Course 513-610 / Hanley
Assignment on sampling. DUE: DECEMBER ___, 1999.

The assignment is to measure some characteristics of studies, and reports on these studies, published in general medical journals. Assume, for the sake of this exercise, that the "universe of interest" comprises the information contained in abstracts of the 1061 original articles published in 1989 in the four journals: British Medical Journal (BMJ), Journal of the American Medical Association (JAMA), The Lancet and the New England Journal of Medicine (NEJM). The following six features are of interest:

1. The number of authors per article
2. Whether the abstract implies that the research concerned intact human beings
3. Whether the study was experimental.
4. The number of subjects studied
5. The number of p-values reported (explicitly, or implicitly using words like 'statistically significant' or 'statistically not significant')
6. The number of confidence intervals reported

Since this universe consists of 1061 units, and since your time and money are limited, you are asked to measure these characteristics using sample survey techniques and to work in a team with 2 other students. Your team is asked to perform and report the results of 4 different methods of sampling. These are

- a simple random sample of size n=10 of the 1061 abstracts
- a stratified random sample (total n=10) of the 1061 abstracts, using the 4 journals as strata.
- a cluster sample of size 2, using each week ("issue") in each journal as a separate cluster (there are, if our counting is correct, 202 issues in all)
- a two-stage sample using systematic samples of 5 articles from each of 2 randomly chosen journals, with the journals chosen with equal probabilities without replacement.

We ask that each member of the team participate in the planning of all 4 surveys so that everyone gets to learn the mechanics of all of the methods of selecting sample elements. We ask you to use published tables of random numbers (not ones you might take out of your head), to describe how each sample element was chosen, and to document your 'trail' so that someone else could replicate your samples (e.g. record how one could reconstruct your sequence of random numbers). Document your trail on the selection sheet itself. Since we want to see sampling variability, we ask that different teams be imaginative in producing different "streams" of random numbers.

Once your team has chosen the abstracts, you may divide up the actual "leg-work" involved (locating the abstracts; recording the data) in whatever way seems most efficient.

Record your team's results on the 3 sheets provided on the web page.
Also, as part of the team's report, make an estimate, for each of the 4 surveys, of the time it would have taken one individual, working alone, to:

- establish the necessary sampling frame*
- make the sample selection
- find the selected sample members
- record the information

*much of this was already done for you on the attached spreadsheet.

Technical Notes on sampling exercise:

For a description of the various sampling methods, see for example

Levy and Lemeshow (Sampling for Health Professionals 1980)
Cochran (Sampling Techniques, 1963 and later editions)

The sampling frame on this same web page will save you some work.

See Armitage&Berry (Statistical methods for medical research 2nd ed 1987), Levy&Lemeshow, Pearson and Hartley's Biometrika Tables for Statisticians, or other texts for a table of random numbers and how to use them. Teams should use them in imaginative ways to avoid choosing the same sequences as other teams.

If you prefer to generate random integers from 1 to K inclusive on a spreadsheet: in Quattro use the function @INT(1+K*@RAND) ; in Excel: =INT(1+K*RAND()) .

Stratified sampling: In general, there is no requirement that the sample sizes $\{n_i\}$ in the stratified sampling be proportional to the sizes $\{N_i\}$ of the strata: the estimates from the stratum-specific samples are always weighted using weights of $\{W_i\}=\{N_i/N\}$, where $N=\sum N_i$. However, the sampling variance of this weighted estimate is a function of the $\{W_i\}$, the $\{n_i\}$ and the$\{\sigma_i\}$, where σ_i is the sd of the variable of interest in the ith stratum. The sampling variance of the overall estimate is minimized by sample sizes that are proportional to $N_i\sigma_i$. In the exercise here, because we have no prior information about variability within the different strata, we are asking you to split the total sample size of 10 into 4:2:2:2 to approximate the ratio of the N_i's {430, 178, 219, 234}.

Cluster sampling: When the clusters contain different numbers of elements, as they do here, there are several ways of sampling, and of estimating totals and means; also, there is the issue of whether the cluster sizes are all known ahead of time or only become available at the time a cluster is chosen. Furthermore, there are important tradeoffs of precision and bias (if interested, see Cochran or Lemeshow).

2-stage sampling: Again,there are some issues of how to sample clusters and how to form estimates, and to calculate sampling variation (see Cochran or Lemeshow). We are not asking you to calculate standard errors (SE's) for your estimates from cluster and 2-stage sampling. They are not likely to be very stable, since they involve estimates of between-cluster variation, and 2 clusters would not give a very trustworthy assessment! However, Cochran and Lemeshow do give formulae.

Systematic sampling: One practical issue is how to be fair when the planned sample size n does not divide evenly into N. Levy page 76 gives a modification that always yields unbiased estimates (it yields different sample sizes). Liddell, in a handout I have somewhere, offers a correction to that given in Kish's book on sampling (the bible to professional survey samplers, that may be the same as Levy's. It will not make a big difference here, so ignore the issue for the sake of the exercise.