# Appendix C: Selecting a Representative Industrial Sample

*Cases, if selected randomly, can be expected to be representative of the larger population of companies, contractors, or programs from which they were drawn.* As a result, statistical relationships found among characteristics of the individuals (e.g., the probability that an establishment is segregated, as a function of annual sales volume) maybe generalized to the population from which the individuals were selected.

Selecting cases randomly requires randomly identifying the establishments, contracts programs, dollars, or other entities about which information is desired, and then obtaining the information desired e.g., by survey. A high response rate is required to assure representativeness of the sample.

The choice of population from which cases are to be selected (e.g., establishments or contracts) will depend in part on the type of information desired but will be limited by the availability of databases or the ability to obtain a high response rate in a survey. For example, congressional researchers may use the Award Contracts (AWCO) database of the House Information System (HIS), which contains data (provided by the Federal Procurement Data Center of the General Services Administration) on all federal contracts of $25,000 or more in the last full year for which data are available. Once a population and database have been found, individuals may be selected from the population by using the pseudorandom-number generation capabilities of common software products. Pseudorandomly generated numbers may be tested for acceptable independence, randomness, and uniformity.[1]

For each DOD contract or prime contractor in the sample, a random sample of subcontracts or subcontractors may be selected. The sampling may be uniform or weighted by contract value. The former would be more efficient if one wished to make inferences about the number of contractors that are integrated or would be affected by some proposed change; the latter would be more efficient if one wished to make inferences about the fraction of the defense budget that would be affected (e.g., potential cost aversion). Sampling could be restricted to particular sectors of interest.

[1] D.A. Darling, "The Kolmogorov-Simimov, Cramer-von Mises Tests," *Annals of Mathematical Statistics, vol. 28,* pp. 823-838, 1957.

After the cases to be studied have been identified and the desired data obtained, statistical analysis can identify the combination of characteristics that best distinguishes the integrated cases from the segregated cases.[2] It can estimate the probability that a company, for example, in the sample is integrated, based on other characteristics (e.g., annual sales) that are known about the companies in the sample.[3]

A model obtained in such a manner describes relationships e.g., between integration of a company and other variables within the sample to which it was fitted. It may be used to predict integration in a larger population, provided 1 ) the incidence (i.e., unconditional probability) of integration in the new population is known, and 2) there is no reason to believe that the new population differs significantly from the sample in any aspect that may influence integration. Drawing the sample randomly from the population ensures this and, moreover, allows the incidence of integration in the population to be estimated from that in the sample.

2 J.A Anderson, "Logistic Discrimination," pp. 169-191 in *Handbook of Statistics*, P.R. Krishnaiah and L.N. Kanal, eds., vol. 2, No. 1, 1982.

[3]Gary G. Koch and Suzanne Edward\, "Logistic Regression," pp. 128-133 in *Encyclopedia of Statistical Sciences*, Samuel Kotz and Norman L. Johnson, eds., vol. 5 (New York, NY: John Wiley & Sons, 1985).