3.3 (a) $\bar{y}_U = 71.83333$, $S^2 = 86.16667$.

(b) $\binom{6}{4} = 15$.

(c) The 15 samples are:

Units in sample				y values in sample				Sample Mean
1	2	3	4	66	59	70	83	69.50
1	2	3	5	66	59	70	82	69.25
1	2	3	6	66	59	70	71	66.50
1	2	4	5	66	59	83	82	72.50
1	2	4	6	66	59	83	71	69.75
1	2	5	6	66	59	82	71	69.50
1	3	4	5	66	70	83	82	75.25
1	3	4	6	66	70	83	71	72.50
1	3	5	6	66	70	82	71	72.25
1	4	5	6	66	83	82	71	75.50
2	3	4	5	59	70	83	82	73.50
2	3	4	6	59	70	83	71	70.75
2	3	5	6	59	70	82	71	70.50
2	4	5	6	59	83	82	71	73.75
3	4	5	6	70	83	82	71	76.50

Using (2.9),

$$V(\bar{y}) = \left(1 - \frac{4}{6}\right) \frac{86.16667}{4} = 7.180556.$$

(d) $\binom{3}{2}\binom{3}{2} = 9$.

(e) You cannot have any of the samples from (c) which contain 3 units from one of the strata. This eliminates the first 3 samples, which contain $\{1, 2, 3\}$ and the three samples containing students $\{4, 5, 6\}$. The stratified samples are

Units in $\mathcal{S}_1$		Units in $\mathcal{S}_2$		y values in sample				$\bar{y}_{\text{str}}$
1	2	4	5	66	59	83	82	72.50
1	2	4	6	66	59	83	71	69.75
1	2	5	6	66	59	82	71	69.50
1	3	4	5	66	70	83	82	75.25
1	3	4	6	66	70	83	71	72.50
1	3	5	6	66	70	82	71	72.25
2	3	4	5	59	70	83	82	73.50
2	3	4	6	59	70	83	71	70.75
2	3	5	6	59	70	82	71	70.50

$$V(\bar{y}_{\text{str}}) = \left(1 - \frac{2}{3}\right)\left(\frac{3}{6}\right)^2 \frac{31}{2} + \left(1 - \frac{2}{3}\right)\left(\frac{3}{6}\right)^2 \frac{44.33333}{2} = 3.14.$$

The variance is smaller because the extreme samples from (c) are excluded by the stratified design. The variances $S_1^2 = 31$ and $S_2^2 = 44.33$ are much smaller than the population variance S^2.

3.7 (a) Here are summary statistics for each stratum:

	Stratum			
	Biological	Physical	Social	Humanities
average	3.142857	2.105263	1.230769	0.4545455
variance	6.809524	8.210526	4.358974	0.8727273

Since we took a simple random sample in each stratum, we use

$$(102)(3.142857) = 320.5714$$

to estimate the total number of publications in the biological sciences, with estimated variance

$$(102)^2\left(1 - \frac{7}{102}\right)\frac{6.809524}{7} = 9426.327.$$

The following table gives estimates of the total number of publications and estimated variance of the total for each of the four strata:

Stratum	Estimated total number of publications	Estimated variance of total
Biological Sciences	320.571	9426.33
Physical Sciences	652.632	38982.71
Social Sciences	267.077	14843.31
Humanities	80.909	2358.43
Total	1321.189	65610.78

We estimate the total number of refereed publications for the college by adding the totals for each of the strata; as sampling was done independently in each stratum, the variance of the college total is the sum of the variances of the population stratum totals. Thus we estimate the total number of refereed papers as 1321.2, with standard error $\sqrt{65610.78} = 256.15$.

(b) From Exercise 2.6, using an SRS of size 50, the estimated total was $\hat{t}_{\text{srs}} = 1436.46$, with standard error 296.2. Here, stratified sampling ensures that each division of the college is represented in the sample, and it produces an estimate with a smaller standard error than an SRS with the same number of observations. The sample variance in Exercise 2.8 was $s^2 = 7.19$. Only Physical Sciences had a sample variance larger than 7.19; the sample variance in Humanities was only 0.87. Observations within many strata tend to be more homogeneous than observations in the population as a whole, and the reduction in variance in the individual strata often leads to a reduced variance for the population estimate.

(c)

	N_h	n_h	$\hat{p}_h$	$\frac{N_h}{N}\hat{p}_h$	$\left(1-\frac{n_h}{N_h}\right)\frac{N_h^2}{N^2}\frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$
Biological Sciences	102	7	$\frac{1}{7}$	.018	.0003
Physical Sciences	310	19	$\frac{10}{19}$	.202	.0019
Social Sciences	217	13	$\frac{9}{13}$	.186	.0012
Humanities	178	11	$\frac{8}{11}$	.160	.0009
Total	807	50		.567	.0043

$$\begin{aligned} \hat{p}_{\text{str}} &= 0.567 \\ \text{SE}[\hat{p}_{\text{str}}] &= \sqrt{0.0043} = 0.066. \end{aligned}$$

3.25 (a) We substitute $n_{h,\text{Neyman}}$ for n_h in (3.4):

$$\begin{aligned}
V_{\text{Neyman}}(\hat{t}_{\text{str}}) &= \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h} \\
&= \sum_{h=1}^{H}\left(1-\frac{N_hS_hn}{N_h\sum_{l=1}^{H}N_lS_l}\right)N_h^2\frac{S_h^2\sum_{l=1}^{H}N_lS_l}{n_hN_hS_hn} \\
&= \sum_{h=1}^{H}\left(1-\frac{S_hn}{\sum_{l=1}^{H}N_lS_l}\right)\frac{N_hS_h\sum_{l=1}^{H}N_lS_l}{n} \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}N_hS_h\right)^2-\sum_{h=1}^{H}N_hS_h^2.
\end{aligned}$$

(b)

$$\begin{aligned} V_{\text{prop}}(\hat{t}_{\text{str}}) - V_{\text{Neyman}}(\hat{t}_{\text{str}}) &= \frac{N}{n}\sum_{h=1}^{H} N_h S_h^2 - \sum_{h=1}^{H} N_h S_h^2 \\ &\quad -\frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 + \sum_{h=1}^{H} N_h S_h^2 \\ &= \frac{N}{n}\sum_{h=1}^{H} N_h S_h^2 - \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 \\ &= \frac{N^2}{n}\left[\sum_{h=1}^{H}\frac{N_h}{N} S_h^2 - \left(\sum_{h=1}^{H}\frac{N_h}{N} S_h\right)^2\right] \\ &= \frac{N^2}{n}\sum_{h=1}^{H}\frac{N_h}{N}\left[S_h^2 - S_h\sum_{l=1}^{H}\frac{N_l}{N} S_l\right] \end{aligned}$$

But

$$\begin{aligned} \sum_{h=1}^{H}\frac{N_h}{N}\left[S_h - \sum_{l=1}^{H}\frac{N_l}{N} S_l\right]^2 &= \sum_{h=1}^{H}\frac{N_h}{N}\left[S_h^2 - 2S_h\sum_{l=1}^{H}\frac{N_l}{N} S_l + \left(\sum_{l=1}^{H}\frac{N_l}{N} S_l\right)^2\right] \\ &= \sum_{h=1}^{H}\frac{N_h}{N} S_h^2 - \left(\sum_{l=1}^{H}\frac{N_l}{N} S_l\right)^2, \end{aligned}$$

proving the result.

(c) When $H = 2$, the difference from (b) is

$$\begin{aligned} &\frac{N^2}{n}\sum_{h=1}^{2}\frac{N_h}{N}\left(S_h - \sum_{l=1}^{2}\frac{N_l}{N} S_l\right)^2 \\ &= \frac{N^2}{n}\left[\frac{N_1}{N}\left(S_1 - \frac{N_1}{N} S_1 - \frac{N_2}{N} S_2\right)^2 + \frac{N_2}{N}\left(S_2 - \frac{N_1}{N} S_1 - \frac{N_2}{N} S_2\right)^2\right] \\ &= \frac{N^2}{n}\left[\frac{N_1}{N}\left(\frac{N_2}{N} S_1 - \frac{N_2}{N} S_2\right)^2 + \frac{N_2}{N}\left(\frac{N_1}{N} S_2 - \frac{N_1}{N} S_1\right)^2\right] \\ &= \frac{N^2}{n}\frac{N_1}{N}\frac{N_2}{N}\left[\frac{N_1}{N} + \frac{N_2}{N}\right](S_1 - S_2)^2 \\ &= \frac{N_1 N_2}{n}(S_1 - S_2)^2. \end{aligned}$$

3.34

(a) In the data step, define the variable one to have the value 1 for every observation. Then $\sum_{i \in \mathcal{S}} w_i 1 = N$. Here, $\sum_{i \in \mathcal{S}} w_i 1 = 85174776$. The standard error is zero because this is a stratified sample. The weights are N_h/n_h so the sum of the weights in stratum h is N_h exactly. There is no sampling variability.

Here is the code used to obtain these values:

```
proc surveymeans data=vius mean clm sum clsum;
   weight tabtrucks;
   stratum stratum;
   var  one miles_annl mpg;
```

(b) The estimated total number of truck miles driven is 1.115×10^{12}; the standard error is 6492344384 and a 95% CI is $[1.102 \times 10^{12}, 1.127 \times 10^{12}]$.

(c) Because these are stratification variables, we can calculate estimates for each truck type by summing $w_{hj} y_{hj}$ separately for each h. We obtain:

```
proc sort data=vius;
   by trucktype;
proc surveymeans data=vius sum clsum;
   by trucktype;
   weight tabtrucks;
   stratum stratum;
   var miles_annl;
   ods output Statistics=Mystat;
proc print data=Mystat;
run;
```

Obs	VarName	VarLabel	Sum
1	MILES_ANNL	Number of Miles Driven During 2002	428294502082
2	MILES_ANNL	Number of Miles Driven During 2002	541099850893
3	MILES_ANNL	Number of Miles Driven During 2002	41279084490
4	MILES_ANNL	Number of Miles Driven During 2002	31752656137
5	MILES_ANNL	Number of Miles Driven During 2002	72301789843

Obs	LowerCLSum	StdDev	UpperCLSum
1	4.19064E11	4708839922	4.37525E11
2	5.32459E11	4408042207	5.4974E11
3	4.05032E10	395841910	4.2055E10
4	3.107E10	348294378	3.24353E10
5	7.12861E10	518195242	7.33175E10

(d) The estimated average mpg is 16.515427 with standard error 0.039676; a 95% CI is [16.4377, 16.5932]. These CIs are very small because the sample size is so large.