Exercise 1: We want to know what proportion of ponds in a region have fish in them. However, the ponds occur in groups so that if we visit one pond we ought to just sample all ponds in the group. So we divide (using an aerial photo) the area into $N = 100$ clusters, each consisting of from one to five ponds. We will select (SRS) $n = 10$ of the clusters. We get the following results:

* Cluster 1: $\tau_1 = 3$ ponds have fish, out of $M_1 = 4$ ponds.
* Cluster 2: $\tau_2 = 2$ ponds have fish, out of $M_2 = 3$ ponds.
* Cluster 3: $\tau_3 = 1$ pond has fish, out of $M_3 = 4$ ponds.
* Cluster 4: $\tau_4 = 1$ pond has fish, out of $M_4 = 3$ ponds.
* Cluster 5: $\tau_5 = 3$ ponds have fish, out of $M_5 = 5$ ponds.
* Cluster 6: $\tau_6 = 2$ ponds have fish, out of $M_6 = 4$ ponds.
* Cluster 7: $\tau_7 = 4$ ponds have fish, out of $M_7 = 5$ ponds.
* Cluster 8: $\tau_8 = 2$ ponds have fish, out of $M_8 = 2$ ponds.
* Cluster 9: $\tau_9 = 4$ ponds have fish, out of $M_9 = 5$ ponds.
* Cluster 10: $\tau_{10} = 5$ ponds have fish, out of $M_{10} = 5$ ponds.

Find a 95 percent confidence interval for the proportion of ponds that have fish, in the region.

Solution:
Since we do not know the total number of ponds in the in the region (no M) we have to use the one stage cluster sample ratio estimator to compute the proportion. Using the ratio estimator in r we get,
Code:

```
> Mi
     [1] 4 3 4 3 5 4 5 2 5 5
> pi
     [1] 3 2 1 1 3 2 4 2 4 5
> n = 10
> N = 100

> p_est = sum(pi)/sum(Mi)
     [1] 0.675
> mse = sum((pi - p_est*Mi)^2)/(n-1)
     [1] 0.9451389
> se = sqrt(((N-n)/N)*(n/sum(Mi))^2*(mse/n))
     [1] 0.07291369
> CI = c(p_est + 2*se, p_est - 2*se)
     [1] 0.8208274 0.5291726
```

Exercise 2: Read the paper Mortality Before and After the 2003 Invasion of Iraq (lancet.pdf). How did they conduct the sample? Why was cluster sampling the only reasonable way to conduct this survey?

Solution:
The paper describes a modified two-stage cluster sample. There were 33 clusters and in each cluster there sampled 30 households. The 33 clusters where randomly divided among the 18 governorates, then a second stage of reassignment was used to group the clusters in 6 pairs of governorates for safety and travel reasons. Without conducting an entire census the two-stage cluster sample gave the researchers the flexibility to actually conduct the experiment. Begin able to reassign clusters for safety, and take a reasonable sample in each cluster made this study possible.

Exercise 3: We want to know the proportion of nests in a forest that are in current use. What we'll do is divide the forest into $N = 10000$ 20m by 20m plots and count the number of nests M_i in each of n plots that we obtain from a simple random sample of plots. Once we get our n plots, we'll select (SRS) m_i nests in the ith plot and examine these. The number being used is τ_i. This is similar to problem 1, but differs in one crucial way.

Plot M_i m_i τ_i
1, 52, 20, 12
2, 40, 20, 11
3, 35, 20, 15
4, 38, 20, 5
5, 51, 25, 12
6,14, 20, 8
7, 50, 25, 13
8, 42, 20, 9
9, 37, 20, 7
10, 30, 20, 14
11, 22, 20, 16
12, 12, 10, 4

a. Find a 95 percent confidence interval for the true proportion p.

Solution:
This problem is similar to the first since we will be using the ratio cluster sample estimator but this time we are working with a two-stage cluster sample. Computing the two-stage cluster sample estimator in r we get,
Code:

```
> N = 10000
> n = 12
> Mi <- c(52, 40, 35, 38, 51, 14, 50, 42, 37, 30, 22, 12)
> mi <- c(20, 20, 20, 20, 25, 10, 25, 20, 20, 20, 20, 10)
> tau_i <- c(12, 11, 15, 5, 12, 8, 13, 9, 7, 14, 16, 4)
> pi = tau_i/mi
  [1] 0.60 0.55 0.75 0.25 0.48 0.80 0.52 0.45 0.35 0.70 0.80 0.40

> p_est = sum(Mi*pi)/sum(Mi)
  [1] 0.5339953
> mse = sum(Mi^2(pi - p_est)^2)/(n-1)
> clusterVar = ((N-n)/N)*(n/sum(Mi))^2*(mse/n)
  [1] 0.00244256
> SRSVar = (1/(n*N))*(n/sum(Mi))^2*sum(Mi^2*((Mi - mi)/Mi)*((pi*(1 -
  [1] 6.37466e-07

> se = sqrt(clusterVar + SRSVar)
  [1] 0.04942871

> CI = c(p_est + 2*se, p_est - 2*se)
  [1] 0.6328527 0.4351378
```

b. When would you sample fewer clusters (plots) and more inside each cluster (plot)?

Solution:
We can sample fewer clusters and more inside each clusters when we know there is little variance between clusters and more variance inside of each cluster. This can occur if clusters in a sample vary wildly, but each cluster seems to exhibit the same behavior. Suppose we are estimating grass cover and each cluster exhibits the same level of patchiness, and the sample in each cluster have high variance. We can see this by computing the terms in the variance independently, in our case the cluster to cluster variability is a lot higher so we would want to visit more clusters. In fact the variability inside each cluster seems very low and it might be better to sample less in each cluster.

Exercise 4: Here's another two-stage cluster sampling problem.

Suppose I want to know the total value of an inventory. There are $N = 30$ warehouses, and I don't want to go to all of them. Also, each warehouse is too large to sample everything, though the items (should) be somewhat similar in value. I'll select $n = 5$ warehouses

as a SRS, then I'll count the number of items (M_i), take a sample of $m_i = 20$ items from each warehouse, then get an average value and standard deviation from each warehouse:

* warehouse 1: $M_1 = 2100$, $m_1 = 20$, $\overline{x}_1 = \$35.00$, $s_1 = \$10.00$.
* warehouse 2: $M_2 = 850$, $m_2 = 20$, $\overline{x}_2 = \$42.00$, $s_2 = \$12.00$.
* warehouse 3: $M_3 = 1500$, $m_3 = 20$, $\overline{x}_3 = \$40.00$, $s_3 = \$8.00$.
* warehouse 4: $M_4 = 2200$, $m_4 = 20$, $\overline{x}_4 = \$38.00$, $s_4 = \$5.00$.
* warehouse 5: $M_5 = 500$, $m_5 = 20$, $\overline{x}_5 = \$18.00$, $s_5 = \$6.00$.

a/b. Find an estimate for the total value of items in all warehouses and obtain a 95 percent confidence interval. Note that we only get an item count for warehouses we sample.

Solution:
Since we don't have the total number of items in the inventory M we will use the unbiased estimator of the total. Computing the estimator in r we get,
Code:

```
> N = 30
> n = 5
> Mi
     [1] 2100  850 1500 2200  500
> m_i
     [1] 20 20 20 20 20
> xbar
     [1] 35 42 40 38 18
> si
     [1] 10 12  8  5  6

## Estimated Total
> tau_est = (N/n)*sum(Mi*xbar)
     [1] 1570800

> mse = sum((Mi*xbar - (tau_est/N))^2)/(n-1)
     [1] 909713000

## Cluster vs Sample Variance
> clusterVar = ((N-n)/N)*(N)^2*(mse/n)
     [1] 1.36457e+11
> SRSVar = (N/n)*sum(Mi^2*((Mi - m_i)/Mi)*(si^2/m_i))
     [1] 242703600

> se = sqrt(clusterVar + SRSVar)
     [1] 369729.2
```

```
# 95 Confidence Interval
> CI = c(tau_est + 2*se, tau_est - 2*se)
    [1] 2310258.3  831341.7
```