1. In the Exxon-Valdez oil spill trial, one important issue was the amount of lost harvest of fish, shellfish, seals, etc. suffered by native subsistence harvesters in the region. In particular, it was desired to estimate the per capita harvest loss and the total loss. Some data had been collected over the years on such harvests, both prespill and postspill. You will analyze data from just one community for one year (1989, the spill year) and estimate total harvest and per capita harvest for this community in 1989. A random sample of 42 households out of 67 was selected from the community. The number of individuals living in each household and the total household harvest (in pounds, estimated through detailed questioning about harvests of many individual species) were recorded.

The data are available on the website in the file `harvest.csv`. The resulting data frame has two variables: `size` (the number of people in household) and `harvest` (in pounds).

(a) Estimate the per capita harvest for this community. Attach a standard error to your estimate and calculate a 95 % confidence interval.

Solution:

Denote the harvest per household y_i and the size of each household M_i. Since we do not (at this point) know the total population, if we assume that the amount harvested per household is roughly linear with the amount of people in the household, we can use a ratio estimator to estimate the population per capita harvest

$$\widehat{\mu}_r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} \approx 282.4 \text{ pounds.}$$

If we denote the average household size $\overline{M}$, the linearized estimate for the variance of $\widehat{\mu}_r$ is given by

$$\text{Var}(\widehat{\tau}) = \left(\frac{N-n}{\overline{M}^2 N}\right)\frac{s_u^2}{n},$$

from which we can estimate the standard error of $\widehat{\mu}_r$ by estimating $\overline{M}^2$ with $(\frac{1}{n}\sum M_i)^2$. We obtain

$$\text{SE}(\widehat{\tau}) \approx 24.36$$

A normal based confidence interval for $\widehat{\mu}_r$ is given by $\big(234.6,\, 330.1\big)$ pounds.

□

(b) Estimate the total harvest for the community. Attach a standard error to your estimate and calculate a 95 % confidence interval.

Solution:

We estimate the total harvest with

$$\widehat{\tau} = N\overline{y} \approx 58\,109 \text{ pounds with } \text{SE}(\widehat{\tau}) = \sqrt{N(N-n)\frac{s_u^2}{n}} \approx 5\,682 \text{ pounds.}$$

A normal based confidence interval for $\widehat{\tau}$ is given by $\big(46\,972,\, 69\,246\big)$ pounds.

□

(c) Now suppose the total number of individuals in the community is known to be 190 (do not use this information in part b). Estimate total harvest using both ratio and regression estimation and calculate SE's. Compare these results with your result in (b); is there much of an improvement? Which seems more appropriate based on a plot of the data: ratio or regression estimation?

Solution:

For the ratio estimator, we merely scale the answers from part (a) by the total population.

$$\widehat{\tau}_r = Mr = M\frac{\sum y_i}{\sum M_i} \approx 53\,652 \text{ pounds}$$

$$\text{with SE}(\widehat{\tau}_r) \approx M\sqrt{\left(\frac{N-n}{\overline{m}^2 N}\right)\frac{s_u^2}{n}} \approx 4\,628 \text{ pounds}$$

where $s_r^2 = \frac{1}{n-1}\sum(y_i - rM_i)^2$. Although, since we now know the true population, M, we can estimate more precicely the standard error with

$$\text{with SE}(\widehat{\tau}_r) \approx \sqrt{N\left(\frac{N-n}{N}\right)\frac{s_u^2}{n}} \approx 5012 \text{ pounds.}$$

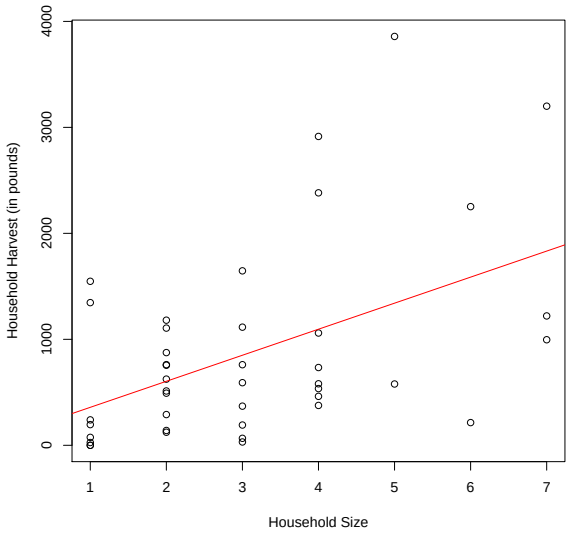
Note that this is *larger* than the estimate that does not take M into account. This may be due to the bias of that estimator.

The regression estimator is given by

$$\widehat{\tau}_{reg} = N\widehat{\mu}_{reg} = Na + Mb \approx 54\,230 \text{ pounds}$$

$$\text{with SE}(\widehat{\tau}_{reg}) \approx M\sqrt{\frac{N-n}{N}\frac{s_{reg}^2}{n}} \approx 5\,058 \text{ pounds}$$

where a and b are estimates of the linear least-squares fit for the population household size versus the population household harvest and $s_{reg}^2 = \frac{1}{n-2}\sum(y_i - a - bx_i)^2$.



It appears that there is a moderate relationship between household size and harvest, and there is an appreciable reduction in the standard errors from part b. It appears that estimating the extra parameter in the regression estimate doesn't offer a reduction in the standard error (although this may be due to the quality of the estimator for $\text{Var}(\widehat{\tau}_{reg})$). From the plot to the left, there doesn't seem to be much of an indication for estimating an intercept.

□

2. Kruuk et al. (1989) used a stratified sample to estimate the number of otter dens (called holts) along the coastline of Shetland, UK. The coastline (except for parts that were predominantly buildings) was divided into 237 5 km sections and each section was assigned to one of four terrain types. A random sample of sections within each stratum were chosen for counting. In each section chosen, researchers counted the number of otter dens in a 110 m wide strip along the coast. The data are in the file `otters.csv`. The population and sample sizes for the strata are given in the table below. Estimate the total number of otters along the coast of Shetland, along with a standard error and a 95 % confidence interval. Note: use the `tapply` command in `R` to compute the variance (or any other function) of the observations by stratum.

Stratum	Total Sections	Sections Counted
1 Cliffs over 10 m	89	19
2 Agriculture	61	20
3 Not 1 or 2, peat	40	22
4 Not 1 or 2, nonpeat	47	21

Solution:

Denote each stratum with $h = 1, 2, 3, 4$, then an estimate of the total number of holts is

$$\widehat{\tau} = \sum_{h=1}^{4} N_h \overline{y}_h \approx 985 \text{ holts with SE}(\widehat{\tau}) = \sqrt{\sum_{h=1}^{4} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}} \approx 74 \text{ holts.}$$

Using a normal approximation, a 95 % confidence interval is given by $\big(840,\ 1130\big)$. On the other hand (thanks to Grant Swicegood for pointing this out to me), if we take estimation of each s_i into account and use the Satterthwaite d.f. approximation to use a t distribution, then the confidence interval is $\big(837,\ 1132\big)$.

□

3. Using aerial photographs, a forester stratifies a 400 acre forest into three strata, and plans to sample each stratum using 0.1 acre plots to estimate the total cubic foot volume of timber. From past surveys, he has information on the range of volumes (largest minus smallest volumes likely to occur per plot) and cost in dollars for surveying a plot in each stratum. These data can be used in planning the survey and are summarized below.

Stratum	Acreage	Range	Cost (c_i)
1. Small Pines (trees 50-70 feet tall)	180	80	20
2. Large Pines (trees 70-90 feet tall)	70	120	25
3. Mixed Pine – hardwood swamp	150	200	35

Suppose he wants to estimate the total cubic foot volume on the tract with an allowable error of 20,000 cubic feet at the 95 % confidence level. (On a per plot basis, the allowable error would be 5 cubic feet.) Compute equal, proportional, optimal equal cost (Neyman), and optimal unequal cost allocations and compute the total cost of each allocation. Briefly discuss.

Solution:

Based on the Finite Central Limit Theorem, let us assume that the stratified mean estimator $\overline{y}_{st}$ is normally distributed, centered at the mean cubic foot volume per plot μ, , with variance (derived in the notes)

$$\text{Var}(\overline{y}_s t) = \sum_{h=1}^{3} \left(\frac{N_h}{N}\right)\left(\frac{N_h - nw_h}{N_h}\right)\frac{\sigma_h^2}{w_h}$$

where $N = 4000$ possibly sampled plots, and w_h is the proportion of N_h (the acreage $\times$ 10) to be sampled. We assume that within each stratum the standard deviation is given by $3\sigma_h = \frac{R_h}{2}$, where R_h is the range of volumes given (based on the fact that $P(|X| < 2\sigma) \approx .997$ for X normally distributed with standard deviation σ). Solving the margin of error equation for n where z is the .975 quantile of the standard normal distribution, we have

$$n = \frac{\displaystyle\sum_{h=1}^{3} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{w_h}}{\dfrac{d^2}{z^2} + \dfrac{1}{N}\displaystyle\sum_{h=1}^{3}\left(\frac{N_h}{N}\right)\sigma_h^2}.$$

For each statification scheme, aside from optimal unequal cost, we can substitute w_h associated with a given allocation scheme into this formula. In the case of unequal cost, the weights are given implicitly in terms of the total added cost $c^* = c - c_0 = \sum n_h c_h$. It turns out that c^* has a similar formula for the optimal n in the other cases, so after solving for c^* we obtain $n = \sum n_h = \sum c^* w_h$.

In the following table we summarize the total sample size calculation for various statified weighting schemes (see notes).

Scheme	w_h	n_1	n_1	n_1	n	c
Equal Weights	$w_h = \dfrac{1}{L}$	31	31	31	93	\$2480
Proportional Weights	$w_h = \dfrac{N_h}{N}$	38	15	32	85	\$2255
Optimal Equal Cost	$w_h = \dfrac{N_h\sigma_h}{\sum N_k\sigma_k}$	20	12	41	73	\$2135
Optimal Unequal Cost	${}^*w_h = \dfrac{c^*}{n}\dfrac{N_h\sigma_h/\sqrt{c_h}}{\sum N_k\sigma_k/\sqrt{c_k}}$	24	12	38	74	\$2110

Note that as more information is included in the analysis, the total cost of the study goes down. For the optimal variance allocation, note that the mixed pine is sampled more heavily due to a wider range, but when cost constraints are taken into account less of them are sampled in favor of the cheaper strata.

□

4. A nursery manager wants to estimate the average height of seedlings in a large field that is divided into 25 plots that vary slightly in size. She believes the heights are fairly constant throughout each plot, but may vary considerable from plot to plot. Although this seems to indicate the need for a stratified random sample, for time and logistical reasons, she decides to use a two-stage sample, where with the low variability within a plot, she decides to sample 10 % of the trees within each of 5 plots. The data are given in the table below. Estimate the average height of seedlings in the field, give the standard error, and find an approximate 95 % confidence interval for the mean. [Problem taken from page 304, Scheaffer, Mendenhall, & Ott.]

Plot	Number of Seedlings	Number of Seedlings Sampled	Heights of Seedlings (in inches)					
1	52	5	12	11	11	10	13	
2	56	6	10	9	7	8	8	10
3	60	6	6	5	7	5	6	4
4	46	5	7	8	6	7	6	
5	49	5	10	11	13	12	12	

Solution:

In this two-stage sample, we do not know the total number of seedlings, hence we cannot use the unbiased estimator for the average height per seedling. If we denote y_i as the total height of seedlings in the ith group of $n = 5$, we can estimate the average height per seedling with the ratio of an estimate of the average height per plot with an estimate of the average number of seedlings per plot. I.e.

$$\widehat{\mu}_r = \frac{\widehat{\mu}_1}{\overline{m}} = \frac{\widehat{y}_i/n}{\frac{1}{n}\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \overline{y}_i}{\sum_{i=1}^{n} M_i} \approx 8.705 \text{ inches.}$$

Denote N as the total number of plots, then the linearized variance is

$$\text{Var}(\widehat{\mu}_r) \approx \left(1 - \frac{n}{N}\right) \frac{1}{\overline{M}^2 n} \left(\frac{1}{N-1} \sum_{i=1}^{N} (y_i - M_i \widehat{\mu})^2 \right) + \frac{1}{N\overline{M}^2 n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{\sigma_i^2}{m_i},$$

from which we can estimate a standard error by estimating y_i with $\widehat{y}_i$, $\overline{M}$ with $\overline{m} = \frac{1}{n}\sum M_i$, μ with $\widehat{\mu}_r$ and σ_i^2 with $s_i^2 = \frac{1}{n-1}\sum (y_{ij} - \overline{y}_i)^2$. The result is

$$\begin{aligned} \text{SE}(\widehat{\mu}_r) &\approx \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{\overline{M}^2 n} \left(\frac{1}{n-1} \sum_{i=1}^{n} (y_i - M_i \widehat{\mu})^2 \right) + \frac{1}{N\overline{M}^2 n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{\sigma_i^2}{m_i}} \\ &\approx \sqrt{1.108362 + 0.00792} \\ &\approx 1.11 \text{ inches.} \end{aligned}$$

(Continued on next page)

Note that the computation is dominated by the first "between-group" variance term. This is not surprising due to the small number of primary units sampled $n = 5$. Due to the small sample size, we use a conservative 95 % confidence interval based on a t distribution with 4 degrees of freedom: $\big(5.63,\ 11.78\big)$. □

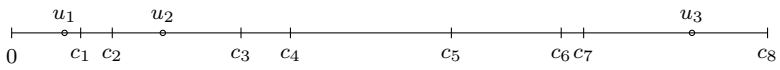
5. An auditor wishes to sample sick-leave records of a large firm in order to estimate the average number of days of sick leave per employee over the past quarter. The firm has eight divisions, with varying numbers of employees per division. Since the number of days of sick leave used within each division should be highly correlated with the number of employees, the auditor decides to sample $n = 3$ divisions with probabilities proportional to the number of employees. The data are given in the table below.

Division	1	2	3	4	5	6	7	8
# of Employees	1200	550	2240	860	2800	1910	390	3200

(a) Explain how, using a random number table or generator, such a sample could be taken. You do not need to do it – just explain it.

Solution:

Let p_i be the proportion of employees in the ith division. If we let D be the number of the division choosen so that the probability of selection is equal to p_i, then the cummulative probability function for D is $c_n := P(D \leq n) = \sum_{i=1}^{n} p_i$. Assuming we have u_1, u_2 and u_3 realizations from a uniform distribution over $[0, 1]$, select i such that $c_{n-1} < u_i \leq c_n$ for $i = 1, 2, 3$. Using a theorem from probability, $n(i)$ are realizations from the distribution of D. Pictorally,



□

(b) Suppose now that such a sample is taken where divisions 1, 3, and 8 are selected, and that the total number of sick days used by the three sampled divisions during the past quarter are, respectively,

$$y_1 = 2410, \quad y_2 = 4320, \quad y_3 = 5790.$$

Estimate the average number of sick days used per person for the entire firm, and give the corresponding standard error.

Solution:

Denote the division size M_i, and y_i the number of sick days, we can estimate using the Hansen-Hurwitz estimator:

$$\widehat{\mu}_p = \frac{1}{M}\widehat{\tau}_p = \frac{1}{Mn}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{M_i} = \frac{1}{3}\left\{\frac{2410}{1200} + \frac{4320}{2240} + \frac{5790}{3200}\right\} \approx 1.92 \text{ hours.}$$

We estimate the standard error with $(\overline{y}_i = y_i/M_i)$

$$\mathrm{SE}(\widehat{\mu}_p) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\overline{y}_i - \widehat{\mu}_p)^2} \approx 0.0578 \text{ hours.}$$

□

6. Show that in two-stage sampling with equal-sized primary units, that if m secondary units will be sampled within each primary unit, then the number of primary units n required to estimate the population mean per secondary unit to within d with probability $100(1-\alpha)\,\%$ is given as

$$n = \frac{\sigma_b^2 + \left(\frac{\overline{M} - m}{\overline{M}}\right)\frac{\sigma_w^2}{m}}{\frac{d^2}{z^2} + \frac{\sigma_b^2}{N}}.$$

Solution:

It was shown in the notes that the variance of this estimator for the population mean per secondary unit was given as

$$\mathrm{Var}(\widehat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_b^2}{n} + \left(\frac{\overline{M} - m}{\overline{M}}\right)\frac{\sigma^2 w}{m \cdot n}.$$

Based on the Finite Population Central Limit Theorem, if we assume that $\widehat{\mu}$ is normally distributed, then the margin of error, say d, for a $100(1-\alpha)\%$ is given by

$$d = z\sqrt{\mathrm{Var}(\widehat{\mu})}$$

where z is the $1 - \frac{\alpha}{2}$ quantile for the standard normal distribution. Then,

$$\begin{aligned}\frac{d^2}{z^2} &= \frac{\sigma_b^2}{n} - \frac{\sigma_b^2}{N} + \frac{1}{n}\left(\frac{\overline{M} - m}{\overline{M}}\right)\frac{\sigma^2 w}{m} \\ &= \frac{1}{n}\left(\sigma_b^2 + \left(\frac{\overline{M} - m}{\overline{M}}\right)\frac{\sigma^2 w}{m}\right) - \frac{\sigma_b^2}{N}\end{aligned}$$

so

$$n = \left(\frac{d^2}{z^2} + \frac{\sigma_b^2}{N}\right)^{-1}\left(\sigma_b^2 + \left(\frac{\overline{M} - m}{\overline{M}}\right)\frac{\sigma^2 w}{m}\right)$$

□

7. Use conditional expectations to solve the following problems.

(a) Derive the variance of a geometric random variable (we derived the expected value by conditioning in class).

Let X be a geometrically distributed random variable with probability p. That is, for $T_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, X is the first i such that $T_i = 1$. If we condition on T_1, then

$$\text{Var}(X|T_1 = 0) = \text{Var}(X + 1) = \text{Var}(X)$$

since X "has no memory" of the first trial. On the other hand,

$$\text{Var}(X|T_1 = 1) = 0$$

since $(X|T_1 = 1) = 1$ is completely determined. Now, thinking of $g(t) = \text{Var}(X|T_1 = t)$, we can calculate

$$E\Big[\text{Var}(X|T_1)\Big] = E(g(T_1)) = (1-p)\text{Var}(X) + \cancel{p \cdot 0}.$$

Thinking along similar lines, $E(X|T_1)$ is a random variable whose mean is $E(X) = 1/p$ that takes on $1/p + 1$ with probabilty $1 - p$ and 1 with probability p (this was done in class). Hence

$$\text{Var}\Big[E(X|T_1)\Big] = (1-p)(1/p + 1 - 1/p)^2 + p(1 - 1/p)^2 = (1-p) + p\frac{(1-p)^2}{p^2} = \frac{1-p}{p}.$$

The "EVVE" law for conditional variance gives

$$\text{Var}(X) = E\Big[\text{Var}(X|T_1)\Big] + \text{Var}\Big[E(X|T_1)\Big] = (1-p)\text{Var}(X) + \frac{1-p}{p} \iff \text{Var}(X) = \frac{1-p}{p^2}.$$

(b) What is the expected number of rolls of a fair die until you have obtained two *consecutive* 6's?

Let X be the number of rolls until two consecutive sixes are rolled. If we denote event of obtaining a six on the ith roll T_i, then it can be thought of as a Bernoulli trial with probability $\frac{1}{6}$. Conditioning on T_1, we have

$$E(X|T_1 = 0) = E(X + 1) = E(X) + 1$$

To evaluate $E(X|T_1 = 1)$, we further condition on T_2 so that

$$E(X|T_1 = 1,\, T_2 = 0) = E(X + 2) = E(X) + 2$$

and

$$E(X|T_1 = 1,\, T_2 = 1) = 2$$

since that event is completely determined to be 2. Now, thinking of $g(t) = E(X|T_1 = 1,\, T_2 = t)$, we can calcuate the expectation

$$E(X|T_1 = 1) = E\Big[E(X|T_1 = 1,\, T_2)\Big] = \frac{1}{6} \cdot 2 + \frac{5}{6}(E(X) + 2) = 2 + \frac{5}{6}E(X).$$

Using similar reasoning,

$$E(X) = E\Big[E(X|T_1)\Big] = \frac{5}{6}(E(X) + 1) + \frac{1}{6}\left(\frac{5}{6}E(X) + 2\right) \iff E(X) = 42.$$

```
> source('problem1.r',echo=T)
> ###################################
> # Problem 1
> ###################################
> d = read.csv('harvest.csv')
> N = 67
> M = 190
> n = nrow(d)

> #(muhat = N/M*mean(d$harvest))
> #(se_muhat = 1/M*sqrt(N*(N-n)*var(d$harvest)/n))
>
> # part (a)
> (murhat = sum(d$harvest)/sum(d$size))
[1] 282.377

> sr2 = sum((d$harvest - murhat*d$size)^2)/(n-1)
> (se_murhat = sqrt( (1-n/N)/mean(d$size)^2 * sr2/n ))
[1] 24.35694

> (murhat_ci = murhat + c(-1,1)*qnorm(.975)*se_murhat)
[1] 234.6382 330.1157

> # part (b)
> (tauhat = N*mean(d$harvest))
[1] 58109.14

> (se_tauhat = sqrt(N*(N-n)*var(d$harvest)/n))
[1] 5682.249

> (tauhat_ci = tauhat + c(-1,1)*qnorm(.975)*se_tauhat)
[1] 46972.14 69246.15

> # part (c)
> plot(d$size,d$harvest,xlab="Household Size",ylab="Household Harvest (in pounds)"
    )
> out = lm(d$harvest ~ d$size)
> abline(out,col="red")

> (taurhat = M*murhat)
[1] 53651.62

> (se_taurhat = M*se_murhat)
[1] 4627.819

> (se2_taurhat = sqrt(N^2*(1-n/N) * sr2/n ))
[1] 5012.311

> (taureghat = N*out$coefficients[1] + M*out$coefficients[2])
(Intercept)
   54230.49

> sreg2 = sum(out$residuals^2)/(n-2)
> (se_taureghat = N*sqrt((1-n/N) * sreg2/n))
[1] 5058.067

> source('problem2.r',echo=T)
> ######################################
```

```
> # Problem 2
> ######################################
> otters = read.csv('otters.csv')
> nh = table(otters$Habitat)
> Nh = c('1'=89,'2'=61,'3'=40,'4'=47)
> ybarh = tapply(otters$Holts,otters$Habitat,mean)
> sh2 = tapply(otters$Holts,otters$Habitat,var)
> (tauhat = sum(Nh*ybarh))
[1] 984.7142

> (se_tauhat = sqrt(sum(Nh^2*(1-nh/Nh)*sh2/nh)))
[1] 73.92099

> (tauhat_ci = tauhat + c(-1,1)*se_tauhat*qnorm(.975))
[1]  839.8317 1129.5967

> ah = Nh*(Nh-nh)/nh
> satterthwaite_df = sum(ah*sh2)^2/sum((ah*sh2)^2/nh-1)
> (tauhat_sat_ci = tauhat + c(-1,1)*se_tauhat*qt(.975,satterthwaite_df))
[1]  837.3809 1132.0476

> source('problem3.r',echo=T)
> ######################################
> # Problem 3
> ######################################
> N = 4000
> Nh = c(1800,700,1500)
> Rh = c(80,120,200)
> ch = c(20,25,35)
> sigh = (Rh/6) # Using P(|X|< 3 sig) = .997
> d = 5
> total_n = function(wh) { # Use the general equation and just input wh
+   sum( (Nh*sigh)^2/wh ) / ( (N*d/qnorm(.975))^2 + sum(Nh*sigh^2) )
+ }

> equal_wh = rep(1/3,3)
> equal_n = total_n(equal_wh)
> equal_nh = equal_n * equal_wh
> prop_wh = Nh/N
> prop_n = total_n(prop_wh)

> prop_nh = prop_n * prop_wh
> equalcost_wh = Nh * sigh / sum( Nh * sigh )
> equalcost_n = total_n(equalcost_wh)
> equalcost_nh = equalcost_n * equalcost_wh
> unequalcost_wh = Nh*sigh/sqrt(ch) / sum( Nh*sigh * sqrt(ch) )
> cstar = total_n(unequalcost_wh)
> unequalcost_nh = cstar*unequalcost_wh
> unequalcost_n = sum(unequalcost_nh)
> (t(data.frame(
+ 'equal'      = c( round(equal_nh)      , sum(round(equal_nh))     , sum(round(
    equal_nh     )*ch)),
+ 'prop'       = c( round(prop_n .... [TRUNCATED]

            [,1] [,2] [,3] [,4] [,5]
equal         31   31   31   93 2480
prop          38   15   32   85 2255
equalcost     20   12   41   73 2135
unequalcost   24   12   38   74 2110
```

```
> source('problem4.r',echo=T)
> ####################################
> # Problem 4
> ####################################
> seedlings = data.frame(
'plot' = c(rep(1,5),rep(2,6),rep(3,6),rep(4,5),rep(5,5)),
'height' = c( 12 , 11 , 11 , 10 , 13 ,
              10 ,  9 ,  7 ,  8 ,  8 , 10,
               6 ,  5 ,  7 ,  5 ,  6 ,  4,
               7 ,  8 ,  6 ,  7 ,  6 ,
              10 , 11 , 13 , 12 , 12 ))

> Mi = c(52,56,60,46,49)
> mi = table(seedlings$plot)
> hatyi = Mi*tapply(seedlings$height,seedlings$plot,mean)
> (mur = sum(hatyi) / sum(Mi))
[1] 8.704689

> n = 5
> mbar = mean(Mi)
> N = 25
> si2 = tapply(seedlings$height,seedlings$plot,var)
> (varbetween = (1 - n/N) * 1/(mbar^2 * n) * 1/(n -1) * sum((hatyi - Mi*mur)^2) )
[1] 1.220548

> (varwithin = 1/(N*mbar^2*n) *sum(Mi*(Mi - mi)*si2/mi))
[1] 0.007918242

> (se_mur = sqrt(varbetween + varwithin))
[1] 1.108362

> (mur_ci = mur + c(-1,1)*se_mur*qt(.975,n-1))
[1]  5.627383 11.781995

> source('problem5.r',echo=T)
> ##################################
> # Problem 5
> ##################################
> m = c(1200,550,2240,860,2800,1910,390,3200)
> p = m/sum(m)
> y = c(2410,4320,5790)
> i = c(1,3,8)
> n = 3
> (muhatp = mean(y/m[i]))
[1] 1.915427

> (se_muhatp = sqrt( var(y/m[i])/n ))
[1] 0.05780915
```