1. A city council of a small city wants to know the proportion of eligible voters that oppose having a incinerator of Phoenix garbage opened just outside of the city limits. They randomly select 100 residential numbers from the city's telephone book that contains 3,000 such numbers. Each selected residence is then called and asked for (a) the total number of eligible voters and (b) the number of voters opposed to the incinerator. A total of 157 voters were surveyed; of these, 23 refused to answer the question. Of the remaining 134 voters, 112 opposed the incinerator, so the council estimates the proportion by

$\hat{p}$ = 112/134 = .83582 with $\hat{V}$ ($\hat{p}$) = .83582(1 − .83582)/134 = 0.00102. Are these estimates valid? Why, or why not?

Solution: No, these estimates are not valid. The people who live inside the same residence are much likely to have similar opinions. When 23 refused to answer the question, the remaining 134 voters could have come from the residences who share the similar stand on the issue. Surveying a large number of people living in the same residence will not give us much more information and thus in this case two stage cluster sampling would be a better idea.

2. Senturia et al. (1994) describe a survey taken to study how many children have access to guns in their households. Questionnaires were distributed to all parents who attended selected clinics in the Chicago area during a one-week period for well or sick child visits.

a.Suppose that the quantity of interest is percentage of the households with guns. Describe why this is a cluster sample. What is the psu? The ssu? Is it a one-stage or two-stage cluster sample? How would you estimate the percentage of households with guns, and the standard error of your estimate?

Solution: The reason this is cluster sampling is because the study is conducted in whole of the Chicago area, but for selection, only those parents are selected who attended certain clinics in certain part of Chicago during a one-week period. So, we can see that clinics are the clusters.

The primary sampling units(psus) are selected clinics in the city area during a one-week period for well or sick child visit.

The secondary sampling units (ssus) are parents who attended selected clinics during a one-week period for well or sick child visit.

This is an example of one-stage cluster sampling because first clinics were randomly selected and as all parents got the questionnaire we can see there was no sampling in the second stage.

This scenario is unlikely to give us total secondary sampling units. So, we have to use ratio estimator . We would use $\widehat{\bar{y}_r}=\frac{\sum_{i\epsilon s} M_i \bar{y}_i}{\sum_{i\epsilon s} M_i}$, to estimate $\overline{y_U}$, where M_i represents the cluster size and $\bar{y}_i$ represents the average of the ssu from each cluster. We would create an indicator random variable y_{ij}, where y_{ij} represents either 1 if the household has gun and 0 if not for the j^{th}observation from $i^{th}\ cluster$. Then we would calculate the standard error by

SE=$\sqrt{[\left(1-\frac{n}{C}\right)\frac{1}{n\bar{M}^2}\sum_{i\epsilon S}\frac{{M_i}^2(\bar{y}_i-\widehat{\bar{y}_r})^2}{n-1}]}$, where n is the number of sampled clusters $and\ \bar{M}$ is the average of cluster sizes and C is the total number of clusters.

b.What is the sampling population for this study? Do you think this sampling procedure results in a representative sample of households with children? Why, or why not?

The sampling population for this study is number of households which have children and visited clinics during one-week for well and sick. This sampling procedure does not result in a representative sample of households with children as it only includes those parents who visited the selected clinics in the certain area during one week-period. It excludes those parents visiting other clinics in other part of the Chicago area as well as those parents who are in no need to visit the clinics.

3.Kleppel et al. (2004) report on a study of wetlands in upstate NewYork. Four wetlands were selected for the study: Two of the wetlands drain watersheds from small towns and the other two drain suburban watersheds. Quantities such as pH were measured at two to four randomly selected sites within each of the four wetlands.

a. Describe why this is a cluster sample. What are the psus? The ssus? How would you estimate the average pH in the suburban wetlands?

This is a cluster sample as we can see from the question that first wetlands were randomly selected and then sites were then randomly selected within each selected wetlands.

The primary sampling units are wetlands selected.

The secondary sampling units are sites within each wetland selected.

In order to estimate the average pH in the suburban wetlands, we first need to stratify the regions into small towns and suburbans. Once, we stratify then we need to separately estimate the $\overline{y_U}$ for

suburban region. If total number of ssu(sites) in the population is known say N, then we can estimate $\overline{y_U}$ by $\widehat{\bar{y}_{unb}} = \frac{\widehat{t_{unb}}}{N}$, where $\widehat{t_{unb}}$ represents the population total and can be estimated as $\widehat{t_{unb}} = \frac{C}{n}\sum_{i\in S}\hat{t}_i$. $Here, \hat{t}_i\ represents$ the cluster totals. If total number of ssu in the population is not known then we employ ratio estimator to estimate $\overline{y_U}$. So, instead we use $\hat{\bar{y}}_r = \frac{\sum_{i\in s} M_i \bar{y}_i}{\sum_{i\in s} M_i}$.

b. The authors used Student's two-sample t test to compare the average pH from the sites in the suburban wetlands with the average pH from the sites in the small town wetlands, treating all sites as independent. Is this analysis appropriate? Why, or why not?

For two-sampled t-test, we assume that the observations are independent of each other. Here, the probability of inclusion of ssu's (sites) depend upon the probability of the specific clusters being selected in the first stage. Due to this, it violates the condition of independence for the observations in the sample. So, the analysis is not appropriate.