University of Toronto

STA304/1003 H1F - Summer 2014

Instructor: Dr. Ramya Thinniyam

Midterm Test -
May 29, 2014

Last Name (print):	SOLUTIONS
First Name (print):	
Student Number:	
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						

[10 marks - 1 each]

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

(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.

[2 marks]

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

The researcher could not use SRS since a sampling frame (list of all users of the UTM Shuttle Bus service) is not available. A list of all UTM and St George students is obtainable but not all of them use the service; a list of users who buy bus passes may be available but not for those who buy tickets.

[2 marks]

b) Explain sampling error in this survey.

Sampling error occurs in this survey because it is not a census. Different samples will yield different results and estimates (inevitable even for 'good' samples).

[6 marks]

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

Any **2** of these answers with correct statistical terminology and explanation/discussion:

1) **Undercoverage:** the survey was conducted during the summer session. There are far less courses offered in the summer so we would expect less students to use the bus service during the summer. (The target population is all users of the UTM Shuttle Bus).

2) **Selection Bias:** only 2 days and 1 timing on each day were selected. Certain days/times will have higher usage than others, so the days/times should be sampled to be representative. For example, typically there are no lectures on Fridays so less people will use the bus and less service times are available.

3) **Multiple Listings:** The same person may be selected more than once if they use the bus at more than one of the selected times.

4) **Non-response:** the questionnaire was given to people as they exit the bus and at this time they will likely be in a rush to get to class/appointments and hence refuse to participate.

5) **Measurement Bias:** people may estimate, round, or forget how often they use the bus.

[5 marks]

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).

-The survey should be conducted during the regular academic year, rather than the summer.
-Amongst the days of the week that the bus operates, one of each day will be randomly selected and one/two of the scheduled times will be randomly selected on each of the selected days. This will ensure that each day is represented in the sample.
- The questionnaire will be given to all the passengers while they are in line to board the bus or even better questionnaires will be given for passengers to fill out while travelling on the bus and collected at the destination. This will decrease non-response and allow for passengers to take more time/give more accurate responses.
- The questionnaire will first ask "Have you already taken this survey?" to eliminate multiple responses.

[Other answers such as the below sampling methods or other possible answers will earn marks if explanations are given.] Other sampling methods:
- Cluster sample of the days - systematic sample of passengers or all passengers sampled
-Stratified sample using the days as strata (as above) - systematic sample of passengers afterwards

[*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.

[6 marks - 1 each]

a) Identify the following for this survey:

Target Population - All (1000) books owned by the library

Sampling Frame - (List of) locations on the book shelves

Sampling Unit - One location (on a book shelf)

Observation Unit - One book

Sample - The selected 100 books

Variable - Whether or not the book needs rebinding

[2 marks]

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

Any **1** of these answers with proper explanation/discussion:

- Some of the books may be checked out or used by library patrons and hence not on the shelves. These books are likely to be more heavily used and therefore more likely to need rebinding.
- If the locations are selected with probabilities proportional to size, thicker books will have a higher probability of being selected than thin books. For example, if thin books are more likely to need rebinding then this will underestimate the true proportion of books that need rebinding. Hence the unequal probability of selection should be accounted for.
- The definition of 'needs rebinding' must be made precise to avoid subjectivity in determining whether or not a book needs rebinding. The librarian could be lazy or inaccurate in measuring this for each book.

[3 marks]

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.***

$$z_{0.025} = 1.96\,, \qquad e = 0.02\,, \qquad N = 1000$$

Use $S^{2*} = 0.25$ (i.e. $p = 0.5$ to maximize S^2 for conservative estimate of the variance)

$$n_0 = \left(\frac{1.96\,(0.5)}{0.02}\right)^2 = 2401 \quad \text{so} \quad n = \frac{n_0}{1 + {}^{n_0}/_{N}} = \frac{2401}{1 + {}^{2401}/_{1000}} \cong 705.9688$$

Sample size required is $\boxed{\boldsymbol{n = 706}}$ books at minimum.

[5 marks]

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.

Let p be the proportion of books in the library that need rebinding.

$$z_{0.025} = 1.96\,, \qquad n = 120\,, \qquad N = 1000\,, \qquad \hat{p} = \frac{30}{120} = 0.25$$

The 95% CI for p is:

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

$= \boxed{[\mathbf{17.70\%}\,,\mathbf{32.30\%}]}$

The upper bound of the CI is much less than 50% so (with 95% confidence) the library need not be concerned with problems at the upcoming inspection.

[1 mark]

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.

$$\tau = N\,\hat{p} = 1000\,(0.25) = 250$$

[2 marks]

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

There are $\binom{1000}{120}$ possible simple random samples of size 120 from this library and each has equal probability of selection. So the probability of selection is: $\frac{1}{\binom{1000}{120}}$.

[1 mark]

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

If the *i*th book is in the sample, then 119 out of the remaining 999 books from the library will be selected as the remaining units for the sample:

$\binom{999}{119}$ different samples contain the *i*th book

[10 marks - 1 each blank]

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 __51st__ unit in the population.

(c) Each restaurant in the sample represents itself + __3__ restaurants that were not sampled.

(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.)

[*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.*

First note that since y_i are binary, $y_i = {y_i}^2$.

$s^2 = \frac{1}{n-1}\ \sum_{i\,\in S}(y_i - \bar{y})^2$	starting with s^2 from formula sheet
$= \frac{1}{n-1}\ \sum_{i\,\in S}({y_i}^2 - 2\bar{y}y_i + \bar{y}^2)$	by expanding the square
$= \frac{1}{n-1}\ \sum_{i\,\in S}(y_i - 2\bar{y}y_i + \bar{y}^2)$	since $y_i = {y_i}^2$
$= \frac{1}{n-1}\ \sum_{i\,\in S}(y_i(1-2\bar{y}) + \bar{y}^2)$	by factoring
$= \frac{1}{n-1}\ (1-2\bar{y})\sum_{i\,\in S} y_i + n\bar{y}^2$	by applying the summation
$= \frac{n}{n-1}\ (1-2\bar{y})\,\bar{y} + \bar{y}^2$	using definition of $\bar{y}$
$= \frac{n}{n-1}\ (\bar{y} - \bar{y}^2)$	by simplifying
$\boldsymbol{s^2 = \frac{n}{n-1}\ \hat{p}\,(1-\hat{p}\,)}$	since $\hat{p} = \bar{y}$.