Remarks About Homework Problems

```
 HOMEWORK SET 1.                                    9/21/05

Ch2 #1, basic, computational

#5  > plot(9:20, c(13,35,44,69,36,24,7,3,2,5,1,1), type="h",
           ylab="# Children",main="Problem 2.5")
## Not too normal-looking because very skewed: long
## upper tail and no lower tail

> Pr2.5 <- rbind(age=9:20, nmbr=c(13,35,44,69,36,24,7,3,2,5,1,1))
> Pr2.5
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
age     9   10   11   12   13   14   15   16   17    18    19    20
nmbr   13   35   44   69   36   24    7    3    2     5     1     1
> sum(Pr2.5[1,]*Pr2.5[2,])/240
[1] 12.07917
> sum(Pr2.5[1,]*(Pr2.5[2,]-12.07917)^2)/239
[1] 309.2037
So assuming N is so large that we can ignore the finite
population correction, the SE is :
> sqrt(309.2037/240)
[1] 1.135055

    #8 similar issues as in #5

    #12 Sample size done separately in each of the 5 cities,
as a foreshadowing of stratified sampling ideas.

    #19. This problem was a little tricky because part (a)
gives a probability about a single SRS sample, (b) about
2000 such samples, being regarded as 2000 Bernoulli
(Binomial) trials with the same success probability p found
in (a) for each one, and you are asked for the probability
of NO succcesses in the 2000 trials which is (1-p)^2000.
Finally, (c) replaces 2000 by an unknown n, and we are asked
to solve (1-p)^n <= .5 (n chosen to make this as close as
possible), leading to n  the smallest integer > =
log(2)/log(1/(1-p)).
```

#20.(a) This is what multinomial means, and can be reasoned directly using combinatorial probability on the sample space of N^n equiprobable outcomes:

$$P(Q_1 = q_1, ..., Q_N = q_N) \;=\; \frac{N!}{q_1! \cdots q_N!}\, N^{-n}$$

For multinomial, it is easy to argue using indicators that $E(Q_j) = n/N$, $\mathrm{Var}(Q_j) = n(N-1)/N^2$, and that for j not equal to k, $\mathrm{Cov}(Q_j, Q_k) = -n/N^2$.

To verify the last assertions about variances and covariances with indicators, recall that the notation w_j for $j = 1, \ldots, n$ stood for the independent identically distributed with-replacement draws, which took values $i = 1, \ldots, N$, such that for each i,

$$Q_i \;=\; \sum_{j=1}^{n} I_{[w_j = i]}$$

Then, using the facts that covariances are bilinear, that covariances of independent variables are 0, and that $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$, we calculate

$$\mathrm{Cov}(Q_i, Q_k) \;=\; \sum_{j=1}^{n} \sum_{j'=1}^{n} \mathrm{Cov}(I_{[w_j=i]}, I_{[w_{j'}=k]}) \;=\; \sum_{j=1}^{n} \mathrm{Cov}(I_{[w_j=i]}, I_{[w_{j'}=k]})$$

$$= \sum_{j=1}^{n} \Big(E(I_{[w_j=i]} \cdot I_{[w_j=k]}) - E(I_{[w_j=i]} \cdot E(I_{[w_j=k]}) \Big) \;=\; \sum_{j=1}^{n} \Big(I_{[i=k]} \, \frac{1}{N} \;-\; \frac{1}{N^2} \Big)$$

and the final expression gives the variance $n(N-1)/N^2$ when $j = k$, and the covariance $-n/N^2$ when $j \neq k$.

(b)-(c). Thus $E(\hat{t}) = (N/n) \sum_{i=1}^{N} y_i \cdot (n/N) = t = N\,\overline{Y}$, and

$$\mathrm{Var}(\hat{t}) \;=\; (\frac{N^2}{n^2} \sum_{i=1}^{N} \sum_{k=1}^{N} Q_i \, Q_k \, y_i \, y_k \;=\; \frac{N^2}{n^2} \Big\{ \sum_{i=1}^{N} y_i^2 \, \frac{n(N-1)}{N^2} \;-\; \sum_{i \neq j} y_i \, y_j \, \frac{n}{N^2} \Big\}$$

$$= \; \frac{1}{n} \, (N \sum_i y_i^2 - (N\,\overline{Y})^2) \;=\; \frac{N(N-1)}{n} \, S_Y^2$$

```
   #24. The best way to rule out this sampling plan is to
understand why the inclusion probabilities are not balanced !
In particular, since the districts 51--75 have different numbers
of houses in them, numbers ranging from 525 to 1313, the
inclusion probability for each house in a sample drawn as in
(a)-(c) is proportional to 1/(# houses in the house's district)
and therefore not equal for all houses as it must be for SRS.
```

```
HOMEWORK SET 2.                                    10/11/05
```

3.2(d) While we know from general theory that the precision using regression is always at least as good versus ratio estimation with the same auxiliary variable, `farms87` in this case, that does not tell whether the ratio estimate

with acres87 or the regression estimator with farms87 gives smaller (estimated MSE). As it turns out, farms87 is much **less** closely correlated than acres87 with farms92: the estimated MSE's tell the story that ratio estimation with acres87 gives about one-tenth as large SE as regression estimation with farms87.

3.3

```
> names(agsrs.fr)
 [1] "COUNTY"   "STATE"    "ACRES92"  "ACRES87"  "ACRES82"  "FARMS92"
 [7] "FARMS87"  "FARMS82"  "LARGEF92" "LARGEF87" "LARGEF82" "SMALLF92"
[13] "SMALLF87" "SMALLF82"
> summary(agsrs$ACRES92[small])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   52150  110700  283800  350600 2234000
> summary(agsrs$FARMS92[small])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   197.5   310.0   326.4   450.5   599.0
> summary(agsrs$FARMS92[!small])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  600.0   720.0   855.0   960.5  1113.0  2760.0

## "small" contains True/False for sampled counties to be in
## the domain (ie to have < 600 farms in 92): 171 "True" values.

## estimated acres devoted to farming
(a)
>  totest1 <- sum(agsrs.fr$ACRES92[small])*3078/300
[1] 497939808
> totest2 <- sum(agsrs.fr$ACRES92[!small])*3078/300
[1] 418987302

(b)
>  SE1 <- 3078*sqrt((1-300/3078)*var(agsrs.fr$ACRES92*small)/300)
[1] 55919525
> SE2 <- 3078*sqrt((1-300/3078)*var(agsrs.fr$ACRES92*!small)/300)
[1] 38938277
```

The portions of the formula in the s_u^2 portion of SE (the things respectively inside the two "var" functions in the R code) are simply the sample variances based on the vector of 300 items consisting of the y-attributes (ACRES92) respectively changed to 0 in all counties where there were 600 or more [respectively, fewer than 600] farms.

One way to view the formula, say for the items in domain D consisting of counties with fewer than 600 farms:

```
(sum of squares of ACRES92 in the D counties minus
    300*(totest1/300)^2)/299
```

3.10(c) I suppose that the book was looking here for comments about whether the best-fitting lines to the scatter plots have a nonzero intercept (indicating regression would work better) or not (indicating ratio is OK). But we know that regression estimation always has MSE at least as good as ratio when data-samples as large; so if you said that and answered "regression" without regard to the plots, I would have accepted that answer.

3.15. From (3.6) and the definitions and approximations preceding it ,

$$V(\hat{\overline{y}}_r) \;=\; V(\overline{x}_U(\hat{B}-B)) \;\approx\; E(\overline{y}-B\overline{x})^2 \;=\; (1-\frac{n}{N})\,\frac{1}{n}(S_y^2-2BRS_xS_y+B^2S_x^2)$$

while from (3.12) and preceding definitions and approximations, **and also** (3.13) which we establish in Problem 3.17 below,

$$V(\hat{\overline{y}}_{reg}) \;=\; (1-\frac{n}{N})\,\frac{1}{n}\,S_d^2 \;=\; (1-\frac{n}{N})\,\frac{1}{n}\,S_y^2\,(1-R^2)$$

It follows immediately that (continuing the same approximations)

$$V(\hat{\overline{y}}_r - V(\hat{\overline{y}}_{reg}) \;=\; (1-\frac{n}{N})\,\frac{1}{n}\,(R^2S_y^2-2BRS_xS_y+B^2S_x^2) \;=\; (1-\frac{n}{N})\,\frac{1}{n}\,(RS_y-BS_x)^2$$

and the last term is obviously nonnegative, with value strictly positive unless $B - RS_y/S_x = B_1$.

Another way to do this, more directly, is to use the idea that the sum of squares appearing in the ratio-estimation residuals is just like the regression-residual sum of squares but with the intercept constrained to be 0. Since the regression coefficients including intercept are chosen to minimize this sum of squared residuals, the regression-residuals sum of squares must be smaller.

3.17. The first line of (3.13) is obtained by substituting the definition of d_i in

$$S_d^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y}_U-B_1(x_i-\overline{x}_U))^2$$

Next we complete the square to obtain the right-hand side equal to

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y}_U)^2 \;+\; \frac{1}{N-1}\sum_{i=1}^{N}B_1^2(x_i-\overline{x}_U)^2 \;-\; 2\,B_1\sum_{i=1}^{N}(y_i-\overline{y}_U)(x_i-\overline{x}_U)$$

$$=\; S_y^2 \;-\; 2B_1S_{xy} \;+\; B_1^2\,S_x^2$$

After substituting $S_{xy} = RS_xS_y$ and $B_1 = RS_y/S_x$, we reduce the last expression to

$$S_y^2 - 2(RS_y/S_x)(RS_xS_y) + (RS_y/S_x)^2S_x^2 \;=\; S_y^2 - 2R^2S_y^2 + R^2S_y^2 \;=\; (1-R^2)S_y^2$$

as was to be proved.

 # 4.3. The numbers N_h are $5704, 1270, 1286, 5064$, where for example $5704 = 222.81/.039$. The problem straightforwardly applies stratified estimator and SE formulas.

#4.5. (a) Here we are given exactly the information needed to apply the Neyman allocation formula (with all costsassumed equal). (b) Here we are meant to use the stratified versus unstratified SRS mean-square errors, where in the stratified case we allocate n observations proportionately to N_h, i.e., $n_h = n(N_h/N)$. Since we are given guesses about the proportions p_h using energy conservation, we can also use $S_h^2 = p_h(1-p_h)$ because the attrbitue 'uses energy conservation' is binary. (*You can use* $n = 900$ *if you want, but the answer to the problem does not actually require knowing* n*, since we take ratios of MSE's and the value of* n *cancels out.*)

#4.6(a) This is a Neyman-allocation problem since all costs are equal, with $n = 15000/30 = 500$. But since the variances are also assumed equal across strata, Neyman allocation and proportional allocation are the same, and $n_1 = 450, n_2 = 50$. (b) Now we have a case where costs vary by stratum, while variances do not. So our general formula becomes

$$n_h \;=\; (\text{Budgeted cost} - \text{fixed cost})\,(N_h/\sqrt{c_h}) \Big/ \sum_k N_k\,\sqrt{c_k}$$

Here $N_1 = .9N$, $N_2 = .1N$, where N does not have to be specified because it cancels out, and $c_1/c_2 = 10/40 = 1/4$.

```
R Script for #21 of Ch.4.
> StratSRS <- function(p1, N1frac, p=.05) {
### Note that p1 and N1frac = N1/N are inputs
### p = (N1/N)*p1+(1-N1/N)*p2 is input, and
###   then p2 is derived from the inputs.
  p2 <- (p-p1*N1frac)/(1-N1frac)
  Varh <- c(p1*(1-p1),p2*(1-p2))
  Nh <- c(N1frac,1-N1frac)  ## these are Nh/N
  nhopt <- Nh*sqrt(Varh)/sum(Nh*sqrt(Varh)) ## these are nh/n
  Vstrat <- sum(Nh^2 * Varh/nhopt) ### must still divide by n
  VSRS <- p*(1-p)
### variances ignore finite-pop correction
  list(nhopt=nhopt, Vstrat=Vstrat, VSRS=VSRS,
      VRatio= Vstrat/VSRS) }
> StratSRS(.1, .4, .058)
$nhopt
[1] 0.539684 0.460316   ## mult by 2000 to answer part (a)
$Vstrat
```

```
[1] 0.04944056          ## divide by 2000 to answer part (b)
$VSRS
[1] 0.054636
$VRatio
[1] 0.9049081

### Now we create an array of the ratios for
##     p1 = .05 to .5 in increments of .05 and
##   N1frac = .02 to .5 in increments of .02
> Vfrac <- array(0, dim=c(25,10), dimnames=list(
        seq(.02,.5,.02), seq(.05,.5,.05)))
### Note that p1*Nfrac must be < p = .05
     for(i in 1:25) for (j in 1:10)
     if (j*.05 *i*.02 < .05)
         Vfrac[i,j] <- StratSRS(j*.05,i*.02)$VRatio
> round(Vfrac,2)   ### Rows give N1/N, columns p1
     0.05  0.1 0.15  0.2 0.25  0.3 0.35  0.4 0.45  0.5
0.02    1 1.00 0.99 0.98 0.96 0.95 0.93 0.91 0.89 0.88
0.04    1 0.99 0.97 0.95 0.92 0.89 0.85 0.82 0.78 0.74
0.06    1 0.99 0.96 0.92 0.88 0.82 0.77 0.71 0.65 0.58
0.08    1 0.98 0.94 0.89 0.83 0.75 0.67 0.58 0.49 0.39
0.1     1 0.98 0.93 0.86 0.77 0.67 0.56 0.43 0.29 0.00
0.12    1 0.97 0.91 0.82 0.71 0.57 0.42 0.21 0.00 0.00
0.14    1 0.97 0.89 0.77 0.63 0.45 0.19 0.00 0.00 0.00
0.16    1 0.96 0.86 0.73 0.54 0.27 0.00 0.00 0.00 0.00
0.18    1 0.95 0.84 0.67 0.42 0.00 0.00 0.00 0.00 0.00
0.2     1 0.95 0.81 0.60 0.00 0.00 0.00 0.00 0.00 0.00
0.22    1 0.94 0.78 0.51 0.00 0.00 0.00 0.00 0.00 0.00
0.24    1 0.93 0.74 0.38 0.00 0.00 0.00 0.00 0.00 0.00
0.26    1 0.92 0.70 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.28    1 0.91 0.65 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.3     1 0.90 0.58 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.32    1 0.89 0.48 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.34    1 0.87 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.36    1 0.86 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.38    1 0.84 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.4     1 0.82 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.42    1 0.79 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.44    1 0.76 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.46    1 0.72 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.48    1 0.65 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.5     1 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
### Biggest advantage when ratio is smallest: this occurs
### at last (nonzero) entry in each column, with smallest
### value when N1/N is .12 or .14 and p1=.35 or .4
### In these cases p2 is close to 0. These are the cases
```

```
### where p1 and p2 are as different as possible subject
### to p=.05.
```

5.3 Now all $M_i = m_i = 215$. (a) $\bar{y}_s = 37/85 = .43529$ is the error rate *per claim*, with the sample variance for the number of errors per claim given by

$$s_y^2 = \left\{(4 - .4353)^2 + (3 - .4353)^2 + 4(2 - .4353)^2 + 22(1 - .4353)^2 + 57(-.4353)^2\right\}/84$$

which gives $s_y^2 = 0.55826$. From this, it follows that

$$SE = \left\{(1 - \frac{85}{828})\,\frac{1}{85}\,0.55826\right\}^{1/2} = 0.07677$$

(b) Now we can compare the per-claim error rate by multiplying a per-field error rate by 215, giving

$$SE(\hat{p}_{SRS}) = \left\{(215)^2\,(1 - \frac{85 \cdot 215}{828 \cdot 215})\,\frac{1}{85 \cdot 215}\,\frac{(85 \cdot 215)}{(85 \cdot 215 - 1)}\,\left(\frac{\bar{y}_s}{215}\right)\,\left(1 - \frac{\bar{y}_s}{215}\right)\right\}^{1/2}$$

and after some cancellations, we find that the last expression is

$$\approx \left\{(1 - \frac{85}{828})\,\frac{1}{85}\,\bar{y}_s\right\}^{1/2} = .06779$$

This confirms our general finding that a simple random sample is always at least as efficient as a cluster sample of the same overall sample size.