Sampling With Unequal Weights (G05NEF)

The term sampling can have a number of different meanings ranging from converting from a continuous signal to a discrete one in signal processing, re-using parts of different sound recordings to produce a new record or selecting subjects for study in a controlled experiment. Here we are using it to mean randomly selecting one or more observations or records from a particular dataset.

The need to sample from a dataset appears in many areas, it forms the basis for bootstrapping algorithms, allocating individuals to a particular arm of a designed experiment or reducing the size of a large database. Sampling can be performed in two ways:

- **With replacement:**
  When sampling with replacement each data point in the original dataset can appear multiple times in the sample. The sample can therefore be larger than the original dataset.

- **Without replacement:**
  When sampling without replacement each data point in the original dataset can appear at most once in the sample. The sample is therefore no larger than the original dataset.

Each of these sampling methods can be further divided into two categories:

- **With equal weights:**
  When sampling with equal weights each observation in the original dataset has the same probability of appearing in the sample as every other observation.

- **With unequal weights:**
  When sampling with unequal weights the probability of an observation from the original dataset appearing in the sample is proportional to the weight assigned to that observation.

Prior to Mark 23 routines were available in the NAG library to facilitate three of these sampling regimes, with the addition of **G05NEF** sampling without replacement using unequal weights, which forms the basis of any importance resampling algorithm, is also possible.

Each of the four sampling methods described above can be carried out using the following NAG routines:

- **G05TLF**: Sampling with replacement, equal weights.
- **G05TDF**: Sampling with replacement, unequal weights.
- **G05NDF**: Sampling without replacement, equal weights.
- **G05NEF**: Sampling without replacement, unequal weights.

An example program showing how each of these routines can be used to obtain a sample is available from NAG

Rather than drawing a sample from the whole dataset it is sometimes desirable to take samples from different strata or subpopulations within that dataset, often referred to as stratified sampling. Whilst the NAG library does not include any routines for performing stratified sampling as such, it is straightforward to order the original dataset by the strata and apply one of the above routines to each in turn.