IV. One-Stage Cluster Sample
V. Two-Stage Cluster Sample
VI. Ratio Estimation
VII. Regression Estimation
VIII. Systematic Sample
IX. Convenience Sample
X. Experimental Design

For each of the following, <u>write the Roman Numeral</u> corresponding to the method that is most appropriate and will give the best results for the scenario. If given, identify the appropriate parameters of the sampling method. If you choose Cluster Sampling/STRS indicate how you would choose the clusters/strata and why.
For ex, if you choose SRS - write 'I' and identify N and n.
If you choose STRS - write 'III' and identify all N_h, n_h, n, N and explain how you would choose strata and why.

(a) A researcher is interested in studying opinions of Catholic Church members. Suppose the country has 400 churches. 50 churches from the country are randomly selected and 20 members from each are randomly chosen and interviewed.

(b) You wish to estimate the mean GPA of students at a particular university. You have a list of the student population by student numbers, but no other information about the students. There are 10,000 students in total and you have enough resources to sample 200 of them.

(c) We wish to estimate the mean blood pressure level of patients in a hospital.

(d) We wish to study the effectiveness of a new toothpaste in reducing the number of cavities in adults.

(e) You wish to estimate the mean number of sales of a certain new book. There are 252 bookstores of which you randomly sample 40. You measure each bookstore's size and the number of sales of the new book.

[20 marks]

3. In order to estimate the total inventory of its products being held, a tire company conducts a proportional Stratified Random Sample of its dealers, stratified according to the inventory being held from the previous year. For a sample of size 100, the following data was obtained:

Previous year's inventory	Total number of dealers	Sample Mean of current inventory	Sample Variance of current inventory
I. (0-99)	400	105	400
II. (100-199)	1000	180	900
III. (200 +)	600	282	2500

Show your work and then circle your final answers.

[2m]

a) Discuss the advantages of using STRS in this case?

[2m]

b) Find the allocation used for this sample. ie find the sample sizes for each stratum.

[4m]

c) Find a 95% CI for this year's total inventory.

[2m]

d) Find a 95% CI for this year's mean inventory.

[3m]

e) Find a 95% CI for this year's mean inventory for all dealers who have 200+ inventory being held from last year.

[4m]

f) Suppose a follow-up study is to be conducted. What would the optimal allocation of the sample size *n=100* be if sampling one dealer from stratum III is four times as costly as sampling from stratum I and the cost of sampling one dealer from stratum II is twice that of I. Use the above sample as a preliminary survey.

[3m]

(g) Ignoring sampling costs, do you think STRS with proportional allocation will be significantly better than SRS? Do you think the optimal allocation will be significantly better than proportional allocation? Justify your reasons (without doing variance calculations)?

[20 marks]

4. <u>FILL IN THE BLANKS</u>: *You may do rough work on scrap paper, but only answers filled in the blanks will be marked.*

We are interested in making inferences on test scores in a population of 12 algebra classes. A SRS of 5 classes is taken, then some students in the selected classes are randomly selected and given an algebra test and the scores are recorded. Below is the description of the data and 'R' output:

'class' - class number
'Mi' - number of students in class
'score' - test score
'ID_unit' - student ID number

```
> algebra <- read.csv("algebra.csv")

> mean(algebra$score)
[1] 62.56856

> var(algebra$score)
[1] 316.6287

> cl = mstage(data=algebra, stage=c("cluster", "stratified"), varnames=c("class","score"),
            size=list(5,c(2,2,3,3,3)),method="srswor")

> mysample=getdata(algebra,cl)

> mysample

[[1]]

     Mi score class ID_unit Prob_ 1 _stage
53  24    37    38      53     0.4166667
57  24    65    38      57     0.4166667
52  24    60    38      52     0.4166667
.
132 28    65    44     132      0.4166667
133 28    60    44     133      0.4166667
136 28    52    44     136      0.4166667
.
159 19    34    46     159      0.4166667
163 19    71    46     163      0.4166667
160 19    42    46     160      0.4166667
.
220 17   100    58     220      0.4166667
221 17    43    58     221      0.4166667
222 17    48    58     222      0.4166667
.
234 21    71    62     234      0.4166667
227 21    31    62     227      0.4166667
.
```

```
[[2]]
     class Mi score ID_unit  Prob_ 2 _stage  Prob
53     38   24    37     53     0.08333333       0.03472222
68     38   24    68     68     0.08333333       0.03472222
141    44   28    75    141    0.07142857       0.02976190
155    44   28    75    155    0.07142857       0.02976190
160    46   19    42    160    0.15789474       0.06578947
170    46   19    64    170    0.15789474       0.06578947
169    46   19    34    169    0.15789474       0.06578947
220    58   17   100     220   0.17647059       0.07352941
226    58   17    49    226    0.17647059       0.07352941
219    58   17    49    219     0.17647059      0.07352941
247    62   21    61    247     0.14285714      0.05952381
239    62   21    63    239     0.14285714      0.05952381
246    62   21    51    246     0.14285714      0.05952381

> Mi = tapply(mysample[[1]]$score,mysample[[1]]$class,length)
> Mi
38 44 46 58 62
24 28 19 17 21

> mean(Mi)
[1] 21.8

> attach(mysample[[2]])


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

> a
38 44 46 58 62
 2  2  3  3  3

> mean(a)
[1] 2.6

> b = tapply(score,class,mean)

> b

  38       44       46     58      62
52.50 75.00 46.67 66.00 58.33

> c = sum(Mi*b)/sum(Mi)
> c
[1] 60.49235

> d <-  tapply(score,class,var)

> d

   38       44       46         58         62
480.50  0.00  241.33  867.00  41.33

> sum(Mi^2 * (b - c)^2)
[1] 281630.7

> sum( Mi^2 *(1-a/Mi)*(d/a) )
[1] 225297.1


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

> e

    38       44       46         58       62
105.00 150.00 140.01 198.00 174.99

> mean(e)
[1] 153.6

> var(e)
[1] 1247.125
```

[1m]
(a) This is a _____sample. *[name the type of sampling method - be specific]*.

[2m]
(b) The psus are the _____ and the ssus are the _____ . *[name the variables]*.

[1m]
(c) The total sample size is _____ .

[5m]
(d) Fill in the correct numbers to identify the parameters: n = _____ , N = _____ ,
$M_1 =$ _____ , $M_2 =$ _____ , $M_3 =$ _____ , $M_4 =$ _____ , $M_5 =$ _____ ,
$m_1 =$ _____ , $m_2 =$ _____ , $m_3 =$ _____ , $m_4 =$ _____ , $m_5 =$ _____.

[2m]
(e) The following classes were selected in the sample: ____, ____, ____, ____, and ____. *[write the class numbers]*

[2m]
(f) The mean algebra test score is estimated as _____ with a standard error of _____ .

[1m]
(g) The estimate for the mean test score has expected value _____ .

[1m]
(h) A student from Class #46 has an inclusion probability of _____.

[1m]
(i) The probability of choosing any class to be in the sample is _____ .

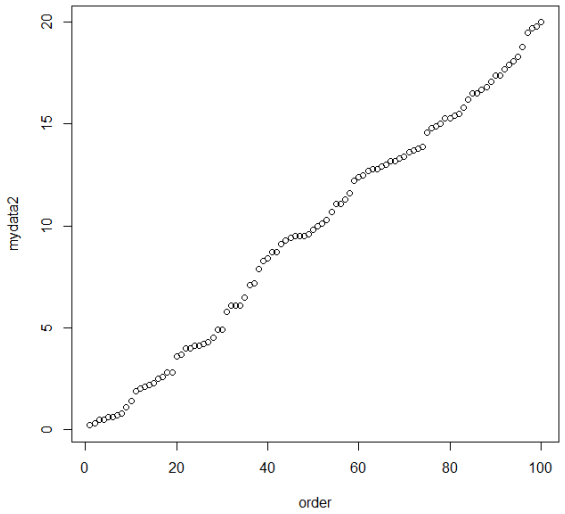
[2m]
(j) You are now informed that there are a total of 299 students in the population. A 95% CI for the mean test score is [_____ , _____] .

[2m]
(k) A point estimate for the mean score in Class #58 is _____ with estimated an variance of _____.

[10 marks]

5. Suppose we have a population with 100 elements and wish to take a Systematic Sample of size 20. The variable '*mydata*' represents the measurements for the whole population. 'R' output is given below:

```
order<- seq(1,100,1)
>plot(order,mydata)
```




```
>start = sample(1:5,1)

> start
[1] 2

> mysample = mydata[seq(from=start,to=100,by=5)]

> mean(mysample)
[1] 9.535

var(mysample)
[1] 35.26869

>cluster <-  rep(c(1,2,3,4,5), 20)

>cluster
[1]  1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
.
.
[90] 1 2 3 4 5 1 2 3 4 5

> lin.reg <- lm(mydata ~ as.factor(cluster))

> anova(lin.reg)
```

Analysis of Variance Table

Response: mydata

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(cluster)	4	7.6	1.89	0.0535	0.9946
Residuals	95	3356.3	35.33		

For each of the following, show your work and then circle your final answers (use proper notation and then you may copy from output directly where possible):

[1m]
(a) What is the sampling interval length?

[1m]
(b) Give a point estimate for the mean.

[2m]
(c) Calculate the theoretical variance of the estimate for the mean using Systematic Sampling.

[3m]
(d) Use SRS to estimate the variance of the point estimate from (b). Why do you have to use SRS to estimate the variance?

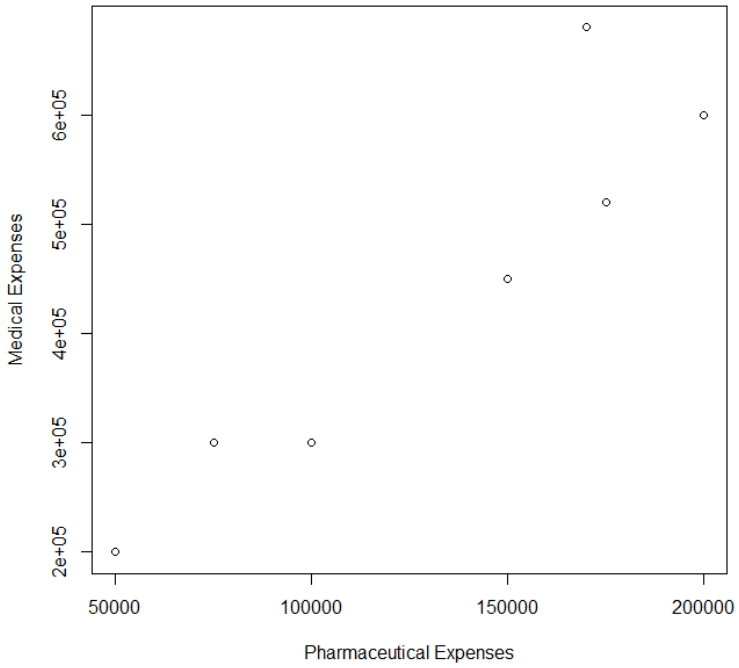
[3m]
(e) Do you expect your SRS variance estimate from (d) to under/over-estimate/ or approximately equal the theoretical variance from (c)? Explain your reasoning by commenting on the structure of the sampling frame, calculating ICC/other relevant measures, etc.

[15 marks]

6. Suppose a community is divided into 20 different areas and it is known that the mean pharmaceutical expenses is $120,000 for the whole community. We wish to make inferences about the medical expenses in this community. We take a random sample of 8 areas and for each area measure the amount of pharmaceutical expenses and total medical expenses. Below is a summary of the data:

	Total Medical Expenses ($)	**Pharmaceutical Expenses ($)**	$(\textbf{Medical} - \widehat{\boldsymbol{B}} * \textbf{Pharmaceutical})^2$
Total:	3, 500, 000	1, 070, 000	29, 402, 323, 347

$s^2_{medical}$ = *26, 650,000,000* ; $s^2_{pharmaceutical}$ = *2,791,071,429* ; $r_{medical,pharmaceutical}$ = *0.9272*



Using proper notation, show your work and then circle the final answers for the following:

[1m]

(a) What are the sampling units in this case?

[2m]

(b) What type of estimation is reasonable to use in this case? Justify.

[1m]
(b) Give a point estimate for the total amount of medical expenses in this community.

[2m]
(d) Give a point estimate for the mean amount of medical expenses per dollar of pharmaceutical expenses.

[4m]
(e) Find a 95% CI for the total medical expenses in this community.

[3m]
(f) Using SRS, find a 95% CI for the total amount of medical expenses.

[2m]
(g) Which CI would you expect to be wider: the one from (e) or the one using SRS from (f)? Explain your reasoning - be specific.

[15 marks]

7. <u>PROOFS/DERIVATIONS</u>: ***Use proper notation, show all your work, and <u>justify each step</u>. Clearly label any random variables or parameters you introduce to prove/derive the following:***

[6m]

(a) In a Stratified Sample, prove that $\bar{y}_{str}$ is unbiased for the population mean, $\bar{y}_U$, and that $\bar{y}_{str}$ has variance $\sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{S_h^2}{n_h}$.

You may use results from SRS without proving them - state them clearly and use them.

[9m]

(b) A population has 3500 units partitioned into 5 groups: the first 3 groups have 500 elements each and the last 2 groups have 1000 elements each. You take a SRS of size 50 from each group and measure the response variable *y* (which gives you a total of 250 measurements). In order to estimate the population mean, you calculate the sample mean for each group and then average these 5 averages.

In your answers, use i to index the groups and j to index the elements within the groups. Clearly label any other notation you introduce. Answer the following:

- Call your estimator of the mean, $\tilde{y}$. Express $\tilde{y}$ as a weighted sum of the 250 observations.
- Express the population mean as a weighted sum of the group means.
- Prove that $\tilde{y}$ is a biased estimator of the population mean. (*you may use any SRS results without proof - state them clearly and use them*).
- Why is $\tilde{y}$ biased and how can it be modified to make it unbiased?
- Under what circumstances will $\tilde{y}$ be unbiased?