*mwp-079*

# サーベイデータの扱い　【 評価版 】

標本データに基づき母集団に関する統計的な推論を行う場合、標本がどのような標本抽出デザインに基づき抽出されたものであるかについて注意を払う必要があります。特に大規模なサーベイデータの場合には単純無作為抽出 (simple random sampling) を仮定できない場合が往々にしてあります。Stata では標本抽出デザインに関連して

- 抽出ウェイト (sampling weight)
- クラスタ抽出 (cluster sampling)
- 層化抽出 (stratification)

に関する特性を設定でき、かつそれを反映した形で推定が行える機能が用意されています。本 whitepaper では複雑なデザインに基づくサーベイデータを分析する上で必要となる基本的な事項について解説します。

# 1. サーベイデータの分析

サーベイデータに対する分析機能は主として svy プリフィックスコマンドによって提供されます。ただし“svy:”というプリフィックスを付けた形で推定コマンドを実行するためには、それに先立つ形で svyset コマンドを発行し、サーベイデータの特性を定義しておく必要があります。

サーベイデータの特性は次の 3 つの側面によって規定されます。

- 抽出ウェイト (sampling weights) — 確率ウェイト (pweight: probability weights) とも呼ばれる
- クラスタ抽出 (cluster sampling)
- 層化抽出 (stratification)

これらの特質はデータ収集プロセスのデザインやその詳細からもたらされるものです。これらの特質がデータの分析方法にどのような影響を及ぼすかについて簡単に記すと次のようになります。

(1) *Sampling weights*
標本調査 (sample surveys) において観測データはランダムなプロセスを通じて選択されるわけですが、選択される確率は観測データごとに異なったものとなることがあります。その場合の重みは標本抽出される確率[*1]の逆数（またはそれに比例した値）に等しいものとなります。この重みに対してさらなる調整 (postsampling adjustments) が施されるケースもあります。$j$ 番目の観測データに対する重みが $w_j$ である場合、（多少厳密さは欠けますが）$j$ 番目の観測データは母集団中における $w_j$ 個の要素を代表しているととらえることができます。
重みを無視して分析を行った場合には推定結果にバイアスが生じることになります。Sampling weights は標準誤差の推定にも関係してきます。

(2) *Clustering*
ほとんどのサーベイデザインにおいて個人は独立に標本抽出されるわけではありません。個人の集合体（例えば郡や世帯、等）がグループ（クラスタ）として抽出されるのが普通です。
クラスタ内でさらに subsampling が行われることもあります。例えば郡 (counties) が sampling された後、郡内の街区、街区内の世帯が次々に sampling され、最終的に世帯内の個人に到達するといったケースが考えられます。Sampling の第 1 レベルに位置するクラスタは PSU (primary sampling unit) と呼ばれます。この例で言うなら郡が PSU となります。クラスタリングが行われない場合には個人 — サイズが 1 のクラスタ — が PSU として定義されます。
クラスタを伴う sampling は個人を直接 sampling する場合に比べ、標本間の変動 (sample-to-sample variability) がより大きなものとなるのが普通です。この変動の増分は標準誤差の推定や仮説検定等の推論に際して考慮されなくてはなりません。

(3) *Stratification*
サーベイにおいてクラスタを構成するグループが異なる形で sampling される場合があります。これらのグループは層 (strata) と呼ばれます。例えばある州における 254 の郡を都市部と農村部の 2 層に区分して扱い、都市層からは 10 個の郡を、農村層からは 15 個の郡を抽出するといったケースがそれに該

[*1] 標本抽出される確率は stratification と clustering の構成によって導かれます。

当します。
Sampling は層間で独立に、また層の区分はあらかじめ固定された形で行われます。従って層は統計的に独立なものとして分析することができます。個々の層が全体としての母集団に比べより均質 (homogeneous) である場合にはそれを利用して、標準誤差のより小さな推定値を得ることが可能です。

以上を要約すると次のようになります。

- Sampling weights を用いることは正しい点推定値を得る上で大切なことです。
- 標準誤差を正しく推定するためにはサーベイデザインに由来する重み付け、クラスタリング、層化について考慮する必要があります。
- サーベイデザインに伴うクラスタリングを無視した場合には本来よりも小さめな標準誤差推定値が得られる公算が強くなります。
- 層化抽出は与えられた標本サイズを前提としたときにより小さな標準誤差推定値を得ることを可能にします。

サーベイデータの分析については Cochran (1977); Heeringa, West, and Berglund (2017); Kish (1965); Levy and Lemeshow (2008); Scheaffer et al. (2012); Skinner, Holt, and Smith (1989); Stuart (1984); Thompson (2012); Williams (1978) を参照ください。

## 2. サーベイデザインツール

svy による推定を行うためには最初に `svyset` を実行する必要があります。`svyset` コマンドはサーベイデザインの特質を規定する変数を特定すると共に、標準誤差を推定するためのデフォルトの手法を設定します。

### ▷ Example 1: 単段階デザイン

単段階のサーベイデザイン (single-stage survey design) の場合には複数の層 (strata) を横断する形でクラスタリングを伴うサンプリング[*2]が行われます。このデザインに基づくデータセットの場合には

- 層 (strata)
- PSU（クラスタ）
- 抽出ウェイト (sampling weights))
- 有限母集団修正 (FPC: finite population correction)

を規定する変数名を平板的に指定します。次に示すのは Example データセット `stage5a.dta` を用いたときの例です。

```
. use http://www.stata-press.com/data/r16/stage5a.dta
```
[*3]

このデータセットの場合、層は変数 `strata` により、PSU は変数 `su1` によって規定されるわけですが、それらの組合せによって 2 元の度数分布表を作成してみると次のようになります。

[*2] 非復元抽出 (sampling without replacement) によるサンプリング。
[*3] メニュー操作：File ▷ Example Datasets ▷ Stata 16 manual datasets と操作、Survey Data Reference Manual [SVY] の Survey の項よりダウンロードする。

```
. tabulate strata su1
```
[*4]

```
. tabulate strata su1

           |                    su1
    strata |         1         2         3         4 |     Total
-----------+--------------------------------------------+----------
         1 |     1,188     1,192         0     1,258 |     3,638
         2 |     1,289     1,277     1,201         0 |     3,767
         3 |         0     1,268     1,152     1,214 |     3,634
-----------+--------------------------------------------+----------
     Total |     2,477     3,737     2,353     2,472 |    11,039
```

これよりそれぞれの層ごとに 3 つのクラスタが抽出されていることがわかります。strata と su1 のそれぞれの組合せごとに約 1,200 の観測データが記録されているわけですが、ここでは参考までに各組ごとの先頭レコードをリスト出力しておきます。なお、変数 pw は sampling weight(probability weight) を規定する変数であり、fpc1 は FPC を規定する変数です。

```
. by strata su1: generate flag=1 if _n==1
. list strata su1 pw fpc1 if flag==1, sepby(strata)
```

```
     +-------------------------------------+
     | strata   su1         pw   fpc1 |
     |-------------------------------------|
  1. |      1     1   61.11111      4 |
1189. |      1     2   55.55556      4 |
2381. |      1     4   54.01234      4 |
     |-------------------------------------|
3639. |      2     1    40.6746      4 |
4928. |      2     2   34.97942      4 |
6205. |      2     3   74.07407      4 |
     |-------------------------------------|
7406. |      3     2   56.66667      4 |
8674. |      3     3       37.5      4 |
9826. |      3     4   42.85714      4 |
     +-------------------------------------+
```

このデータセットに対する svyset の指定は次のようになります。

- Statistics ▷ Survey data analysis ▷ Setup and utilities ▷ Declare survey design for dataset と操作

[*4] メニュー操作： Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Two-way table with measures of association

- Main タブ: Number of stages: `1`
  Stage 1: Primary sampling units: `su1`
  Strata: `strata`
  Finite pop. correction: `fpc1`

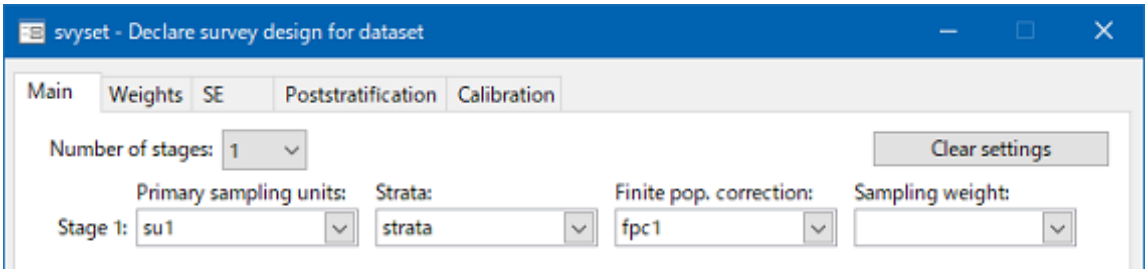


図 1 `svyset` ダイアログ - Main タブ

- Weights タブ: Sampling weight variable: `pw`

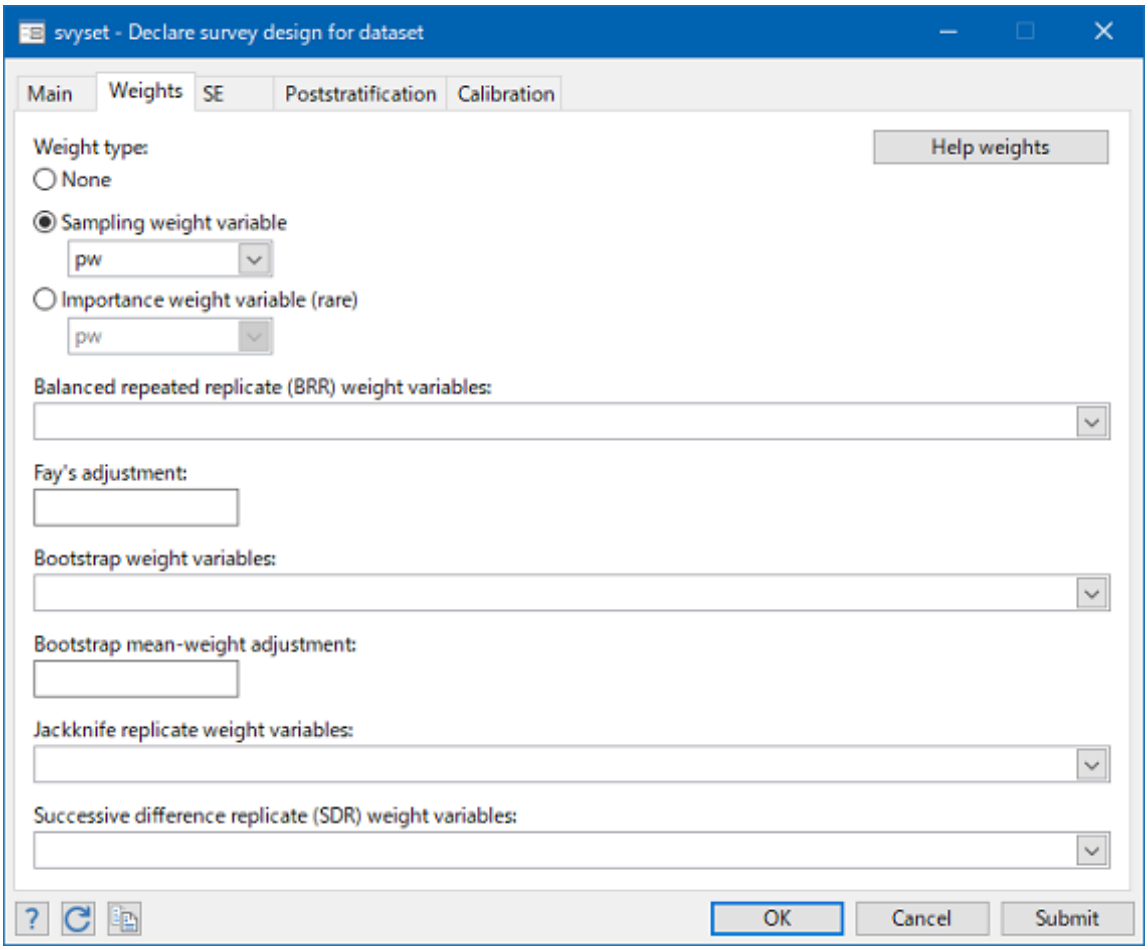


図 2 `svyset` ダイアログ - Weights タブ

```
. svyset su1 [pweight=pw], strata(strata) fpc(fpc1) vce(linearized) singleunit
> (missing)

      pweight: pw
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: su1
        FPC 1: fpc1
```

ダイアログ上で指定したのは 4 つの変数名のみであったわけですが、生成されたコマンド中には次のようなデフォルト設定も含まれています。

- 標準誤差の推定には Taylor linearization を用いる。
- 層中に 1 つの sampling unit しか存在しなかった場合（singleton strata とも呼ばれる）には標準誤差として欠損値を応答する。

◁

▷ Example 2: **多段階デザイン**

評価版では割愛しています。

▷ Example 3: svydescribe

評価版では割愛しています。

## 3. サーベイデータ分析ツール

評価版では割愛しています。

▷ Example 4: **母集団平均の推定**

評価版では割愛しています。

▷ Example 5: **線形回帰**

評価版では割愛しています。

▷ Example 6: Cox **の比例ハザードモデル**

評価版では割愛しています。

▷ Example 7: 2 **元分布表**

評価版では割愛しています。

▷ Example 8: **平均値の比較**

評価版では割愛しています。

▷ Example 9: **デザイン効果**

評価版では割愛しています。

▷ Example 10: BRR **と複製重み変数**

評価版では割愛しています。

## 4. **サーベイデータに関する留意点**

評価版では割愛しています。

▷ Example 11: **部分母集団推定**

評価版では割愛しています。

▷ Example 12: **割合の標準化**

評価版では割愛しています。

▷ Example 13: **事後層化**

評価版では割愛しています。

## 5. **プログラマ向け機能**

本セクションについては英文マニュアルをご参照ください。

■