University of Toronto

STA304/1003 H1F - Summer 2014

Instructor: Dr. Ramya Thinniyam

Midterm Test -
May 29, 2014

Last Name (print):	Qiu
First Name (print):	Rui
Student Number:	999292509
Enrolled in (circle one):	(STA304) STA1003

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

INSTRUCTIONS:
-There are 5 questions – answer all questions.
-There are 7 pages total and a separate formula sheet. 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.
-You may copy/use numbers from R output as needed.
-Simplify answers and round to **4 decimal places** where appropriate.
-Recall: **SRS**=Simple Random Sample without replacement
-**SRSWR**=Simple Random Sample With Replacement

BEST WISHES! ☺

Question	1.	2.	3.	4.	5.	TOTAL
Value	10	15	20	10	5	60
Mark Earned	9	15	17.5	10	5	56.5

[10 marks]

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

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

(b) A bank asks this on their survey: "In order to improve wait times and offer our customers better service, we have recently implemented the new feature *X*. On a scale of 1 to 10 (1 being the lowest and 10 being the highest), how would you rate your satisfaction with *X*?" This is an example of a leading question. (T) F select F

(c) A 99% CI for the population proportion yields [0.65,1.00]. There is only a 1% chance that the true proportion is below 0.65. T (F)

(d) A census will eliminate/reduce sampling error. (T) F

(e) In probability sampling, each sampling unit has a known probability of selection. (T) F

(f) Sample results can be generalized to the population as long as the sample size is large. T (F)

(g) If a sample is selected in such a way that all units in the population have the same inclusion probability, the sample will be self-weighting. (T) F

(h) Undercoverage is a type of non-sampling error. (T) F

(i) Online surveys tend to have selection bias. (T) F

(j) A tree farm contains 100 trees of which 60 are classified as tall (height of at least 40 feet) and 40 are classified as short (under 40 feet). A sample of 9 tall trees and 6 short trees were taken. Since each tree has a 15% chance of being selected in the sample, this is an SRS. T (F)

[15 marks]

2. A researcher wishes to make inferences about the opinions of all users of the UTM Shuttle Bus (a bus service that is offered between the St. George and UTM campuses). During this summer session, the researcher randomly selects 2 days of the week and randomly selects 1 scheduled timing on each of the selected days. At the selected timings, when the bus arrives at the destination, the researcher samples every 5th person who gets off the bus. Each sampled person is given a questionnaire that asks about their opinions about the shuttle bus service such as how often they use the bus, their satisfaction level with the service, if they think more timings should be added to the existing schedule, etc.

a) Explain why the researcher was not able to use Simple Random Sampling and instead had to take a Systematic Sample.

And more important, since we don't have a specific sampling frame, we cannot do SRS.

Because if ~~use~~ using Simple Random Sampling, it would be hard to conduct the survey, ~~especially~~ ~~saying we randomly selected 2 person, but one on the first date of the summer session & one on the last day, it would be extremely time consuming and costly~~. specifically speak it would be ~~or~~ time consuming & costly.

b) Explain sampling error in this survey.

Sampling error is due to the randomness generated by 'sampling'. Since it's a sample, it cannot have the the full information of all population, in other words, if we choose different samples, the information would be different. However, sampling errors are inevitable even if we have a good sample.

c) Briefly discuss 2 sources of non-sampling error in this survey. Use the correct statistical terminology and then explain them in plain English.

- ~~Undercoverage (Selection Bias)~~ ~~The following situation would happen~~
  Those people on the same bus, say the 1st, 2nd, 3rd, etc. person who gets off the bus would not be interviewed. They are target population, but not ~~inclu~~ included in the Sampled population.

- No response (measurement errors)
  The following situation would happen:
  People taking the bus could be in ~~the~~ a hurry to go to class, So they probably would refuse to complete the interview due to personal reasons.

d) Propose a different sampling procedure to improve the survey design and obtain more accurate results. Be specific in your description and make sure to address how your proposed procedure would reduce the non-sampling errors identified above from part c).

My procedure is similar, instead I suggest that we can do the interview on bus, (*) or to say during the trip, so that the no-response problem would be solved. And we can increase the sample size, say we select the first 3 people get on the bus, such that ~~the~~ more information would be included in the survey.

(*) Let the bus driver give out questionaires before leaving starting station, and collect them at the terminal station.

[20 marks]

3. A library contains a total of 1000 books. If the majority of the books in the library need to be rebound, the library will face problems during the upcoming library inspection. To estimate the proportion of books that need rebinding, a librarian uses a random number table to randomly select 100 locations on library shelves. The librarian then walks to each location, looks at the book that resides at that spot, and records whether the book needs rebinding or not.

a) Identify the following for this survey:

Target Population - the total of 1000 books in the library

Sampling Frame - ~~a list of [illegible]~~ the random number table (location from 1 to 100)

Sampling Unit - ~~the location~~ a number corresponding to a specific location

Observation Unit - the book at that location.

Sample - the 100 books at the 100 selected locations

Variable - whether the book at that location needs rebinding or not.

b) Discuss any possible sources of selection bias or inaccuracy of responses.

① Undercoverage, ~~not~~ some books (could ~~be~~ need rebinding) are not included in the sampled population.

② Since the number of sample is not small (for a single librarian doing this job), if he goes to one after one location continuously, he ~~#~~ COULD be pretty tired as the books are located in the different locations, such that he COULD check for a wrong book. (just because tired body misleads his decision. etc)

c) At minimum, how many books should be sampled to estimate the percentage of interest within 2% of the true value using 95% confidence? ***Show your work and circle the final answer.***

$e = 0.02,\ \alpha = 0.05,\ S^* = \frac{1}{2} = 0.5$ (since $S^{*2} = p(1-p)$ => when $p = \frac{1}{2}$, S^{*2} max)

$$n_0 = \left(\frac{Z_{\alpha/2}\, S^*}{e}\right)^2 = \left(\frac{1.96 \times 0.5}{0.02}\right)^2 = 2401$$

$$n = \frac{Z_{\alpha/2}^2 \cdot S^{*2}}{e^2 + \frac{Z_{\alpha/2}^2 \cdot S^{*2}}{N}} = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{2401}{1 + \frac{2401}{1000}} \doteq 705.9688 \doteq \boxed{706}$$

d) Now suppose the librarian takes a SRS of 120 books and finds that 90 of them are in good condition (and do not need rebinding). Find a 95% CI for the true percentage of books in this library that need rebinding. ***Show your work and circle your final answer.*** Also, comment on whether there is evidence that the library will face problems during the upcoming inspection. Justify your answer.

4

$\alpha = 0.05$, $n = 120$, $N = 1000$

$\hat{p} = \frac{120 - 90}{120} = 0.25$

$\hat{p} \pm z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.25 \pm 1.96\sqrt{\left(1 - \frac{120}{1000}\right)\frac{0.25 \times 0.75}{119}}$

$\doteq (0.1770, 0.3230)$

in percentage it's (17.70, 32.30)

They will have problems ✗ In a word, the sample is not very confident, have to choose larger sample size in order to get a more accurate number of books which need rebinding. ✗

e) Using the information from this question and the sample in part d), give a point estimate for the total number of books in this library that need rebinding.

1

$\hat{t} = N\hat{p} = 1000 \times 0.25 = 250$

f) What is the probability of selection for an SRS of size 120 from this library?
(You do not need to simplify the answer.)

1

$P(\text{a book selected in a SRS of size 120}) = \frac{\binom{999}{119}}{\binom{1000}{120}}$ ✗

g) How many simple random samples of size 120 from this library will contain the ith book?

1

$\binom{999}{119} = \frac{999!}{119!\,880!}$ (probably should just leave it here)

[*10 marks*]

4. Fill in the blanks: *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.*

We are interested in taking a SRS from the population of all restaurants in downtown Toronto and estimating the mean rating (rating is measured out of 100 points). Below is some 'R' output:

```
>restaurantdata <- read.csv("restaurant.csv")
>restpopulationratings <- restaurantdata$rating
> length(restpopulationratings)
[1] 100
> mean(restpopulationratings)
[1] 70.27187
> var(restpopulationratings)
[1] 5.851095
>sample1<-sample(restpopulationratings,25,replace=T)
> [1]  71.6  67.3  67.4  67.4  73.2  69.4  73.4  67.9  66.0  70.5  70.9  70.0  70.0
[14]  71.1  67.3  70.1  73.2  71.2  71.6  65.9  67.3  71.9  69.8  69.9  71.9
> units <- sample(1:100,25,replace=F)
> units
 [1] 100  78  87  30  53  63  51  47  36  76  14  33  52  70  83  68  46  25  91  10  88  21  56  80  58
> sample2 <- restpopulationratings[units]
 [1]  67.6  72.0  67.4  67.9  71.2  69.4  73.4  67.9  66.0  72.5  70.9  70.2  70.0
[14]  71.1  67.3  70.1  72.5  70.7  71.6  73.2  67.3  65.9  69.8  69.9  71.9
> mean(sample1)
[1] 69.848
> var(sample1)
[1] 5.0226
> mean(sample2)
[1] 69.908
> var(sample2)
[1] 4.8916
```

(a) The population size is **100** and the sample size is **25**.

(b) The 7th selected rating measurement for the sample is **73.4** which corresponds to the **57st** unit in the population.

(c) Each restaurant in the sample represents itself + **3** restaurants that were not sampled.

$E(\bar{y}) = \bar{y}_u$ $\quad \sqrt{(1-\frac{25}{100})\frac{s^2}{25}}$ $\quad s^2 = 4.8916$

(d) The expected value of the sample mean is **70.2719** with a standard error of **0.3831**.

(e) The expected value of the sample variance is **5.8511**.

(f) An approximate 99% CI for the population mean rating is [**68.9197** , **70.8963**].

(assume the sample size is large enough and that you do not know any of the population parameters even if they are given in the output.)

$69.908 \pm 2.58\sqrt{(1-\frac{25}{100})\frac{4.8916}{25}}$

[5 marks]

5. Prove that for binary data, $s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$.

Hint: *Begin with the definition of s^2 from the formula sheet and then prove that it is equal to the above expression. Define any terms you introduce and justify all the steps.*

Proof:

note: $y_i = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob } (1-p) \end{cases}$

$$s^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})^2$$

$$= \frac{1}{n-1}\sum_{i\in S}(y_i^2 - 2y_i\bar{y} + \bar{y}^2) \quad \text{(computation)}$$

$$= \frac{1}{n-1}\sum_{i\in S}(y_i^2 - 2y_i\hat{p} + \hat{p}^2) \quad (\hat{p} = \bar{y})$$

$$= \frac{1}{n-1}\left[\sum_{i\in S}(y_i^2 - 2y_i\hat{p}) + n\hat{p}^2\right] \quad (\sum_{i\in S}\hat{p}^2 = n\hat{p}^2)$$

$$= \frac{1}{n-1}\left[\sum_{i\in S}(y_i^2) - 2\hat{p}\sum_{i\in S}y_i + n\hat{p}^2\right] \quad \text{(linearity)}$$

$$= \frac{1}{n-1}\left[\sum_{i\in S}y_i^2 - 2\hat{p}\cdot\hat{p}\cdot n + n\hat{p}^2\right] \quad (\sum_{i\in S}y_i = n\hat{p})$$

$$= \frac{1}{n-1}\left[\sum_{i\in S}y_i^2 - n\hat{p}^2\right] \quad \text{(computation)}$$

$$= \frac{1}{n-1}\left[\sum_{i\in S}y_i - n\hat{p}^2\right] \quad (y_i = y_i^2 \text{ for binary data i.e. } 1^2=1,\ 0^2=0)$$

$$= \frac{1}{n-1}\left[n\hat{p} - n\hat{p}^2\right]$$

$$= \frac{1}{n-1}\, n\hat{p}(1-\hat{p})$$

$$= \frac{n}{n-1}\hat{p}(1-\hat{p})$$