

**University of Toronto Mississauga**

**STA304H5F - Fall 2012**

**Instructor: Ramya Thinniyam**

**Term Test #1 - October 11th, 2012**

**Version 1**

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

**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. 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
- SRSWR**=Simple Random Sample With Replacement

BEST WISHES! ☺

Question	1. (/5)	2. (/15)	3. (/10)	4. (/10)	5. (/10)	TOTAL:(/50)
Marks						

**[5 marks - 1 each]**

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

- (a) A SRS is a self-weighting sample. **T** F
- (b) Sample results can be generalized to the population as long as the sample size is large. T **F**
- (c) A bank asks this on their survey: "How high would you rate our customer service?"  
This is an example of a leading question. **T** F
- (d) A 95% CI for the population proportion yields [0.78,1.00]. There is only a 5% chance  
that the true proportion is below 0.78. T **F**
- (e) The sampled population is always a subset of the target population. T **F**

**[15 marks]**

**2.** A survey is conducted to determine if test marks in STA304 is associated with major. There are 80 students in the class - 50 Statistics majors and 30 from other majors. 25 Statistics majors and 15 majors from other disciplines attend the STA304 study group this week. The investigator attends this study group meeting and randomly selects 5 of the Statistics majors and then chooses at random 3 of the other majors and asks the selected 8 students to report their test mark.

**[3 marks]**

**a)** The investigator has given every student the same chance of being selected (20% chance) so this is a Simple Random Sample. TRUE/FALSE?

*If true, write 'TRUE' and find  $P(S)$  and  $\pi_i$ . If false, write 'FALSE' and explain why the above does not meet the conditions of a SRS by showing the necessary calculations.*

**FALSE.**

$(\pi_i = 0$  if student  $i$  didn't attend study group and  $\pi_i = 0.2$  if student  $i$  attended study group - not required for the solution!)

A SRS in this case would mean that each sample of size 8 would have the same probability of selection. But,

$$P(4 \text{ Stats}, 4 \text{ Other}) = 0$$

$$P(\text{any sample with at least one person that did not attend study group}) = 0$$

$$P(\text{any 5 Stats and 3 Other that attended study group}) = \frac{\binom{25}{5} \binom{15}{3}}{\binom{40}{8}} \neq 0$$

So different samples of size 8 have different probabilities and therefore it is not a SRS.

**[4 marks]**

b) Identify the following for this survey:

**Target Population** - STA304 students

**Sampling Frame** - list of students who attend study group this week

**Observation Unit** - a student

**Sample** - the selected 8 students (5 Stats, 3 Other)

**[4 marks]**

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

Any 2 of these answers:

1) **Undercoverage:** not all students attended the study group so they did not have a chance of being selected.

2) **Selection Bias / Convenience Sampling:** students from study group were easily accessible at once to survey but they are not a representative sample of the STA304 class. Typically students attending study groups are either strong students (and want to do even better) or weak students (who want to take any extra help to improve their mark) and will also tend to be more hardworking, so marks will not be average.

3) **Non-response:** students may refuse to tell their marks (personal / sensitive information)

4) **Measurement Bias:** students may lie, estimate, forget, or round when reporting their test marks (mostly overestimating their marks)

**[2 marks]**

d) Explain sampling error in this survey.

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

**[2 marks]**

e) Briefly explain how the survey design could be improved to obtain more accurate results in order to answer the question of interest. [ie. how you would reduce non-sampling errors]

-take a SRS from the entire class (not study group) or better yet a stratified sample using the two different Majors as the strata.

- obtain test marks from portal/instructor's records/registrar's office, etc. to get accurate measurements (rather than asking students)

**[10 marks - 1 each blank]**

**3. 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 golf courses in Toronto and estimating the mean course rating. Below is some 'R' output:

```
>golfddata <- read.csv("golfsrs.csv")
>golfpopulationratings <- golfddata$rating
> length(golfpopulationratings)

[1] 120
> mean(golfpopulationratings)

[1] 70.27187
> var(golfpopulationratings)

[1] 5.851095
>golfsample1<-sample(golfpopulationratings,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(1:120,25,replace=F)
> units

[1] 17 28 60 57 66 89 25 47 104 44 108 107 109 82 27 71 97 91 68 101 87 99 96 118 55
> golfsample2 <- golfpopulationratings[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(golfsample1)

[1] 69.848
> var(golfsample1)

[1] 5.0226
> mean(golfsample2)

[1] 69.908
> var(golfsample2)

[1] 4.8916
```

- (a) The population size is 120 and the sample size is 25.
- (b) The fifth selected rating measurement for the sample is 71.2 which corresponds to the 66 th unit in the population.
- (c) The population mean is 70.2719.
- (d) The expected value of the sample mean is 70.2719 with a standard error of 0.3936.
- (e) The expected value of the sample variance is 5.8511.
- (f) An approximate 95% CI for the population mean rating is [ 69.1366 , 70.6794 ].
- (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.)

[10 marks]

4. We wish to take a sample of 2 from a population of size 10. Assume the sampling units equal the elements. Show your work and then circle the final answer for the following questions:

[3 marks]

a) If we take a SRSWR, what is the probability that both  $i$  and  $j$  are selected in the sample?  $i \neq j$

$N=10$  and  $n=2$ : there are a total of  $10^2 = 100$  samples.

$$P(\{i, j\}) = P((i, j)) + P((j, i)) = \frac{1}{100} + \frac{1}{100} = \frac{1}{50} = \boxed{0.02}$$

[3 marks]

b) In a SRS, what is the probability that both  $i$  and  $j$  are selected in the sample?

There are a total of  $\binom{10}{2} = 45$  samples.

$$P(\{i, j\}) = \frac{\binom{8}{0} \binom{2}{2}}{\binom{10}{2}} = \boxed{1/45}$$

[2 marks]

c) For a SRS, how many different samples contain the  $i$ th population unit?

If  $i$  is in the sample then one of the remaining 9 population units must be the remaining unit for the sample:

$$\binom{9}{1} = \boxed{9 \text{ different samples}}$$

[2 marks]

d) In a SRS, what is the probability that units  $i, j$ , and  $k$  are included in the sample?

Impossible to include  $i, j$ , and  $k$  since sample size is only 2:

$$P(\{i, j, k\}) = \boxed{0}$$

[10 marks]

5. A library suspects that the percentage of patrons with overdue books has increased to at least 40%. They plan to take a SRS from their 580 current patrons and record whether or not the person had at least one overdue book this year.

[3 marks]

a) At minimum, how many patrons should be sampled to estimate the percentage of interest within 5% of the true value using 95% confidence? Show your work and circle the final answer.

$$z_{0.025} = 1.96, \quad e = 0.05, \quad N = 580$$

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

$$n_0 = \left( \frac{1.96 (0.5)}{0.05} \right)^2 = 384.16$$

$$n = \frac{n_0}{1 + n_0/N} = \frac{384.16}{1 + 384.16/580} \cong 231.0953$$

Sample size required is  $\boxed{n = 232}$  at minimum.

**[3 marks]**

b) Now suppose the library takes a SRS of 120 patrons from their library, of which 93 had no overdue books.

Find a 95% CI to answer the question of interest and comment on the library's suspicion.

**Show your work and circle your final answer.**

Let  $p$  be the proportion of library patrons with overdue books.

$$z_{0.025} = 1.96, \quad n = 120, \quad N = 580, \quad \hat{p} = \frac{27}{120} = 0.225$$

The 95% CI for  $p$  is:

$$\begin{aligned} \hat{p} \pm 1.96 \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}} &= 0.225 \pm 1.96 \sqrt{\left(1 - \frac{120}{580}\right) \frac{0.225(0.775)}{119}} \\ &= \boxed{[0.1582, 0.2918]} \end{aligned}$$

The upper bound of the CI is much less than 0.4 so I would not believe the library's suspicion (with 95% confidence).

**[4 marks]**

c) Prove that for binary data like this,  $s^2$  is unbiased for  $S^2$ . **Show your work and justify steps.**

**Hint:** Look at formulas for  $E(\hat{p})$  and  $V(\hat{p})$  and recall that in this case,  $s^2 = \frac{n}{n-1} \hat{p} (1 - \hat{p})$ . (You may use the above  $s^2$  formula without proof).

Using hint, start with  $s^2 = \frac{n}{n-1} \hat{p} (1 - \hat{p})$  and use expectation and variance formulas for  $\hat{p}$ .

Show that  $E(s^2) = S^2$

$$\begin{aligned} E(s^2) &= E\left(\frac{n}{n-1} \hat{p} (1 - \hat{p})\right) \\ &= \frac{n}{n-1} E(\hat{p} - \hat{p}^2) && \text{by linearity of expectation and expanding} \\ &= \frac{n}{n-1} [E(\hat{p}) - E(\hat{p}^2)] && \text{by linearity of expectation} \\ &= \frac{n}{n-1} [E(\hat{p}) - V(\hat{p}) - (E(\hat{p}))^2] && \text{since } V(X) = E(X^2) - [E(X)]^2 \\ &= \frac{n}{n-1} \left[ p - \frac{(N-n)}{N-1} \frac{p(1-p)}{n} - p^2 \right] && \text{using formula sheet} \\ &= \frac{n}{n-1} p(1-p) \left[ 1 - \frac{(N-n)}{n(N-1)} \right] && \text{by factoring} \\ &= \frac{N}{N-1} p(1-p) && \text{by algebra/simplification} \\ &= S^2 \quad \blacksquare \end{aligned}$$

Thus,  $s^2$  is unbiased for  $S^2$ .