The whSample Package

whSample helps analysts quickly generate statistical samples from Excel or Comma Separated Value (CSV) files and write them to a new Excel workbook. Users have a choice of Simple Random or Stratified Random samples, and a third choice of having each stratum included in a separate worksheet.

ssize

The workhorse function is *sampler*. A helper function, *ssize*, estimates the minimum sample size necessary to achieve statistical requirements using a Normal Approximation of the Hypergeometric Distribution. This distribution spans the probabilities of yes/no-type responses without replacement. These parameters are:

- **N**, the population size.
- **ci**, the required confidence interval. The default is 95%.
- **me**, the required level of precision, or margin of error. The default is +/- 7%.
- **p**, the anticipated rate of occurrence. The default is 50%.

ssize(N, ci=0.95, me=0.07, p=0.50) (showing the defaults) only requires the **N** argument. Used as a standalone, it can be used to explore sample sizes under other conditions. For example, a probe sample may suggest that a 50-50 probability isn't realistic. A revised sample size can be estimated with the observed success probability (p=0.6, for example).

sampler

The *sampler* function calls *ssize* to get its sample size estimate. Therefore, it requires the **ci**, **me**, and **p** argments, which it passes to *ssize*.

sampler also takes two additional arguments:

- **backups** provides a buffer for use if necessary to replace samples found to be invalid for some reason, and
- **seed**, is used to seed the internal random number generator.

The defaults for these arguments are *backups=5* and *seed=NULL*. The default seed will tell *sampler* to use the current system time in milliseconds (a common seeding approach).

To override any of these defaults, enter *name=value* as an argument.

sampler will pop up a file chooser to allow the user to navigate to the source file. It then will pop up a menu to let the user pick the preferred sampling type.

Output

sampler generates two files:

- an Excel spreadsheet with the requested sample, and
- a CSV file called *Sampling Report.csv*. This file lists the:
  - path and name of the source file
  - size (in rows) of the source file
  - sample type (Simple Random Sample, Stratified Random Sample, or Tabbed Stratified Sample)

    - sample size
    - number of backups requested (this number is applied to every stratum in a stratified sample)
    - random number seed used, for documentation and reproducibility
    - date-time stamp of when the sample was generated

Installation

You can install the latest version of whSample from the R console with:

```r
devtools::install_github("km4ivi/R-whSample")
```

Other necessary packages

sampler depends on several external packages to run properly. Ensure these packages are installed:

- tidyverse (or individually: magrittr, dplyr, purrr)
- openxlsx
- data.table
- tools

Examples

ssize(5000): N=5000, other arguments use defaults

ssize(5000, p=0.60): N=5000, with a 60% expected rate of occurrence

sampler(): Uses all defaults, gets N from the source data.

sampler(backups=0, seed=12345): Overrides specific defaults