Stat 5201, April 6, 2012 Name ____________________

This test is closed book but you may use both sides of one 8 by 11 formula sheet. You may not use a calculator. It is enough to express any numerical answer as a formula which can easily be evaluated

Each part of the exam was worth 12 points except problem 2 which was worth 12. There were 51 students who took the exam. There was 1 score in the 90's, 2 in the 80's, 2-70's, 7-60's, 5-50's, 16-40's, 9-30's, 6-20's and 2 in the 10's.

1. From a population of their 300 truck drivers a company selected a random sample of 60 drivers and for each one found the number of moving violations they had in the past two years. The results are given in the table below.

number of violations	0	1	2	3
number of drivers	40	12	6	2

i) Find an estimate of the proportion of drivers who received at least one moving violation during the past two years and give an estimate of the variance of your estimate.

ii) Using the results in the table estimate the total number of moving violations for the 300 drivers and give the variance of your estimate.

iii) Suppose it is know from another source that 180 of the 300 drivers have not received any moving violations during the past two years. Use this information to find a new estimate of the total number of moving violations for the past two years.

ANS i) The estimate is 20/60 with estimated variance

$$(1-60/300)(1/59)(2/6)(1-2/6)$$

ii) The estimate is

$$N\bar{y} = 300\frac{(0\times 40 + 1\times 12 + 2\times 6 + 3\times 2)}{60} = 150$$

and since $\sum_{i\in smp} y_i = 30$ and $\sum_{i\in smp} y_i^2 = 54$ the estimated variance is

$$300^2(1-60/300)(1/60)\frac{54-(30)^2/60}{59}$$

iii) This is a domain estimation problem where the domain is the total number of drivers who received at least one moving violation and where the size of the domain is known. The new estimate is

$$120\frac{1\times 12 + 2\times 6 + 3\times 2}{20} = 180$$

with variance

$$120^2(1-20/120)(1/20)\frac{54-(30)^2/20}{19}$$

2. Consider srs **with replacement** of size n from a population of size N. For $i = 1, \ldots, N$ let W_i be the number of times that unit i appears in the sample. Then

$$\hat{t} = \frac{N}{n} \sum_{i=1}^{N} W_i y_i$$

is a possible estimator of the population total. Show that $\hat{t}$ is in fact an unbiased estimator of the population total under this sampling design.

ANS: Now W_i is a binomial$(n, 1/N)$ random variable with expectation n/N. So we have

$$E(\hat{t}) = E(\frac{N}{n} \sum_{i=1}^{N} W_i y_i) = \frac{N}{n} \sum_{i=1}^{N} E(W_i) y_i = \sum_{i=1}^{N} y_i$$

3. Consider a stratified population consisting of 3 strata where there are good prior guesses for the true strata variances. In addition the cost of sampling units from each strata is also known approximately. The information is given in the table below.

strata	1	2	3
size	300	500	200
est var	20	10	30
cost	\$4	\$12	\$8

i) Find the optimal allocation of the strata sample sizes if you can spend \$1,000.

ii) Find the minimum total cost necessary to obtain an estimate whose estimated variance will be the value v.

ANS i) Let N_h, σ_h, c_h and n_h be the size, standard deviation, cost and sample size of stratum h. Then for some constant λ the optimal allocation must satisfy

$$n_h = \lambda \frac{N_h \sigma_h}{\sqrt{c_h}}$$

where λ is found from the equation

$$1000 = \sum_h c_h n_h = \lambda \sum_h \sqrt{c_h} N_h \sigma_h$$

ii) Let $W_h = N_h/(N_1 + N_2 + N_3)$ then the minimum total cost is given by

$$c = (\sum_h W_h \sigma_h \sqrt{c_h})^2 / (v + \sum_h W_h^2 \sigma_h^2 / N_h)$$

4. An administrator was interesting in estimating the total number of prescription drugs taken by the 2,000 individuals living in the 20 nursing homes. she was managing. She selected 4 homes at random and found the total number of prescriptions for all the people in each of the 4 homes These numbers and along with the total number of patients in each home are in the table just below.

number of prescriptions	150	100	190	160
number of residents	50	75	125	30

ANS i) This is a cluster sample where the sizes of the cluster vary quite a bit so we should use the ratio estimator.

$$\hat{t} = 2000\,\hat{R} = 2000\frac{150+100+190+160}{50+75+125+30} = 2000\frac{600}{280} = 2000\frac{15}{7}$$

ii)

$$V(\hat{t})\hat{=}\frac{20^2(1-4/20)}{4\times 3}\sum_{i=1}^{4}(y_i - \frac{15}{7}M_i)^2$$

where the y_i's and M_i's are given in the table.

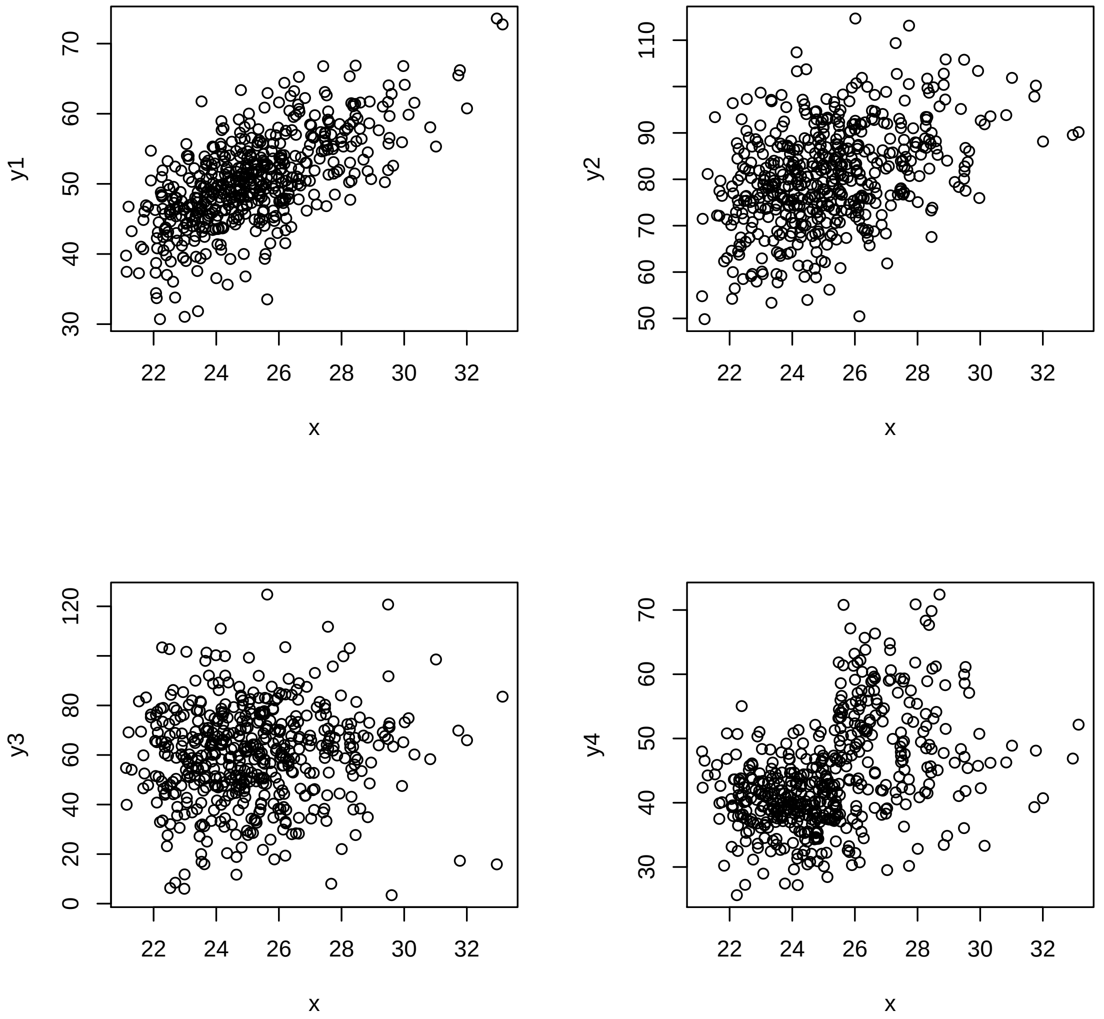
5. On the next page there are plots of four populations with the same auxiliary variable x but four different choices for the characteristic of interest y which are denoted by y_1, y_2, y_3 and y_4. For each of the four populations suggest a sampling plan and an estimator of the population total. Briefly justify your answer.

ANS i) In this case $y_i \propto x_i$ so we should use srs and the ratio estimator.

ii) In this case we should use srs and the regression estimator since the relationship is approximately linear but with a non-zero intercept.

iii) Here there is no relationship between y and x so we should use srs and the population size times the sample mean of the y's as the estimator.

iv) Here there are two strata so we should take a stratified sample and use the usual estimator. Note we should not use proportional allocation but sample with a higher rate from the strata containing the larger x values.

y1
30
40
50
60
70
22
24
26
28
30
32
x
y2
50
60
70
80
90
110
22
24
26
28
30
32
x
y3
0
20
40
60
80
120
22
24
26
28
30
32
x
y4
30
40
50
60
70
22
24
26
28
30
32
x