Exercise 1: In a study of sick leave at a large company we want to sample several branch offices. When we sample a branch office we will check the number of days that employees were sick but did not take sick leave. One way to do this is via sampling with probability proportional to size. Suppose the following is a list of offices and their size...

a What are some advantages of sampling proportional to size, in general?

Solution:
In many sampling plans, the property that the probability of a sample is proportional to it's size is built-in. For example any sort or geo-spatial application where you need to cluster transects, the probability of sampling a cluster is proportional to the size of the cluster(number of transects contained in the cluster.) Beyond that we get all the flexibility afforded to us with unequal probability sampling schemes like adaptive sampling, while only using an estimate for each probability.

b. With the list, select a sample of $n = 5$ clusters (offices) proportional to M_i. In this problem we will use the Hansen-Hurwitz estimator, so you should sample with replacement. If you know the selection probabilities you want (the P_i), you can put them in a vector called p...

Solution:
Computing the probability vector p using the size of each office, and computing an office cluster sample of size 5 with probability p,
Code:

```
> df = data.frame(office, employees)
> p = df$employees/sum(df$employees)
    0.05117271 0.07462687 0.04264392 0.11087420
    0.03837953 0.03198294 0.05117271 0.13432836
    0.04477612 0.03624733 0.04904051 0.05543710
    0.09168443 0.10234542 0.03198294 0.05330490

> Office_Sample  = sample(df$office, size = 5, replace = TRUE,prob = p)
[1]  4  3  2  7 16

> df$employees[Office_Sample]
[1] 52 20 35 24 25
```

c. After you select your sample proportional to size, get the total sick days from the table below. Then compute the 95 percent confidence interval for the true total sick days at work for the entire company. Did it contain the actual total, which is $\tau = 706$?

Solution:
Recall that the Hansen-Hurwitz estimator for τ when sampling proportional to size becomes,

$$\hat{\tau} = \frac{M}{n}\sum_{i=1}^{n}\frac{x_i}{m_i}$$

$$V(\hat{\tau}) = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(\frac{Mx_i}{m_i} - \hat{\tau}\right)^2}{n-1}$$

Recall that M is the total number of employees in the 16 offices, m_i is the number of employees in the ith office, and n is our cluster sample size, and x_i is the total sick days used at the ith office. Computing the estimator and confidence interval for our sample in r, we get the following,
Code:

```
#Assigning Variables
x_i = total_sick_days[Office_Sample]
m_i = df$employees[Office_Sample]
M = sum(df$employees)
n = 5

#Estimating tau
est_tau = (M/n)*sum(x_i/m_i)
[1] 695.4669

#Computing Variance and CI
> est_Var = (1/n)*(sum((((M*x_i)/m_i) - est_tau)^2)/(n-1))
[1] 1822.809

> se = sqrt(est_Var)
[1] 42.69437

> CI = c(est_tau + 2*se, est_tau - 2*se )
[1] 780.8556 610.0781
```

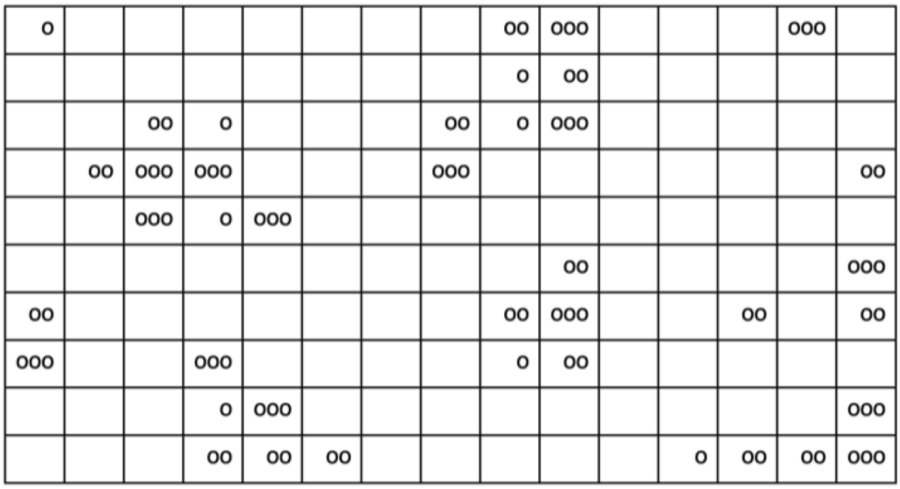
Exercise 2: Read the article by Kraft, Johnson, Samuelson and Allen (1995). The authors compared both simple random samples and systematic samples with and without stratification, and sampling proportional to size to estimate pronghorn populations. What techniques worked best? Which did not work as well? What did they think make some techniques work better than others?

Solution:
Of the techniques used the the Kraft article, it was found that a stratified SRS performed the best as measured by the average coefficient of variation among all three experiments. In general it was found that sampling with probability proportional to size(PPS), without stratification performed worse than all other sampling methods and the stratified PPS sample outperformed the SRS and SYS(systematic sample) with no stratification. The paper mentions that the correlation coefficients between sample area and pronghorn counts ranged from .003 to .46 stating that as the reason PPS didn't perform better. In the discussion it is stated that when sampling intensities are less than 10 percent there is little to no difference between sampling methods, when that is the case it is best to choose the sampling technique with the lowest cost.

Exercise 3: Suppose we wish to estimate the total number of birds on a sample of a large number of islands ($N = 50$ islands). We suppose that there will be more birds on larger islands, so we would like to somehow use the (known, for all islands) size of the island to guide us to more efficient sampling and/or analysis. Obviously (from this assignment), one way is to sample the islands with probability proportional to size. Can you think of two (or more) alternative approaches that we have already worked with that will use the island size to get an accurate estimate of bird total? Describe how you would apply these approaches.

Solution:
If we have reason to believe that island size is correlated with bird count, than we can use a ratio or regression estimator to improve our estimate. In both cases, ratio and regression we need to sample island area and bird count in pairs. We could also consider a cluster or stratified sampling scheme that puts similar size islands in the same group. This should reduce our variance inside each sample(cluster or strata).

Exercise 4: Consider the following plot,



a. Select a simple random sample of size $n = 8$ quadrats **with replacement**. (NOT THAT YOU'LL DO THIS IN REAL LIFE, but if you only get empty clusters, you can draw the sample again. In normal situations, you'd either give up or else get a larger sample size.

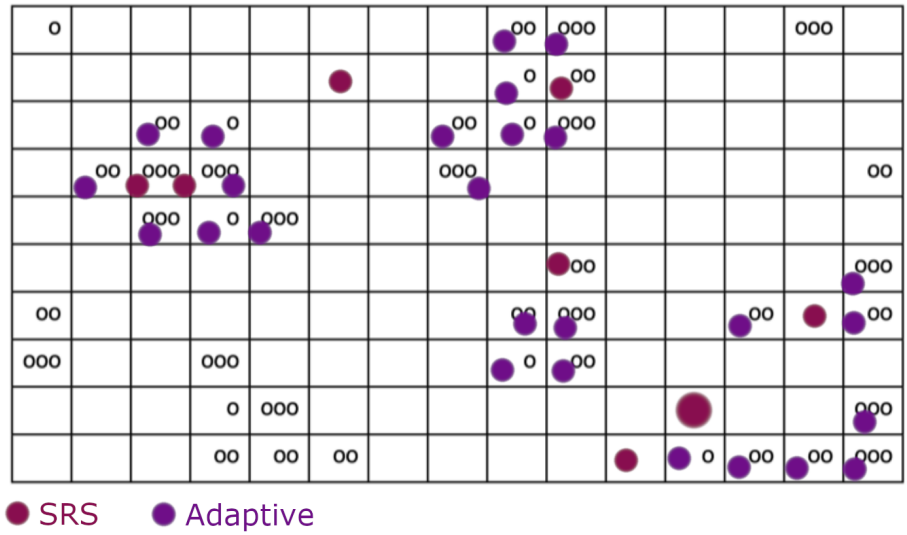
Solution:
We can use r to perform a simple random sample with replacement using the coordinates of each quadrat.
Code:

```
Coordinates <- matrix( nrow = 150, ncol= 2)

for (i in 1:10) {
    for (j in 1:15){
        Coordinates[((i - 1)*15 + j),1] = j
        Coordinates[((i - 1)*15 + j),2] = i
        }
}

SRS_plot = Coordinates[sample(nrow(Coordinates),size=8,replace=TRUE),
SRS_plot
     [,1] [,2]
[1,]   12    2
[2,]   11    1
[3,]   10    5
```

[4 ,]	10	9
[5 ,]	14	4
[6 ,]	3	7
[7 ,]	6	9
[8 ,]	3	7

b. For each unit in your sample, color in the resulting cluster if the rule is: Examine all squares to the N, S, E, W, NE, NW, SE, SW. Add any occupied quadrats to the cluster. Now repeat, searching around the newly added quadrats. Don't add any empty quadrats. Stop when your cluster is surrounded by empty quadrats (or the boundary).

Solution:
The following are the results of the adaptive sample from the SRS taken in the previous part,



c. Write down the count for each cluster you hit, and the probability of getting that cluster in your sample. This is a type of probability proportional to size sampling. NOTE: If you get the same cluster more than once, you put it in your data as many times as you 'hit' the cluster.

Solution:
The following is a .csv file containing coordinate information, flower count, and probability for each of the sampled clusters under the adaptive sampling scheme.
Code:

```
 x,y,count,probability
12,2,0, 1/150
11,1,0, 1/150
10,5,2, 5/150
10,9,2, 8/150
14,4,0, 1/150
 3,7,3, 8/150
 6,9,0, 1/150
 3,7,3, 8/150
12,1,1, 5/150
13,1,2, 5/150
14,1,2, 5/150
15,1,3, 5/150
15,2,3, 5/150
12,1,1, 5/150
13,1,2, 5/150
14,1,2, 5/150
15,1,3, 5/150
15,2,3, 5/150
 9,4,2, 5/150
 9,3,1, 5/150
10,4,3, 5/150
10,3,2, 5/150
10,0,3, 8/150
 9,0,2, 8/150
 9,9,1, 8/150
 9,8,1, 8/150
10,8,3, 8/150
 8,8,2, 8/150
 8,7,3, 8/150
15,4,2, 2/150
13,4,2, 1/150
15,5,3, 2/150
 3,8,2, 8/150
 4,8,1, 8/150
 2,7,2, 8/150
 4,7,3, 8/150
 3,6,3, 8/150
 4,6,1, 8/150
 5,6,3, 8/150
 3,8,2, 8/150
 4,8,1, 8/150
 2,7,2, 8/150
```

```
4,7,3, 8/150
3,6,3, 8/150
4,6,1, 8/150
5,6,3, 8/150
```

d. Get an estimator for the total number of plants in the region and compute a 95 percent confidence interval for the total number of plants. Is the estimator of the total unbiased?

Solution:
Note that each unit sampled yields an unbiased estimator of τ with $\hat{\tau} = x_i/P_i$ and similarly we can compute the variance. Doing so in r, we get the following,
Code:

```
df <- read.csv('AdaptiveSampling.csv')
Vec_of_tau <- df$count/(df$probability)
Tau_est = mean(Vec_of_tau)
    [1] 57.14674

Var_Tau = var(Vec_of_tau)/length(Vec_of_tau)
    [1] 61.86392

se = sqrt(Var_Tau)
    [1] 7.865362

CI = c(Tau_est + 2*se, Tau_est - 2*se)
    [1] 72.87746 41.41601
```