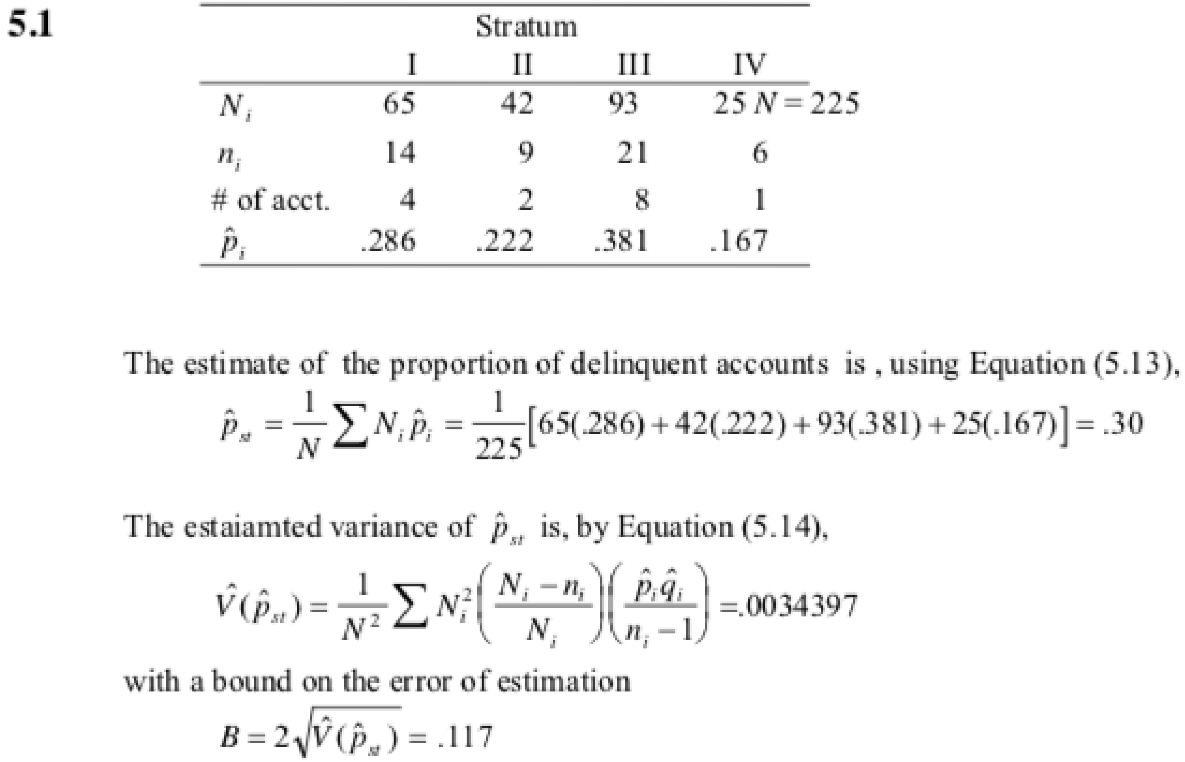
# E2 review key

Stat 422

(1) 5.1



**5.1**

|  | Stratum |  |  |  |
|---|---|---|---|---|
|  | I | II | III | IV |
| $N_i$ | 65 | 42 | 93 | 25 $N = 225$ |
| $n_i$ | 14 | 9 | 21 | 6 |
| # of acct. | 4 | 2 | 8 | 1 |
| $\hat{p}_i$ | .286 | .222 | .381 | .167 |

The estimate of the proportion of delinquent accounts is , using Equation (5.13),

$$\hat{p}_{st} = \frac{1}{N}\sum N_i \hat{p}_i = \frac{1}{225}\left[65(.286) + 42(.222) + 93(.381) + 25(.167)\right] = .30$$

The estaiamted variance of $\hat{p}_{st}$ is, by Equation (5.14),

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{p}_i\hat{q}_i}{n_i - 1}\right) = .0034397$$

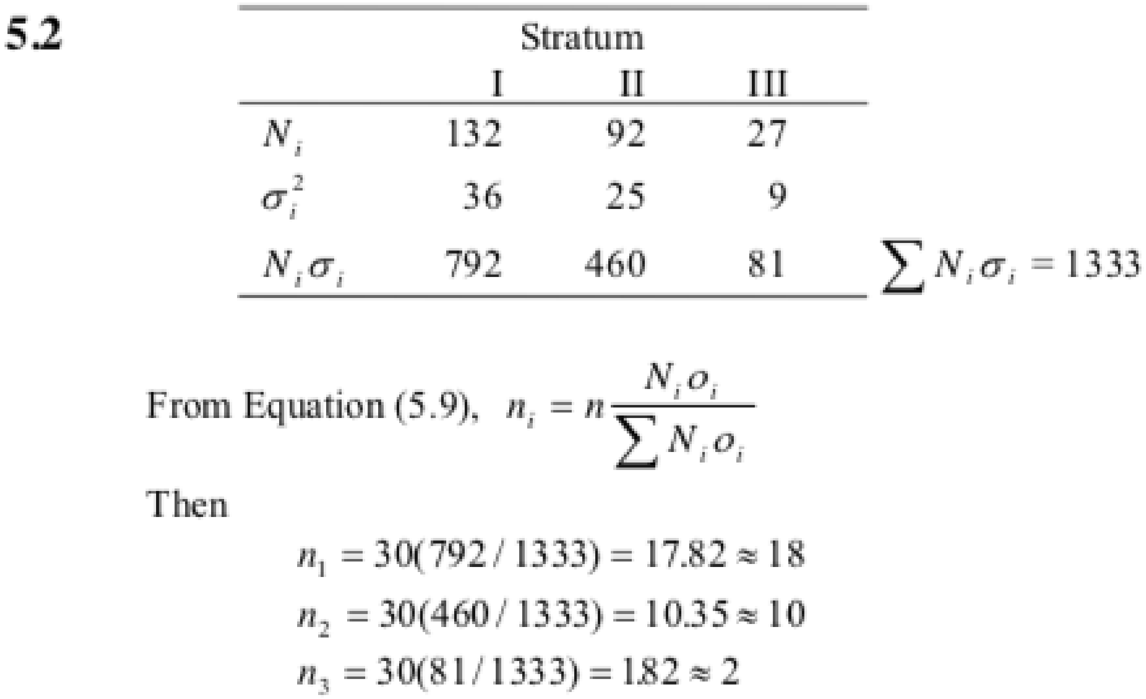
with a bound on the error of estimation

$$B = 2\sqrt{\hat{V}(\hat{p}_{st})} = .117$$

Figure 1:

(2) 5.2

(3) 5.27



**5.2**

| | Stratum | | | |
|---|---|---|---|---|
| | I | II | III | |
| $N_i$ | 132 | 92 | 27 | |
| $\sigma_i^2$ | 36 | 25 | 9 | |
| $N_i\sigma_i$ | 792 | 460 | 81 | $\sum N_i\sigma_i = 1333$ |

From Equation (5.9), $n_i = n\dfrac{N_i o_i}{\sum N_i o_i}$

Then

$$n_1 = 30(792/1333) = 17.82 \approx 18$$

$$n_2 = 30(460/1333) = 10.35 \approx 10$$

$$n_3 = 30(81/1333) = 1.82 \approx 2$$

Figure 2:

**5.27** **(a)**

| | Stratum | | Total |
|---|---|---|---|
| | I | II | |
| $N_i$ | 20 | 26 | |
| $\sigma_i$ | 25 | 47.5 | |
| $N_i\sigma_i$ | 500 | 1235 | 1735 |
| $w_i$ | .29 | .71 | |
| $N_i\sigma_i^2$ | 12500 | 58662.5 | 71162.5 |

where

$\frac{\text{range}}{4}$ is used to estimate $\sigma_i$.

$\sigma_1 \approx \frac{100-0}{4} = 25$ for small plants (stratum I)

$\sigma_2 \approx \frac{200-10}{4} = 47.5$ for large plants (stratum II)

$$a_i = \frac{N_i\sigma_i}{\sum N_i\sigma_i}$$

$$a_1 = \frac{500}{1735} = .29 \qquad a_2 = \frac{1235}{1735} = .71$$

**(b)** $B = 100$

$$N^2D = \frac{B^2}{4} = \frac{(100)^2}{4} = 2500$$

$$n = \frac{\left(\sum N_i\sigma_i\right)^2}{N^2D + \sum N_i\sigma_i^2} = \frac{(1735)^2}{2500 + 71162.5} = 40.87 \approx 41$$

$n_1 = na_1 = 41(.29) = 11.9 \approx 12$

$n_2 = na_2 = 41(.71) = 29.1 \approx 29$

Since there is only 26 “large” plants, we allocate $n_1 = 15$, $n_2 = 26$.

(4) 5.35

**5.35** (a) This analysis requires the use of the Chapter 4 method for estimating a mean and finding a two-standard deviation margin of error. The method is used four times, once for each region, to produce the results in the table below. This shows, for example, that the plausible values for the true mean farm acreage per county in the South is 140.9±56.5, or (84.4, 197.4) thousands of acres.

| | n | Sample Mean | Sample Standard Deviation | Standard Deviation of Mean | Margin of Error |
|---|---|---|---|---|---|
| S:ACRES | 22 | 140.9 | 133.6 | 28.257 | 56.514 |
| W:ACRES | 22 | 726.0 | 518.0 | 107.519 | 215.038 |
| NC:ACRES | 22 | 410.2 | 375.2 | 79.155 | 158.310 |
| NE:ACRES | 22 | 75.7 | 63.8 | 12.865 | 25.729 |

(b) The estimated total farm acreage for a region is the sample mean per county times the number of counties in the region. The margin of error for the estimated total is the margin of error for the mean times the population size. The results are given below. This shows that the estimate of the total farm acreage in the South is 193,867±77,763 thousands of acres. Notice that the margins of error are large compared to the estimated totals. This is due to the large amount of variation from county to county and the small sample size.

| | N | Sample Mean | Estimated Total | Margin of Error |
|---|---|---|---|---|
| S:ACRES | 1376 | 140.9 | 193,867 | 77,763 |
| W:ACRES | 418 | 726.0 | 303,269 | 89,886 |
| NC:ACRES | 1052 | 410.2 | 431,546 | 166,542 |
| NE:ACRES | 210 | 75.7 | 15,891 | 5,403 |

**(c)** The estimate of the difference in population means is the difference in sample means, and the estimated variance of that difference is the sum of the variances of the parts. Comparing North Central to South, the result is

$$(410.2 - 140.9) \pm 2\sqrt{79.155^2 + 28.257^2} \text{ or}$$
$$269.3 \pm 168.1$$

The North Central counties average at least 128.2 thousand acres more than the counties of the South in farm acreage.

**(d)** Using the same reasoning as in part (c), the estimated difference in mean farm acreage per county when comparing the West with the North East is

$$(726.0 - 75.7) \pm 2\sqrt{107.519^2 + 12.865^2} \text{ or}$$
$$650.3 \pm 216.6$$

As might be expected, the counties of the West have much greater mean farm acreage, but the margin of error with these data is quite large.

**(e)** This is a stratified random sample. The estimated mean acreage per county for the four regions together is given by

$$\overline{y}_{st} = \frac{1}{N}\left[N_1\overline{y}_1 + N_2\overline{y}_2 + N_3\overline{y}_3 + N_4\overline{y}_4\right] = 309.1$$

The variance of this estimate is given by

$$\hat{V}(\overline{y}_{st}) = \frac{1}{N^2}\left[N_1^2\,\hat{V}(\overline{y}_1) + N_2^2\,\hat{V}(\overline{y}_2) + N_3^2\,\hat{V}(\overline{y}_3) + N_4^2\,\hat{V}(\overline{y}_4)\right] = 1121.41$$

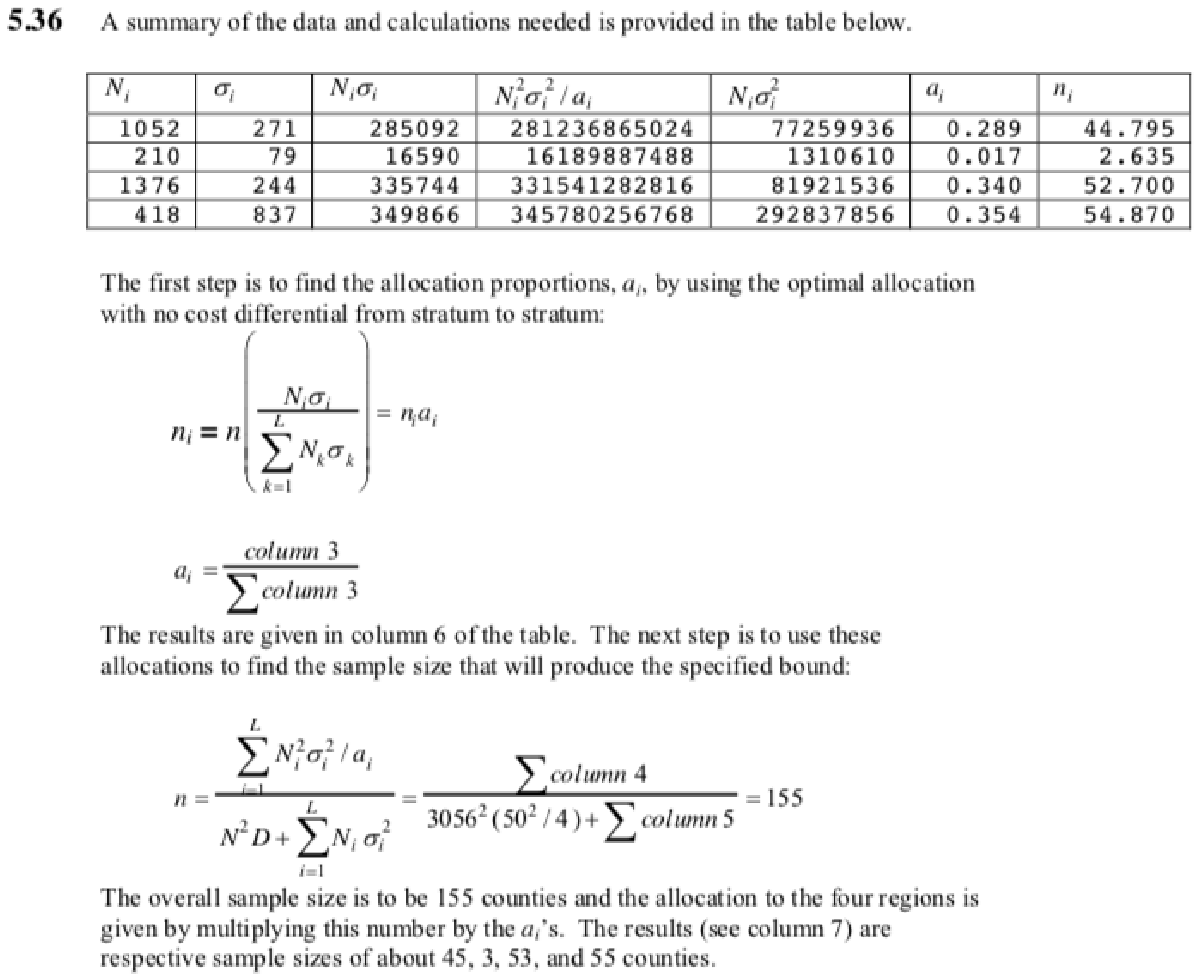
The margin of error is then

$$2\sqrt{\hat{V}(\overline{y}_{st})} = 2\sqrt{1121.41 = 67.0}$$

The plausible values for the mean farm acreage per county across the United States are those in the interval 309±70 or (239, 379) thousand acres. Notice that his margin of error is smaller than three out of four of the individual margins of error for the regions.

(5) 5.36

(6) 6.1



**5.36** A summary of the data and calculations needed is provided in the table below.

| $N_i$ | $\sigma_i$ | $N_i\sigma_i$ | $N_i^2\sigma_i^2/a_i$ | $N_i\sigma_i^2$ | $a_i$ | $n_i$ |
|---|---|---|---|---|---|---|
| 1052 | 271 | 285092 | 281236865024 | 77259936 | 0.289 | 44.795 |
| 210 | 79 | 16590 | 16189887488 | 1310610 | 0.017 | 2.635 |
| 1376 | 244 | 335744 | 331541282816 | 81921536 | 0.340 | 52.700 |
| 418 | 837 | 349866 | 345780256768 | 292837856 | 0.354 | 54.870 |

The first step is to find the allocation proportions, $a_i$, by using the optimal allocation with no cost differential from stratum to stratum:

$$n_i = n\left(\frac{N_i\sigma_i}{\sum_{k=1}^{L} N_k\sigma_k}\right) = n_i a_i$$

$$a_i = \frac{column\ 3}{\sum column\ 3}$$

The results are given in column 6 of the table. The next step is to use these allocations to find the sample size that will produce the specified bound:

$$n = \frac{\sum_{i=1}^{L} N_i^2\sigma_i^2/a_i}{N^2D + \sum_{i=1}^{L} N_i\sigma_i^2} = \frac{\sum column\ 4}{3056^2(50^2/4) + \sum column\ 5} = 155$$

The overall sample size is to be 155 counties and the allocation to the four regions is given by multiplying this number by the $a_i$'s. The results (see column 7) are respective sample sizes of about 45, 3, 53, and 55 counties.

Figure 3:

**6.1** A scatter plot of the data shows evidence of a positive linear association (correlation) between $y$ and $x$, which is a good for ratio estimation. The following data table gives ($y_i$ - $rx_i$) column along with $x_i$ and $y_i$ column, where

$$r = \frac{\sum y_i}{\sum x_i} = \frac{142}{6.7} = 21.194$$

An estimate of $\tau_y$ is, using Equation (6.4),

$$\hat{\tau}_y = r\tau_x = 21.194(75) = 1589.55$$

The standard deviation $s_r$ is simply the sample standard deviation of the values for ($y_i$ - $rx_i$). Then the estimated variance of $\hat{\tau}_y$ is, from Equation (6.5),

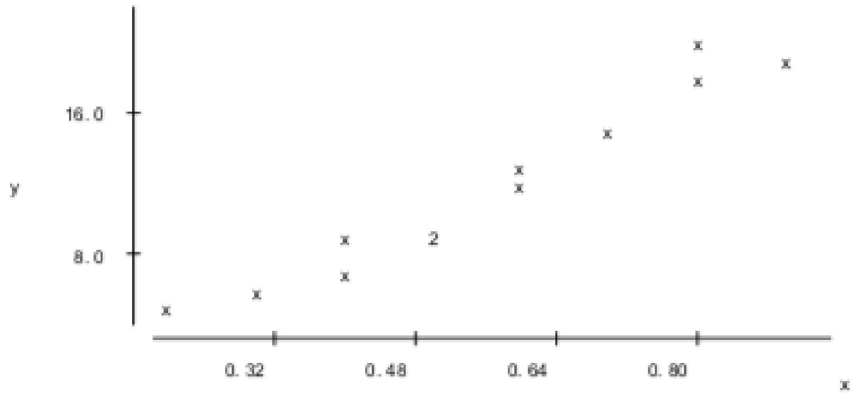
$$\hat{V}(\hat{\tau}_y) = \tau_x^2\left(\frac{N-n}{nN}\right)\left(\frac{1}{\mu_x^2}\right)s_r^2 = (N\mu_x)^2\left(\frac{N-n}{nN}\right)\left(\frac{1}{\mu_x^2}\right)s_r^2 = N\left(\frac{N-n}{n}\right)s_r^2$$

$$B = 2\sqrt{\hat{V}(\hat{\tau}_y)} = 2\sqrt{\frac{250(250-12)}{12}(1.323)} = 186.32$$

Data summary for Exercise 6.1

| | $n$ | sum | st dev |
|---|---|---|---|
| $x_i$ | 12 | 6.7 | 0.2151 |
| $y_i$ | 12 | 142 | 5.1845 |
| $y_i$ - $rx_i$ | 12 | 0 | 1.323 |

Scatter plot of volume versus basal area:



(7) 6.33 (assume $N = 10000$ and also include bound with estimations of means)

(a) ratio estimator of mean
(b) regression estimation of mean
(c) difference estimation of mean
(d) estimate $n$ of ratio estimator of mean with $B = 35$ *lbs*
(e) calculate REs for:
   (i) ratio vs. regression
   (ii) ratio vs. difference
   (iii) regression vs. difference
   (iv) srs to ratio
   (v) which one looks 'better'?

```
head(gators)
```

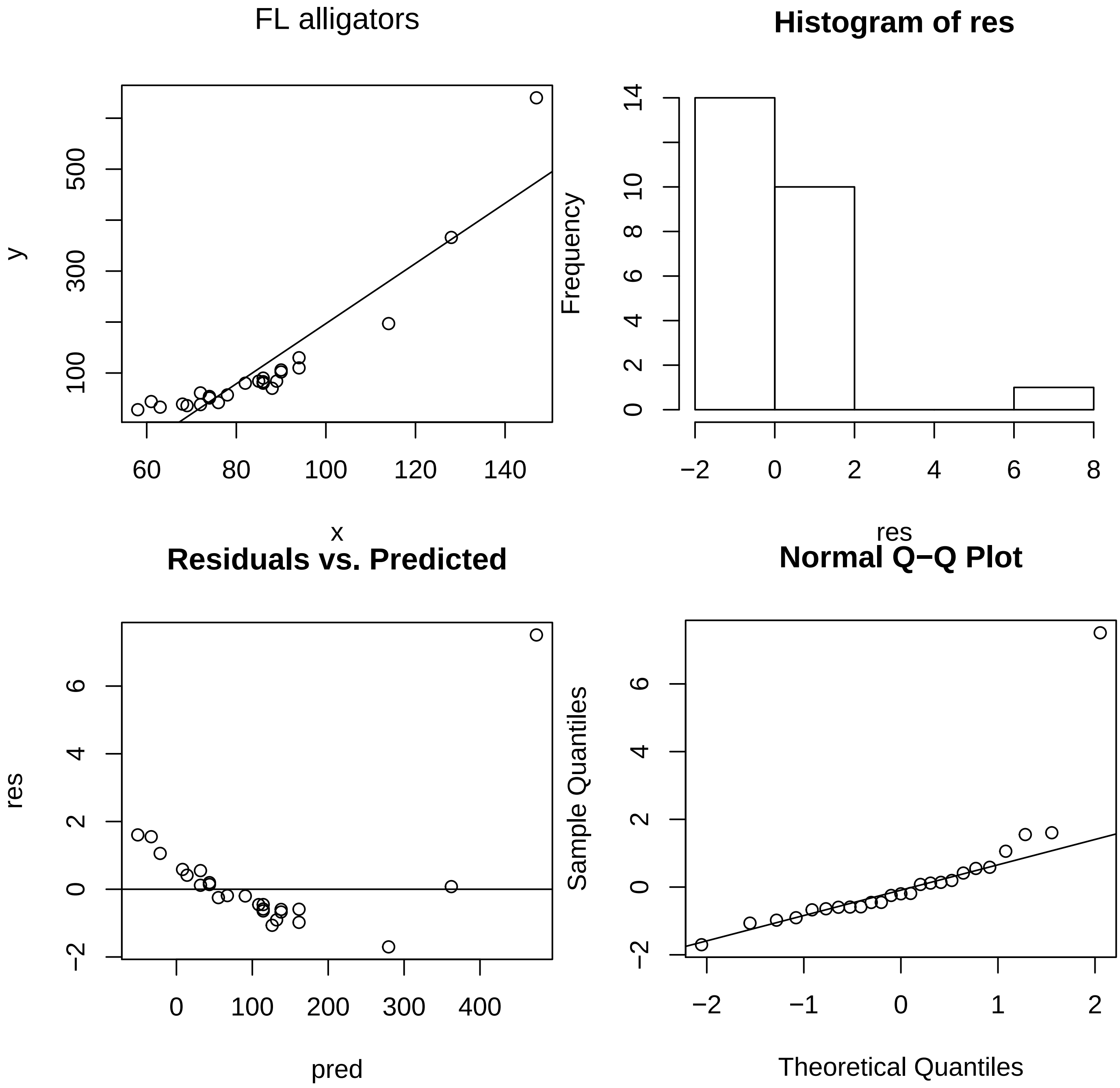
```
  length weight
1     94    130
2     74     51
3    147    640
4     58     28
5     86     80
6     94    110
```

```
with(gators,srs.muhaty('FL alligators',length,weight,25,10000,100))
```

```
 Results from SRS: Data = FL alligators
 Estimation method = Ratio for mean y
 N = 10000 n = 25
 FPC = 0.9975
 Ratio r = 1.27354
 Mu y = 127.354
 Vhat mu.y = 461.59
 Bound = 42.96929
```

```
Lower Bound = 84.38476 Upper Bound = 170.3233
```

```r
with(gators,reg.mean('FL alligators',length,weight,10000,100))
```



```
Results from SRS: Data = FL alligators
Estimation method = Regression for mean y
N = 10000 n = 25 FPC = 0.9975
Slope b = 5.902354  mux = 100
MuyL = 196.9714
Correlation = 0.9144007
```

```
Mean square error = 2917.238
Vhat muhat.yL = 116.3978
Bound = 21.57756
Lower Bound = 175.3938 Upper Bound = 218.549
```

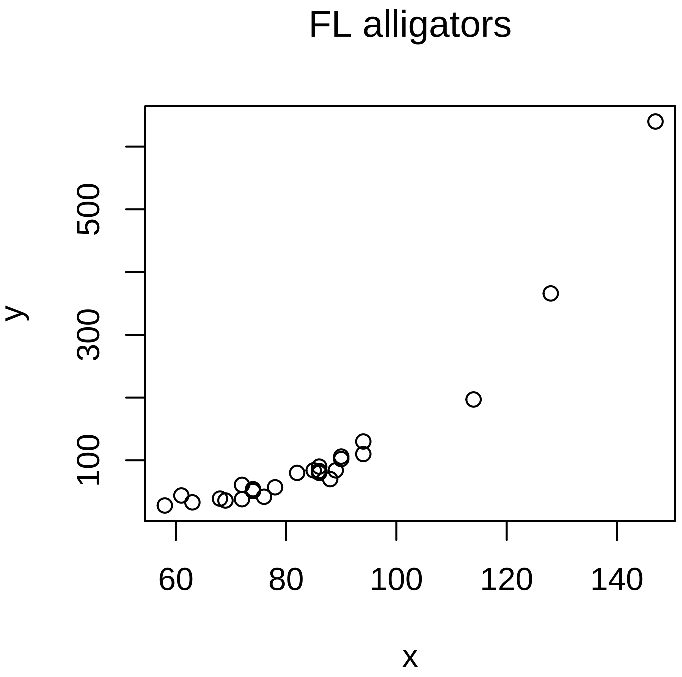
```
with(gators,diff.mean('FL alligators',length,weight,10000,100))
```

```
Results from SRS: Data = FL alligators
Estimation method = Difference estimator of muy
N = 10000 n = 25
FPC = 0.9975
MuyD = 123.24 Variance of differences = 12636.19
Vhat muhat.yD = 504.184
Bound = 44.90808
Lower Bound = 78.33192 Upper Bound = 168.1481
```

```
with(gators,srsratio.n('FL alligators',10000,11568.67,100,100*10000,'mean',35))
```

```
Sample size estimation for SRS data  FL alligators
Ratio estimation for  mean
N = 10000 Required bound = 35
Estimated variance = 11568.67
Mean of x (mu.x) = 100
Total of x (taux) 1e+06
Estimated sample size = 38
```

```
with(gators,RE('FL alligators',length,weight,'ratio','regr'))
```



```
Estimated Relative Efficiency
Data:  FL alligators
```

```
Estimation methods:  ratio and regr
Var1 = 11568.67 , Var2 = 2795.686
RE( ratio , regr ) = V( regr )/V( ratio ) = 0.2417
```

```
with(gators,RE('FL alligators',length,weight,'ratio','diff'))
```

```
Estimated Relative Efficiency
Data:  FL alligators
Estimation methods:  ratio and diff
Var1 = 11568.67 , Var2 = 12636.19
RE( ratio , diff ) = V( diff )/V( ratio ) = 1.0923
```

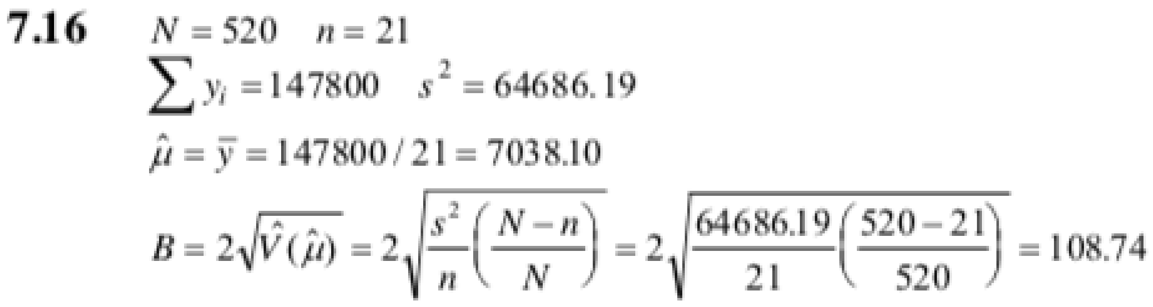
```
with(gators,RE('FL alligators',length,weight,'regr','diff'))
```

```
Estimated Relative Efficiency
Data:  FL alligators
Estimation methods:  regr and diff
Var1 = 2795.686 , Var2 = 12636.19
RE( regr , diff ) = V( diff )/V( regr ) = 4.5199
```

```
with(gators,RE('FL alligators',length,weight,'srs','ratio'))
```

```
Estimated Relative Efficiency
Data:  FL alligators
Estimation methods:  srs and ratio
Var1 = 17060.25 , Var2 = 11568.67
RE( srs , ratio ) = V( ratio )/V( srs ) = 0.6781
```

(8) 7.16



**7.16** $N = 520 \quad n = 21$

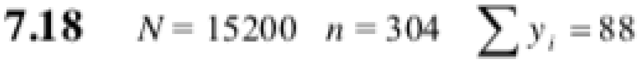
$$\sum y_i = 147800 \quad s^2 = 64686.19$$

$$\hat{\mu} = \bar{y} = 147800/21 = 7038.10$$

$$B = 2\sqrt{\hat{V}(\hat{\mu})} = 2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)} = 2\sqrt{\frac{64686.19}{21}\left(\frac{520-21}{520}\right)} = 108.74$$

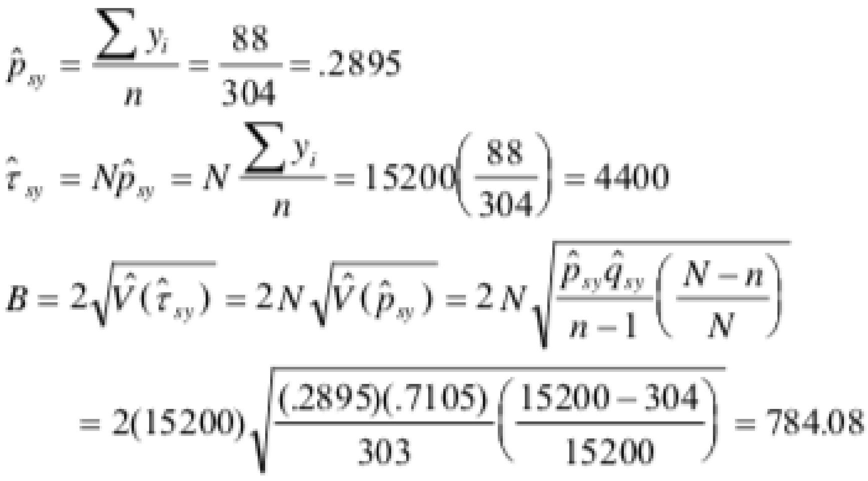
Figure 4:

(9) 7.18



**7.18** $N = 15200 \quad n = 304 \quad \sum y_i = 88$

88



$$\hat{p}_{sy} = \frac{\sum y_i}{n} = \frac{88}{304} = .2895$$

$$\hat{\tau}_{sy} = N\hat{p}_{sy} = N\frac{\sum y_i}{n} = 15200\left(\frac{88}{304}\right) = 4400$$

$$B = 2\sqrt{\hat{V}(\hat{\tau}_{sy})} = 2N\sqrt{\hat{V}(\hat{p}_{sy})} = 2N\sqrt{\frac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}\left(\frac{N-n}{N}\right)}$$

$$= 2(15200)\sqrt{\frac{(.2895)(.7105)}{303}\left(\frac{15200-304}{15200}\right)} = 784.08$$

Figure 5: