mwp-471

splitsample - 標本データの分割　【 評価版 】

`splitsample` はデータをいくつかのランダムなサンプルに分割します。

1. 機能概要と用例

`splitsample` はデータをトレーニング用のサンプルとテスト用のサンプルに分割する際に有用な機能を提供します。

▷ Example 1: 観測データ単位の分割

本用例では観測データ (observation) 単位の分割例を紹介します。最初に 101 個の観測データからなるデータセットを生成します。

```
. set obs 101
Number of observations (_N) was 0, now 101.
```

この操作によって 101 個の観測データを含むデータセットが生成されたわけですが、この段階では変数を全く持たないデータセットとなっています。それでも `splitsample` コマンドの実行には何ら支障がありません。最初に標本識別用の変数名のみを指定する形で `splitsample` を実行してみます。

- Data ▷ Create or change data ▷ Other variable-creation commands
  ▷ Split data into random samples と操作
- Main タブ: Generate sample ID variable: `svar`
  Split into # random samples of equal size: • （デフォルト）
  2 （デフォルト）

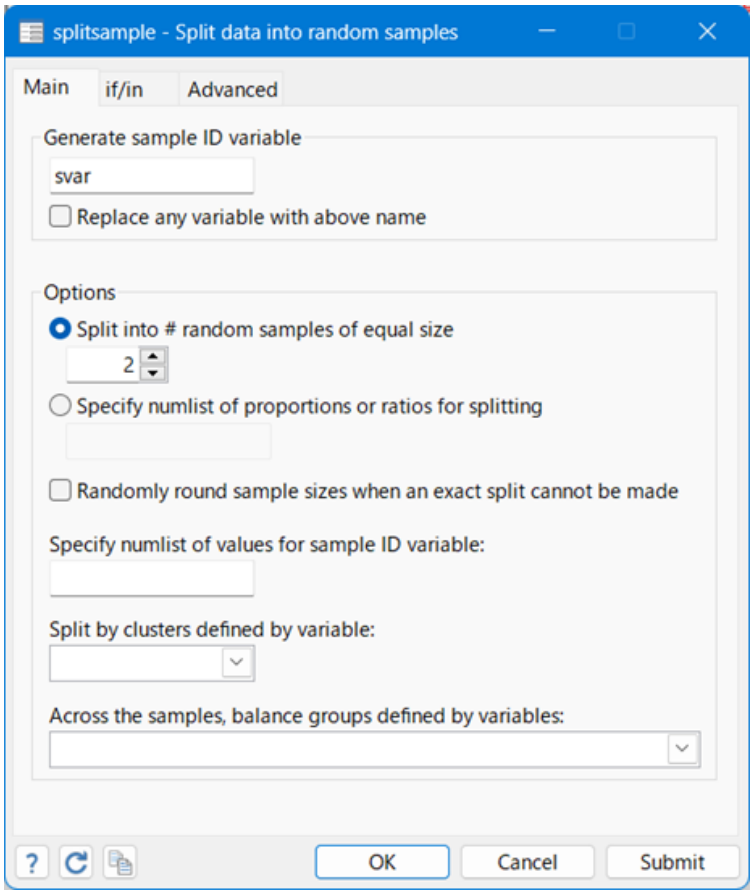

図 1 `splitsample` ダイアログ - Main タブ

```
. splitsample, generate(svar)
```

生成された標本識別用変数 svar の内容（度数）を確認しておきます。

. tabulate svar [*1]

```
. tabulate svar

      svar |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |         51       50.50       50.50
         2 |         50       49.50      100.00
-----------+-----------------------------------
     Total |        101      100.00
```

デフォルトの場合、splitsample はデータを二等分します。この例では観測データ数が奇数であるため、データセットは svar = 1 という識別子を持つデータ 51 件と svar = 2 という識別子を持つデータ 50 件に分割されたことがわかります。

nsplit(#) オプションを使用すると任意の数のサンプルに分割することができます。ここでは 3 分割を行ってみます。

- splitsample ダイアログ:
  - Main タブ: Generate sample ID variable: svar
    - Replace any variable with above name: ✓
    - Split into # random samples of equal size: •（デフォルト）
      - 3

```
. splitsample, generate(svar, replace) nsplit(3)
```

. tabulate svar

```
. tabulate svar

      svar |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |         34       33.66       33.66
         2 |         33       32.67       66.34
         3 |         34       33.66      100.00
-----------+-----------------------------------
     Total |        101      100.00
```

[*1] メニュー操作： Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ One-way table

`split(`*numlist*`)` オプションを使うと分割するサンプルの割合（相対的サイズ）を指定することができます。例えばサンプル 1, 2 はそれぞれ 25% 、サンプル 3 は 50% の大きさとするには次のように操作します。

評価版では割愛しています。

▷ Example 2: **クラスタ単位の分割**

評価版では割愛しています。

▷ Example 3: balance() **オプション**

評価版では割愛しています。

▷ Example 4: **欠損値の扱い**

評価版では割愛しています。

■