Stat407-607, fall 2011
Exam, 10/17/11

Name: Key

Student ID #: _____

A) You should have 5 pages (including this one)

C) Take your time, if you can't do a problem go on to the next one and come back to it.

Points for each Problem:

Multiple Choice:

1) 4

2) 4

3) 4

4) 4

5) 6

6) 6

7) 36

8) 36

Total=100

1) What was one main reason that the Literary Digest predicted Al Landon's victory over Roosevelt in the 1932 presidential election?

a) They didn't take large enough samples.

b) The people asked were different from people who were not asked.

c) People were not honest in reporting their choice.

d) Roosevelt wanted the Digest to incorrectly predict he would be defeated.

2) It was desired to estimate the average income in a town with 600 individuals. We can afford to take a simple random sample of 20 people from this population. Unknown to us, the average income is \$36,000. What can we say about our sample average of the 20 incomes we sample?

a) It is likely to be somewhat larger than \$36,000.

b) It is likely to be somewhat smaller than \$36,000.

c) It is likely to be approximately \$36,000.

d) It will be exactly \$36,000.

$\bar{X}$ unbiased but has some variability.

3) In forming a 95% confidence interval for the number of unemployed people in the U.S. we estimate $10,000,000 \pm 2,000,000$ from a sample of 400 individuals sampled randomly. In the future they want an interval which is the estimated number $\pm 1,000,000$. Approximately how many individuals do they need to sample next time?

a) 100

b) 200

c) 800

d) 1600

half width $= \frac{S}{\sqrt{n}} 1.96$, to make this half of current value, need to sample 4 times as much. I.e., $400 \times 4 = 1600$

4) A CEO wants to gauge employee satisfaction in his company. To do this he samples 10 factories out of a total of 200. Then, 30 employees are asked if they are pleased with their job from each factory. It is known that overall satisfaction tends to be higher in some factories than others. What is true about the the correct variance for a sample proportion relative to the regular (naive) variance (with $n = 300$) in this case?

a) In this case, the two should be about the same.

b) The regular variance will be too big.

c) The regular variance will be too small.

d) It is impossible to tell without more information.

Positive correlation within factory makes $n=300$ too optimistic. I.e., variance is too small

5a) For 2 strata, with population sizes 200 and 400, which of the following is true for a total sample size of n=30 under proportional allocation?

a) n1=15, n2=15.

b) n1=20, n2=10.

c) n1=10, n2=20.

d) n1=5, n2=25.

$n_1 = n\frac{N_1}{N} = 30\frac{200}{600} = 10$

So $n_2 = n - n_1 = 30 - 10 = 20$

5b) For the populations in 4a), it is estimated from a small pilot study, that the variance is twice as large in the second stratum as in the first. What are the sample sizes to get the best estimation?

a) n1=4, n2=26.

b) n1=8, n2=22.

c) n1=12, n2=18.

d) n1=15, n2=15.

$n_1 = n\frac{N_1 S_1}{N_1 S_1 + N_2 S_2} = 30\frac{200 \cdot 1}{200 \cdot 1 + 400\sqrt{2}}$

$= 30(.261) = 7.84$

$\cong 8$

6) We are interested in estimating the total number of turtles in two strata. Each stratum is broken up into 100 regions and we can afford to sample 5 regions using SRS in each stratum. The response is the number of turtles counted in the region. There are 3 possible situations we wish to consider and want a good sampling plan in each.

Let SRS denote simple random sampling, PR denotes proportional allocation, and OP denotes optimal allocation. Use the notation "=" if two methods are approximately equally variable, and "<" if the first is much less variable than the second. E.G. $OP < SRS = PR$ means that SRS and PR are approximately equally variable, but that OP is much less variable than either. Say what the relationships are for each situation. [$s_i$ =sample from stratum i],:

Situation 1: $s_1$=(1,5,9,13,17), $s_2$=(9,21,32,44,56) $OP < PR < SRS$

Optimal is best. $s_2^2 \neq s_1^2$ and $\bar{x}_1 \neq \bar{x}_2$

Situation 2: $s_1$=(1,5,9,13,17), $s_2$=(7,8,9,10,11) $OP < PR = SRS$

Proportional is no better than SRS $s_1^2 \neq s_2^2$ but $\bar{x}_1 = \bar{x}_2$

Situation 3: $s_1$=(1,5,9,13,17), $s_2$=(13,17,21,26,30) $OP = PR < SRS$

Proportional is optimal. $s_1^2 = s_2^2$ but $\bar{x}_1 \neq \bar{x}_2$

7) A group dedicated to decent treatment of pets wants to estimate the total number of dogs in the U.S. Using random digit dialling they contacted 4 houses, and received the following response: $3, 0, 2, 1$, as the number of dogs in the 4 houses. It is assumed known that there are 100,000,000 homes in the U.S.

a) Estimate the total number of dogs in the U.S.

$N = 100,000,000$

$n = 4$

$\bar{y} = \frac{6}{4} = 1.5$, so $\hat{t} = N\bar{y} = 150,000,000$ dogs

b) Give the estimated standard deviation of the estimate in a)

$Var(\bar{y}) \cong \frac{s^2}{n} = \frac{5/3}{4} = \frac{5}{12} \Rightarrow Var(\hat{t}) = N^2 \frac{5}{12}$

fpc $\cong 1$

and $\widehat{SE}(\hat{t}) = N\sqrt{\frac{5}{12}} = 100,000,000\sqrt{\frac{5}{12}}$

$= 64,550,000$ dogs

c) Give an approximate 95 % confidence interval for the true total number of dogs in the U.S. [the 95th percentile of the t-distr. with 3 d.f. is 2.35, with 4 d.f. is 2.13, the 97.5th percentile of the t-distr. with 3 d.f. is 3.08, with 4 d.f. is 2.78.]

$\hat{t} \pm 3.08(\widehat{SE}(\hat{t})) = 150,000,000 \pm 3.08(64,550,000)$

$= (-48.8 \text{ million}, 348 \text{ million})$

or practically $(0, 348 \text{ million})$

The small sample sizes make results uninformative.

d) What specific assumption justifies the interval in d)? How could you assess this assumption based on this data, or a slightly larger data set?

assumption: # dogs per home follow a normal distr.

assess: q-q plot or histogram of data

For $n=4$ we hope distr. is not too nonnormal.

e) What is the approximate probability that the estimated total is within 10% of the true unknown total? What values would we need to look up on what table to obtain the answer?

$$P\left[-.1 \le \frac{\hat{t} - T}{T} \le .1\right] = P\left[\frac{-.1T}{SE(\hat{t})} \le \frac{\hat{t} - T}{SE(\hat{t})} \le \frac{.1T}{SE(\hat{t})}\right]$$

$$\cong P\left[\frac{-.1\hat{t}}{SE(\hat{t})} \le t_3 \le \frac{.1\hat{t}}{SE(\hat{t})}\right] = P\left[\frac{-15}{64.5} \le t_3 \le \frac{15}{64.5}\right]$$

use t-table

answer: 0.16, very unlikely, SE is too big

8) In a survey to assess the damage to crop yield from pollution a sample of 3 plots from a a total of 1000 is chosen. The weights of the crop (in 100's of lbs.) in the plots are 6, 8, 13, with respective pollutant levels (in PPM) of 9, 7, 5. The average pollution level for the 1000 plots is 5PPM.

The goal is to estimate the average crop yield per plot.

a) Use the pollutant levels to help estimate the true average crop yield.

Note: Relationship between x (pollutants) & y (crop yield) is negative. So, need regression, not ratio estimation

$\hat{B}_1 = \frac{(9-7)(6-9)+(5-7)(13-9)}{(9-7)^2+(5-7)^2} = -1.75$, $\hat{B}_0 = \bar{y} - \hat{B}_1\bar{x}$

$= 9 + 1.75(7) = 21.25$

$\bar{y}_{regr} = 21.25 - 1.75(5) = 12.5$ (100s pounds)

b) Estimate the variance of the estimator in a)

$s_e^2 = \sum_{i=1}^{3}(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2$

$= [6-21.25+1.75(9)]^2 + [8-21.25+1.75(7)]^2 + [13-21.25+1.75(5)]^2$

$= 1.5$ and then $SE \cong \frac{s_e}{\sqrt{n}} = \frac{\sqrt{1.5}}{\sqrt{3}} = .707$

variance $= \frac{1}{2}$

c) Consider the usual confidence interval for the total crop weight, of the form $\hat{t} \pm 1.96SE(\hat{t})$. Give two reasons why this may not be appropriate here.

1) 1.96 from normal, need t-distr.
2) sample size is small, so no CLT is applicable

d) How many plots would be need to choose using SRS to get the same information as obtained from the estimator in a)?

$s_y^2 = \frac{(6-9)^2+(8-9)^2+(13-9)^2}{3-1} = 13$

effective sample size: $n\frac{s_y^2}{s_e^2} = \frac{3(13)}{1/2} =$ 78

using fpc $n = \frac{78}{1+\frac{78}{1000}}$

Slightly smaller