NAME: STUDENT NO:

STA 304/1003F Test 1 October 16, 2009 SS 2117 1.10 to 2 p.m.

Aids: One side of handwritten notes (8 1/2 x 11) and one non-programmable calculator.

There are **five** questions total, and **six pages.**
Please answer all questions **on the question paper**

1. **20 marks** "Survey Shows Pull of the U.S. Is Still Strong Inside Mexico", NY Times, September 24. From the article:

> In spite of high unemployment in the United States and strict border enforcement, one-third of Mexicans say they would move to this country if they could, and more than half of those would move even if they did not have legal immigration documents, according to a survey published Wednesday by the Pew Research Center.
>
> The United States still exerts a powerful attraction for Mexicans, the survey found, with 57 percent saying that those who leave home to settle here have better lives, while only 14 percent say life is worse in the United States.

Later in the article we read: "The survey was conducted through face-to-face interviews from May 26 to June 2 with 1,000 adults in Mexico, with a margin of sampling error of plus or minus three percentage points.", and on the Pew Research Center's report we have more information on the survey details

Sample Design: Probability
Mode: Face-to-face adults 18 plus
Languages: Spanish
Fieldwork dates: May 26 to June 2, 2009
Sample size: 1,000
Margin of error: 3 percentage points
Representative: Adult Population

(a) Describe the target population, sampling frame, sampling unit and observation unit.[4]
Target population – adult population of Mexico
sampling frame – ? not clear ? census? list of households?
sampling unit – seems to be adult residents of Mexico, not clear – could be households
observation unit – adult Mexicans

(b) Why is the margin of error equal to 3 percentage points? [4]
because $n = 1000$ *and* $1/\sqrt{n} \simeq 0.03$
OR margin of error $= \pm\sqrt{1-(n/N)}\sqrt{p(1-p)N/(n(N-1))}1.96$*; we ignore* $(N-1)/N$ *and fpc, use* $p = 1/2$

(c) Can you think of possible sources of selection bias? [4]
hard to know without more details – perhaps not all willing to respond – perhaps rural communities hard to access

(d) Questions about the possibility of illegal immigration might be viewed as sensitive. The full report gives the exact wording of all the questions in the appendix. The questions around immigration were:

> Q68 If, at this moment you had the means and opportunity to go to live in the United States, would you go?
> Q69 ASK IF RESPONDENT WANTS TO GO TO LIVE IN THE UNITED STATES: And would you be inclined to go work and live in the U.S. without authorization?

Do you think these questions are reasonably free from questionnaire bias? Explain. [4]
Yes, they don't seem like leading questions to me
[NO is okay if justified by discussion of the wording]

(e) On balance, do you think that the survey is a reliable indication of Mexican opinion on the issues questioned? [4]
Yes, – reliable source (Pew), –details provided on how survey conducted, – questions disclosed

NAME: STUDENT NO:

2. **20 marks** The following is an excerpt from an R session. The vector `marks` has the final marks of 138 students that I have taught in Applied Statistics over the past three years. There were 45, 52 and 41 students in the three years, respectively, and the marks vector is ordered by year. (The first 45 marks are for year 1, etc.).

```
> sample1 = sample(45,10,marks[1:45])  ## take a SRS of size 10 from year 2006
> sample2 = sample(52, 10, marks[46:97]) ## take a SRS of size 10 from year 2007
> sample3 = sample(41, 10, marks[98:138]) ## take a SRS of size 10 from year 2008

# c(x,y,z) joins x, y and z into a vector of length 3

> c(mean(sample1), mean(sample2), mean(sample3))

[1] 77.55708 81.1798 87.5115
> c(var(sample1), var(sample2), var(sample3))
[1] 7.940174 35.05466 8.614929
> sqrt(.Last.value/10) *sqrt(1-10/138)
# .Last.value is the output in the line just above
[1] 0.8581747 1.803057 0.8936491
> 1.96* .Last.value
[1] 1.68168 3.53388 1.75224
```

(a) Give an approximate 95% confidence interval for the class average, in each of the three years. [15]

$77.56 \pm 1.68 = (75.88, 79.24)$, $81.18 \pm 3.53 = (77.65, 84.71)$, $87.51 \pm 1.75 = (85.36, 89.26)$

NOTE: there is a mistake in the R code, the fpc should be `(1-10/45)`, `(1-10/52)`, `(1-10/41)` not `(1-10/138)`.

(b) Is there evidence that the class average increased over the three years in question? Explain. [5]

Yes, because confidence intervals for 2007 and 2008 don't overlap. Also the sample means are increasing.

NAME: STUDENT NO:

3. **20 marks** There are 15 graduate classes in statistics, with enrolments as follows:

class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
# students	3	10	45	6	2	33	4	14	24	10	8	9	12	4	40	224(total)

(a) Explain how you could select a simple random sample without replacement (SRS) from this population. [5]
NOTE: Question wasn't completely clear, should have said SRS of classes.
Use a random number table, choose a random start point, choose sets of 2 digit numbers, discard those ¿ 15
OR some similar scheme OR use a calcuator to get 5 numbers in (0,1), and multiply them by 15, OR use R or another program

(b) I selected an SRS of size 5, and the classes chosen were 1, 7, 10, 13, 15. Give a 95% confidence interval for the population total. [10]
$\bar{y} = (3 + 4 + 10 + 12 + 40)/5 = 13.8; \quad \hat{t} = 207$
$s^2 = (1/4)[(3-13.8)^2+(4-13.8)^2+(10-13.8)^2+(12-13.8)^2+(40-13.8)^2] = 229.2$
$SE(\hat{t}) = 15SE(\bar{y}) = 15 * 5.52811 = 82.92$
CI for t is $(44.475, 369.53)$

(c) Is the true value contained in your 95% confidence interval? [5]
Yes

4. **20 marks** In ratio estimation we use an auxiliary variable x whose population mean is known to improve the sample estimate of the population mean of the variable of interest y:

$$\widehat{\bar{y}}_r = \widehat{B}\bar{x}_U = \frac{\bar{y}}{\bar{x}}\bar{x}_U.$$

(a) We know that $\bar{y}$ is unbiased for $\bar{y}_U$. Explain why $\widehat{\bar{y}}_r$ is **not** unbiased for $\bar{y}_U$. [10]
$E(\widehat{\bar{y}}_r) = E\{\bar{y}\bar{x}_U/\bar{x}) = \bar{x}_U E(\bar{y}/\bar{x}) \neq \bar{x}_U E(\bar{y})/E(\bar{x}) = \bar{y}_U$ because $E(\bar{y}/\bar{x}) \neq E(\bar{y})/E(\bar{x})$

(b) If $\widehat{\bar{y}}_r$ is not unbiased, why is it sometimes used to estimate $\bar{y}_U$? [5]
Because is sometimes has smaller mean squared error.

(c) Suppose we wanted to estimate the number of families in Toronto who are planning to take a March break flight to Mexico, and we have a sample of 100 clients of a travel agency who have enquired about flights for March break. In addition, based on surveys taken at Pearson airport, we have good information on the number of people in Toronto who flew out of Pearson during March break last year. What information should we collect from our sample of size 100, and how could we use this to estimate the number of families in Toronto who are planning to take a March break flight to Mexico?
This question was not very clear, especially for a test. My answer was: assume we have t_x, the number of people who flew to Mexico last year and t_y, the number who flew anywhere last year. Then by asking our sample of 100 their destination for this year, we could multiply up using N from last year as a proxy.

5. **20 marks**

(a) Explain the difference between simple random sampling without replacement (SRS) and simple random sampling with replacement (SRSWR). [4]
in an SRS no unit can be sampled more than once

(b) Explain the difference between the sampled population and the target population. [4]
the population we want to ask is the target; often different from the sample we can ask, which is the sampled.

(c) The population of Hamilton is 504, 560 (as of 2006). Suppose we want to estimate the percentage of the population that has been immunized against swine flu, using an SRS. What should the sample size be to have a margin of error of 1%?[8]
$n_0 = z^2S^2/e^2 = 1.96^2p(1-p)/(.01^2) \simeq 1.96^2 * .25/.01^2 = 9604$
$n = n_0/(1 + n_0/N) = 9604/(1 + 9604/504560) = 9424$

(d) We can be fairly sure that the percentage that have been immunized is at most 10%. How would this change your answer to (c)? [4]
instead of $p = 0.5$ *we could put* $p = 0.1$*; this will make* n_0 *and* n *smaller (3457 and 3433, respectively).*