

## Sample size estimation – dichotomous example

Dimitrios Mavridis, Colin Aitken

Consider a population or consignment, which consists of discrete units, such as individual tablets in a consignment of tablets or individual computer disks in consignment of disks. Each unit may, or may not, contain something illegal, such as drugs or pornographic images. It is of interest to an investigating scientist to determine the proportion of the consignment which contains something illegal. This may be done exactly (assuming no mistakes are made) by examination of every unit in the consignment. Such an examination can be extremely costly. Considerable resources can be saved if information, sufficient to satisfy the needs of the investigators, may be gained from examination of a sample from the consignment. Uncertainty is introduced when inference is made from the sample to the population, because the whole population is not inspected. However, this uncertainty may be quantified probabilistically. It is shown that if two numbers are specified in advance of the inspection of the consignment, then a sample size may be specified. The first of these numbers is the minimum proportion of units in the consignment which contain something illegal that it is desired to find. The second is the probability with which the true proportion of illegal units exceeds this minimum proportion.

With reasonable assumptions, a probability distribution for the true proportion of units in the consignment is derived, based on the scientist's prior beliefs prior to the inspection of individual units) and the outcome of the inspection of the sample. It is possible to choose a function to represent the strength of the scientist's prior beliefs. It may be thought inappropriate that the scientist's prior beliefs should have any effect on the decision to be made regarding the sample size. In such a case, it is possible to choose the function in such a way that the effect is very small. (It is also possible to choose the function such that the effect is very large.) It is not possible for the scientist's prior beliefs to have no effect on the analysis. For example, the choice of the model which is used to represent the uncertainty introduced by the sampling process is a subjective choice. The binomial model described here requires assumptions about independence of the probability for each unit being illegal and the choice of a constant value for this probability.