

**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! ☺**

<b>Question</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>TOTAL</b>
Value	10	10	20	20	10	15	15	<b>100</b>
Mark								

**[10 marks - 1 each]**

**1. TRUE/FALSE:** 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. SRS T F
- (h) In STRS, Neyman and Optimal allocation are the same when the variances in each stratum are the same. cost. T F
- (i) Ratio estimates have smaller variances than SRS estimates. T F
- (j) A company wants to assess their customers 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. NAME THE APPROPRIATE METHOD:** 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, write the Roman Numeral 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.

Two-stage cluster

$psus = \text{churches}$

$ssus = \text{person}$

$N=400$

$n=50$

$M_i = \text{unknown}$

$M = \text{unknown}$

$m_i = 20 \quad \forall i \quad \text{churches provide convenient access to sample}$

- (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.

Assume students #'s have no pattern with GPA

I. SRS / VIII systematic

$N=10000$

$N=1000$

$n=200$

$k=50$

$n=200$

$$= \frac{N}{n} .$$

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

III. STRS

stratify by : age/weight/health conditions/etc.

Blood pressure varies by age/weight/etc. so we can increase precision by stratifying.

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

X Experimental Design

Survey design not appropriate

Since this isn't observational

data.

- (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.

VI. Ratio Estimation

$y_i = \text{sales of book}$

$x_i = \text{size of bookstore}$

Appropriate to fit through origin

$x \& y$  are pos. correlated ,  $N=252$ ,  $n=40$ .

[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:

Stratum	$N_i$	$\bar{y}_i$	$s_i^2$
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
	total 2000		

Show your work and then circle your final answers.

$$n=100 \quad N=2000$$

[2m]

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

- expect current inventory to vary by previous years  
 $\Rightarrow$  increase precision
- easy to access strata (sort previous year's inventory).

[2m]

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

proportional allocation

$$n_I = \frac{40}{2000} (100) = 20$$

$$n_{\text{III}} = \frac{600}{2000} (100) = 30$$

$$n_{\text{II}} = \frac{1000}{2000} (100) = 50$$

$$n = 20 + 50 + 30 = 100$$

[4m]

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

$$\begin{aligned} \hat{T}_{\text{str}} &= \sum_{i=1}^3 N_i \bar{y}_i = 400(105) + 1000(180) + 600(282) = 371,200 \\ \text{se}(\hat{T}_{\text{str}}) &= \sqrt{\sum_{i=1}^3 (1 - \frac{n_i}{N_i}) \frac{s_i^2}{n_i} N_i} = \sqrt{(1 - \frac{20}{400}) \frac{400}{20} + (1 - \frac{50}{100}) \frac{1000}{50} + (1 - \frac{30}{60}) \frac{600}{30}} \\ &= 6974.2383 \end{aligned}$$

CI is

$$\begin{aligned} \hat{T}_{\text{str}} &\pm 1.96 \text{se}(\hat{T}_{\text{str}}) \\ &= (377530.49, 404869.51) \end{aligned}$$

[2m]

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

$\bar{y}_{\text{u}}$

Divide CI by  $N=2000$

$$(188.77, 202.43)$$

[3m]

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

$$\bar{Y}_{3u}$$

stratum III only:

SRS of size  $n_3 = 30$  pop size  $N_3 = 600$

$$95\% \text{ CI: } \bar{Y}_3 \pm 1.96 \sqrt{(1 - \frac{n_3}{N_3}) \frac{s_3^2}{n_3}} = 282 \pm 1.96 \sqrt{(1 - \frac{30}{600}) \frac{2500}{30}} = (264.56, 299.44)$$

[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.

$$\begin{array}{lll} n=100 & C_3=4C_1 & C_2=2C_1 \\ S_1=20 & S_2=30 & S_3=50 \end{array}$$

optimal allocation

$$\sum_{i=1}^3 \frac{N_i S_i}{\sqrt{C_i}} = \frac{400(20)}{\sqrt{C_1}} + \frac{1000(30)}{\sqrt{2C_1}} + \frac{600(50)}{2\sqrt{C_1}} = \frac{44213.2}{\sqrt{C_1}}$$

$$\text{using } n_i = \frac{N_i S_i}{\sum_{j=1}^3 \frac{N_j S_j}{\sqrt{C_j}}} = \dots$$

$$\left. \begin{array}{l} n_1 = 18.09 \approx 18 \\ n_2 = 47.98 \approx 48 \\ n_3 = 33.93 \approx 34 \end{array} \right\} \Rightarrow n = 100$$

[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)?

①

① Yes: Since strata large, STRS with proportional allocation is better than SRS

② Yes. Since variance differ. Optimal better than proportional allocation.

**[20 marks]**

**4. FILL IN THE BLANKS:** 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:

2-stage

'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 =  $y_i$  not  $N_i \bar{y}_i$

```
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 two-stage cluster sample. [name the type of sampling method - be specific].

[2m]

- (b) The psus are the classes and the ssus are the students. [name the variables].

[1m]

- (c) The total sample size is 13.

$$2+2+3+3+3$$

[5m]

- (d) Fill in the correct numbers to identify the parameters:  $n = \underline{5}$ ,  $N = \underline{12}$ ,  $M_1 = \underline{24}$ ,  $M_2 = \underline{28}$ ,  $M_3 = \underline{19}$ ,  $M_4 = \underline{17}$ ,  $M_5 = \underline{21}$ ,  $m_1 = \underline{2}$ ,  $m_2 = \underline{2}$ ,  $m_3 = \underline{3}$ ,  $m_4 = \underline{3}$ ,  $m_5 = \underline{3}$ .

[2m]

- (e) The following classes were selected in the sample: 38, 44, 46, 58, and 62. [write the class numbers]

[2m]

- (f) The mean algebra test score is estimated as 60.4924 with a standard error of 5.0186.

$$\begin{aligned} f &\rightarrow \text{Ratio Est. (M unknown)} \\ \hat{y}_r &= \frac{\sum_{i \in s} M_i \bar{y}_i}{\sum_{i \in s} M_i} = \hat{y} = 60.4924 \end{aligned}$$

$$\begin{aligned} SE(\hat{y}_r) &= \sqrt{\frac{1}{nM} \left( \frac{1}{N} \sum_{i \in s} M_i^2 \right) S_p^2 + \frac{1}{nN} \sum_{i \in s} M_i^2 \left( \frac{1-m_i}{M_i} \right) S_p^2} \\ &= \sqrt{\frac{1}{5(218)}} \end{aligned}$$

[1m]

- (g) The estimate for the mean test score has expected value 62.5686.

[1m]

- (h) A student from Class #46 has an inclusion probability of 0.0658.

[1m]

- (i) The probability of choosing any class to be in the sample is  $\frac{5}{12} = 0.4167$ .

4.j).

[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 [39.6491, 66.2029].

$$\begin{aligned} \hat{y}_{umb} &= \frac{1}{M} \sum_{i \in s} M_i \bar{y}_i \quad M=299 \\ &= \frac{1}{299} \frac{12}{5} \left( \frac{1}{5} \sum_{i \in s} M_i \right) \end{aligned}$$

$$= \frac{1}{299} \frac{12}{5} (60.4924 - 21.8 \times 5) = 52.9257$$

[2m]

- (k) A point estimate for the mean score in Class #58 is 66 with estimated an variance of 238. use SRS

$$\begin{aligned} \hat{y}_{Li} &= M_i \bar{y}_i = (1260, 2100, 886.73, 102, 1224.93) \\ S_b^2 &= \frac{1}{n-1} \left[ \sum_{i \in s} \hat{y}_{Li}^2 - \frac{1}{n} (\sum_{i \in s} \hat{y}_{Li})^2 \right] = 21198.52 \end{aligned}$$

$$\begin{aligned} \hat{y}_{umb} &= 45.8839 \\ 95\% \text{ CI for } \bar{y}_u &: [52.9257 \pm 1.96 \cdot 45.8839] \end{aligned}$$

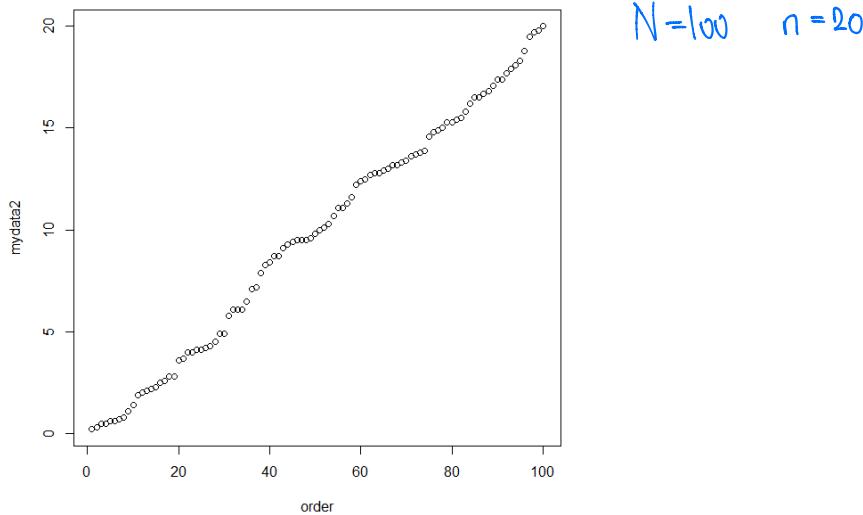
$$S_b^2 = \frac{1}{n-1} \sum_{i \in s} (M_i \bar{y}_i - \bar{M} \hat{y}_{umb})^2 : S_b^2 \text{ is the sample var of } \hat{y}_i = M_i \bar{y}_i \text{ terms}$$

$$\hat{y}_i = y_i \text{ NOT } M_i \bar{y}_i$$

**[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?

$$k = \frac{N}{n} = \frac{100}{20} = 5$$

[1m]

(b) Give a point estimate for the mean.

$$\hat{\bar{y}}_{\text{sys}} = \bar{y} = 9.535$$

[2m]

(c) Calculate the theoretical variance of the estimate for the mean using Systematic Sampling.

$$V(\hat{\bar{y}}_{\text{sys}}) = (1 - \frac{1}{k}) \frac{MSB}{n} = (1 - \frac{1}{5}) \frac{1.89}{20} = 0.0756$$

$$ICC = 1 - \frac{M}{m-1} \left( \frac{SSW}{SSTO} \right) = 1 - \frac{20}{19} \frac{3356.3}{3356.3 + 7.5} = -0.0503$$

usually it's "m": cluster size      ↗ now notation changed

$$V(\hat{\bar{y}}_{\text{sys}}) = \frac{s^2}{n} [1 + (n-1) ICC] = 0.0767$$

[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?

This is a one-stage cluster sample with only 1 psu  $\rightarrow$  cannot est. var. using only 1 total.

$$\text{using SRS: } V_{\text{SRS}}(\bar{y}_{\text{sys}}) = (1 - \frac{20}{100}) \frac{s^2}{20} = (1 - \frac{20}{100}) \frac{35.26869}{20} \div 14107 \quad \boxed{\sqrt{141071}}$$

[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.

Since SF is in increasing order, we expect higher precision than SRS since values in our sample (ie in each cluster) every 5th element will be spread apart.  $\Rightarrow$  good rep. of population  
 $ICC = -0.0503 < 0 \Rightarrow$  heterogeneity within clusters.

$$\text{Lowest } ICC = \frac{1}{n-1} = -0.0526$$

Slightly higher than minimal ICC.

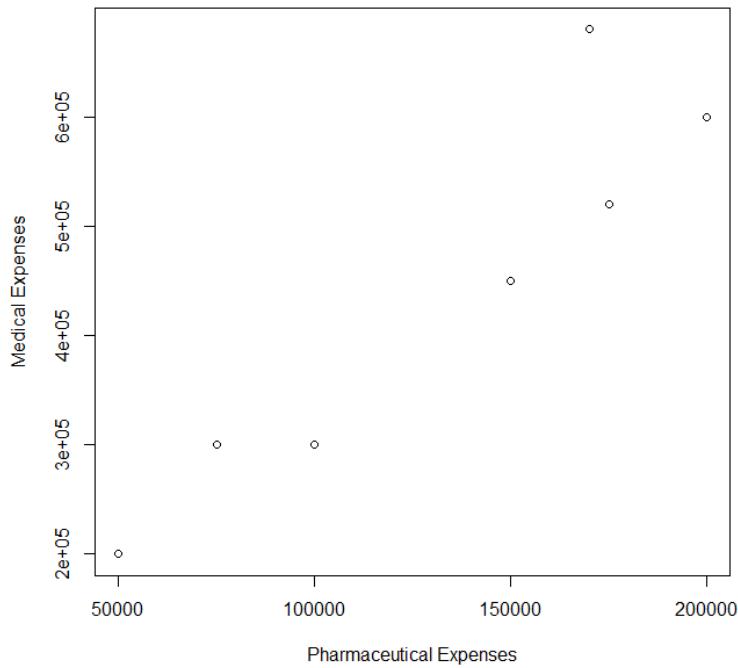
Better precision with systematic. SRS will over estimate variance

**[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 (\$)	$(\text{Medical} - \hat{B} * \text{Pharmaceutical})^2$
<b>Total:</b>	3, 500, 000	1, 070, 000	29, 402, 323, 347

$$s_{\text{medical}}^2 = 26,650,000,000 ; s_{\text{pharmaceutical}}^2 = 2,791,071,429 ; r_{\text{medical, pharmaceutical}} = 0.9272$$



**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?

?

Area

**[2m]**

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

ratio estimate  
"strong correlation"

[1m]

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

$$Y_i = \text{medical} \quad X_i = \text{pharma}$$

$$\text{res}_i^2 = (\text{Med} - \hat{\beta} \rightarrow \text{pharma})^2$$

$$\bar{X}_u = 120000$$

$$\bar{x} = \frac{3070000}{8} = 133750$$

$$\text{Est: } T_{\text{med}} \quad \hat{T}_{y_r} = r\bar{X}_u = 3.27 \times 20 \times 120000 = 7858467.29$$

$$\hat{T}_{y_r} = \frac{\sum y_i}{r}$$

[2m]

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

$$r = \frac{\hat{T}_{y_r}}{\bar{X}_u} = \frac{3500000}{1070000} = 3.27 \rightarrow \$3.27 \text{ per dollar of pharm expense}$$

[4m]

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

$$\hat{T}_{y_r} = 7858467.29 \quad V(\hat{T}_{y_r}) = (1 - \frac{8}{20}) \frac{(120000)^2 \times 29402323347}{133750 \times (8-1) \cdot 8}$$

95% CI for  $\hat{T}_{y_r}$ :

$$\hat{T}_{y_r} \pm 1.96 \text{ se}(\hat{T}_{y_r}) = (\$7027255.70, \$7889678.89)$$

[3m]

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

f). Use SRS.  $\bar{T} = N\bar{y} = 20 \frac{3500000}{8} = 8750000$

$$\begin{aligned} V(\bar{T}) &= N^2(1 - \frac{n}{N}) \frac{s_y^2}{n} \\ &= 20^2(1 - \frac{8}{20}) \frac{26650000000}{8} \Rightarrow 95\% \text{ CI for } \bar{T}: \\ &(\$6997470.63, \$10502529.37) \end{aligned}$$

[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.

(SRS)

since highly correlated,  
use ratio est.  
var ↓ & CI ↑

[15 marks]

7. **PROOFS/DERIVATIONS:** Use proper notation, show all your work, and justify each step.

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.

proved in lecture

[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?

$$N=3500$$

$$N_1=N_2=N_3=500$$

$$N_4=N_5=1000$$

$$n_i=50 \quad (i=1, \dots, 5)$$

$$n=250$$

$$\Rightarrow \tilde{y} = \frac{1}{5} \sum_{i=1}^5 \bar{y}_i = \frac{1}{5} \sum_{i=1}^5 \sum_{j=1}^{50} \frac{y_{ij}}{50}$$

$$= \frac{1}{250} \sum_{i=1}^5 \sum_{j=1}^{50} y_{ij}$$

$$\cdot \bar{y}_u = \frac{1}{3500} \sum_{i=1}^5 \sum_{j=1}^{N_i} y_{ij} = \frac{1}{3500} \sum_{i=1}^5 N_i \bar{y}_{iu}$$

$$\cdot E(\tilde{y}) = E\left(\frac{1}{5} \sum_{i=1}^5 \bar{y}_i\right) = \frac{1}{5} \sum_{i=1}^5 E(\bar{y}_i) = \frac{1}{5} \sum_{i=1}^5 \bar{y}_{iu} \quad \text{since SRS from each group}$$

&  $\bar{y}_i$  unbiased..

$$= \frac{1}{3500} \sum_{i=1}^5 700 \bar{y}_{iu} \neq \bar{y}_u$$

not all  $N_i=700$

$E(\tilde{y}) \neq E(y) \Rightarrow$  biased.

$$\cdot \text{modify: } \hat{y}^* = \frac{1}{N} \sum_{i=1}^5 N_i \bar{y}_i = \hat{y}_{str} \quad \text{weights} = \frac{N_i}{N} \quad (\text{not equal weights!})$$

$\uparrow$   
when.  
group size proportional to population.