第 1.0 版：2009.08.25

『概説 標本調査法』 付 録

朝 倉 書 店

『概説 標本調査法』付録

土屋 隆裕 著

朝倉書店　第 1.0 版：2009.08.25

本資料は『概説 標本調査法』（土屋隆裕著，2009 年，朝倉書店）の付録として利用されることを目的としています．目的外の利用はご遠慮ください．

また，本資料中のプログラム等の実行により万が一損失や損害，障害等が発生しても，著者および出版社は一切の責任を負いません．読者自身の責任において本資料を利用されるようお願いいたします．

目 次

第1章　はじめに

1.a　この資料について

この資料 (以下，本資料) は『概説 標本調査法 (土屋隆裕著, 2009 年, 朝倉書店)』(以下，本書) の補足資料であり，主に R (`http://www.r-project.org/`) を用いた標本調査データの分析法を解説したものである[1]．本資料の章立ては，この第 1 章と第 14 章を除けば本書の章立てに対応している (ただし章の中の節は対応していない)．また例題番号や表番号，数式番号も本書における番号であり，p. で表すページ数も本書のページである．

本資料には，本書で扱うデータよりも大きなサイズの標本を用いた演習問題とその解答例も含まれている．本資料の例題をなぞってみた後には，独力で演習問題にも取り組んでみるとよい．

1.b　表記

本資料では，R の関数やプログラム・出力はタイプライタ体のフォント (`svydesign()` など) で示してある．関数の指定方法の説明などでは，変数名やオブジェクト名といった分析者が必要に応じて変える部分をスモールキャップス (DATA など) で示してある．

1.c　分析のための準備

R を用いて標本調査データを分析するには以下の準備が必要である．本資料では R のインストール方法や基本操作に関しては説明していない．他の文献や R のヘルプを参照のこと．

`survey` パッケージ

複雑な標本抽出デザインに従って得られた調査データを扱うために，本資料では `survey` パッケージ[2]を利用する．標準では `survey` パッケージはインストールされないため，CRAN から別途インストールしておく必要がある．インストール方法は他の文献や R のヘルプを参照のこと．本資料はバージョン `3.13` に基づく[3]．

[1] 今後さらに SUDAAN (Research Triangle Institute) など市販のソフトウェアの解説も追加する予定である．
[2] Lumley, T. (2009). “survey: analysis of complex survey sample.” R package version 3.13.
[3] 現在のバージョンは `3.16` である．追加された機能に関しては，本資料にも追って加えていく予定である．

分析対象のデータ

本資料の例題では，関数 data.frame() を用いてデータフレームを作成しているが，データはファイルから読み込むのが一般的であろう．例えば第 14 章には本資料の演習問題で用いる 3 つのデータフレーム kigyo，kenko，otona の内容を説明してある．データは本資料とともに配付しているファイル exercise.rda に保存されている．ファイルを置いたディレクトリに移動し[4]，load('exercise.rda') とすることで読み込むことができる．その他の読み込み方法は他の文献等を参照のこと．

例題 1.1-1

```
> load('exercise.rda')
```

分析プログラム

分析は R GUI を用いて対話的に行うことができる．あるいはエディタなどを使って分析プログラムをあらかじめファイル ABC.R などに保存しておき，source('ABC.R') とすることで分析を行うこともできる．

1.d 分析の流れ

1. survey **パッケージを読み込む**

   パッケージを読み込むには library() を用いる．パッケージは各分析セッションのはじめに一度だけ読み込めばよい．

   例題 1.1-2

   ```
   > library(survey)
   ```

2. **データを用意する**

   load() などを用いて分析対象のデータを用意する．以下の例では，データを読み込む前にワークスペースを消去した上で，本資料の演習問題で用いる 3 つのデータフレームを読み込んでいる．

   例題 1.1-3

   ```
   ##### ワークスペースの消去 #####
   > rm(list=ls())

   ##### データの読み込み #####
   > load('exercise.rda')
   ```

[4] Windows 版ではメニューの「ファイル」–「ディレクトリの変更...」を選ぶ．Mac 版ではメニューの「その他」–「作業ディレクトリの変更...」を選ぶ．

3. **標本抽出デザインを指定する**

標本抽出デザインの指定には関数 svydesign() を用いる．以下の例では，データフレーム kigyo に対して非復元単純無作為抽出法を指定している．標本抽出デザインの具体的な指定方法は次章以降で説明する．

分析・集計のたびに，svydesign() を用いて抽出デザインを指定することも可能である．しかし分析対象のデータはある特定の標本抽出デザインに従って得られたものであるため，指定結果を survey.design (正確には survey.design2) クラスのオブジェクトとして残しておき，これを繰り返し利用する方が効率的である．以下の例では，svydesign() の結果を des に代入している．

得られた survey.design クラスのオブジェクトは，save() を使ってファイル (以下の例では kigyo.rda) に保存しておくと，後日再利用するときに便利である．その際，元のデータフレーム (以下の例では kigyo) は survey.design クラスのオブジェクト中にコピーされているため，必ずしも同時に保存しておく必要はない．des だけで十分である．

例題 1.1-4

```
##### 標本抽出デザインの指定 #####
> des <- svydesign(ids=~1, fpc=~N, data=kigyo)

##### desの保存 #####
> save(des, file='kigyo.rda')
```

4. **分析方法を指定する**

母集団総計の推定は svytotal()，母集団平均の推定は svymean() など目的に応じた関数を用いて分析結果を得る．その際，関数の引数として survey.design クラスのオブジェクトと分析に用いる変数などを指定する．以下の例では survey.design クラスのオブジェクト des を用いて，変数 shihon と uriage の母集団総計を推定している．分析結果をオブジェクトとして保存しておき，さらに分析を加えたり適当な形式で出力することも可能である．

例題 1.1-5

```
##### 母集団総計の推定 #####
> svytotal(~shihon + uriage, des)
         total    SE
shihon 1236645 29481
uriage 1008365 21488
```

第2章　線形推定量

一般に HT 推定量では，その分散を推定するために標本の要素間の全ての組み合わせについて二次の包含確率 π_{ij} が必要である．survey パッケージは，π_{ij} を全て指定する推定方法には未だ対応していない[1]．そこでこの第 2 章では復元抽出法に基づく HH 推定量のみを取り上げ，svydesign() における抽出ウェイト w_i の指定方法と，母集団総計 τ_y の線形推定量の求め方を解説する．

2.a　抽出ウェイトの指定

抽出ウェイトの指定

```
DES <- svydesign(ids=~1, weights=~W, data=DATA)
```

標本抽出デザインを指定するには関数 svydesign() を用いる．引数 data の DATA にはデータが含まれるデータフレーム，引数 weights の W には抽出ウェイト w_i の変数を指定する．引数 ids は抽出単位を指定するためのものであり，集落抽出法を扱う第 8 章や多段抽出法を扱う第 9 章で説明する．引数 weights の指定がなく，引数 ids と data しか指定されていないと，自動的に全ての要素の抽出ウェイトが $w_i = 1$ とされる[2]．DES は survey.design クラスのオブジェクトである．

survey.design オブジェクトの確認

```
summary(DES)
```

svydesign() によって指定した標本抽出デザインの内容を確認するには関数 summary() を用いる．引数 DES には survey.design クラスのオブジェクトを指定する．複雑な標本抽出デザインでは特に，意図した指定がなされているか summary() を利用して確認するとよい．

[1] 本資料が基づくバージョン 3.13 では対応していなかったが，最新のバージョン 3.16 では対応している．この点については今後解説を追加していく予定である．なお，SUDAAN の UNEQWOR では π_{ij} を全て指定した推定が可能となっている．

[2] $w_i = 1$ であっても引数 weights は明示的に指定する方がよい．

ウェイトの取り出し

```
weights(DES)
```

survey.design クラスのオブジェクトから標本の各要素に与えられているウェイトを取り出すには，関数 weights() を用いる．

例題 2.8　復元抽出における抽出ウェイト

本書の表 2.4 の標本 (p.34) に対して復元抽出法を指定してみよう．以下の例では，まず data.frame() を用いて表 2.4 のデータから成るデータフレーム data を作成している．データフレームの変数 y は各企業の売上高 y_i であり，変数 w は各企業の抽出ウェイト w_i である．表 2.4 の標本は，表 1.1 の母集団から資本金 x_i で確率比例抽出されたものであるので，各企業の抽出確率は $p_i = x_i/\tau_x$ であり，抽出ウェイトは $w_i = 1/(np_i) = 1/(nx_i/\tau_x)$ となる．

例題 2.8-1

```
##### データの作成 #####
> (data <- data.frame(y=c(576, 576, 74), w=1/(3*c(47/159, 47/159, 25/159))))
    y        w
1 576 1.127660
2 576 1.127660
3  74 2.120000
```

次に関数 svydesign() を用いて標本抽出デザインを指定する．引数 weights には作成した抽出ウェイトの変数 w を指定し，引数 data にはデータフレーム data を指定している．svydesign() による指定結果は survey.design クラスのオブジェクト wr に代入している．

例題 2.8-2

```
##### 復元抽出法の指定 #####
> wr <- svydesign(ids=~1, weights=~w, data=data)
```

以下の例では，上記の syvdesign() による指定結果 wr を関数 summary() を用いて確認している．表示される結果の一行目には (with replacement) とあり，復元抽出法が指定されていることが分かる．Probabilities:は抽出ウェイト w_i の逆数の分布を示すものである．

例題 2.8-3

```
##### survey.designオブジェクトの確認 #####
> summary(wr)
Independent Sampling design (with replacement)
svydesign(ids = ~1, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4717  0.6792  0.8868  0.7484  0.8868  0.8868
Data variables:
[1] "y" "w"
```

weights() を用いて抽出ウェイト w_i の値を表示させると，例えば最初の企業は 1.127660 となっている．本書の表 2.4 (p.34) に示された抽出ウェイトの値 $w_i = 1.13$ は，この値を四捨五入したものであることに注意すること．

例題 2.8-4

```
##### 抽出ウェイト #####
> weights(wr)
       1        2        3
1.127660 1.127660 2.120000
```

2.b 母集団総計の線形推定

母集団総計の線形推定

```
STAT <- svytotal(x=~Y, design=DES, na.rm=FALSE, deff=FALSE)
```

母集団総計の線形推定値を求めるには関数 svytotal() を用いる．引数 x の Y には推定対象の変数 y を指定する．複数の変数を同時に指定するときには変数名を+でつないで並べればよい．引数 design の DES には survey.design クラスのオブジェクトを指定する．syytotal() は DES に含まれるウェイト w_i と変数 y_i を使って，本書の (2.21) 式に従い母集団総計の推定値を求める．また推定量の分散は，一般的な抽出ウェイト w_i を用いた本書の (3.7) 式による．変数 Y に欠損値 (NA) が含まれていると推定値も NA となる．欠損値を除いて推定するには na.rm=TRUE と指定する．詳細は**演習 3.6** を参照のこと．引数 deff については 4.b 節を参照のこと．STAT は svystat クラスのオブジェクトである．

推定値の取り出し

```
coef(STAT)
```

svytotal() によって得られた svystat クラスのオブジェクト STAT には様々な情報が含まれている．その中から推定値だけを取り出すには関数 coef() を用いる．関数 coef() は母集団総計を推定する svytotal() の結果だけでなく，母集団平均を推定する svymean() の結果などにも用いることができる．

標準誤差の取り出し

SE(STAT)

推定値を取り出すのと同様に，svystat クラスのオブジェクト STAT から標準誤差だけを取り出すには関数 SE() を用いる．

例題 2.10　HH 推定量の分散の推定

例題 2.8-2 で復元抽出法を指定した wr を用いて，売上高 y の母集団総計の推定値 $\hat{\tau}_y$ とその分散の推定値 $\hat{V}(\hat{\tau}_y)$ を求めてみよう．以下の例では，svytotal() の最初の引数に目的とする変数 y，二番目の引数に survey.design クラスのオブジェクト wr を指定している．母集団総計の推定値は $\hat{\tau}_y = 1455.9$，標準誤差の推定値は $\widehat{SE}(\hat{\tau}_y) = 492.65$ となる．これらは本書の (2.22) 式と (2.26) 式に対応する．なお本書の値は四捨五入したものであることに注意すること．

例題 2.10-1

```
##### 母集団総計の推定 #####
> svytotal(~y, wr)
   total     SE
y 1455.9 492.65
```

svytotal() の結果にはいろいろな情報が含まれるが，その中から推定値のみを取り出すには関数 coef() を用いればよい．

例題 2.10-2

```
##### 母集団総計の推定値 #####
> coef(svytotal(~y, wr))
       y
1455.944
```

同様に標準誤差のみを取り出すには関数 SE() を用いる．

例題 2.10-3

```
##### 推定値の標準誤差 #####
> SE(svytotal(~y, wr))
       y
492.6519
```

2.b.1　演習問題

演習 2.1　復元抽出法の指定と母集団総計の線形推定

本資料の 14.a 節に示すデータフレーム kigyo が復元抽出されたものとして，資本金 shihon と売上高 uriage の母集団総計を線形推定してみよう．ただし抽出ウェイトはどの企業も $w_i = 5$ とする．

ヒント：データフレーム kigyo は本資料とともに配付しているファイル exercise.rda に含まれている．まず抽出ウェイトの値が代入された変数をデータフレーム kigyo に追加する．例えば変数名を w とするのであれば，`kigyo$w <- 5` とすればよい．

2.c　演習問題解答例

演習 2.1 解答例

まずファイル exercise.rda を読み込み，必要なデータフレームを用意する．なお，この**演習 2.1** 以外の演習問題の解答例では，同様にして既に必要なデータフレームが読み込まれたことを前提とする．また 1.d 節で説明したように，既に survey パッケージは読み込まれているものとする．

演習 2.1 解答例 -1

```
##### データフレームの読み込み #####
> load('exercise.rda')
```

まず抽出ウェイト $w_i = 5$ が代入された変数 w を用意する．以下の例では，データフレーム kigyo に全ての企業が 5 という値の変数 w が作成される．

演習 2.1 解答例 -2

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- 5
```

以下の例では関数 svydesign() を用いて標本抽出デザインを指定し，その結果を wr に代入している．関数 svydesign() の最初の引数 ids には~をつけて 1 を指定する．引数 weights には抽出ウェイトの変数 w を指定する．最後に引数 data にはデータフレーム kigyo を指定する．

演習 2.1 解答例 -3

```
##### 復元抽出法の指定 #####
> wr <- svydesign(ids=~1, weights=~w, data=kigyo)
```

関数 summary() を用いて wr の内容を確認すると，復元抽出法が指定されていることが分かる．Probabilities:の値は全て 0.2 であり，これは抽出ウェイト $w_i = 5$ の逆数である．

演習 2.1 解答例 -4

```
##### survey.designオブジェクトの確認 #####
> summary(wr)
Independent Sampling design (with replacement)
svydesign(ids = ~1, weights = ~w, data = kigyo)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.2     0.2     0.2     0.2     0.2     0.2
Data variables:
 [1] "obs"            "area"           "gyoshu"         "shihon"
 [5] "uriage"         "uriage.na"      "N"              "n"
 [9] "N.h"            "n.h"            "total.shihon"   "total.shihon.h"
[13] "w"
```

関数 weights() を用いると，survey.design クラスのオブジェクトからウェイトを取り出すことができる．以下の例では wr に含まれる抽出ウェイトの分布を調べている．全ての企業の抽出ウェイトが $w_i = 5$ となっている．

演習 2.1 解答例 -5

```
##### 抽出ウェイトの分布 #####
> summary(weights(wr))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5       5       5       5       5       5
```

資本金 shihon と売上高 uriage の母集団総計を推定するには，関数 svytotal() の最初の引数にこれらの変数名を+でつないで並べればよい．また第二の引数には survey.design クラスのオブジェクト wr を指定する．資本金の母集団総計の推定値は $\hat{\tau}_{\text{資本金}} = 1236645$ となり，標準誤差の推定値は $\widehat{SE}(\hat{\tau}_{\text{資本金}}) = 32960$ となる．以下の例では，推定結果を tau.y に代入している．

演習 2.1 解答例 -6

```
##### 母集団総計の線形推定 #####
> (tau.y <- svytotal(~shihon + uriage, wr))
         total    SE
shihon 1236645 32960
uriage 1008365 24024
```

関数 svytotal() の結果 tau.y から推定値だけを取り出すには，関数 coef() を用いればよい．

演習 2.1 解答例 -7

```
##### 母集団総計の推定値 #####
> coef(tau.y)
 shihon  uriage
1236645 1008365
```

また関数 svytotal() の結果 tau.y から標準誤差だけを取り出すには，関数 SE() を用いればよい．

演習 2.1 解答例 -8

```
##### 母集団総計の推定値の標準誤差 #####
> SE(tau.y)
  shihon   uriage
32960.48 24024.00
```

第3章　単純無作為抽出法

3.a　単純無作為抽出法の指定

単純無作為抽出法の指定

```
DES <- svydesign(ids=~1, fpc=~N, weights=~W, data=DATA)
```

単純無作為抽出法では，抽出ウェイトは全ての要素について $w_i = N/n$ となる．そこで引数 weights の W には，データフレーム DATA の変数のうち，抽出ウェイト $w_i = N/n$ が入った変数を指定すればよい．

復元単純無作為抽出法と非復元単純無作為抽出法とでは，母集団総計の推定量 $\hat{\tau}_y$ は同じであるが，その分散の推定量 $\hat{V}(\hat{\tau}_y)$ は異なり，非復元単純無作為抽出法では有限母集団修正項 $1 - f = 1 - n/N$ が乗じられる．そこで非復元単純無作為抽出法の場合には引数 fpc の N には母集団サイズ N が代入されている変数を指定する[1]．復元単純無作為抽出法の場合には引数 fpc は指定しない[2]．つまり引数 fpc の指定の有無で非復元と復元とを区別する．なお，引数 fpc に指定された母集団サイズ N と引数 weights に指定された抽出ウェイトの合計 $\sum_s w_i$ とが異なっていたとしても，特にメッセージは表示されない．引数 fpc は基本的に有限母集団修正項 $1 - f = 1 - n/N$ の計算のためだけに用いられる．

引数 fpc が指定され，かつ引数 weights が無指定の場合には，(層ごとに) 全ての要素に対して等しい抽出ウェイト $w_i = N/n$ が自動的に与えられる．N は引数 fpc で指定された変数値であり，n は DATA に含まれるデータの件数である．この機能は慣れてきたら使うと便利である．

例題 3.3　抽出ウェイトによる推定

表 3.2 (p.44) のデータに対して非復元単純無作為抽出法を指定し，売上高の母集団総計の推定を行ってみよう．以下の例では，まず data.frame() を用いて表 3.2 のデータに対応するデータフレーム data を作成している．変数 y は各企業の売上高 y_i である．変数 N と変数 n

[1] あるいは抽出率 $f = n/N$ が代入されている変数を指定してもよい．指定された変数の値が 1 よりも大きい場合には N が指定されたものと解釈され，1 以下の場合には $f = n/N$ が指定されたものと解釈される．両者が混ざっているとエラーとなる．$n > N > 1$ となる N が指定された場合にもエラーとなる．

[2] あるいは値が Inf である変数を指定する．

は，それぞれ母集団サイズ $N = 20$ と標本サイズ $n = 3$ を表す変数であり，3 つの企業について同じ値が入っている．

例題 3.3-1

```
##### データの作成 #####
> (data <- data.frame(y=c(380, 639, 209), N=20, n=3))
    y  N n
1 380 20 3
2 639 20 3
3 209 20 3
```

次に抽出ウェイト $w_i = N/n$ を求め，これを変数 w に代入している．単純無作為抽出法であるので，抽出ウェイトはどの企業も同じ $w_i = 20/3 = 6.666667$ である．当然のことながら，関数 data.frame() を使ってデータフレームを作成するときに，w=20/3 などとして抽出ウェイトを作成しておいてもよい．

例題 3.3-2

```
##### 抽出ウェイトの作成 #####
> (data$w <- data$N / data$n)
[1] 6.666667 6.666667 6.666667
```

以下の例では，上記で作成したデータフレーム data に対して関数 svydesign() を用いて標本抽出デザインを指定している．非復元単純無作為抽出法なので引数 fpc には母集団サイズ N の値が代入された変数 N を指定する．また引数 weights には抽出ウェイトの変数 w を指定する．svydesign() による標本抽出デザインの指定結果は survey.design クラスのオブジェクト si に代入している．

例題 3.3-3

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, weights=~w, data=data)
```

関数 summary() を使って survey.design オブジェクト si の内容を確かめると，**例題 2.8-3** とは異なって (with replacement) と表示されず，非復元抽出法が指定されていることが分かる．Probabilities:に示される値は抽出ウェイトの逆数 $1/w_i = n/N = 3/20 = 0.15$ である．Population size (PSUs):に示された値 20 は，syvdesign() の引数 fpc に指定した母集団サイズ $N = 20$ である．

例題 3.3-4

```
##### survey.designオブジェクトの確認 #####
> summary(si)
Independent Sampling design
svydesign(ids = ~1, fpc = ~N, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.15    0.15    0.15    0.15    0.15    0.15
Population size (PSUs): 20
Data variables:
[1] "y" "N" "n" "w"
```

以下の例では関数 syvtotal() を用いて，売上高の母集団総計の推定値 $\hat{\tau}_y$ とその標準誤差の推定値 $\widehat{SE}(\hat{\tau}_y)$ を求めている．これらは本書の (3.9) 式と (3.10) 式に対応する．繰り返すが，本書では四捨五入により丸めた値を示していることに注意すること．

例題 3.3-5

```
##### 母集団総計の推定 #####
> svytotal(~y, si)
   total     SE
y 8186.7 2304.8
```

本書の**例題 3.3** (p.44) では，表 3.2 のデータが復元単純無作為抽出されたものとして売上高の推定値とその分散を求めている．ここでも同じことをしてみよう．以下の例ではまず，引数 fpc を指定せずに svydesign() を使って標本抽出デザインを指定している．用いたデータ data は非復元単純無作為抽出法を指定したときと同じものであり，抽出ウェイトの変数 w にも $w_i = 20/3$ が入っている．指定結果は sir に代入している．

例題 3.3-6

```
##### 復元単純無作為抽出法の指定 #####
> sir <- svydesign(ids=~1, weights=~w, data=data)
```

次に関数 summary() によって，復元単純無作為抽出法を指定した sir の内容を確かめると，(with replacement) と表示され，確かに復元抽出法が指定されていることが分かる．また Population size (PSUs): は表示されない．引数 fpc を指定しなかったためである．

例題 3.3-7

```
##### survey.designオブジェクトの確認 #####
> summary(sir)
Independent Sampling design (with replacement)
svydesign(ids = ~1, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.15    0.15    0.15    0.15    0.15    0.15
Data variables:
[1] "y" "N" "n" "w"
```

以下の例では復元単純無作為抽出法を指定した sir を用いて，売上高の母集団総計とその標準誤差の推定値を求めている．推定値 $\hat{\tau}_y = 8186.7$ は非復元単純無作為抽出法のときと同じ値となる．しかし標準誤差は $\widehat{SE}(\hat{\tau}_y) = 2499.9$ となって，非復元単純無作為抽出法のときよりも大きくなる．この結果は本書の (3.11) 式に対応する．

例題 3.3-8

```
##### 母集団総計の推定 #####
> svytotal(~y, sir)
   total     SE
y 8186.7 2499.9
```

3.a.1　演習問題

演習 3.1　復元単純無作為抽出法の指定と母集団総計の推定

データフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団総計を推定してみよう．ただし標本は “復元” 単純無作為抽出されたものとする．

ヒント：まず抽出ウェイト $w_i = N/n$ が代入された変数をデータフレーム kigyo に追加する．母集団サイズは $N = 10,000$ であり，標本サイズは $n = 2,000$ である．

演習 3.2　非復元単純無作為抽出法の指定と母集団総計の推定

同じくデータフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団総計を推定してみよう．ただし標本は “非復元” 単純無作為抽出されたものとする．推定値 $\hat{\tau}_y$ や標準誤差 $\widehat{SE}(\hat{\tau}_y)$ の大きさを，“復元” 単純無作為抽出を指定した場合と比べてみよう．また標本抽出デザインを指定した survey.design オブジェクトから抽出ウェイトを取り出し，その標本総計を求めて，この値が何を意味するのか確認すること．

ヒント：“非復元” 抽出なので，関数 svydesign() で標本抽出デザインを指定するときには引数 fpc が必要である．引数 fpc が指定されていれば引数 weights は必ずしも必要ない．また survey.design オブジェクトから抽出ウェイトを取り出すには関数 weights() を用いればよい．

演習 3.3　復元単純無作為抽出法の指定

データフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団総計を推定してみよう．ただし標本は “復元” 単純無作為抽出されたものとし，関数 svydesign() では引数 fpc を指定すること．**演習 3.1** の結果と標準誤差を比較してみよう．この方法は，層化抽出法において復元抽出をした層と非復元抽出をした層が混在しているとき役立つ．

ヒント：引数 fpc には全ての企業の値が Inf である変数を指定すればよい．

3.b　部分母集団に関する推定

部分母集団の指定

```
DES.SUB <- subset(DES, SUBSET, ...)
```

部分母集団に関する推定を行うときには，まず標本全体に対して関数 svydesign() を用いて標本抽出デザインを指定し，次に関数 subset() を用いて目的とする部分母集団を指定した survey.design クラスのオブジェクトを作成する．そして例えば関数 svytotal() などを用いて推定を行うのが一つの方法である．

subset() は部分母集団を指定した新たな survey.design クラスのオブジェクトを作成する関数である．ある特定の部分母集団に関心があるときに用いる．第一の引数である DES には標本全体に対して標本抽出デザインを指定した survey.design クラスのオブジェクトを指定する．すなわち svydesign() の結果を指定する．SUBSET には部分母集団を定義する表現を指定する．具体的な指定方法は例題を参照のこと．結果の DES.SUB は survey.design クラスの新たなオブジェクトである．

部分母集団ごとの推定

```
STAT <- svyby(~Y, by=~SUBSET, design=DES, FUN=FUN)
```

svyby() を用いると，互いに排反な複数の部分母集団に関して同時に推定を行うことができる．例えば男性と女性のそれぞれについて部分母集団総計を推定する場合などである．つまり svyby() を用いれば，subset() を繰り返し用いる必要はない．引数 by の SUBSET には部分集団を表す変数を指定する．引数 design の DES には svydesign() の結果などを指定する．引数 FUN の FUN には svytotal や svymean などの関数名，Y には FUN に渡す変数名を指定する．

部分集団ごとのウェイト合計

```
TAB <- svytable(~SUBSET, design=DES)
```

本来 svytable() はクロス表を推定するための関数である．しかし SUBSET に部分集団を表す変数を指定することで，部分集団ごとのウェイトの標本合計を求めることができる．つまり次式を使って，部分母集団のサイズ N_d の推定値 $\hat{N}_d$ を求めることができる．

$$\hat{N}_d = \sum\nolimits_s w_i \delta_{d,i}$$

ただし $\delta_{d,i}$ は本書の (3.17) 式のとおりである．

例題 3.4　部分母集団総計の推定

非復元単純無作為抽出された表 3.5 (p.50) の標本を使って，市部の企業の売上高総計 $\tau_{y,\,市}$ を推定してみよう．以下の例では表 3.5 のデータに対応して，各企業の売上高が代入された変数 y，母集団サイズが代入された変数 N，企業の所在地が代入された変数 area から成るデータフレーム data を作成している．

例題 3.4-1

```
##### データの作成 #####
> (data <- data.frame(y=c(380, 639, 209), N=20, area=c('市', '郡', '郡')))
    y  N area
1 380 20   市
2 639 20   郡
3 209 20   郡
```

次に関数 svydesign() を用いて非復元単純無作為抽出法を指定している．非復元抽出となるのは引数 fpc が指定されているからである．引数 weights を指定していないが，引数 fpc に指定した変数 N には 20 が代入されているので，この値とデータの件数 3 から抽出ウェイトは自動的に $w_i = N/n = 20/3$ と計算される．

例題 3.4-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

関数 summary() を用いて suvey.design オブジェクト si の内容を確認すると，抽出ウェイトの逆数は確かに $1/w_i = n/N = 3/20 = 0.15$ となっている．

例題 3.4-3

```
##### survey.designオブジェクトの確認 #####
> summary(si)
Independent Sampling design
svydesign(ids = ~1, fpc = ~N, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.15    0.15    0.15    0.15    0.15    0.15
Population size (PSUs): 20
Data variables:
[1] "y"    "N"    "area"
```

以下の例では市部の企業に関して推定を行うため，まず subset() を用いて si. 市を作成している．第一の引数に指定するのは上記で標本抽出デザインを指定した si であり，第二の引数では変数 area の値が' 市' である企業を部分集団として指定している．

例題 3.4-4

```
##### 市部の企業に限定 #####
> si.市 <- subset(si, area=='市')
```

このsi.市を用いて関数svytotal()により売上高の総計を推定すると，推定値は$\hat{\tau}_{y,\,市} = 2533.3$となり，その標準誤差の推定値は$\widehat{SE}(\hat{\tau}_{y,\,市}) = 2335.6$となる．これらは本書の(3.22)式と(3.23)式に対応する．Warning message:が表示されるのは，市部の企業が1社しかないためである．

例題 3.4-5

```
##### 市部の売上高総計の推定 #####
> svytotal(~y, si.市)
   total     SE
y 2533.3 2335.6
Warning message:
In onestrat(x[index, , drop = FALSE], clusters[index], nPSU[index][1],  :
  Stratum (1) has only one PSU at stage 1
```

以下の例では関数subset()によって市部の企業を指定する代わりに，関数svyby()を用いて市部の企業と郡部の企業のそれぞれについて部分母集団総計の推定を行っている．関数svyby()の第一の引数に指定するのは目的とする変数yであり，第二の引数に指定するのは部分集団を表す変数areaである．第三の引数には，標本全体に対して標本抽出デザインを指定した結果のsiを指定し，最後の引数には母集団総計を推定するための関数svytotalを指定する．市部については推定値が$\hat{\tau}_{y,\,市} = 2533.333$となり，郡部については$\hat{\tau}_{y,\,郡} = 5653.333$となる．

例題 3.4-6

```
> svyby(~y, ~area, si, svytotal)
   area        y     se.y
市    市 2533.333 2335.618
郡    郡 5653.333 3468.483
Warning message:
In onestrat(x[index, , drop = FALSE], clusters[index], nPSU[index][1],  :
  Stratum (1) has only one PSU at stage 1
```

3.b.1 演習問題

演習 3.4 部分母集団総計の推定

データフレーム kigyo を用いて，業種 gyoshu が 1 という企業の資本金 shihon と売上高 uriage の部分母集団総計を推定してみよう．ただし標本は非復元単純無作為抽出されたものとする．

ヒント：関数 subset() を用いて部分母集団を指定し，その結果を関数 svytotal() の引数として用いればよい．

演習 3.5 互いに排反な部分母集団総計の推定

データフレーム kigyo を用いて，5 つの業種 gyoshu ごとの資本金 shihon と売上高 uriage の部分母集団総計を推定してみよう．ただし標本は非復元単純無作為抽出されたものとする．

さらに 3 つの所在地 area と 5 つの業種 gyoshu の 15 の組み合わせのそれぞれについて，資本金 shihon と売上高 uriage の総計を推定してみよう．

ヒント：関数 subset() を用いるのではなく，関数 svyby() を用いればよい．

演習 3.6 欠測値を除いた推定

この演習では，関数 svytotal() において引数 na.rm=TRUE と指定することの意味を確かめることにする．まずデータフレーム kigyo が非復元単純無作為抽出されたものとして，欠測を含む変数 uriage.na の母集団総計を関数 svytotal() を用いて推定してみよう．ただし引数 na.rm=TRUE は指定しない．次に関数 svytotal() の引数に na.rm=TRUE を追加し，推定値を確かめてみよう．

さらに関数 subset() を用いて，変数 uriage.na が欠測 NA ではない部分母集団の総計を推定してみよう．

ヒント：関数 subset() において，変数 uriage.na が欠測 NA ではない集団を指定するには !is.na(uriage.na) とすればよい．

3.c　演習問題解答例

演習 3.1 解答例

標本は復元単純無作為抽出されたものとするので，あらかじめ抽出ウェイト w_i を用意しておく必要がある．以下の例ではデータフレーム kigyo に抽出ウェイトの変数 w を追加している．

演習 3.1 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- 10000 / 2000
```

復元抽出なので関数 svydesign() では引数 fpc を指定せず，ids と weights を指定する．

演習 3.1 解答例 -2

```
##### 復元単純無作為抽出法の指定 #####
> sir <- svydesign(ids=~1, weights=~w, data=kigyo)
```

以下の例では念のため，関数 weights() を用いてどの企業の抽出ウェイトも $w_i = 5$ であることを確かめている．

演習 3.1 解答例 -3

```
##### 抽出ウェイトの確認 #####
> summary(weights(sir))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5       5       5       5       5       5
```

資本金 shihon と売上高 uriage の母集団総計を推定するには，関数 svytotal() の最初の引数に目的とする変数 shihon と uriage を+でつないで指定すればよい．例えば資本金の母集団総計の推定値は $\hat{\tau}_{\text{資本金}} = 1236645$ となり，標準誤差は $\widehat{SE}(\hat{\tau}_{\text{資本金}}) = 32960$ となる．

演習 3.1 解答例 -4

```
##### 資本金と売上高の母集団総計の推定 #####
> svytotal(~shihon + uriage, sir)
         total    SE
shihon 1236645 32960
uriage 1008365 24024
```

演習 3.2 解答例

標本は “非復元” 単純無作為抽出されたものとするので，以下の例では関数 svydesign() の引数 fpc に母集団サイズの変数 N を指定し，引数 weights は指定していない．引数 fpc に指定した変数 N に代入されている値 $10,000$ とデータフレームから求めた標本サイズ $n = 2,000$ から自動的に $w_i = N/n = 5$ と計算される．

演習 3.2 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

念のため関数 weights() を用いて確かめてみると，確かにどの企業も抽出ウェイトは $w_i = 5$ となっていることが分かる．

演習 3.2 解答例 -2

```
##### 抽出ウェイトの確認 #####
> summary(weights(si))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5       5       5       5       5       5
```

関数 svytotal() で求めた推定値は復元抽出のときと同じであるが，その標準誤差は**演習 3.1 解答例 -4** に示す復元抽出のときと比べ若干小さくなる．

演習 3.2 解答例 -3

```
##### 資本金と売上高の母集団総計の推定 #####
> svytotal(~shihon + uriage, si)
         total    SE
shihon 1236645 29481
uriage 1008365 21488
```

標本抽出デザインを指定した survey.design オブジェクト si から抽出ウェイト w_i を取り出すには関数 weights() を用いればよい．以下の例では関数 sum() を用いて抽出ウェイトの標本総計を求めている．結果は 10000 となり，母集団サイズ $N = 10,000$ に一致する．

演習 3.2 解答例 -4

```
##### 抽出ウェイトの標本総計 #####
> sum(weights(si))
[1] 10000
```

演習 3.3 解答例

まず単純無作為抽出法なので，抽出ウェイトは $w_i = N/n$ となる．

演習 3.3 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- kigyo$N / kigyo$n
```

さらに，全ての企業の値が Inf である変数 inf を作成している．

演習 3.3 解答例 -2

```
##### 値がInfである変数の作成 #####
> kigyo$inf <- Inf
```

以下の例では関数 svydesign() の引数 fpc に，全ての企業の値が Inf である変数 inf を指定している．そのため有限母集団修正項は $1 - n/N \to 1$ となる．

演習 3.3 解答例 -3

```
##### 復元単純無作為抽出法の指定 #####
> sir <- svydesign(ids=~1, fpc=~inf, weights=~w, data=kigyo)
```

関数 svytotal() を用いて推定した標準誤差は，関数 svydesign() の引数 fpc を指定しなかった**演習 3.1 解答例 -4** の結果と等しくなる．

演習 3.3 解答例 -4

```
##### 母集団総計の推定 #####
> svytotal(~shihon + uriage, sir)
         total    SE
shihon 1236645 32960
uriage 1008365 24024
```

演習 3.4 解答例

部分母集団に関する推定を行う場合であっても，まず関数 svydesign() を用いて標本全体に対して標本抽出デザインを指定する．以下の例では関数 svydesign() による指定結果を si としている．

演習 3.4 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

次に関数 subset() の第一の引数には si，第二の引数には部分母集団を特定する gyoshu==1 を指定することで，目的とする部分母集団を指定する．以下の例ではこれを関数 svytotal() の引数として指定している．なお例えば si.sub <- subset(si, gyoshu==1) などとして一度指定結果を保存した上で，svytotal(~shihon + uriage, si.sub) などとしてもよい．

演習 3.4 解答例 -2

```
##### gyoushuが1の企業の部分母集団総計の推定 #####
> svytotal(~shihon + uriage, subset(si, gyoshu==1))
        total    SE
shihon 247725 16551
uriage 200390 12439
```

演習 3.5 解答例

まず関数 svydesign() を用いて非復元単純無作為抽出法を指定する.

演習 3.5 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

互いに排反な部分母集団のそれぞれについて推定を行うには，関数 svyby() を用いればよい．関数 svyby() の第一の引数に指定するのは目的とする変数 shihon と uriage である．第二の引数には部分母集団を表す変数 gyoshu を指定する．第三の引数には標本抽出デザインを指定した結果 si を指定し，最後に関数 svytotal を指定する.

演習 3.5 解答例 -2

```
##### gyoshuごとの部分母集団総計の推定 #####
> svyby(~shihon + uriage, ~gyoshu, si, svytotal)
  gyoshu shihon uriage se.shihon se.uriage
1      1 247725 200390  16551.47  12439.33
2      2 253210 201285  16843.98  12671.72
3      3 240405 205220  16090.79  12834.09
4      4 245635 198260  16161.99  12222.60
5      5 249670 203210  16760.13  12563.21
```

所在地 area と業種 gyoshu の全ての組み合わせごとに推定を行うには，関数 svyby() の第二の引数に 2 つの変数を+でつないで指定すればよい.

演習 3.5 解答例 -3

```
##### areaとgyoshuの組み合わせごとの部分母集団総計の推定 #####
> svyby(~shihon + uriage, ~area + gyoshu, si, svytotal)
    area gyoshu shihon uriage se.shihon se.uriage
1.1    1      1  61790  46735  8307.275  5959.358
2.1    2      1  89490  72065 10315.756  7985.136
3.1    3      1  96445  81590 10706.516  8118.772
1.2    1      2  66355  50525  9442.540  6713.506
2.2    2      2  84800  67315  9494.490  7283.968
3.2    3      2 102055  83445 11011.977  8546.220
1.3    1      3  58430  50090  8099.455  6441.676
2.3    2      3  82525  72890  9624.632  8087.225
3.3    3      3  99450  82240 10759.098  8296.564
1.4    1      4  62125  50545  8517.911  6460.437
2.4    2      4  88595  68530 10087.628  7292.827
3.4    3      4  94915  79185 10136.592  8049.153
1.5    1      5  61220  49960  8275.516  6316.929
2.5    2      5  88325  73885 10636.762  7944.662
3.5    3      5 100125  79365 10751.253  8101.518
```

演習 3.6 解答例

まず関数 svydesign() を用いて非復元単純無作為抽出法を指定する．

演習 3.6 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

関数 svytotal() の引数に欠測 NA が含まれた変数 uriage.na を指定すると，推定値は NA となる．変数 y_i に欠測が含まれていると，線形推定値 $\hat{\tau}_y = \sum_s w_i y_i$ は求まらないからである．

演習 3.6 解答例 -2

```
##### 母集団総計の推定 #####
> svytotal(~uriage.na, si)
          total SE
uriage.na    NA NA
```

次に関数 svytotal() の引数として na.rm=TRUE とすると推定値は求まる．これは標本 s 全体ではなく，欠測値のない標本企業についてのみ $w_i y_i$ を合計するからである．

$$\hat{\tau}_{y,\text{ 非欠測}} = \sum_{i\in \text{ 欠測値のない標本企業}} w_i y_i$$

演習 3.6 解答例 -3

```
##### na.rm=TRUEを指定した母集団総計の推定 #####
> svytotal(~uriage.na, si, na.rm=TRUE)
           total    SE
uriage.na 868170 21045
```

なお上式で「欠測値のない標本企業」とは，関数 svytotal() で指定した全ての変数において欠測値がない企業のことである．したがって欠測値のない変数 uriage を同時に指定すると，変数 uriage についても「欠測値のない標本企業」の $w_i y_i$ の合計が求められる．以下の例では変数 uriage と uriage.na の推定値は等しく 868170 となる．**演習 3.2 解答例 -3** の結果では $\hat{\tau}_{\text{売上高}} = 1008365$ となっていたことに注意すること．

演習 3.6 解答例 -4

```
##### na.rm=TRUEを指定した母集団総計の推定 #####
> svytotal(~uriage + uriage.na, si, na.rm=TRUE)
           total    SE
uriage    868170 21045
uriage.na 868170 21045
```

以下の例では，関数 subset() を用いて変数 uriage.na が欠測 NA でない部分母集団を指定した上で，関数 svytotal() で引数 na.rm=TRUE を指定せずに部分母集団総計を推定している．推定値は $\hat{\tau}_{y,\,\text{非欠測}} = 868170$ となり，この結果は svytotal() において引数 na.rm=TRUE を指定したときの**演習 3.6 解答例 -3** の結果と同一である．すなわち引数 na.rm=TRUE を指定するということは,「変数の値が欠測とならない部分母集団」総計を推定することに相当する．

演習 3.6 解答例 -5

```
##### uriage.naが欠測でない部分母集団総計の推定 #####
> svytotal(~uriage.na, subset(si, !is.na(uriage.na)))
           total    SE
uriage.na 868170 21045
```

第4章　確率比例抽出法

4.a　確率比例抽出法の指定

確率比例抽出法の指定

```
DES <- svydesign(ids=~1, weights=~W, data=DATA)
```

確率比例抽出法では復元抽出法を前提とするため (p.58)，引数 fpc は指定しない．仮に引数 fpc を指定しても警告やエラーは表示されない．引数 weights には抽出ウェイト $w_i = \tau_x/(nx_i)$ が代入された変数 W を指定する．

例題 4.2　母集団総計の推定

復元確率比例抽出された表 4.3 (p.55) の標本を使って，売上高の母集団総計を推定してみよう．以下の例ではまず data.frame() を用いて，売上高 y_i の値が代入された変数 y と，資本金 x_i の値が代入された変数 x の 2 つの変数を持つデータフレーム data を作成している．

例題 4.2-1

```
##### データの作成 #####
> (data <- data.frame(y=c(636, 465, 65), x=c(57, 51, 19)))
    y  x
1 636 57
2 465 51
3  65 19
```

次に抽出ウェイトを $w_i = \tau_x/(nx_i)$ により求め，変数 w に代入している．ただし $\tau_x = 663$ は表 4.1 (p.51) にある母集団から分かる値である．また $n = 3$ である．

例題 4.2-2

```
##### 抽出ウェイトの作成 #####
> (data$w <- 663 / (3 * data$x))
[1]  3.877193  4.333333 11.631579
```

以下の例では関数 svydesign() を用いて復元確率比例抽出法を指定している．復元抽出法なので引数 fpc は指定しない．指定結果は survey.design オブジェクト pps に代入している．

例題 4.2-3

```
##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=data)
```

関数 summary() を用いて pps の内容を確認すると，Probabilities:が一定ではなく，抽出ウェイトも一定ではないことが分かる．

例題 4.2-4

```
##### survey.designオブジェクトの確認 #####
> summary(pps)
Independent Sampling design (with replacement)
svydesign(ids = ~1, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.08597 0.15840 0.23080 0.19160 0.24430 0.25790
Data variables:
[1] "y" "x" "w"
```

以下の例では**例題 4.2-3** で作成した survey.design オブジェクト pps を用いて，売上高 y と資本金 x の母集団総計を推定している．売上高の母集団総計の推定値は $\hat{\tau}_y = 5237$ であり，その標準誤差の推定値は $\widehat{SE}(\hat{\tau}_y) = 1534.9$ となる．これらは本書の (4.9) 式と (4.10) 式に対応する．資本金の母集団総計の推定値は $\hat{\tau}_x = 663$ となり，真の母集団総計 $\tau_x = 663$ に一致する．その標準誤差の推定値は 3.481e-14 と出力されているが，これは 0 とみなしてよい．

例題 4.2-5

```
##### 母集団総計の推定 #####
> svytotal(~y + x, pps)
  total        SE
y  5237    1534.9
x   663 3.481e-14
```

4.a.1　演習問題

演習 4.1　復元確率比例抽出法の指定と母集団総計の推定

データフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団総計を推定してみよう．ただし標本は資本金 shihon で復元確率比例抽出されたものとする．さらに抽出ウェイトの標本総計を求めてみよう．

ヒント：資本金の母集団総計 $\tau_x = 1,725,000$ は変数 total.shihon に代入されており，標本サイズ $n = 1,000$ は変数 n に代入されている．

4.b　デザイン効果の推定

デザイン効果の推定

```
STAT <- svytotal(x=~X, design=DES, na.rm=FALSE, deff=DEFF)
```

デザイン効果を推定するには，関数 svytotal() 中で引数 deff を指定する．デフォルトでは deff=FALSE となっており，デザイン効果は計算されない．deff=TRUE とすると (4.33) 式に対応した Deff の推定値が求められ，deff="replace"とすると (4.34) 式に対応した Deft の二乗の推定値が求められる．ただし結果はいずれも DEff と表示される．

デザイン効果の取り出し

```
deff(STAT)
```

推定結果の svystat クラスのオブジェクト STAT からデザイン効果の値だけを取り出すには deff() を用いる．ただし STAT を求める際に，あらかじめ引数 deff を指定しておく必要がある．

例題 4.7　デザイン効果の推定

表 4.3 (p.55) の標本を使って，復元確率比例抽出法のデザイン効果を推定してみよう．以下の例では，**例題 4.2-3** で作成した survey.design オブジェクト pps を再利用している．関数 svytotal() で引数 deff に TRUE を指定すると deff $= 0.2387$ が得られ，'replace' を指定すると $\text{deft}^2 = 0.2026$ が得られる．これらは本書の (4.41) 式に対応する．

例題 4.7-1

```
##### デザイン効果の推定（deff） #####
> svytotal(~y, pps, deff=TRUE)
   total     SE   DEff
y 5236.9 1534.9 0.2387

##### デザイン効果の推定（deft^2） #####
> svytotal(~y, pps, deff='replace')
   total     SE   DEff
y 5236.9 1534.9 0.2026
```

4.b.1　演習問題

演習 4.2　復元確率比例抽出法のデザイン効果

データフレーム kigyo が資本金 shihon で復元確率比例抽出されたものとして，売上高 uriage の母集団総計を求めたときのデザイン効果 Deff および Deft を推定してみよう．さらに有効標本サイズ n_{EFF} (p.67) を求めてみよう．

ヒント：有効標本サイズを求めるためにデザイン効果の推定値を取り出すには，関数 deff() を用いればよい．

4.c 演習問題解答例

演習 4.1 解答例

まず確率比例抽出法の抽出ウェイト $w_i = \tau_x/(nx_i)$ を用意する．資本金の母集団総計 τ_x の値は変数 total.shihon，標本サイズ n の値は変数 n，各企業の資本金 x_i は変数 shihon に代入されている．以下の例では抽出ウェイトの式にしたがって計算した値を変数 w に代入している．

演習 4.1 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- kigyo$total.shihon / (kigyo$n * kigyo$shihon)
```

関数 svydesign() で標本抽出デザインを指定するときには，“復元” 抽出なので引数 fpc は指定しない．引数 weights には抽出ウェイトの変数 w を指定する．

演習 4.1 解答例 -2

```
##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=kigyo)
```

関数 svytotal() を用いて資本金 shihon の母集団総計 τ_x を推定すると，$\hat{\tau}_x = 1725000$ となって真の母集団総計に一致する．また，その標準誤差は 0 となる．

演習 4.1 解答例 -3

```
##### 母集団総計の推定 #####
> svytotal(~shihon + uriage, pps)
         total        SE
shihon 1725000 2.776e-12
uriage 1908696    32918
```

さらに関数 weights() と関数 sum() を用いて抽出ウェイトの標本総計を求めると $\sum_s w_i = 47305.37$ となる．この値は母集団サイズの推定値 $\hat{N}$ であるが，真の母集団サイズ $N = 10,000$ には一致しない．

演習 4.1 解答例 -4

```
##### 抽出ウェイトの標本総計 #####
> sum(weights(pps))
[1] 47305.37
```

演習 4.2 解答例

資本金 shihon で復元確率比例抽出なので，標本抽出デザインの指定は**演習 4.1** と同様に行えばよい．

演習 4.2 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- kigyo$total.shihon / (kigyo$n * kigyo$shihon)

##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=kigyo)
```

分母を “非復元” 単純無作為抽出法としたデザイン効果 Deff の推定値を求めるには，関数 svytotal() において引数 deff=TRUE を指定すればよい．また分母を “復元” 単純無作為抽出法としたデザイン効果 Deft^2 を求めるには，引数 deff='replace' を指定すればよい．デザイン効果を推定するときの分母は “非復元” 抽出である Deff の方が小さくなるので，デザイン効果の推定値は Deff の方が大きくなる．

演習 4.2 解答例 -2

```
##### デザイン効果の推定（deff） #####
> svytotal(~uriage, pps, deff=TRUE)
         total      SE   DEff
uriage 1908696   32918 0.3703

##### デザイン効果の推定（deft^2） #####
> svytotal(~uriage, pps, deff='replace')
         total      SE   DEff
uriage 1908696   32918 0.3546
```

有効標本サイズ n_{EFF} を求めるには，実際の標本サイズ n をデザイン効果で割ればよい．deff の方を用いると $n_{\text{EFF}} = 5401.16$ が得られる．この値は，確率比例抽出した標本のサイズは $n = 2,000$ であるが，推定値の精度という観点から見れば，単純無作為抽出した $n_{\text{EFF}} = 5401.16$ の標本と同等であることを意味する．

演習 4.2 解答例 -3

```
##### 有効標本サイズ #####
> 2000 / deff(svytotal(~uriage, pps, deff=TRUE))
 uriage
5401.16
```

第5章　比推定量

5.a　比推定

母集団比の推定

```
STAT <- svyratio(numerator=~Y, denominator=~X, design=DES,
                                                    separate=FALSE)
```

svyratio() は母集団比 $R = \tau_y/\tau_x$ の推定値 $\hat{R} = \hat{\tau}_y/\hat{\tau}_x$ を求める関数である．引数 numerator の Y には比の分子の変数 y，denominator の X には分母の変数 x を指定する．引数 separate を TRUE とすると個別比推定を行うことができる．詳細は 6.b 節を参照のこと．結果の STAT は svyratio クラスのオブジェクトである．

母集団総計の比推定

```
predict(STAT, total=TOTAL, se=TRUE, ...)
```

関数 predict() は，第一の引数 STAT に指定された svyratio() の結果に対して，引数 total の TOTAL に指定された値を乗ずる．そのため引数 total の TOTAL に母集団総計 τ_x の値を指定すれば，母集団総計の比推定値 $\hat{\tau}_{y,\mathrm{R}} = \tau_x \hat{R}$ を求めることができる．引数 se を TRUE (デフォルト) とすると標準誤差を表示し，FALSE とすると表示しない．

例題 5.2　復元確率比例抽出法と比推定

表 4.3 (p.55) の標本を使って，資本金 x を補助変数とした売上高 y の母集団総計の比推定値を求めてみよう．

例題 5.2-1

```
##### データの作成 #####
> (data <- data.frame(y=c(636, 465, 65), x=c(57, 51, 19)))
    y  x
1 636 57
2 465 51
3  65 19
```

以下の例では標本が復元確率比例抽出されたものとして抽出ウェイトを作成している．

例題 5.2-2

```
##### 抽出ウェイトの作成 #####
> (data$w <- 663 / (3 * data$x))
[1]  3.877193  4.333333 11.631579
```

関数 svydesign() によって復元確率比例抽出を指定した結果は pps である．

例題 5.2-3

```
##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=data)
```

以下の例では関数 svyratio() を用いて，売上高の変数 y を分子とし，資本金の変数 x を分母とした比の推定値 $\hat{R} = \hat{\tau}_y/\hat{\tau}_x = 7.898865$ を求めている．その標準誤差は $\widehat{SE}(\hat{R}) = 2.315078$ である．得られた結果は R.hat に代入している．

例題 5.2-4

```
##### 比の推定 #####
> (R.hat <- svyratio(~y, ~x, pps))
Ratio estimator: svyratio.survey.design2(~y, ~x, pps)
Ratios=
         x
y 7.898865
SEs=
         x
y 2.315078
```

以下の例では関数 predict() を用いて，比の推定値 $\hat{R}$ に資本金の母集団総計 $\tau_x = 663$ を乗じている．predict() の最初の引数に指定するのは，svyratio() の結果オブジェクト R.hat であり，二番目の引数に指定するのは補助変数である資本金の母集団総計値 $\tau_x = 663$ である．売上高の母集団総計の比推定値は $\hat{\tau}_{y,\mathrm{R}} = 5236.947$ となる．この結果は本書の (5.17) 式に対応する．また標準誤差は $\widehat{SE}(\hat{\tau}_{y,\mathrm{R}}) = 1534.897$ となる．

例題 5.2-5

```
##### 売上高の母集団総計の比推定 #####
> predict(R.hat, 663)
$total
         x
y 5236.947

$se
         x
y 1534.897
```

例題 5.4　補助変数の母集団総計の比推定

表 5.3 (p.77) の非復元単純無作為抽出標本を用いて，売上高 y と資本金 x の母集団総計の比推定値を求めてみよう．

例題 5.4-1

```
##### データの作成 #####
> (data <- data.frame(y=c(380, 639, 209), x=c(31, 60, 28), N=20))
    y  x  N
1 380 31 20
2 639 60 20
3 209 28 20
```

以下の例では関数 `svydesign()` を用いて，非復元単純無作為抽出法を指定した `survey.design` オブジェクト `si` を作成している．

例題 5.4-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

以下の例では，まず関数 `svyratio()` を用いて売上高 `y` あるいは資本金 `x` を分子，資本金 `x` を分母とした比の推定値を求めている．当然のことながら，分子・分母ともに資本金 `x` としたときの比の推定値は `1.00000` となり，標準誤差は `0.0000000` となる．

例題 5.4-3

```
##### 比の推定値 #####
> (R.hat <- svyratio(~y + x, ~x, si))
Ratio estimator: svyratio.survey.design2(~y + x, ~x, si)
Ratios=
         x
y 10.31933
x  1.00000
SEs=
          x
y 0.9674943
x 0.0000000
```

以下の例では関数 `predict()` を用いて，資本金の母集団総計 $\tau_x = 663$ を比の推定値に乗じることで，母集団総計の比推定値を求めている．売上高については $\hat{\tau}_{y,\mathrm{R}} = 6841.714$ となり，これは本書の (5.21) 式に対応する[1]．また資本金については $\hat{\tau}_{x,\mathrm{R}} = 663.000$ となり，これは本書の (5.22) 式に対応する．なお売上高の母集団総計の比推定値は，標準誤差が $\widehat{SE}(\hat{\tau}_{y,\mathrm{R}}) = 641.4487$ となる．これは本書の (5.36) 式に対応する．

[1] 本書の推定値 6,844 は 6,842 の誤りである．

例題 5.4-4

```
##### 母集団総計の比推定値 #####
> predict(R.hat, 663)
$total
         x
y 6841.714
x  663.000

$se
         x
y 641.4487
x   0.0000
```

例題 5.6　サイズを用いた部分母集団総計の比推定

表 5.4 (p.79) の非復元単純無作為抽出標本を用いて，市部の売上高総計の比推定値を求めてみよう．ただし補助変数は $x_i = 1$ とする．すなわちサイズを用いた比推定値 $\hat{\tau}_{y,\,\text{市,N}}$ を求めることにする[2]．

以下の例では，まず変数 y を各企業の売上高，変数 x を全ての企業が 1 という値を持つ変数としてデータフレーム data を作成している．

例題 5.6-1

```
##### データの作成 #####
> (data <- data.frame(y=c(380, 639, 209), x=1, N=20, area=c('市', '郡', '郡')))
    y x  N area
1 380 1 20   市
2 639 1 20   郡
3 209 1 20   郡
```

次に関数 svydesign() を用いて非復元単純無作為抽出法を指定している．

例題 5.6-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

以下の例では，まず地域 area が市の企業に限定した上で，売上高 y と補助変数 x との比の推定値 $\hat{R}_{\text{市}} = \hat{\tau}_{y,\,\text{市}}/\hat{N}_{\text{市}} = 380$ を求めている．なお標準誤差が $\widehat{SE}(\hat{R}_{\text{市}}) = 0$ となり，Warning messages:が表示されているが，これは市部の企業が 1 社しかないためであり，本書の**例題 5.9** (p.84) で述べているとおりである．

[2] ここで示す方法よりも関数 svymean() を用いる方がより簡単である．**演習 5.5** を参照のこと．

例題 5.6-3

```
##### サイズとの比の推定 #####
> (R.hat <- svyratio(~y, ~x, subset(si, area==’市’)))
Ratio estimator: svyratio.survey.design2(~y, ~x, subset(si, area == "市"))
Ratios=
    x
y 380
SEs=
  x
y 0
Warning messages:
1: In onestrat(x[index, , drop = FALSE], clusters[index], nPSU[index][1],  :
  Stratum (1) has only one PSU at stage 1
2: In onestrat(x[index, , drop = FALSE], clusters[index], nPSU[index][1],  :
  Stratum (1) has only one PSU at stage 1
```

次に，比の推定値 $\hat{R}_{\text{市}}$ に市部の母集団サイズ $N_{\text{市}} = 8$ を乗じることで，サイズを用いた比推定値 $\hat{\tau}_{y,\,\text{市,N}} = N_{\text{市}}\hat{R}_{\text{市}} = 3040$ を求めている．これは本書の (5.26) 式に対応する．

例題 5.6-4

```
##### サイズを用いた部分母集団総計の比推定 #####
> predict(R.hat, 8)
$total
     x
y 3040

$se
  x
y 0
```

5.a.1　演習問題

演習 5.1　単純無作為抽出法と比推定

データフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団総計の比推定を行ってみよう．ただし標本は非復元単純無作為抽出されたものとし，比推定の補助変数 x は資本金 shihon とする．

ヒント：資本金の母集団総計は変数 total.shihon に代入されており，$\tau_x = 1,725,000$ である．

演習 5.2　確率比例抽出法と比推定

データフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団総計を推定してみよう．ただし標本は資本金 shihon で復元確率比例抽出されたものとする．推定値として線形推定値 $\hat{\tau}_y$ と，資本金 shihon を補助変数 x_i とした比推定値 $\hat{\tau}_{y,\mathrm{R}}$ との両方を求め，比較すること．

5.b　母集団平均の推定

本書では母集団平均の推定量として線形推定量 $\hat{\mu}_y$，比推定量 $\hat{\tau}_{y,\mathrm{R}}$，サイズとの比の推定量 $\hat{\mu}_{y,\mathrm{N}}$ の 3 つを紹介している (p.84)．以下ではそれらの推定量を順に紹介する．

母集団平均の線形推定

```
coef(svytotal(x=~Y, design=DES, na.rm=FALSE, deff=FALSE))/N
SE(svytotal(x=~Y, design=DES, na.rm=FALSE, deff=FALSE))/N
```

母集団平均の線形推定を行う関数はない．そのため母集団平均の線形推定値 $\hat{\mu}_y$ を求めるには母集団総計 τ_y の線形推定値 $\hat{\tau}_y$ を求め，それを母集団サイズ N で割るという作業を手ずから行う必要がある．推定値を求めるには `svytotal()` の結果から母集団総計の推定値 $\hat{\tau}_y$ を `coef()` によって取り出し，それを `N` で割る．標準誤差を求めるには同様に `SE()` によって取り出した値を `N` で割ればよい．

母集団平均の比推定

```
STAT <- svyratio(numerator=~Y, denominator=~X, design=DES,
                                                    separate=FALSE)
predict(STAT, total=MEAN, se=TRUE, ...)
```

母集団平均の比推定を行うには，まず関数 `svyratio()` によって母集団比 $R = \tau_y/\tau_x$ の推定値 $\hat{R}$ を求める．次に関数 `predict()` の二番目の引数に補助変数の母集団平均 μ_x を指定し，$\hat{\mu}_{y,\mathrm{R}} = \mu_x \hat{R}$ として比推定値を求めればよい．

サイズとの比の推定

```
STAT <- svymean(x=~Y, design=DES, na.rm=FALSE, deff=FALSE)
```

補助変数を $x_i = 1$ としたときの，母集団平均の比推定値 $\hat{\mu}_{y,\mathrm{N}} = \hat{\tau}_y/\hat{N}$ つまり加重標本平均を求めるには関数 `svymean()` を用いる．引数 `na.rm` と引数 `deff` の指定方法は `svytotal()` と同様である．

以下に各推定量とそれに応じた関数を整理しておく．ただし変数 one は全ての要素の値が 1 という変数である．

	母集団総計	母集団平均・割合
線形推定量	$\hat{\tau}_y = \sum_s w_i y_i$	$\hat{\mu}_y = \frac{1}{N}\sum_s w_i y_i$
	svytotal(Y)	svytotal(Y)/N
比推定量	$\hat{\tau}_{y,\mathrm{R}} = \tau_x \frac{\hat{\tau}_y}{\hat{\tau}_x}$	$\hat{\mu}_{y,\mathrm{R}} = \mu_x \frac{\hat{\tau}_y}{\hat{\tau}_x}$
	predict(svyratio(Y, X), TOTAL)	predict(svyratio(Y, X), MEAN)
サイズを用いた比推定量	$\hat{\tau}_{y,\mathrm{N}} = N\frac{\hat{\tau}_y}{\hat{N}}$	$\hat{\mu}_{y,\mathrm{N}} = \frac{\hat{\tau}_y}{\hat{N}}$
	predict(svyratio(Y, ONE), N)	svymean(Y)
	svymean(Y)*N	

例題 5.10　単純無作為抽出と母集団平均の推定

表 5.6 (p.85) の非復元単純無作為抽出標本を用いて，売上高の母集団平均 μ_y を推定してみよう．以下の例ではまず関数 data.frame() を用いてデータフレーム data を作成している．変数 y は売上高，変数 x は資本金であり，変数 N は母集団サイズである．

例題 5.10-1

```
##### データの作成 #####
> (data <- data.frame(y=c(380, 639, 209), x=c(31, 60, 28), N=20))
    y  x  N
1 380 31 20
2 639 60 20
3 209 28 20
```

以下の例では非復元単純無作為抽出を指定した survey.design オブジェクト si を作成している．引数 fpc を指定しているので，引数 weights の指定は省略している．

例題 5.10-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

以下の例ではまず線形推定値 $\hat{\mu}_y$ を求める．そのため関数 svytotal() によって母集団総計 τ_y の線形推定を行う．関数 coef() によって推定値のみを取り出した後に，母集団サイズ $N = 20$ で割ることで，線形推定値 $\hat{\mu}_y = 409.3333$ を得ている．この結果は本書の (5.47) 式に対応する．

例題 5.10-3

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, si))/20
       y
409.3333
```

同様に標準誤差は関数 SE() を用いることで $\widehat{SE}(\hat{\mu}_y) = 115.2386$ となる．これは本書の (5.50) 式に対応する．

例題 5.10-4

```
##### 標準誤差の推定値 #####
> SE(svytotal(~y, si))/20
       y
115.2386
```

次に資本金 x との比推定値 $\hat{\mu}_{y,\mathrm{R}}$ を求める．以下の例ではまず関数 svyratio() によって比 $R = \tau_y/\tau_x$ の推定を行った後，これに資本金の母集団平均 $\mu_x = \tau_x/N = 663/20$ を乗じることで，比推定値 $\hat{\mu}_{y,\mathrm{R}} = 342.0857$ とその標準誤差 $\widehat{SE}(\hat{\mu}_{y,\mathrm{R}}) = 32.07244$ を得ている．これらは本書の (5.48) 式と (5.51) 式に対応する．

例題 5.10-5

```
##### 母集団平均の比推定 #####
> predict(svyratio(~y, ~x, si), 663/20)
$total
         x
y 342.0857

$se
         x
y 32.07244
```

最後にサイズとの比の推定値 $\hat{\mu}_{y,\mathrm{N}}$ を求める．関数 svymean() を用いると推定値は $\hat{\mu}_{y,\mathrm{N}} = 409.33$ となり，標準誤差は $\widehat{SE}(\hat{\mu}_{y,\mathrm{N}}) = 115.24$ となる．これらは本書の (5.49) 式と (5.52) 式に対応する．

例題 5.10-6

```
##### サイズとの比の推定 #####
> svymean(~y, si)
    mean     SE
y 409.33 115.24
```

例題 5.11　復元確率比例抽出と母集団平均の推定

表 5.7 (p.86) の復元確率比例抽出標本を用いて，売上高の母集団平均 μ_y を推定してみよう．標本は資本金 x で確率比例抽出されている．以下の例では関数 data.frame() によってデータフレームを作成した後に，抽出ウェイト $w_i = \tau_x/(nx_i)$ を代入した変数 w を作成している．

例題 5.11-1

```
##### データの作成 #####
> (data <- data.frame(y=c(636, 465, 65), x=c(57, 51, 19)))
    y  x
1 636 57
2 465 51
3  65 19

##### 抽出ウェイトの作成 #####
> (data$w <- 663 / (3 * data$x))
[1]  3.877193  4.333333 11.631579
```

標本抽出デザインを指定した結果は pps である．復元抽出法なので引数 fpc は指定しない．

例題 5.11-2

```
##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=data)
```

以下の例では，まず母集団平均の線形推定値 $\hat{\mu}_y$ を求める．関数 svytotal() を用いて母集団総計の線形推定を行い，その結果を母集団サイズ $N = 20$ で割ればよい．推定値は $\hat{\mu}_y = 261.8474$ であり，標準誤差は $\widehat{SE}(\hat{\mu}_y) = 76.74483$ となる．これらは本書の (5.53) 式に対応する．

例題 5.11-3

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, pps)) / 20
       y
261.8474

##### 標準誤差の推定値 #####
> SE(svytotal(~y, pps)) / 20
       y
76.74483
```

以下の例では，資本金 x による比推定値を求める．関数 svyratio() を用いて母集団比 $R = \tau_y/\tau_x$ を推定した後に，関数 predict() を用いて母集団比の推定値 $\hat{R}$ に資本金の母集団平均 $\mu_x = 663/20$ を乗じる．推定値は $\hat{\mu}_{y,\mathrm{R}} = 261.8474$ であり，標準誤差は $\widehat{SE}(\hat{\mu}_{y,\mathrm{R}}) = 76.74483$ となる．これらは本書の (5.54) 式に対応する．

例題 5.11-4

```
##### 母集団平均の比推定 #####
> predict(svyratio(~y, ~x, pps), 663/20)
$total
         x
y 261.8474

$se
         x
y 76.74483
```

以下の例では関数 svymean() を用いて，サイズとの比の推定値を求めている．$\hat{\mu}_{y,\mathrm{N}} = 263.93$ と $\widehat{SE}(\hat{\mu}_{y,\mathrm{N}}) = 176.69$ が得られ，これらは本書の (5.55) 式に対応する．

例題 5.11-5

```
##### サイズとの比の推定 #####
> svymean(~y, pps)
    mean     SE
y 263.93 176.69
```

5.b.1 演習問題

演習 5.3 確率比例抽出法と母集団平均の推定

データフレーム kigyo を用いて，企業の資本金 shihon と売上高 uriage の母集団平均をそれぞれ推定してみよう．ただし標本は資本金 shihon で復元確率比例抽出されたものとする．線形推定値 $\hat{\mu}_y$ とサイズとの比の推定値 $\hat{\mu}_{y,\mathrm{N}}$ の両方を求め，両者の違いが生じた理由を考えてみること．

ヒント：母集団サイズは $N = 10,000$ である．資本金の母集団総計は変数 total.shihon に代入されており，$\tau_x = 1,725,000$ である．

演習 5.4 単純無作為抽出法と母集団平均の推定

データフレーム kenko を用いて，生徒の身長 height と体重 weight の母集団平均を推定してみよう．ただし標本は非復元単純無作為抽出されたものととする．

さらに性 gender 別の平均身長と平均体重も推定してみよう．ただし線形推定値 $\hat{\mu}_{y,d}$ と，サイズとの比の推定値 $\hat{\mu}_{y,d,\mathrm{N}}$ の両方を求め，標準誤差を比較すること．

ヒント：互いに排反な部分母集団に関して推定を行うには関数 svyby() を用いればよい．男子の部分母集団サイズは $N_{男} = 375,574$ であり，女子の部分母集団サイズは $N_{女} = 373,995$ である．

演習 5.5 単純無作為抽出法と部分母集団総計の比推定

データフレーム kigyo を用いて，所在地 area が 1 であり，かつ業種 gyoshu が 1 である企業の資本金 shihon と売上高 uriage の部分母集団総計を推定してみよう．ただし標本は非復元単純無作為抽出されたものとし，部分母集団サイズ $N_{1,1} = 400$ を利用した比推定とすること．

ヒント：関数 svymean() を用いて目的とする部分母集団平均を推定し，その結果に部分母集団サイズを乗じればよい．

5.c　母集団割合の推定

母集団割合 p_y は，1 または 0 という値をとる二値変数 y_i の母集団平均 μ_y である (p.88). そのため母集団平均の推定方法をそのまま用いればよい．ただしデータフレームの変数が二値変数でない場合には，factor() を用いることで変数がカテゴリカルなものであることを指示する必要がある．factor() を用いることで，各カテゴリに対応した二値変数が用意されると考えればよい．factor である変数に対して関数 svytotal() を用いると，母集団において各カテゴリに該当する要素の数が推定できる．

例題 5.12　母集団割合の推定

表 5.9 (p.89) の非復元単純無作為抽出標本を用いて，“男に” 生まれ変わりたい人の母集団割合を推定してみよう．以下の例では回答を変数 y とし，回答者の性別を変数 gen としたデータフレーム data を作成している．母集団サイズは N=20 である．

例題 5.12-1

```
##### データの作成 #####
> (data <- data.frame(y=c('女に', '男に', '男に', '女に', '男に'),
                      gen=c('女性', '女性', '男性', '女性', '男性'), N=20))
     y  gen  N
1 女に 女性 20
2 男に 女性 20
3 男に 男性 20
4 女に 女性 20
5 男に 男性 20
```

次に関数 svydesign() を用いて非復元単純無作為抽出法を指定している．

例題 5.12-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

以下の例ではまず関数 svymean() を用いて，生まれ変わりたい性別の母集団割合を推定している．ただし最初の引数に指定する変数を factor(y) とすることで，カテゴリカルな変数として扱うよう指示している[3]．“男に” 生まれ変わりたい人の母集団割合の推定値は $\hat{p}_y = 0.6$ であり，その標準誤差は $\widehat{SE}(\hat{p}_y) = 0.2121$ となる．これらは本書の (5.61) 式と (5.62) 式に対応する．

例題 5.12-3

```
##### サイズとの比の推定 #####
> svymean(~factor(y), si)
               mean     SE
factor(y)女に   0.4 0.2121
factor(y)男に   0.6 0.2121
```

[3] この例では，data.frame() でデータフレームを作成した時点で既に変数 y は factor となっているので，factor(y) の代わりに y でも十分である．

例題 5.13　部分母集団割合の推定

例題 5.12 に引き続き，“男に” 生まれ変わりたい女性の母集団割合を推定してみよう．部分母集団に関して推定を行うには関数 `subset()` を利用すればよい．以下の例では**例題 5.12** で作成した `survey.design` オブジェクト `si` を用いて，“女に” あるいは “男に” 生まれ変わりたい女性の部分母集団割合を推定している．推定値は $\hat{p}_{\text{男に, 女,N}} = 0.33333$ であり，標準誤差は $\widehat{SE}(\hat{p}_{\text{男に, 女,N}}) = 0.2635$ となる．これらは本書の (5.63) 式と (5.64) 式に対応する．

例題 5.13-1

```
##### 女性の中での推定 #####
> svymean(~factor(y), subset(si, gen==’女性’))
                  mean     SE
factor(y)女に 0.66667 0.2635
factor(y)男に 0.33333 0.2635
```

5.c.1　演習問題

演習 5.6　単純無作為抽出法と母集団割合の推定

データフレーム otona を用いて，質問項目 Q1 と Q3 の各カテゴリを選ぶ児童の母集団割合を推定してみよう．ただし標本は非復元単純無作為抽出されたものとする．

さらに質問項目 Q1 について，市郡 city 別の推定値も求めてみよう．線形推定値 $\hat{p}_{y,d}$ とサイズとの比の推定値 $\hat{p}_{y,d,\mathrm{N}}$ の両方を求めること．

ヒント：変数をカテゴリカルな変数として扱うときには factor() を用いる．市郡別の部分母集団サイズは $N_1 = 225,634$ と $N_2 = 279,618$ である．

5.d　母集団分散の推定

母集団分散の推定

```
STAT <- svyvar(x=~X, design=DES, na.rm=FALSE)
```

母集団分散 $\hat{\sigma}_y^2$ の推定を行うには関数 svyvar() を用いる．本書の (5.68) 式を用いた母集団分散の推定が行われる．引数 na.rm の指定は svytotal() や svymean() と同様だが，svyvar() には引数 deff はなく，デザイン効果の推定はできない．

例題 5.14　母集団分散の推定

表 5.11 (p.92) の非復元単純無作為抽出標本を用いて，身長の母集団分散を推定してみよう．

例題 5.14-1

```
##### データの作成 #####
> (data <- data.frame(y=c(161, 171, 154), N=20))
    y  N
1 161 20
2 171 20
3 154 20
```

以下の例ではまず survey.design オブジェクトとして si を作成している．

例題 5.14-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

次に関数 svyvar() を用いて，身長の母集団分散の推定値 $\hat{\sigma}_{y,\mathrm{n}}^2 = 73$ とその標準誤差 $\widehat{SE}(\hat{\sigma}_{y,\mathrm{n}}^2) = 33.651$ を得ている．これらは本書の (5.71) 式と (5.72) 式に対応する．

例題 5.14-3

```
##### 母集団分散の推定 #####
> svyvar(~y, si)
  variance     SE
y       73 33.651
```

5.d.1　演習問題

演習 5.7　単純無作為抽出法と母集団分散の推定

データフレーム kenko を用いて，身長 height と体重 weight の母集団分散を推定してみよう．ただし標本は非復元単純無作為抽出されたものとする．

さらに性 gender 別の部分母集団分散の推定も行ってみよう．

5.e　母集団分位数の推定

母集団分位数の推定

```
STAT <- svyquantile(x=~X, design=DES, quantiles=Q)
```

母集団分位数 $Q_{y,q}$ を推定するには関数 svyquantile() を用いる．引数 quantiles には 0 と 1 の間の任意の数を指定する．Q を 0.5 とすると母集団中央値が推定される．

例題 5.15　母集団中央値の推定

表 5.13 (p.95) の復元確率比例抽出標本を用いて，売上高の母集団中央値を推定してみよう．以下では表 5.13 に従って売上高 y の順に企業を並べているが，一般には並び替えておく必要はない．

例題 5.15-1

```
##### データの作成 #####
> (data <- data.frame(y=c(74, 209, 479, 660), x=c(25, 28, 42, 52)))
    y  x
1  74 25
2 209 28
3 479 42
4 660 52
```

標本は資本金 x で確率比例抽出されたものである．

例題 5.15-2

```
##### 抽出ウェイトの作成 #####
> data$w <- 663 / (4 * data$x)

##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=data)
```

母集団中央値 $Q_{y,0.5}$ を推定するには関数 svyquantile() の第三の引数に 0.5 を指定すればよい．$\hat{Q}_{y,0.5} = 147.2462$ となる．この結果は本書の (5.84) 式に対応する．

例題 5.15-3

```
##### 母集団中央値の推定 #####
> svyquantile(~y, pps, 0.5)
      0.5
y 147.2462
```

5.f　演習問題解答例

演習 5.1 解答例

まず関数 svydesign() を用いて非復元単純無作為抽出法を指定する．

演習 5.1 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

以下の例では関数 svyratio() を用いて，資本金 shihon あるいは売上高 uriage の母集団総計の推定値を分子とし，資本金 shihon の母集団総計の推定値を分母とした比を求めている．当然ながら，分子・分母ともに資本金 shihon を用いたときの比の推定値は $\hat{R} = \hat{\tau}_x/\hat{\tau}_x = 1.0000000$ となる．

演習 5.1 解答例 -2

```
##### shihonとの比の推定 #####
> svyratio(~shihon + uriage, ~shihon, si)
Ratio estimator: svyratio.survey.design2(~shihon + uriage, ~shihon, si)
Ratios=
          shihon
shihon 1.0000000
uriage 0.8154038
SEs=
           shihon
shihon 0.00000000
uriage 0.01263653
```

以下の例では関数 predict() を用いて，関数 svyratio() の結果に資本金の母集団総計 $\tau_x = 1725000$ を乗じている．売上高 uriage の母集団総計の比推定値は $\hat{\tau}_{y,\mathrm{R}} = 1406572$ となり，その標準誤差は $\widehat{SE}(\hat{\tau}_{y,\mathrm{R}}) = 21798.01$ となる．なお資本金 shihon の母集団総計の比推定値は $\hat{\tau}_{x,\mathrm{R}} = 1725000$ となり，真の母集団総計 τ_x に一致する．

演習 5.1 解答例 -3

```
##### shihonを用いた比推定 #####
> predict(svyratio(~shihon + uriage, ~shihon, si), 1725000)
$total
        shihon
shihon 1725000
uriage 1406572

$se
         shihon
shihon     0.00
uriage 21798.01
```

演習 5.2 解答例

まず抽出ウェイト $w_i = \tau_x/(nx_i)$ を変数 w として作成し，関数 svydesign() を用いて復元確率比例抽出法を指定する．

演習 5.2 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- kigyo$total.shihon / (kigyo$n * kigyo$shihon)

##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weights=~w, data=kigyo)
```

母集団総計の線形推定を行うには関数 svytotal() を用いればよい．資本金 shihon の線形推定値は $\hat{\tau}_x = 1725000$ となり，真の母集団総計 τ_x に一致する．

演習 5.2 解答例 -2

```
##### 母集団総計の線形推定 #####
> svytotal(~shihon + uriage, pps)
         total        SE
shihon 1725000 2.776e-12
uriage 1908696     32918
```

以下の例では資本金 shihon を補助変数とした比推定を行っている．売上高 uriage の母集団総計の比推定値は $\hat{\tau}_{y,\mathrm{R}} = 1908696$ となり，**演習 5.2 解答例 -2** で求めた線形推定値 $\hat{\tau}_y = 1908696$ に一致する．資本金 shihon の比推定値 $\hat{\tau}_{x,\mathrm{R}} = 1725000$ は線形推定値 $\hat{\tau}_x$ に一致するだけでなく，真の母集団総計 τ_x に一致する．

演習 5.2 解答例 -3

```
##### 母集団総計の比推定 #####
> predict(svyratio(~shihon + uriage, ~shihon, pps), 1725000)
$total
        shihon
shihon 1725000
uriage 1908696

$se
         shihon
shihon     0.00
uriage 32917.99
```

演習 5.3 解答例

資本金 shihon で確率比例抽出なので，以下の例ではまず抽出ウェイト $w_i = \tau_x/(nx_i)$ を求め，これを関数 svydesign() の引数 weights に指定している．

演習 5.3 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- kigyo$total.shihon / (kigyo$n * kigyo$shihon)

##### 復元確率比例抽出法の指定 #####
> pps <- svydesign(ids=~1, weight=~w, data=kigyo)
```

まず線形推定値 $\hat{\mu}_y$ は，関数 svytotal() を用いて母集団総計の線形推定値 $\hat{\tau}_y$ を求め，それを真の母集団サイズ $N = 10,000$ で割ればよい．資本金の母集団平均の推定値は $\hat{\mu}_{\text{資本金}} =$ 172.5000 となる．この値は，真の母集団総計 $\tau_{\text{資本金}} = 1,725,000$ を母集団サイズ $N = 10,000$ で割った真の母集団平均 $\mu_{\text{資本金}} = 172.5$ に一致する．売上高の推定値は $\hat{\mu}_{\text{売上高}} =$ 190.8696 となる．

演習 5.3 解答例 -2

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~shihon + uriage, pps)) / 10000
  shihon   uriage
172.5000 190.8696

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~shihon + uriage, pps)) / 10000
      shihon       uriage
2.776144e-16 3.291799e+00
```

サイズとの比の推定値を求めるには関数 svymean() を用いればよい．資本金と売上高のいずれの推定値も，線形推定値に比べ小さな値となる．その理由は母集団サイズの推定値である抽出ウェイトの標本総計が以下に示すように $\hat{N} = \sum_s w_i =$ 47305.37 となって真の母集団サイズ $N = 10,000$ よりも大きくなるためである．

演習 5.3 解答例 -3

```
##### サイズとの比の推定値 #####
> svymean(~shihon + uriage, pps)
         mean     SE
shihon 36.465 0.9474
uriage 40.348 0.9509

##### 抽出ウェイトの標本総計 #####
> sum(weights(pps))
[1] 47305.37
```

演習 5.4 解答例

母集団サイズ N は変数 N に代入されている．そこでまず関数 svydesign() の引数に fpc=N を指定して非復元単純無作為抽出法を指定する．

演習 5.4 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kenko)
```

単純無作為抽出法なので，母集団平均の線形推定値とサイズとの比の推定値は，いずれも関数 svymean() を用いて求められる．例えば平均身長の推定値は $\hat{\mu}_y = 163.575$ となり，標準誤差は $\widehat{SE}(\hat{\mu}_y) = 0.2140$ となる．

演習 5.4 解答例 -2

```
##### 母集団平均の推定 #####
> svymean(~height + weight, si)
          mean     SE
height 163.575 0.2140
weight  58.231 0.2694
```

性別の推定を行うには関数 svyby() を用いる．部分母集団平均の線形推定値 $\hat{\mu}_{y,d}$ を求めるには，まず部分母集団総計の線形推定値 $\hat{\tau}_{y,d}$ を求め，次にそれを部分母集団サイズ N_d で割ればよい．以下の例では，例えば男子の平均身長の線形推定値は $\hat{\mu}_{y,\,男} = 177.3533$ となり，標準誤差は $\widehat{SE}(\hat{\mu}_{y,\,男}) = 5.663987$ となる．

演習 5.4 解答例 -3

```
##### 性別の部分母集団平均の線形推定 #####
> svyby(~height + weight, ~gender, si, svytotal) / c(375574, 373995)
        gender   height   weight se.height se.weight
1 2.662591e-06 177.3533 66.52344  5.663987  2.140154
2 5.347665e-06 149.7383 49.90354  5.307104  1.792924
```

サイズとの比の推定値 $\hat{\mu}_{y,d,\mathrm{N}}$ を求めるには関数 svymean() を用いればよい．例えば男子の平均身長の推定値は $\hat{\mu}_{y,\,男,\mathrm{N}} = 168.9361$ であり，標準誤差は $\widehat{SE}(\hat{\mu}_{y,\,男,\mathrm{N}}) = 0.1364715$ となる．この標準誤差は，**演習 5.4 解答例 -3** に示す線形推定値の標準誤差 $\widehat{SE}(\hat{\mu}_{y,\,男}) = 5.663987$ と比べ明らかに小さい．

演習 5.4 解答例 -4

```
##### 性別のサイズとの比の推定 #####
> svyby(~height + weight, ~gender, si, svymean)
  gender   height   weight se.height se.weight
1      1 168.9361 63.36624 0.1364715 0.2513461
2      2 157.6251 52.53198 0.1427817 0.3127896
```

演習 5.5 解答例

以下の例ではまず非復元単純無作為抽出法を指定している．

演習 5.5 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

以下の例ではまず関数 subset() を用いて所在地 area が 1 であり，かつ業種 gyoshu が 1 である部分母集団を指定し，関数 svymean() を用いることでサイズとの比の推定値 $\hat{\mu}_{y,d,\mathrm{N}} = \hat{\tau}_{y,d}/\hat{N}_d$ を求めている．推定結果は mu に代入している．

演習 5.5 解答例 -2

```
##### 部分母集団サイズとの比の推定値 #####
> (mu <- svymean(~shihon + uriage, subset(si, area==1 & gyoshu==1)))
         mean      SE
shihon 123.58 12.6459
uriage  93.47  8.6962
```

次に上記で求めた部分母集団平均の推定値 $\hat{\mu}_{y,d,\mathrm{N}}$ に真の部分母集団サイズ N_d を乗じることで，部分母集団総計の推定を行っている．**例題 5.6** では補助変数 $x_i = 1$ と関数 svyratio() を利用することで，サイズを用いた比推定値 $\hat{\tau}_{y,d,\mathrm{N}}$ を求めている．この**演習 5.5** で示した方法は**例題 5.6** の方法に代わるものである．

演習 5.5 解答例 -3

```
##### 部分母集団総計の比推定値 #####
> coef(mu) * 400
shihon uriage
 49432  37388

##### 比推定値の標準誤差 #####
> SE(mu) * 400
  shihon   uriage
5058.346 3478.485
```

演習 5.6 解答例

母集団サイズ N は変数 N に代入されている．そこでまず関数 svydesign() の引数に fpc=N を指定して非復元単純無作為抽出法を指定する．

演習 5.6 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=otona)
```

単純無作為抽出法のとき母集団割合 p_y の線形推定を行うには，factor である変数に対して関数 svymean() を用いればよい．質問項目 Q1 に対して 1 と回答する児童の母集団割合の線形推定値は $\hat{p}_{\texttt{Q1=1}} = 0.824150$ であり，その標準誤差は $\widehat{SE}(\hat{p}_{\texttt{Q1=1}}) = 0.0130$ となる．

演習 5.6 解答例 -2

```
##### 母集団割合の推定 #####
> svymean(~factor(Q1) + factor(Q3), si)
                mean     SE
factor(Q1)1 0.824150 0.0130
factor(Q1)2 0.175850 0.0130
factor(Q3)1 0.187573 0.0134
factor(Q3)2 0.548652 0.0170
factor(Q3)3 0.206331 0.0139
factor(Q3)4 0.057444 0.0080
```

市郡別の部分母集団割合を線形推定するには，まず市郡別に各カテゴリを選ぶ児童の人数を関数 subset() と関数 svytotal() を用いて推定し，それを部分母集団サイズで割ればよい．例えば市郡 city が 1 では，質問項目 Q1 に 1 と回答する児童の割合の推定値は $\hat{p}_{\texttt{Q1=1,1}} = 0.9398044$ であり，2 と回答する児童の割合の推定値は $\hat{p}_{\texttt{Q1=2,1}} = 0.1365079$ である．ただし両者の和は 100% とはならない．

演習 5.6 解答例 -3

```
##### 市郡別の部分母集団割合の線形推定 #####
> svyby(~factor(Q1), ~city, si, svytotal) / c(225634, 279618)
          city factor(Q1)1 factor(Q1)2 se.factor(Q1)1 se.factor(Q1)2
1 4.431956e-06   0.9398044   0.1365079     0.03782787     0.01833942
2 7.152615e-06   0.7308243   0.2075965     0.03035628     0.01972394
```

サイズとの比の推定値を求めるには関数 svymean() を用いればよい．市郡 city が 1 では，質問項目 Q1 に 1 と回答する児童の割合の推定値は $\hat{p}_{\texttt{Q1=1,1,N}} = 0.8731707$ であり，2 と回答する児童の割合の推定値は $\hat{p}_{\texttt{Q1=2,1,N}} = 0.1268293$ である．両者の和は 100% となる．

演習 5.6 解答例 -4

```
##### 市郡別の部分サイズとの比の推定 #####
> svyby(~factor(Q1), ~city, si, svymean)
  city factor(Q1)1 factor(Q1)2 se.factor(Q1)1 se.factor(Q1)2
1    1   0.8731707   0.1268293     0.01643067     0.01643067
2    2   0.7787810   0.2212190     0.01971536     0.01971536
```

演習 5.7 解答例

母集団サイズ N は変数 N に代入されている．そこでまず関数 svydesign() の引数に fpc=N を指定して非復元単純無作為抽出法を指定する．

演習 5.7 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kenko)
```

母集団分散を推定するには関数 svyvar() を用いればよい[4]．例えば身長 height の母集団分散の推定値は $\hat{\sigma}^2_{y,\mathrm{n}} = 40.549$ であり，その標準誤差は $\widehat{SE}(\hat{\sigma}^2_{y,\mathrm{n}}) = 1.2018$ である．

演習 5.7 解答例 -2

```
##### heightの母集団分散の推定 #####
> svyvar(~height, si)
       variance     SE
height   40.549 1.2018

##### weightの母集団分散の推定 #####
> svyvar(~weight, si)
       variance     SE
weight   64.223 2.8841
```

性 gender 別に部分母集団分散を推定するには関数 svyby() を利用する．例えば男子の身長の部分母集団分散の推定値は $\hat{\sigma}^2_{\text{身長, 男子,n}} = 8.679468$ となる．

演習 5.7 解答例 -3

```
##### 性別のheightの部分母集団分散の推定 #####
> svyby(~height, ~gender, si, svyvar)
  gender   height se.height
1      1 8.679468 0.6038984
2      2 8.562839 0.5964137

##### 性別のweightの部分母集団分散の推定 #####
> svyby(~weight, ~gender, si, svyvar)
  gender   weight se.weight
1      1 29.44103  1.820353
2      2 41.09390  2.741195
```

[4]第一の引数に height と weight の両方を同時に指定すると，母集団分散共分散行列の推定値が表示されてしまう．そのためここでは変数ごとに svyvar() を用いている．

第6章　層化抽出法

6.a　層化抽出法の指定

層化抽出法の指定

```
DES <- svydesign(ids=~1, strata=~H, fpc=~N, weights=~W, data=DATA)
```

層化抽出法では，関数 svydesign() の引数 strata に層化変数を指定する．層内が復元抽出であれば引数 fpc は指定せず，非復元抽出であれば層サイズ N_h が代入された変数を N に指定する．層の間で復元と非復元が混在するのであれば，引数 fpc に指定する変数の値を復元の層は Inf (無限大) とする[1]．**演習 3.3** を参照のこと．層内で変数 N の値が異なると警告メッセージが表示される．抽出ウェイト W は各層の標本抽出方法に応じて適切に定める．

標本サイズ 1 の対処法の指定

```
options(survey.lonely.psu=LONELY)
```

標本サイズが 1 の層への対処法 (p.107) は，options() で引数 survey.lonely.psu により指定する．この指定は svytotal() や svymean() などを用いる前に行う．デフォルトでは LONELY は"fail"であり，標本サイズ 1 の層があるとエラーとなる．"certainty"あるいは"remove"は除去法，"average"は平均法，"adjust"は調整法[2]の指定となる．融合法を行うにはこの options() で指定するのではなく，データフレームの中の層化変数を適宜修正する．

[1] 抽出率を指定しているのであれば 0 とする．
[2] バージョン 3.13 では (6.29) 式を例えば $w_i^2 y_i^2$ などと計算している．

例題 6.2　層化抽出法における線形推定量

表 6.2 (p.100) の層化抽出標本を用いて，売上高の母集団総計を推定してみよう．以下の例ではまず関数 data.frame() を用いてデータフレーム data を作成している．データフレームの変数 y は売上高であり，変数 h は層化変数である規模である．変数 N.h は各層サイズ N_h であり，変数 n.h は各層の標本サイズ n_h である．

例題 6.2-1

```
##### データの作成 #####
> (data <- data.frame(y=c(158, 65, 380, 74, 236, 636, 465, 565, 660),
    h=c('小', '小', '中', '中', '中', '大', '大', '大', '大'),
    N.h=c(8, 8, 7, 7, 7, 5, 5, 5, 5), n.h=c(2, 2, 3, 3, 3, 4, 4, 4, 4)))
    y  h N.h n.h
1 158 小   8   2
2  65 小   8   2
3 380 中   7   3
4  74 中   7   3
5 236 中   7   3
6 636 大   5   4
7 465 大   5   4
8 565 大   5   4
9 660 大   5   4
```

以下の例では関数 factor() を用いて層化変数 h のカテゴリの順序を' 小', ' 中', ' 大' となるようにしている．この作業は必ずしも必要なく，関数 svydesign() の引数 strata に指定する変数は必ずしも factor でなくともよい．

例題 6.2-2

```
##### 層化変数の順序 #####
> data$h <- factor(data$h, levels=c('小', '中', '大'))
```

どの層も単純無作為抽出なので，抽出ウェイトは $w_i = N_h/n_h$ により求まる．

例題 6.2-3

```
##### 抽出ウェイトの作成 #####
> (data$w <- data$N.h / data$n.h)
[1] 4.000000 4.000000 2.333333 2.333333 2.333333 1.250000 1.250000 1.250000 1.250000
```

以下の例ではまず，関数 svydesign() を用いて層化抽出法を指定している．引数 strata に指定するのは層化変数 h である．どの層内も非復元単純無作為抽出なので，引数 fpc には層サイズの変数 N.h を指定する．作成した survey.design オブジェクトは stsi である．

例題 6.2-4

```
##### 層化抽出法の指定 #####
> stsi <- svydesign(ids=~1, strata=~h, fpc=~N.h, weights=~w, data=data)
```

関数 summary() を用いて stsi の内容を確認すると，一行目に Stratified と表示され，層化抽出法が指定されていることが分かる．Stratum Sizes:を見ると，小から大まで三つの層があることが分かる．それぞれの層の標本サイズは 2，3，4 である．obs と design.PSU, actual.PSU の三つの違いは，集落抽出法の第 8 章や多段抽出法の第 9 章の例題で見ていく．Population stratum sizes (PSUs):に示されるのは各層サイズ N_h である．仮に引数 fpc を指定しなければ，これは表示されない．

例題 6.2-5

```
##### survey.designオブジェクトの確認 #####
> summary(stsi)
Stratified Independent Sampling design
svydesign(ids = ~1, strata = ~h, fpc = ~N.h, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2500  0.4286  0.4286  0.5540  0.8000  0.8000
Stratum Sizes:
            小 中 大
obs          2  3  4
design.PSU   2  3  4
actual.PSU   2  3  4
Population stratum sizes (PSUs):
小 中 大
 8  7  5
Data variables:
[1] "y"   "h"   "N.h" "n.h" "w"
```

以下の例では関数 svytotal() を用いて売上高の母集団総計を線形推定している．推定値は $\hat{\tau}_y = 5409.50$ であり，標準誤差は $\widehat{SE}(\hat{\tau}_y) = 576.28$ となる．これらは本書の (6.10) 式と (6.11) 式に対応する．なおデザイン効果の推定値は $\mathrm{deff} = 0.2832$ であり，層化抽出法を採用することで単純無作為抽出法に比べ推定量の分散は 3 割程度となっていることが分かる．

例題 6.2-6

```
##### 母集団総計の推定 #####
> svytotal(~y, stsi, deff=TRUE)
    total      SE   DEff
y 5409.50  576.28 0.2832
```

関数 svyby() を用いて各層の総計を推定すると，例えば小規模は $\hat{\tau}_{y,\,小} = 892.0$ となる．この結果は本書の (6.9) 式に対応する．

例題 6.2-7

```
##### 層ごとの総計の推定 #####
> svyby(~y, ~h, stsi, svytotal)
    h      y      se.y
小 小  892.0 322.16145
中 中 1610.0 467.69221
大 大 2907.5  97.84554
```

6.a.1 演習問題

演習 6.1 層化抽出法と母集団総計の推定

データフレーム kigyo を用いて，資本金 shihon と売上高 uriage の母集団総計を線形推定してみよう．ただし標本は所在地 area と業種 gyoshu の組み合わせで層化抽出されたものとし，各層内は非復元単純無作為抽出とする．

ヒント：層化変数としてまず所在地 area と業種 gyoshu を組み合わせた変数を作成する．例えば `kigyo$h <- paste(kigyo$area, kigyo$gyoshu, sep='.')` などとすればよい．層サイズ N_h は変数 N.h に代入されている．

演習 6.2 層化確率比例抽出法と母集団総計の推定

データフレーム kigyo を用いて，資本金 shihon と売上高 uriage の母集団総計を線形推定してみよう．ただし標本は所在地 area と業種 gyoshu の組み合わせで層化抽出されたものとし，各層内は資本金 shihon で復元確率比例抽出されたものとする．

ヒント：資本金の層総計 $\tau_{x,h}$ は変数 total.shihon.h に代入されており，層ごとの標本サイズ n_h は変数 n.h に代入されている．

6.b　層化抽出法における比推定

層化抽出法における母集団比の推定

```
STAT <- svyratio(numerator=~Y, denominator=~X, design=DES,
                                                    separate=FALSE)
```

5.a 節で説明したように，母集団比 $R = \tau_y/\tau_x$ の推定を行うには関数 `svyratio()` を用いる．層化抽出法では，結合比推定量とするのであれば引数 `separate` を `FALSE` (デフォルト) とし，個別比推定量とするのであれば `TRUE` とする．`separate=TRUE` とすると層ごとの比 $R_h = \tau_{y,h}/\tau_{x,h}$ の推定値とその標準誤差が求められる．

層化抽出法における母集団総計の比推定

```
predict(STAT, total=TOTAL, se=TRUE, ...)
```

結合比推定では引数 `total` に補助変数の母集団総計 τ_x の値を一つ指定する．個別比推定では `c()` を用いて各層の補助変数の総計 $\tau_{x,h}$ を並べたベクトルを引数 `total` に指定する．

例題 6.5　比推定量の分散

本書の**例題 6.5** (p.110) で求めているのは各推定量の分散だが，ここでは表 6.5 (p.108) の層化抽出標本に基づいて 2 つの比推定値とその標準誤差の推定値を求めてみよう．

例題 6.5-1

```
##### データの作成 #####
> (data <- data.frame(y=c(158, 65, 380, 74, 236, 636, 465, 565, 660),
    x=c(19, 19, 31, 25, 36, 57, 51, 54, 52),
    h=c('小', '小', '中', '中', '中', '大', '大', '大', '大'),
    N.h=c(8, 8, 7, 7, 7, 5, 5, 5, 5), n.h=c(2, 2, 3, 3, 3, 4, 4, 4, 4)))
    y  x  h N.h n.h
1 158 19 小   8   2
2  65 19 小   8   2
3 380 31 中   7   3
4  74 25 中   7   3
5 236 36 中   7   3
6 636 57 大   5   4
7 465 51 大   5   4
8 565 54 大   5   4
9 660 52 大   5   4
```

以下の例では**例題 6.2-2** や**例題 6.2-3** と同様に層化変数 h と抽出ウェイト w を用意している．

例題 6.5-2

```
##### 層化変数の順序 #####
> data$h <- factor(data$h, levels=c('小', '中', '大'))

##### 抽出ウェイトの作成 #####
> (data$w <- data$N.h / data$n.h)
[1] 4.000000 4.000000 2.333333 2.333333 2.333333 1.250000 1.250000 1.250000 1.250000
```

関数 svydesign() では引数 strata によって層化抽出法であることを指定し，引数 fpc によって層内は非復元抽出であることを指定している．

例題 6.5-3

```
##### 層化抽出法の指定 #####
> stsi <- svydesign(ids=~1, strata=~h, fpc=~N.h, weights=~w, data=data)
```

まず求めるのは結合比推定量 $\hat{R}_{y,\mathrm{R}^c}$ である．以下の例ではまず引数 separate を指定せずに (デフォルトの FALSE を指定して)，関数 svyratio() を用いている．得られる値は母集団全体についての比の推定値 $\hat{R} = \hat{\tau}_y/\hat{\tau}_x = 8.530092$ である．

例題 6.5-4

```
##### 母集団比の推定 #####
> svyratio(~y, ~x, stsi)
Ratio estimator: svyratio.survey.design2(~y, ~x, stsi)
Ratios=
        x
y 8.530092
SEs=
         x
y 0.825258
```

次に predict() の二番目の引数に，資本金の母集団総計 $\tau_x = 146 + 243 + 274$ を指定し，結合比推定値 $\hat{\tau}_{y,\mathrm{R}^c} = 5655.451$ を求めている．これは本書の (6.30) 式に対応する．また，その標準誤差は $\widehat{SE}(\hat{\tau}_{y,\mathrm{R}^c}) = 547.146$ となる．

例題 6.5-5

```
##### 結合比推定 #####
> predict(svyratio(~y, ~x, stsir), 146+243+274)
$total
        x
y 5655.451

$se
        x
y 547.146
```

次に求めるのは個別比推定量 $\hat{R}_{y,\mathrm{R}^s}$ である．以下の例では，関数 svyratio() の引数 separate に TRUE を指定することで，層ごとの比の推定値 $\hat{R}_h = \hat{\tau}_{y,h}/\hat{\tau}_{x,h}$ を求めている．例えば一番目の小規模層の比の推定値は $\hat{R}_{y,\text{小}} = 5.868421$ であり，その標準誤差は $\widehat{SE}(\hat{R}_{y,\text{小}}) = 2.119483$ である．

例題 6.5-6

```
##### 層ごとの比の推定 #####
> svyratio(~y, ~x, stsi, separate=TRUE)
Stratified ratio estimate: svyratio.survey.design2(~y, ~x, stsi, separate = TRUE)
Ratio estimator: Stratum == 1L
Ratios=
         x
y 5.868421
SEs=
         x
y 2.119483
Ratio estimator: Stratum == 2L
Ratios=
    x
y 7.5
SEs=
         x
y 1.903912
Ratio estimator: Stratum == 3L
Ratios=
         x
y 10.86916
SEs=
          x
y 0.3212115
```

次に関数 predict() では，第二の引数に補助変数の層総計 $\tau_{x,h}$ を並べる．個々の比 $\hat{R}_h$ にこれらの $\tau_{x,h}$ が乗じられることで個別比推定値 $\hat{R}_{y,\mathrm{R}^s} = 5657.439$ が求まる．これは本書の (6.32) 式に対応する．またその標準誤差は $\widehat{SE}(\hat{R}_{y,\mathrm{R}^s}) = 563.5137$ である．

例題 6.5-7

```
##### 個別比推定 #####
> predict(svyratio(~y, ~x, stsi, separate=TRUE), c(146, 243, 274))
$total
         x
y 5657.439

$se
         x
y 563.5137
```

6.b.1　演習問題

演習 6.3　結合比推定量と個別比推定量

データフレーム kigyo を用いて，売上高 uriage の母集団総計の比推定をしてみよう．標本は所在地 area と業種 gyoshu の組み合わせで層化抽出されたものとし，各層内は非復元単純無作為抽出とする．比推定のための補助変数は資本金 shihon とし，結合比推定と個別比推定を行うこと．

ヒント：資本金 shihon の母集団総計は $\tau_x = 1,725,000$ である．資本金の層総計は変数 total.shihon.h に代入されている．この値を取り出すには，例えば

```
shihon.h <- aggregate(kigyo$total.shihon.h, kigyo['h',drop=F], mean)
```

などとすればよい．shihon.h$x が資本金の層総計となる．

6.c 事後層化推定

事後層化の指定

```
DES.PS <- postStratify(design=DES, strata=~D, population=~N)
```

事後層化を行うには関数 postStratify() を用いるのが一つの方法である[3]．引数 design には svydesign() の結果 DES を指定する．引数 strata には事後層の変数 D を指定し，引数 population には各事後層のサイズ N_d を含むデータフレーム N を指定する．具体的には**例題 6.7** を参照のこと．関数 svytotal() や svymean() の引数として DES.PS を用いればよい．

レイキングの指定

```
DES.RR <- rake(design=DES, sample.margins=X, population.margins=T)
```

レイキングを行うには関数 rake() を用いる[4]．引数 sample.margins の X には補助変数のリストを指定する．引数 population.margins の T には各補助変数の母集団総計を含むデータフレーム (postStratify() の引数 population に指定する形式) のリストを指定する．具体的には**例題 6.8** を参照のこと．

例題 6.7 事後層化推定

表 6.7 (p.111) あるいは表 5.9 (p.89) の非復元単純無作為抽出標本を用いて，“男に” 生まれ変わりたい人の割合の事後層化推定値を求めてみよう．

例題 6.7-1

```
##### データの作成 #####
> (data <- data.frame(y=c('女に', '男に', '男に', '女に', '男に'),
    gen=c('女性', '女性', '男性', '女性', '男性'), N=20))
     y  gen  N
1 女に 女性 20
2 男に 女性 20
3 男に 男性 20
4 女に 女性 20
5 男に 男性 20
```

[3]もう一つの方法は関数 calibrate() を用いることである (第 7 章).
[4]関数 calibrate() を用いてもレイキングを行うことができる．

以下の例では関数 factor() を用いて回答の変数 y と性別の変数 gen のカテゴリの順序を指定している．この作業は必ずしも必要ないが，特に変数 gen ではカテゴリの順序が混乱しないようにするための処置である．

例題 6.7-2

```
##### 回答 #####
> data$y <- factor(data$y, levels=c('男に', '女に'))

##### 性別 #####
> data$gen <- factor(data$gen, levels=c('男性', '女性'))
```

関数 svydesign() を用いて非復元単純無作為抽出法を指定した結果は si である．

例題 6.7-3

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

性別 gen で事後層化を行うために，以下の例ではまず性別の事後層サイズ N_d を変数 N.d としたデータフレーム N.d を作成している．

例題 6.7-4

```
##### 事後層サイズ #####
> (N.d <- data.frame(gen=c('男性', '女性'), N.d=c(10, 10)))
   gen N.d
1 男性  10
2 女性  10
```

次に関数 postStratify() を用いて新たな survey.design オブジェクト si.ps を作成している．第一の引数に指定するのは元となる survey.design オブジェクト si であり，第二の引数は事後層化変数 gen である．第三の引数には事後層サイズが入ったデータフレーム N.d を指定する．

例題 6.7-5

```
##### 事後層化の指定 #####
> si.ps <- postStratify(si, strata=~gen, population=N.d)
```

事後層化した si.ps を用いて “男に” 生まれ変わりたい人の割合を求めると $\hat{p}_{y,\mathrm{PS}} = 0.66667$ となり，標準誤差は $\widehat{SE}(\hat{p}_{y,\mathrm{PS}}) = 0.1318$ となる．これらは本書の (6.51) 式と (6.52) 式に対応する．なお関数 svymean() の引数を factor(y) としていないのは，**例題 6.7-2** で既に factor としてあるからである．

例題 6.7-6

```
##### 事後層化推定 #####
> svymean(~y, si.ps)
         mean     SE
y男に 0.66667 0.1318
y女に 0.33333 0.1318
```

例題 6.8　レイキング比推定

本書の**例題 6.8** (p.118) で示したレイキングを行ってみよう．以下では抽出ウェイトを $w_i = 1$ としてデータフレーム data を作成している．

例題 6.8-1

```
##### データの作成 #####
> tmp <- data.frame(gender=c(1,1,1,2,2,2), age=c(1,2,3,1,2,3),
    N=c(733, 835, 745, 818, 944, 840))
> data <- data.frame(gender=rep(1,tmp$N[1]), age=rep(1,tmp$N[1]))
> for (i in c(2:nrow(tmp))) {
      data <- rbind(data, data.frame(gender=rep(tmp$gender[i],tmp$N[i]),
                                              age=rep(tmp$age[i],tmp$N[i])))
  }
> data$w <- 1
```

以下の例ではデータフレーム data に対して復元単純無作為抽出法を指定している．

例題 6.8-2

```
##### 復元単純無作為抽出法の指定 #####
> sir <- svydesign(ids=~1, weights=~w, data=data)
```

抽出ウェイトの性別・年齢層別標本合計は関数 svytable() を用いて求めることができる．この結果は本書の表 6.9 (p.116) に対応する．

例題 6.8-3

```
##### 部分母集団サイズの推定値 #####
> svytable(~gender + age, sir)
      age
gender   1   2   3
     1 733 835 745
     2 818 944 840
```

レイキングを行うには各部分母集団サイズを指定する必要がある．以下の例では性 gender と年齢 age の各部分母集団サイズが代入されたデータフレームのリスト pop を作成している．

例題 6.8-4

```
##### 部分母集団サイズ #####
> (pop <- list(data.frame(gender=c(1,2), N.d=c(2470, 2420)),
     data.frame(age=c(1,2,3), N.e=c(1523, 1811, 1556))))
[[1]]
  gender  N.d
1      1 2470
2      2 2420

[[2]]
  age  N.e
1   1 1523
2   2 1811
3   3 1556
```

以下の例では関数 rake() を用いてレイキングを行っている．第一の引数に指定しているのは元の標本抽出デザインが指定された survey.design オブジェクト sir である．第二の引数には補助変数のリストを指定し，第三の引数で部分母集団サイズが代入されたデータフレームのリスト pop を指定している．

例題 6.8-5

```
##### レイキング #####
> sir.rake <- rake(sir, list(~gender, ~age), pop)
```

以下では関数 svytable() を用いて，レイキングされたウェイトの標本総計をセルごとに求めている．これは本書の表 6.10 に対応する．

例題 6.8-6

```
##### 部分母集団サイズの推定値 #####
> svytable(~gender + age, sir.rake)
      age
gender        1        2        3
     1 772.3639 912.5433 785.0928
     2 750.6361 898.4567 770.9072
```

補助変数とした性 gender 別あるいは年齢 age 別の部分母集団サイズを関数 svytable() を用いて推定すると，いずれも真の部分母集団サイズに一致していることが分かる．

例題 6.8-7

```
##### 性別の部分母集団サイズの推定値 #####
> svytable(~gender, sir.rake)
gender
   1    2
2470 2420

##### 年齢別の部分母集団サイズの推定値 #####
> svytable(~age, sir.rake)
age
   1    2    3
1523 1811 1556
```

6.c.1　演習問題

演習 6.4　単純無作為抽出法と母集団平均の事後層化推定量

データフレーム kenko が非復元単純無作為抽出されたものとして，身長 height と体重 weight の母集団平均の事後層化推定値 $\hat{\mu}_{y,\mathrm{PS}}$ を求めてみよう．ただし事後層は性 gender とする．事後層化推定量と (事後層化をしない) 線形推定量の標準誤差を比べてみること．

ヒント：性 gender 別の部分母集団サイズは男子が $N_{\text{男}} = 375,574$ であり，女性が $N_{\text{女}} = 373,995$ である．

演習 6.5　単純無作為抽出法と母集団平均のレイキング比推定量

同じくデータフレーム kenko が非復元単純無作為抽出されたものとして，身長 height と体重 weight の母集団平均のレイキング比推定値 $\hat{\mu}_{y,\mathrm{RR}}$ を求めてみよう．ただし事後層は性 gender と所在地 area とする．

ヒント：性 gender 別の部分母集団サイズは男子が $N_{\text{男}} = 375,574$ であり，女性が $N_{\text{女}} = 373,995$ である．また所在地 area 別の部分母集団サイズは 1 が $N_1 = 439,996$ であり，2 が $N_2 = 309,573$ である．

6.d　演習問題解答例

演習 6.1 解答例

以下の例では，まず層化変数として所在地 area と業種 gyoshu を組み合わせた変数 h を関数 paste() を用いて作成している．例えば所在地が 1 で業種が 2 の企業は変数 h の値が 1.2 となる．

演習 6.1 解答例 -1

```
##### 層化変数の作成 #####
> kigyo$h <- paste(kigyo$area, kigyo$gyoshu, sep='.')
```

関数 svydesign() を用いて層化抽出法を指定するには，引数 strata に層化変数 h を指定する．また層内では非復元抽出なので，層サイズ N_h が代入された変数 N.h を引数 fpc に指定する．引数 fpc が指定されているので，抽出ウェイトは自動的に $w_i = N_h/n_h$ とされる．

演習 6.1 解答例 -2

```
##### 層化抽出法の指定 #####
> stsi <- svydesign(ids=~1, strata=~h, fpc=~N.h, data=kigyo)
```

関数 summary() を用いると，stsi では全部で 15 の層が指定されていることが確認できる．

演習 6.1 解答例 -3

```
##### survey.designオブジェクトの確認 #####
> summary(stsi)
Stratified Independent Sampling design
svydesign(ids = ~1, strata = ~h, fpc = ~N.h, data = kigyo)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1111  0.1556  0.2000  0.2591  0.2933  0.8000
Stratum Sizes:
            1.1 1.2 1.3 1.4 1.5 2.1 2.2 2.3 2.4 2.5 3.1 3.2 3.3 3.4 3.5
obs         100 100 100 100 100 140 140 140 140 140 160 160 160 160 160
design.PSU 100 100 100 100 100 140 140 140 140 140 160 160 160 160 160
actual.PSU 100 100 100 100 100 140 140 140 140 140 160 160 160 160 160
Population stratum sizes (PSUs):
 1.1  1.2  1.3  1.4  1.5  2.1  2.2  2.3  2.4  2.5  3.1  3.2  3.3  3.4  3.5
 400  600  300  900  300  600  900  500 1100  400  900  800 1200  900  200
Data variables:
 [1] "obs"            "area"           "gyoshu"         "shihon"         "uriage"
 [6] "uriage.na"      "N"              "n"              "N.h"            "n.h"
[11] "total.shihon"   "total.shihon.h" "h"
```

母集団総計の線形推定を行うには関数 svytotal() を用いればよい．

演習 6.1 解答例 -4

```
##### 母集団総計の線形推定 #####
> svytotal(~shihon + uriage, stsi)
         total    SE
shihon 1238587 32486
uriage 1006839 23478
```

演習 6.2 解答例

演習 6.1 解答例 -1 と同様に，まず層化変数として所在地 area と業種 gyoshu を組み合わせた変数 h を作成する．

演習 6.2 解答例 -1

```
##### 層化変数の作成 #####
> kigyo$h <- paste(kigyo$area, kigyo$gyoshu, sep='.')
```

層内では資本金 shihon で確率比例抽出されているので，抽出ウェイトは $w_i = \tau_{x,h}/(n_h x_i)$ となる．資本金の層総計 $\tau_{x,h}$ は変数 total.shihon.h に，層標本サイズ n_h は変数 n.h に代入されている．

演習 6.2 解答例 -2

```
##### 抽出ウェイトの作成 #####
> kigyo$w <- kigyo$total.shihon.h / (kigyo$n.h * kigyo$shihon)
```

関数 svydesign() では引数 strata に層化変数 h を指定し，引数 weights に抽出ウェイトの変数 w を指定する．各層内は復元抽出なので引数 fpc は指定しない．

演習 6.2 解答例 -3

```
##### 層化確率比例抽出法の指定 #####
> stpps <- svydesign(ids=~1, strata=~h, weights=~w, data=kigyo)
```

母集団総計の線形推定を行うには関数 svytotal() を用いればよい．資本金 shihon で確率比例抽出されているので，資本金 shihon の母集団総計の推定値は真の母集団総計 1725000 に一致する．

演習 6.2 解答例 -4

```
##### 母集団総計の線形推定 #####
> svytotal(~shihon + uriage, stpps)
         total        SE
shihon 1725000 3.133e-12
uriage 1912925     33406
```

演習 6.3 解答例

演習 6.1 と同様に，まず層化変数として所在地 area と業種 gyoshu を組み合わせた変数 h を作成し，関数 svydesign() では引数 strata と fpc を指定する．

演習 6.3 解答例 -1

```
##### 層化変数の作成 #####
> kigyo$h <- paste(kigyo$area, kigyo$gyoshu, sep='.')

##### 層化抽出法の指定 #####
> stsi <- svydesign(ids=~1, strata=~h, fpc=~N.h, data=kigyo)
```

まず結合比推定量では，関数 svyratio() において引数 separate を指定せずに一つの比 $R = \tau_y/\tau_x$ を推定する．次に関数 predict() を用いてこれに資本金の母集団総計 $\tau_x = 1725000$ を乗じればよい．

演習 6.3 解答例 -2

```
##### 結合比推定量 #####
> predict(svyratio(~uriage, ~shihon, stsi), 1725000)
$total
        shihon
uriage 1402241

$se
         shihon
uriage 23305.77
```

個別比推定を行うには，補助変数の層総計 $\tau_{x,h}$ が必要である．以下の例では関数 aggregate() を用いて，変数 total.shihon.h に代入されている層総計の値を取り出している．

演習 6.3 解答例 -3

```
> shihon.h <- aggregate(kigyo$total.shihon.h, kigyo[,'h',drop=F], mean)
```

関数 svyratio() で引数 separate=TRUE を指定すると，層ごとの比の推定値 $\hat{R}_h$ が得られる．そこで関数 predict() を用いて，これに上記で取り出した補助変数の層総計 $\tau_{x,h}$ を乗じればよい．

演習 6.3 解答例 -4

```
##### 個別比推定量 #####
> predict(svyratio(~uriage, ~shihon, stsi, separate=TRUE), shihon.h$x)
$total
        shihon
uriage 1408521

$se
         shihon
uriage 21171.83
```

演習 6.4 解答例

まず標本全体に対して非復元単純無作為抽出法を指定する．

演習 6.4 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kenko)
```

次に部分母集団サイズから成るデータフレームを作成する．以下の例では関数 data.frame() を用いて，事後層を表す変数 gender と事後層サイズを表す変数 N.d から成るデータフレーム N.d を作成している．

演習 6.4 解答例 -2

```
##### 事後層サイズ #####
> (N.d <- data.frame(gender=c(1, 2), N.d=c(375574, 373995)))
  gender    N.d
1      1 375574
2      2 373995
```

事後層化ウェイトを求めるには関数 postStratify() を用いる．以下の例では，第二の引数に事後層化変数 gender を指定し，第三の引数に事後層サイズから成るデータフレーム N.d を指定している．

演習 6.4 解答例 -3

```
##### 事後層化の指定 #####
> si.ps <- postStratify(si, ~gender, N.d)
```

身長 height と体重 weight の母集団平均の事後層化推定値はそれぞれ $\hat{\mu}_{\text{身長,PS}} = 163.293$ と $\hat{\mu}_{\text{体重,PS}} = 57.961$ となる．

演習 6.4 解答例 -4

```
##### 母集団平均の事後層化推定 #####
> svymean(~height + weight, si.ps)
          mean     SE
height 163.293 0.0987
weight  57.961 0.2005
```

線形推定値と比べると，事後層化推定値の標準誤差は特に身長 height では半分以下となっていることが分かる．

演習 6.4 解答例 -5

```
##### 母集団平均の線形推定 #####
> svymean(~height + weight, si)
          mean     SE
height 163.575 0.2140
weight  58.231 0.2694
```

演習 6.5 解答例

まず標本全体に対して非復元単純無作為抽出法を指定する.

演習 6.5 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kenko)
```

次に部分母集団サイズから成るデータフレームを性 gender と所在地 area のそれぞれについて作成し，それらをまとめたリスト pop を作成する.

演習 6.5 解答例 -2

```
##### 事後層サイズ #####
> (pop <- list(data.frame(gender=c(1, 2), N.d=c(375574, 373995)),
+                 data.frame(area=c(1, 2), N.e=c(439996, 309573))))
[[1]]
  gender    N.d
1      1 375574
2      2 373995

[[2]]
  area    N.e
1    1 439996
2    2 309573
```

レイキングを行うには関数 rake() を用いる．第二の引数には 2 つの補助変数のリストを指定し，第三の引数には**演習 6.5 解答例 -2** で作成したデータフレームのリスト pop を指定する.

演習 6.5 解答例 -3

```
##### レイキングの指定 #####
> si.rr <- rake(si, list(~gender, ~area), pop)
```

関数 svymean() を用いて母集団平均のレイキング比推定値を求めると，身長については $\hat{\mu}_{\text{身長,RR}} = 163.288$ となる.

演習 6.5 解答例 -4

```
##### 母集団平均のレイキング比推定 #####
> svymean(~height + weight, si.rr)
          mean     SE
height 163.288 0.0984
weight  57.964 0.2011
```

第7章　回帰推定量

7.a　キャリブレーション推定

キャリブレーションの指定

```
DES.C <- calibrate(design=DES, formula=~X, population=T,
           calfun=c("linear","raking","logit"), bounds=c(-Inf,Inf))
```

キャリブレーションウェイトを求めるには関数 calibrate() を用いる[1]．引数 formula の X には K 個の補助変数を並べる．カテゴリカルな変数として扱いたい場合には factor() を用いる．引数 population の T に補助変数の母集団総計のベクトル $\boldsymbol{\tau}_x$ を指定する．ただし T の最初の値は母集団サイズ N とする．また X がカテゴリカル変数の場合には，最初のカテゴリに対応する値を T から除く．引数 calfun には距離関数を指定する．"linear"は線形関数，"raking"は乗法関数，"logit"はロジット関数の指定となる．ただし本書の (7.47) 式における C は指定できず，$C = 1$ に固定される[2]．デフォルトは"linear"である．引数 bounds には w_i^c/w_i の上限 L と下限 U を指定する．引数 bounds は"linear"と"raking"では必ずしも指定しなくともよい．"logit"では必須である．

例題 7.4　一般化回帰推定

表 7.2 (p.125) の標本を用いて売上高の母集団総計の一般化回帰推定を行ってみよう．

例題 7.4-1

```
##### データの作成 #####
> (data <- data.frame(y=c(380, 639, 209), x=c(31, 60, 28), N=20))
    y  x  N
1 380 31 20
2 639 60 20
3 209 28 20
```

非復元単純無作為抽出法を指定した結果は si である．

[1]引数 formula= X-1 とし，引数 variance=1 を指定すると関数 calibrate() を用いて比推定を行うことも可能であるが，本資料では触れない．

[2]SUDAAN では C を指定できる．

例題 7.4-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

関数 weights() で抽出ウェイトを確かめると，どの企業も $w_i = N/n = 20/3 = 6.666667$ である．

例題 7.4-3

```
##### 抽出ウェイト #####
> weights(si)
       1        2        3
6.666667 6.666667 6.666667
```

以下の例ではまず関数 calibrate() を用いてウェイトのキャリブレーションを行っている．第一の引数に指定するのは svydesign() の結果 si である．第二の引数に指定するのは補助変数である資本金の変数 x である．定数である補助変数 $x_i = 1$ は特に指定しなくとも自動的に含まれる．第三の引数には定数の母集団総計すなわち母集団サイズ $N = 20$ と資本金の母集団総計 $\tau_{x_{(2)}} = 663$ を並べて指定する．キャリブレーションの結果は si.c である．引数 calfun を指定しなかったので線形関数を用いることになり，si.c を用いた推定は一般化回帰推定となる．

例題 7.4-4

```
##### ウェイトのキャリブレーション #####
> si.c <- calibrate(si, ~x, c(20, 663))
```

関数 weights() を用いてキャリブレーションウェイト w_i^c を確かめると，例えば最初の企業については $w_i^c = 8.474920$ となっている．これは抽出ウェイト w_i に g ウェイト $g_i = 1.2712380$ を乗じた値である．これらの結果は本書の表 7.3 (p.127) に対応する．

例題 7.4-5

```
##### キャリブレーションウェイト #####
> weights(si.c)
       1        2        3
8.474920 2.424226 9.100854

##### gウェイト #####
> weights(si.c) / weights(si)
        1         2         3
1.2712380 0.3636339 1.3651281
```

売上高の母集団総計の一般化回帰推定値は，関数 svytotal() の二番目の引数に si.c を指定することで $\hat{\tau}_{y,\mathrm{GREG}} = 6671.6$ となり，標準誤差は $\widehat{SE}(\hat{\tau}_{y,\mathrm{GREG}}) = 953.74$ となる．これらは本書の (7.25) 式と (7.32) 式に対応する．

例題 7.4-6

```
##### 母集団総計のキャリブレーション推定 #####
> svytotal(~y, si.c)
   total     SE
y 6671.6 953.74
```

例題 6.7　事後層化推定

本書の**例題 6.7** (p.115) では事後層化推定を行っている．これをキャリブレーション推定の一つとして，関数 calibrate() を用いて行ってみよう．以下の例ではまずデータフレーム data を作成している．

例題 6.7-1

```
##### データの作成 #####
> (data <- data.frame(y=c('女に', '男に', '男に', '女に', '男に'),
    gen=c('女性', '女性', '男性', '女性', '男性'), N=20))
     y  gen  N
1 女に 女性 20
2 男に 女性 20
3 男に 男性 20
4 女に 女性 20
5 男に 男性 20

##### 回答 #####
> data$y <- factor(data$y, levels=c('男に', '女に'))

##### 性別 #####
> data$gen <- factor(data$gen, levels=c('男性', '女性'))
```

非復元単純無作為抽出法を指定した結果は si である．

例題 6.7-2

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=data)
```

以下の例では関数 calibrate() を用いてウェイトのキャリブレーションを行っている．第二の引数に指定しているのは性別を表す変数 gen であり，既に factor となったものである．第三の引数には母集団サイズ $N = 20$ と女性の部分母集団サイズ $N_{女} = 10$ を指定する．つまり性 gender の最初のカテゴリである男性の部分母集団サイズは指定しない．キャリブレーション結果は si.c に代入している．

例題 6.7-3

```
##### ウェイトのキャリブレーション #####
> si.c <- calibrate(si, ~gen, c(20, 10))
```

キャリブレーションウェイトを確かめると男性については $w_i^c = 5.000000$ となっており，女性については $w_i^c = 3.333333$ となっている．これらは本書の (6.44) 式と (6.45) 式に対応する．

例題 6.7-4

```
##### キャリブレーションウェイト #####
> weights(si.c)
       1        2        3        4        5
3.333333 3.333333 5.000000 3.333333 5.000000
```

関数 svymean() を用いて，各回答のキャリブレーション推定値を求めると “男に” 生まれ変わりたい人の割合は $\hat{p}_{y,\mathrm{C}} = 0.66667$ となり，標準誤差は $\widehat{SE}(\hat{p}_{y,\mathrm{C}}) = 0.1318$ となる．これらは本書の (6.51) 式と (6.52) 式に対応する．

例題 6.7-5

```
##### 母集団割合のキャリブレーション推定 #####
> svymean(~y, si.c)
         mean     SE
y男に 0.66667 0.1318
y女に 0.33333 0.1318
```

例題 6.8　レイキング比推定

本書の**例題 6.8** (p.118) で示したレイキング比推定を関数 calibrate() を用いて行ってみよう．以下では抽出ウェイトを $w_i = 1$ としてデータフレーム data を作成している．

例題 6.8-1

```
##### データの作成 #####
> tmp <- data.frame(gender=c(1,1,1,2,2,2), age=c(1,2,3,1,2,3),
    N=c(733, 835, 745, 818, 944, 840))
> data <- data.frame(gender=rep(1,tmp$N[1]), age=rep(1,tmp$N[1]))
> for (i in c(2:nrow(tmp))) {
      data <- rbind(data, data.frame(gender=rep(tmp$gender[i],tmp$N[i]),
                                     age=rep(tmp$age[i],tmp$N[i])))
  }
> data$w <- 1
```

復元単純無作為抽出法を指定した結果は sir である．

例題 6.8-2

```
##### 復元単純無作為抽出法の指定 #####
> sir <- svydesign(ids=~1, weights=~w, data=data)
```

以下の例では calibrate() を用いてウェイトのキャリブレーションを行っている．二番目の引数には性別の変数 gender と年齢層の変数 age を factor() を使ってカテゴリであることを指示しながら指定している．第三の引数には母集団サイズ $N = 4890$ と女性のサイズ $N_{女性} = 2420$，30 歳代と 40 歳代のサイズ $N_{30\text{歳代}} = 1811$ と $N_{40\text{歳代}} = 1556$ を指定している．つまり各変数の最初のカテゴリである男性と 20 歳代のサイズは指定しない．さらに第四の引数 calfun には"raking"を指定している．レイキング比推定量は距離関数を乗法関数としたキャリブレーション推定量だからである．

例題 6.8-3

```
##### ウェイトのキャリブレーション #####
> sir.c <- calibrate(sir, ~factor(gender) + factor(age), c(4890, 2420, 1811, 1556),
                                                                    calfun="raking")
```

性別あるいは年齢層別のキャリブレーションウェイト総計を関数 svytable() を用いて求めると，指定したサイズに一致していることが分かる．

例題 6.8-4

```
##### 性別のキャリブレーションウェイト総計 #####
> svytable(~gender, sir.c)
gender
   1    2
2470 2420

##### 年齢層別のキャリブレーションウェイト総計 #####
> svytable(~age, sir.c)
age
   1    2    3
1523 1811 1556
```

さらに性と年齢層を組み合わせたセルごとのキャリブレーションウェイト総計も関数 svytable() を使って求められる．これらは本書の表 6.10 に対応する．

例題 6.8-5

```
##### 部分母集団サイズの推定値 #####
> svytable(~gender + age, sir.c)
      age
gender        1        2        3
     1 772.3639 912.5433 785.0928
     2 750.6361 898.4567 770.9072
```

7.a.1 演習問題

演習 7.1 単純無作為抽出法と母集団総計の回帰推定量

データフレーム kigyo が非復元単純無作為抽出されたものとして，売上高 uriage の母集団総計の回帰推定を行ってみよう．母集団値としては資本金 shihon の母集団総計と，所在地 area・業種 gyoshu ごとの部分母集団サイズを用いる．

さらにキャリブレーションウェイトを用いれば，補助変数の母集団総計の推定値が真値に一致することを確かめること．またキャリブレーションウェイトの分布を確認すること．

ヒント：資本金 shihon の母集団総計は $\tau_x = 1,725,000$ であり，所在地 area・業種 gyoshu ごとの部分母集団サイズは変数 N.h に代入されている．関数 calibrate() で母集団値を指定するときには，最初に母集団サイズ $N = 10,000$ を指定する必要がある．

演習 7.2 単純無作為抽出法と母集団平均のレイキング比推定量

データフレーム kenko が非復元単純無作為抽出されたものとして，関数 calibrate() を用いて身長 height と体重 weight の母集団平均のレイキング比推定を行ってみよう．ただし所在地 area 別の部分母集団サイズと性 gender 別の部分母集団サイズを用いてウェイトのキャリブレーションを行うこと．

さらに所在地と性に関しては，キャリブレーションウェイトの標本総計がそれぞれの部分母集団サイズに一致することを確かめてみよう．

ヒント：母集団サイズは $N = 749,569$ であり，所在地 area 別の部分母集団サイズは 1 が $N_1 = 439,996$，2 が $N_2 = 309,573$ である．性 gender 別の部分母集団サイズは 1 が $N_1 = 375,574$ であり，2 が $N_2 = 373,995$ である．レイキング比推定を行うには，関数 calibrate() で引数 calfun="raking"を指定すればよい．

7.b　演習問題解答例

演習 7.1 解答例

以下の例では，まず所在地 area と業種 gyoshu を組み合わせた変数 h を作成している．

演習 7.1 解答例 -1

```
##### 層化変数の作成 #####
> kigyo$h <- paste(kigyo$area, kigyo$gyoshu, sep='.')
```

変数 h の部分母集団サイズはデータフレーム kigyo の変数 N.h に代入されている．そこで以下の例では関数 aggregate() を利用して母集団値を並べたベクトル total を作成している．

演習 7.1 解答例 -2

```
##### 母集団値の作成 #####
> (total <- aggregate(kigyo$N.h, kigyo[,'h',drop=F], mean)[,2])
 [1]  400  600  300  900  300  600  900  500 1100  400  900  800 1200  900  200
```

ただし total の最初の値は不要である．そこで以下の例では total[-1] とすることで，最初の値を取り除いている．total の先頭には母集団サイズ $N = 10000$ と資本金の母集団総計 $\tau_x = 1725000$ を追加し，改めて total としている．

演習 7.1 解答例 -3

```
##### 母集団値の作成 #####
> (total <- c(10000, 1725000, total[-1]))
 [1]   10000 1725000     600     300     900     300     600     900     500    1100     400
[12]     900     800    1200     900     200
```

以下の例では標本抽出デザインとして非復元単純無作為抽出法を指定している．

演習 7.1 解答例 -4

```
> si <- svydesign(ids=~1, fpc=~N, data=kigyo)
```

ウェイトのキャリブレーションを行うには関数 calibrate() を用いる．第一の引数は**演習 7.1 解答例 -4** で作成した si である．第二の引数は補助変数である資本金 shihon と，所在地と業種を組み合わせた変数 h である．ただし変数 h はカテゴリカルな変数として扱うため factor(h) としている．第三の引数には母集団値を並べたベクトルである total を指定する．引数 calfun を指定していないので"linear"が指定されたことになり，si.GREG を用いた推定は一般化回帰推定となる．

演習 7.1 解答例 -5

```
##### ウェイトのキャリブレーション #####
> si.GREG <- calibrate(si, ~shihon + factor(h), total)
```

売上高 uriage の母集団総計の回帰推定を行うには，新たに作成した si.GREG を用いればよい．

演習 7.1 解答例 -6

```
##### 母集団総計の回帰推定 #####
> svytotal(~uriage, si.GREG)
         total    SE
uriage 1279949 19454
```

ウェイトのキャリブレーションに用いた補助変数 shihon と factor(h) に関して母集団総計を推定すると，いずれも真の母集団総計に一致することが確認できる．

演習 7.1 解答例 -7

```
##### 補助変数の母集団総計の推定 #####
> svytotal(~shihon + factor(h), si.GREG)
               total        SE
shihon       1725000 8.983e-11
factor(h)1.1     400 3.546e-13
factor(h)1.2     600 2.405e-14
factor(h)1.3     300 1.121e-13
factor(h)1.4     900 2.806e-13
factor(h)1.5     300 3.916e-14
factor(h)2.1     600 2.724e-13
factor(h)2.2     900 1.471e-14
factor(h)2.3     500 8.950e-14
factor(h)2.4    1100 4.484e-14
factor(h)2.5     400 5.062e-14
factor(h)3.1     900 6.588e-14
factor(h)3.2     800 5.676e-14
factor(h)3.3    1200 4.712e-14
factor(h)3.4     900 5.123e-14
factor(h)3.5     200 9.408e-15
```

ただしキャリブレーションウェイトの分布を確かめると，負のウェイトが得られてしまっていることが分かる．

演習 7.1 解答例 -8

```
##### キャリブレーションウェイトの分布 #####
> summary(weights(si.GREG))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.09659  3.02100  4.76600  5.00000  6.68500 15.96000
```

演習 7.2 解答例

まず関数 svydesign() を用いて標本抽出デザインを指定する．

演習 7.2 解答例 -1

```
##### 非復元単純無作為抽出法の指定 #####
> si <- svydesign(ids=~1, fpc=~N, data=kenko)
```

所在地 area 別の部分母集団サイズは $N_1 = 439,996$ と $N_2 = 309,573$ であり，性 gender 別の部分母集団サイズは $N_1 = 375,574$ と $N_2 = 373,995$ である．母集団サイズは $N = 749,569$ なので，これらを並べたベクトルを total とする．ただし各部分母集団の最初のカテゴリに対応する値は省略する．

演習 7.2 解答例 -2

```
##### 部分母集団サイズの指定 #####
> total <- c(749569, 309573, 373995)
```

レイキング比推定を行うためのウェイトを求めるには，関数 calibrate() において引数 calfun="raking"を指定する．

演習 7.2 解答例 -3

```
##### ウェイトのキャリブレーション #####
> si.c <- calibrate(si, ~factor(area) + factor(gender), total, calfun="raking")
```

関数 svymean() において，ウェイトのキャリブレーションを行った si.c を指定すればレイキング比推定値が求まる．

演習 7.2 解答例 -4

```
##### 母集団平均のレイキング比推定 #####
> svymean(~height + weight, si.c)
          mean     SE
height 163.288 0.0984
weight  57.964 0.2011
```

補助変数とした変数 area と gender の各カテゴリについて，キャリブレーションウェイトの合計を関数 svytotal() を用いて求めると，確かに真の部分母集団サイズに一致していることが分かる．

演習 7.2 解答例 -5

```
##### 補助変数の部分母集団サイズの推定 #####
> svytotal(~factor(area) + factor(gender), si.c)
                  total        SE
factor(area)1   439996 1.133e-10
factor(area)2   309573 7.103e-11
factor(gender)1 375574 1.365e-10
factor(gender)2 373995 1.209e-10
```

第8章　集落抽出法

8.a　集落抽出法の指定

集落抽出法の指定

```
DES <- svydesign(ids=~A, strata=~H, fpc=~M, weights=~W, data=DATA)
```

集落抽出法では，関数 svydesign() の引数 ids に集落を識別する変数 A を指定する．他の引数は要素を抽出単位としたときと同様の指定を行えばよい．例えば層化抽出では引数 strata，非復元抽出では引数 fpc を指定する．ただし引数 fpc の M に指定するのは抽出単位である集落の母集団における総数 M である．要素の母集団サイズ N ではない．

例題 8.2　単純無作為集落抽出法

表 8.2 (p.138) の非復元単純無作為集落抽出標本を用いて，身長の母集団平均を推定してみよう．以下の例では，まず関数 data.frame() を用いてデータフレームを作成している．変数 a は集落である学校を表す変数である．変数 M は母集団における集落の数 $M = 5$ であり，変数 w は抽出ウェイト $w_i = M/m = 5/2$ である．

例題 8.2-1

```
##### データの作成 #####
> (data <- data.frame(a=c(1, 1, 1, 1, 1, 1, 4, 4, 4),
    y=c(162, 170, 172, 172, 161, 155, 172, 154, 152), M=5, w=5/2))
  a   y M   w
1 1 162 5 2.5
2 1 170 5 2.5
3 1 172 5 2.5
4 1 172 5 2.5
5 1 161 5 2.5
6 1 155 5 2.5
7 4 172 5 2.5
8 4 154 5 2.5
9 4 152 5 2.5
```

次に関数 svydesign() を用いて非復元単純無作為集落抽出法を指定している．引数 ids に指定するのは集落を表す変数 a であり，引数 fpc には母集団における集落数の変数 M を指定している．指定結果は sic である．

例題 8.2-2

```
##### 非復元単純無作為集落抽出法の指定 #####
> sic <- svydesign(ids=~a, fpc=~M, weights=~w, data=data)
```

関数 summary() を用いて sic の内容を確認してみると Cluster Sampling design と表示され，集落抽出法が指定されていることが分かる．With (2) clusters. とあるのは，標本となった集落が $m = 2$ であることを表す．また Population size (PSUs): 5 と表示されるのは，母集団における集落の数が $M = 5$ であることを表す．母集団サイズ N を表すわけではない．

例題 8.2-3

```
##### survey.designオブジェクトの確認 #####
> summary(sic)
1 - level Cluster Sampling design
With (2) clusters.
svydesign(ids = ~a, fpc = ~M, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.4     0.4     0.4     0.4     0.4     0.4
Population size (PSUs): 5
Data variables:
[1] "a" "y" "M" "w"
```

以下では身長の母集団平均 μ_y を推定する．まず線形推定値を求めるため，関数 svytotal() を用いて母集団総計の線形推定値 $\hat{\tau}_y$ を求め，それを母集団サイズ $N = 20$ で割っている．推定値は $\hat{\mu}_y = 183.75$ となり，標準誤差は $\widehat{SE}(\hat{\mu}_y) = 49.76784$ となる．これらは本書の (8.18) 式と (8.21) 式に対応する．

例題 8.2-4

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, sic)) / 20
     y
183.75

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~y, sic)) / 20
       y
49.76784
```

次に関数 svymean() を用いてサイズとの比の推定値を求めている．推定値は $\hat{\mu}_{y,\mathrm{N}} = 163.33$ であり，標準誤差は $\widehat{SE}(\hat{\mu}_{y,\mathrm{N}}) = 2.0656$ となる．これらは本書の (8.19) 式と (8.21) 式に対応する．

例題 8.2-5

```
##### サイズとの比の推定値 #####
> svymean(~y, sic)
    mean     SE
y 163.33 2.0656
```

例題 8.3　確率比例集落抽出法

表 8.5 (p.144) の確率比例集落抽出標本を用いて，身長の母集団平均を推定してみよう．以下の例では，まず関数 data.frame() を用いてデータフレーム data を作成している．変数 a は集落である学校を表す変数であり，変数 y は身長である．変数 w は抽出ウェイト $w_i = N/(mN_a)$ である．

例題 8.3-1

```
##### データの作成 #####
> (data <- data.frame(a=c(1, 1, 1, 1, 1, 1, 4, 4, 4),
    y=c(162, 170, 172, 172, 161, 155, 172, 154, 152),
    w=20 / (2 * c(6, 6, 6, 6, 6, 6, 3, 3, 3))))
  a   y        w
1 1 162 1.666667
2 1 170 1.666667
3 1 172 1.666667
4 1 172 1.666667
5 1 161 1.666667
6 1 155 1.666667
7 4 172 3.333333
8 4 154 3.333333
9 4 152 3.333333
```

次に関数 svydesign() を用いて復元確率比例集落抽出法を指定している．引数 ids には集落を表す変数 a を指定し，復元抽出法なので引数 fpc は指定していない．結果は ppsc に代入している．

例題 8.3-2

```
##### 復元確率比例集落抽出法の指定 #####
> ppsc <- svydesign(ids=~a, weights=~w, data=data)
```

以下の例では関数 summary() を用いて ppsc の内容を確認している．一行目には，**例題 8.2-3** では表示されなかった (with replacement) が表示され，復元抽出法が指定されていることが分かる．

例題 8.3-3

```
##### survey.designオブジェクトの確認 #####
> summary(ppsc)
1 - level Cluster Sampling design (with replacement)
With (2) clusters.
svydesign(ids = ~a, weights = ~w, data = data)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.3     0.3     0.6     0.5     0.6     0.6
Data variables:
[1] "a" "y" "w"
```

以下の例では母集団平均の線形推定を行っている．まず関数 svytotal() を用いて身長の母集団総計の線形推定を行った後に，母集団サイズ $N = 20$ で割る．推定値は $\hat{\mu}_y = 162.3333$ であり，標準誤差は $\widehat{SE}(\hat{\mu}_y) = 3$ となる．これらは本書の (8.26) 式と (8.27) 式に対応する．

例題 8.3-4

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, ppsc)) / 20
       y
162.3333

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~y, ppsc)) / 20
y
3
```

これらの線形推定の結果は，関数 svymean() を用いたサイズとの比の推定値に一致する．

例題 8.3-5

```
##### サイズとの比の推定値 #####
> svymean(~y, ppsc)
    mean SE
y 162.33  3
```

8.a.1 演習問題

演習 8.1 層化集落抽出法と母集団平均の推定

データフレーム kenko が層化集落抽出されたものとして，身長 height と体重 weight の母集団平均を推定してみよう．層化変数は所在地 area であり，集落は学校 school とする．また層内で集落は非復元単純無作為抽出されたものとする．母集団平均の線形推定値 $\hat{\mu}_y$ とサイズとの比の推定値 $\hat{\mu}_{y,\mathrm{N}}$ を求めるとともに，デザイン効果も推定してみること．

さらに性 gender と所在地 area の部分母集団サイズを用いてウェイトのレイキングを行ってみよう．

ヒント：非復元抽出であっても集落抽出なので，抽出ウェイトは関数 svydesign() を用いる前に用意しておく必要がある．各層における母集団集落数は変数 M.h にあり，標本集落数は変数 m.h にある．母集団サイズは $N = 749,569$ である．性 gender 別の部分母集団サイズは $N_1 = 375,574$ と $N_2 = 373,995$ であり，所在地 area 別が $N_1 = 439,996$ と $N_2 = 309,573$ である．

演習 8.2 層化確率比例集落抽出法と母集団平均の推定

データフレーム kenko が層化集落抽出されたものとして，身長 height と体重 weight の母集団平均を推定してみよう．ただし集落は変数 n.a で確率比例抽出されたものとする．母集団平均の線形推定値 $\hat{\mu}_y$ とサイズとの比の推定値 $\hat{\mu}_{y,\mathrm{N}}$ を比較してみよう．

ヒント：変数 n.a の層総計である層サイズ N_h は変数 N.h に代入されており，各層の標本集落数は変数 m.h にある．

8.b　演習問題解答例

演習 8.1 解答例

各層において母集団集落数 M_h から標本集落数 m_h を単純無作為集落抽出しているので，抽出ウェイトは $w_i = M_h/m_h$ となる．

演習 8.1 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kenko$w <- kenko$M.h / kenko$m.h
```

集落抽出法なので関数 svydesign() の引数 ids には集落を表す変数 school を指定する．引数 strata は層を表す変数 area である．層内では集落を非復元抽出なので引数 fpc には各層における母集団集落数の変数 M.h を指定する．

演習 8.1 解答例 -2

```
##### 層化集落抽出の指定 #####
> stsic <- svydesign(ids=~school, strata=~area, fpc=~M.h, weights=~w, data=kenko)
```

stsic の内容を確認すると，標本となった集落数は全部で 12 であり，design.PSU に示されるように層 1 は $m_1 = 7$，層 2 は $m_2 = 5$ であることが分かる．obs に示される標本サイズは層 1 は $n_1 = 494$，層 2 は $n_2 = 390$ である．Population stratum sizes (PSUs): に示されるのは母集団における集落数であり，層 1 は $M_1 = 1999$，層 2 は $M_2 = 1577$ である．

演習 8.1 解答例 -3

```
##### survey.designオブジェクトの確認 #####
> summary(stsic)
Stratified 1 - level Cluster Sampling design
With (12) clusters.
svydesign(ids = ~school, strata = ~area, fpc = ~M.h, weights = ~w,
    data = kenko)
Probabilities:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003171 0.003171 0.003502 0.003356 0.003502 0.003502
Stratum Sizes:
             1   2
obs        494 390
design.PSU   7   5
actual.PSU   7   5
Population stratum sizes (PSUs):
   1    2
1999 1577
Data variables:
 [1] "obs"    "area"   "school" "gender" "height" "weight" "N"      "n"      "N.g"    "n.g"
[11] "M"      "M.h"    "m.h"    "N.h"    "N.a"    "n.a"    "N.ag"   "n.ag"   "w"
```

まず母集団平均の線形推定値を求めてみよう．関数 svytotal() によって母集団総計 τ_y を推定し，それを母集団サイズ $N = 749569$ で割ればよい．平均身長の線形推定値は $\hat{\mu}_y = 57.62979$ となる．

演習 8.1 解答例 -4

```
##### 母集団総計の線形推定 #####
> coef(svytotal(~height + weight, stsic)) / 749569
  height   weight
57.62979 20.51409
```

サイズとの比の推定値を求めるには関数 svymean() を用いる．平均身長の推定値は $\hat{\mu}_{y,\mathrm{N}} =$ 163.57841 となり，標準誤差は $\widehat{SE}(\hat{\mu}_{y,\mathrm{N}}) =$ 0.49936 となる．デザイン効果の推定値は deff $=$ 5.4452 であり，同じサイズの標本を単純無作為抽出した場合と比べ，推定量の分散は 5 倍以上となっていることが分かる．

演習 8.1 解答例 -5

```
##### サイズとの比の推定 #####
> svymean(~height + weight, stsic, deff=TRUE)
            mean        SE   DEff
height 163.57841   0.49936 5.4452
weight  58.22790   0.49294 3.3598
```

ウェイトのレイキングを行うには，まず性 gender と所在地 area の部分母集団サイズを用意する．以下の例では**演習 7.2 解答例 -2** と同様にベクトル total を用意している．

演習 8.1 解答例 -6

```
##### 部分母集団サイズの指定 #####
> total <- c(749569, 309573, 373995)
```

関数 calibrate() において引数 calfun="raking"を指定することで，ウェイトのレイキングを行うことができる．

演習 8.1 解答例 -7

```
##### ウェイトのキャリブレーション #####
> stsic.rake <- calibrate(stsic, ~factor(area) + factor(gender), total, calfun="raking")
```

平均身長のレイキング比推定値は $\hat{\mu}_{y,\mathrm{RR}} =$ 163.28841 となる．ウェイトのキャリブレーションを行わないときの推定値 $\hat{\mu}_{y,\mathrm{N}} =$ 163.57841 と大きくは異ならない．ただしキャリブレーション後の標準誤差は $\widehat{SE}(\hat{\mu}_{y,\mathrm{RR}}) =$ 0.12008 であり，キャリブレーションを行わない場合の $1/4$ 程度となっている．

演習 8.1 解答例 -8

```
##### 母集団平均のレイキング比推定 #####
> svymean(~height + weight, stsic.rake, deff=TRUE)
            mean        SE   DEff
height 163.28841   0.12008 0.3147
weight  57.96365   0.18561 0.4714
```

演習 8.2 解答例

各層において集落は変数n.aで確率比例抽出しているので，抽出ウェイトは $w_i = N_h/(m_h n_a)$ となる．

演習 8.2 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kenko$w <- kenko$N.h / (kenko$m.h * kenko$n.a)
```

関数 svydesign() の引数 ids には学校を表す変数 school を指定し，引数 strata には層を表す変数 area を指定する．復元抽出なので引数 fpc は指定しない．

演習 8.2 解答例 -2

```
##### 層化確率比例集落抽出法の指定 #####
> stppsc <- svydesign(ids=~school, strata=~area, weights=~w, data=kenko)
```

母集団平均の線形推定値は，関数 svytotal() を用いて母集団総計の線形推定値を求め，それを母集団サイズ $N = 749569$ で割る．平均身長の推定値は $\hat{\mu}_y = 163.50947$ となる．

演習 8.2 解答例 -3

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~height + weight, stppsc)) / 749569
   height    weight
163.50947  58.23981

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~height + weight, stppsc)) / 749569
   height    weight
0.5168454 0.5155772
```

サイズとの比の推定値を求めるには関数 svymean() を用いる．例えば身長については $\hat{\mu}_{y,\mathrm{N}} = 163.51$ となり，この値は線形推定値 $\hat{\mu}_y$ に一致する．

演習 8.2 解答例 -4

```
##### サイズとの比の推定値 #####
> svymean(~height + weight, stppsc)
         mean     SE
height 163.51 0.5168
weight  58.24 0.5156
```

第9章　多段抽出法

9.a　多段抽出法の指定

二段抽出法の指定

```
DES <- svydesign(ids=~A+B, strata=~H1+H2, fpc=~M1+M2, weights=~W,
                                                    data=DATA)
```

多段抽出法では各段の抽出単位 ids，層 strata，有限母集団修正項のための母集団における抽出単位の数 fpc をそれぞれ段の順に+でつなげて指定すればよい．例えば二段抽出であれば引数 ids の A には PSU，B には SSU を表す変数を指定する．ある段が復元抽出法のときには，その段以降は引数 fpc に指定しない．抽出単位の間で復元と非復元が混在していれば，適当な要素については Inf を値とする変数を引数 fpc に指定する．抽出ウェイトの変数 W は，何段であっても一つである．

例題 9.1　単純無作為抽出 - 単純無作為抽出

表 9.1 (p.154) の標本が一段目・二段目ともに非復元単純無作為抽出されたものとして，身長の母集団平均を推定してみよう．以下の例ではまず関数 data.frame() を用いてデータフレーム data を作成している．変数 a は PSU である学校を表し，変数 id は SSU である生徒を表す．変数 y は身長である．変数 M は母集団における学校の数 M であり，変数 m は抽出した学校の数 m である．さらに変数 N.a は各学校サイズ N_a であり，変数 n.a は各学校において抽出した生徒の人数 n_a である．

例題 9.1-1

```
##### データの作成 #####
> (data <- data.frame(a=c(rep(2,12), rep(3,6), rep(5,10)), id=c(1:28),
    y=c(162, 172, 170, 160, 172, 168, 172, 154, 161, 161, 155, 159,
        168, 173, 154, 167, 156, 159,
        171, 168, 168, 168, 169, 160, 152, 160, 154, 161), M=336, m=3,
    N.a=c(rep(33,12), rep(14,6), rep(30,10)), n.a=c(rep(12,12), rep(6,6), rep(10,10))))
   a id   y   M m N.a n.a
1  2  1 162 336 3  33  12
2  2  2 172 336 3  33  12
3  2  3 170 336 3  33  12
4  2  4 160 336 3  33  12
5  2  5 172 336 3  33  12
...
```

次に抽出ウェイトを $w_i = M/m \times N_a/n_a$ により求め，変数 w に代入している．

例題 9.1-2

```
##### 抽出ウェイトの作成 #####
> (data$w <- data$M / data$m * data$N.a / data$n.a)
 [1] 308.0000 308.0000 308.0000 308.0000 308.0000 308.0000 308.0000 308.0000 308.0000 308.0000
[11] 308.0000 308.0000 261.3333 261.3333 261.3333 261.3333 261.3333 261.3333 336.0000 336.0000
[21] 336.0000 336.0000 336.0000 336.0000 336.0000 336.0000 336.0000 336.0000
```

二段抽出法を指定するため，関数 svydesign() の引数 ids には PSU を表す変数 a と SSU を表す変数 id を+でつないで順に指定している．また一段目・二段目ともに非復元抽出であるため，引数 fpc には母集団における PSU の数 $M = 336$ が代入された変数 M と各 PSU における SSU 総数 N_a が代入された変数 N.a を順に指定している．引数 weights には抽出ウェイトが代入された変数 w を指定する．指定結果は si.si に代入している．

例題 9.1-3

```
##### 二段抽出法の指定 #####
> si.si <- svydesign(ids=~a + id, fpc=~M + N.a, weights=~w, data=data)
```

si.si の内容を関数 summary() で確認すると，2 - level Cluster Sampling design と表示され，二段抽出法が指定されていることが分かる．With (3, 28) clusters. とあるのは，標本における PSU の数と SSU の数である．Population size (PSUs): 336 は母集団における PSU の数 M であり，母集団サイズ N ではない．

例題 9.1-4

```
##### survey.designオブジェクトの確認 #####
> summary(si.si)
2 - level Cluster Sampling design
With (3, 28) clusters.
svydesign(ids = ~a + id, fpc = ~M + N.a, weights = ~w, data = data)
Probabilities:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002976 0.002976 0.003247 0.003274 0.003247 0.003827
Population size (PSUs): 336
Data variables:
[1] "a"   "id"  "y"   "M"   "m"   "N.a" "n.a" "w"
```

以下の例では，まず関数 svytotal() を用いて身長の母集団総計 τ_y の線形推定を行っている．線形推定値は $\hat{\tau}_y = 1408867$ であり，これは本書の (9.9) 式に対応する．なお標準誤差は $\widehat{SE}(\hat{\tau}_y) = 323831$ となる．

例題 9.1-5

```
##### 母集団総計の線形推定 #####
> svytotal(~y, si.si)
    total     SE
y 1408867 323831
```

次にこの結果を母集団サイズ $N = 3370$ で割ることで，母集団平均 μ_y の線形推定値 $\hat{\mu}_y = 418.0613$ とその標準誤差 $\widehat{SE}(\hat{\mu}_y) = 96.09222$ を求めている．これらは本書の (9.10) 式と**例題 9.3** (p.161) に対応する．

例題 9.1-6

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, si.si)) / 3370
       y
418.0613

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~y, si.si)) / 3370
       y
96.09222
```

関数 svymean() を用いてサイズとの比の推定値を求めると，$\hat{\mu}_{y,\mathrm{N}} = 163.37$ と $\widehat{SE}(\hat{\mu}_{y,\mathrm{N}}) = 0.3147$ が得られる．これらは本書の (9.12) 式と**例題 9.3** (p.161) に対応する．

例題 9.1-7

```
##### サイズとの比の推定 #####
> svymean(~y, si.si)
    mean     SE
y 163.37 0.3147
```

最後に関数 svymean() の結果に母集団サイズ $N = 3370$ を乗じることで，サイズを用いた母集団総計の比推定値 $\hat{\tau}_{y,\mathrm{N}} = 550542.7$ を求めている．これは本書の (9.13) 式に対応する．

例題 9.1-8

```
##### サイズを用いた母集団総計の比推定値 #####
> coef(svymean(~y, si.si)) * 3370
       y
550542.7
```

例題 9.2　確率比例抽出 - 単純無作為抽出

表 9.1 (p.154) の標本が，一段目は学校サイズ N_a で復元確率比例抽出，二段目は非復元単純無作為抽出されたものとして，身長の母集団平均を推定してみよう．以下の例では，**例題 9.1-1** で既にデータフレーム data が作成されたものとする．抽出ウェイト $w_i = N/(mN_a) \times N_a/n_a = N/(mn_a)$ を変数 w に代入する．

例題 9.2-1

```
##### 抽出ウェイトの作成 #####
> (data$w <- 3370 / (3 * data$n.a))
 [1]  93.61111  93.61111  93.61111  93.61111  93.61111  93.61111  93.61111  93.61111  93.61111
[10]  93.61111  93.61111  93.61111 187.22222 187.22222 187.22222 187.22222 187.22222 187.22222
[19] 112.33333 112.33333 112.33333 112.33333 112.33333 112.33333 112.33333 112.33333 112.33333
[28] 112.33333
```

以下の例では関数 svydesign() を用いて二段抽出法を指定している．引数 ids には PSU である学校を表す変数 a と SSU である生徒を表す変数 id を順に指定する[1]．一段目が復元抽出法なので引数 fpc は指定しない．指定結果は pps.si に代入している．

例題 9.2-2

```
##### 二段抽出法の指定 #####
> pps.si <- svydesign(ids=~a + id, weights=~w, data=data)
```

関数 summary() を用いて pps.si の内容を確認すると，一行目に (with replacement) と表示され，一段目が復元抽出法であることが分かる．

例題 9.2-3

```
##### survey.designオブジェクトの確認 #####
> summary(pps.si)
2 - level Cluster Sampling design (with replacement)
With (3, 28) clusters.
svydesign(ids = ~a + id, weights = ~w, data = data)
Probabilities:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.005341 0.008902 0.008902 0.008902 0.010680 0.010680
Data variables:
[1] "a"   "id"  "y"   "M"   "m"   "N.a" "n.a" "w"
```

以下の例ではまず関数 svytotal() を用いて，身長の母集団総計の線形推定を行っている．線形推定値は $\hat{\tau}_y = 550171$ であり，標準誤差は $\widehat{SE}(\hat{\tau}_y) = 1007.5$ となる．これらは本書の (9.15) 式と**例題 9.4** (p.161) に対応する．

例題 9.2-4

```
##### 母集団総計の線形推定 #####
> svytotal(~y, pps.si)
   total     SE
y 550171 1007.5
```

次にその結果を母集団サイズ $N = 3370$ で割り，母集団平均の線形推定値を求めると $\hat{\mu}_y = 163.2556$ となる．この結果は本書の (9.16) 式に対応する．

例題 9.2-5

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, pps.si)) / 3370
       y
163.2556

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~y, pps.si)) / 3370
         y
0.2989694
```

[1]引数 ids に変数 a だけを指定した場合も試みるとよい．結果は同じとなる．これは一段目が復元抽出なので推定量の分散は (9.32) 式を用いて推定され，二段目に対応する分散成分の推定は不要となるためである．

なお関数 svymean() を用いてサイズとの比の推定値を求めると，線形推定値と同じ推定値 $\hat{\mu}_{y,\mathrm{N}} = 163.26$ が得られる．

例題 9.2-6

```
##### サイズとの比の推定 #####
> svymean(~y, pps.si)
    mean    SE
y 163.26 0.299
```

例題 9.5　二段抽出における層化

表 9.1 (p.154) の標本が一段目は非復元単純無作為集落抽出，二段目は性別で層化し各層で非復元単純無作為抽出されたものとして，身長の母集団平均を推定してみよう．以下の例では，まず関数 data.frame() を用いてデータフレーム data を作成している．変数 gen は生徒の性別である．変数 M は母集団における学校数 M であり，変数 m は標本における学校数 m である．変数 N.al は各学校における性別の在籍数 $N_{a\ell}$ であり，変数 n.al は各学校の性別の標本サイズ $n_{a\ell}$ である．

例題 9.5-1

```
##### データの作成 #####
> (data <- data.frame(a=c(rep(2,12), rep(3,6), rep(5,10)), id=c(1:28),
     y=c(162, 172, 170, 160, 172, 168, 172, 154, 161, 161, 155, 159,
         168, 173, 154, 167, 156, 159,
         171, 168, 168, 168, 169, 160, 152, 160, 154, 161),
     gen=c(rep('男',7), rep('女',5), rep('男',2), rep('女',4), rep('男',6), rep('女',4)),
     M=336, m=3, N.al=c(rep(19,7), rep(14,5), rep(5,2), rep(9,4), rep(18,6), rep(12,4)),
     n.al=c(rep(7,7), rep(5,5), rep(2,2), rep(4,4), rep(6,6), rep(4,4))))
   a id   y gen   M m N.al n.al
1  2  1 162  男 336 3   19    7
2  2  2 172  男 336 3   19    7
3  2  3 170  男 336 3   19    7
4  2  4 160  男 336 3   19    7
5  2  5 172  男 336 3   19    7
6  2  6 168  男 336 3   19    7
7  2  7 172  男 336 3   19    7
8  2  8 154  女 336 3   14    5
9  2  9 161  女 336 3   14    5
10 2 10 161  女 336 3   14    5
...
```

一段目・二段目ともに単純無作為抽出なので，抽出ウェイトは $w_i = M/m \times N_{a\ell}/n_{a\ell}$ となる．以下の例では抽出ウェイトを変数 w に代入している．

例題 9.5-2

```
##### 抽出ウェイトの作成 #####
> data$w <- data$M / data$m * data$N.al / data$n.al
```

一段目は層化していないが，関数 svydesign() の引数 strata で二段目の層を指定するためには，一段目の層も指定しなければならない．そこで一段目は層が 1 つであるとする．以下の例では，一段目の層化変数として全ての生徒の値が 1 である変数 one を作成している．

例題 9.5-3

```
##### 一段目の層化変数 #####
> data$one <- 1
```

以下の例では関数 svydesign() を用いて二段抽出法を指定している．引数 ids に指定するのは PSU である学校を表す変数 a と SSU である生徒を表す変数 id である．一段目・二段目ともに非復元抽出なので，引数 fpc には母集団における PSU の数 M に対応する変数 M と，各 PSU の層サイズ $N_{a\ell}$ に対応する変数 N.al を指定している．引数 strata には一段目と二段目の層化変数 one と gen を指定する．ただし変数 one は全ての生徒の値が 1 なので，実質的には層化していないことになる．指定結果は si.st に代入している．

例題 9.5-4

```
##### 二段抽出法の指定 #####
> si.st <- svydesign(ids=~a + id, fpc=~M + N.al, strata=~one + gen, weights=~w, data=data)
```

関数 summary() を用いて si.st の内容を確認すると，一行目に Stratified と表示され，層化抽出法が指定されていることが分かる．

例題 9.5-5

```
##### survey.designオブジェクトの確認 #####
> summary(si.st)
Stratified 2 - level Cluster Sampling design
With (3, 28) clusters.
svydesign(ids = ~a + id, fpc = ~M + N.al, strata = ~one + gen,
    weights = ~w, data = data)
Probabilities:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002976 0.002976 0.003189 0.003277 0.003289 0.003968
First-level Stratum Sizes:
            1
obs        28
design.PSU  3
actual.PSU  3
Population stratum sizes (PSUs):
  1
336
Data variables:
 [1] "a"    "id"   "y"    "gen"  "M"    "m"    "N.al" "n.al" "w"    "one"
```

以下の例では，まず関数 svytotal() を用いて身長の母集団総計の線形推定値 $\hat{\tau}_y = 1409016$ を求めている．

例題 9.5-6

```
##### 母集団総計の線形推定 #####
> svytotal(~y, si.st)
    total     SE
y 1409016 323233
```

母集団総計の線形推定値 $\hat{\tau}_y$ を母集団サイズ $N = 3370$ で割ると母集団平均の線形推定値 $\hat{\mu}_y = 418.1056$ が得られる．

例題 9.5-7

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~y, si.st)) / 3370
       y
418.1056

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~y, si.st)) / 3370
       y
95.91486
```

関数 svymean() を用いてサイズとの比の推定値を求めると $\hat{\mu}_{y,\mathrm{N}} = 163.38$ となる．この結果は本書の (9.49) 式に対応する．

例題 9.5-8

```
##### サイズとの比の推定 #####
> svymean(~y, si.st)
    mean     SE
y 163.38 0.2529
```

なおサイズを用いた母集団総計の比推定値を求めるには，関数 svymean() の結果に母集団サイズ $N = 3370$ を乗じればよい．

例題 9.5-9

```
##### サイズを用いた母集団総計の比推定 #####
> coef(svymean(~y, si.st)) * 3370
       y
550601.1

##### サイズを用いた母集団総計の比推定の標準誤差 #####
> SE(svymean(~y, si.st)) * 3370
       y
852.2412
```

9.a.1　演習問題

演習 9.1　層化二段抽出法と母集団平均の推定

データフレーム kenko が層化二段抽出されたものとして，身長 height と体重 weight の母集団平均を推定してみよう．一段目の層化変数は所在地 area である．PSU は学校 school であり層内で PSU を非復元単純無作為抽出，SSU は生徒 obs であり PSU 内で非復元単純無作為抽出とする．母集団平均の線形推定値 $\hat{\mu}_y$ とサイズとの比の推定値 $\hat{\mu}_{y,\mathrm{N}}$ を求めるとともに，デザイン効果も推定してみること．

さらに関数 svydesign() の全ての引数において，あえて二段目を指定しない場合の結果を確認すること．

ヒント：一段目は各層において変数 M.h にある PSU 数から変数 m.h にある PSU 数を非復元単純無作為抽出し，二段目は各 PSU において変数 N.a にある SSU 数から変数 n.a にある SSU 数を非復元単純無作為抽出している．

演習 9.2　層化二段抽出法と母集団割合の推定

データフレーム otona が層化二段抽出されたものとして，質問項目 Q1 と Q3 の各カテゴリの母集団割合とデザイン効果を関数 svymean() を用いて推定してみよう．ただし一段目は市郡 city で層化し，学校 school を在籍児童数 N.a で復元確率比例抽出したものとする．二段目は性 gender で層化し，児童 obs を非復元単純無作為抽出したものとする．標本抽出デザインに従い引数 fpc を指定する場合と，引数 fpc を指定せず，かつ他の引数についても二段目を指定しない場合との結果を比較すること．

ヒント：在籍児童数の層総計は変数 N.h にあり，層ごとの標本学校数は変数 m.h にある．各学校の性別在籍数は変数 N.ag とし，標本となった児童数は変数 n.ag にある．

演習 9.3　層化三段抽出法と母集団割合の推定

データフレーム otona が層化三段抽出されたものとして，質問項目 Q1 の各カテゴリの母集団割合とデザイン効果を関数 svymean() を用いて推定してみよう．ただし一段目は市郡 city で層化し，学校 school を非復元単純無作為抽出したものとする．二段目は一学級を単純無作為抽出したものとする．三段目は児童 obs を非復元単純無作為抽出したものとする．標本抽出デザインに従って引数を三段目まで指定する場合と，一段目だけを指定する場合とを比較すること．

ヒント：市郡別の母集団学校数は変数 M.h にあり，標本学校数は変数 m.h にある．各学校の学級数は変数 N.class にある．各学級の在籍児童数は変数 N.a にあり，標本児童数は変数 n.a にある．

9.b　演習問題解答例

演習 9.1 解答例

一段目・二段目ともに非復元単純無作為抽出なので，抽出ウェイトは $w_i = M_h/m_h \times N_a/n_a$ となる．

演習 9.1 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kenko$w <- kenko$M.h / kenko$m.h * kenko$N.a / kenko$n.a
```

関数 svydesign() では，引数 ids には PSU である変数 school と SSU である変数 obs を指定する．層化は一段目のみで行われているため，引数 strata には一段目の層化変数である area だけを指定する．一段目・二段目ともに非復元抽出なので，引数 fpc には母集団における PSU の数である変数 M.h と SSU の数である変数 N.a を指定する．

演習 9.1 解答例 -2

```
##### 層化二段抽出法の指定 #####
> si.si <- svydesign(ids=~school + obs, strata=~area, fpc=~M.h + N.a, weights=~w, data=kenko)
```

関数 summary() を用いて si.si の内容を確認すると，各層における標本 PSU の数は 7 と 5 であり，標本サイズは 494 と 390 となっていることが分かる．なお母集団における PSU の数は 1999 と 1577 である．

演習 9.1 解答例 -3

```
##### survey.designオブジェクトの確認 #####
> summary(si.si)
Stratified 2 - level Cluster Sampling design
With (12, 884) clusters.
svydesign(ids = ~school + obs, strata = ~area, fpc = ~M.h + N.a,
    weights = ~w, data = kenko)
Probabilities:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0003752 0.0007138 0.0013870 0.0015650 0.0022900 0.0032390
First-level Stratum Sizes:
             1   2
obs        494 390
design.PSU   7   5
actual.PSU   7   5
Population stratum sizes (PSUs):
   1    2
1999 1577
Data variables:
 [1] "obs"    "area"   "school" "gender" "height" "weight" "N"      "n"      "N.g"    "n.g"
[11] "M"      "M.h"    "m.h"    "N.h"    "N.a"    "n.a"    "N.ag"   "n.ag"   "w"
```

身長 height と体重 weight の母集団平均の線形推定値はそれぞれ $\hat{\mu}_{\text{身長}} = 189.02103$ と $\hat{\mu}_{\text{体重}} = 66.91491$ となる．またその標準誤差は $\widehat{SE}(\hat{\mu}_{\text{身長}}) = 40.59866$ と $\widehat{SE}(\hat{\mu}_{\text{体重}}) = 14.19197$ となる．

演習 9.1 解答例 -4

```
##### 母集団平均の線形推定値 #####
> coef(svytotal(~height + weight, si.si)) / 749569
   height    weight
189.02103  66.91491

##### 母集団平均の線形推定値の標準誤差 #####
> SE(svytotal(~height + weight, si.si)) / 749569
  height   weight
40.59866 14.19197
```

関数 svymean() を用いて求めたサイズとの比の推定値は $\hat{\mu}_{\text{身長,N}} = 163.311$ となり，標準誤差は $\widehat{SE}(\hat{\mu}_{\text{身長,N}}) = 0.6249$ となる．

演習 9.1 解答例 -5

```
##### サイズとの比の推定値 #####
> svymean(~height + weight, si.si)
          mean     SE
height 163.311 0.6249
weight  57.813 0.5304
```

以下の例では関数 svydesign() の全ての引数において二段目を指定していない．ただし抽出ウェイト w はこれまでと同じであり，二段目を考慮して求めている．

演習 9.1 解答例 -6

```
##### 二段目の指定を省略した指定 #####
> si <- svydesign(ids=~school, strata=~area, fpc=~M.h, weights=~w, data=kenko)
```

si を用いたときの結果と si.si を用いたときの結果とを比較すると，推定値については違いはない．標準誤差についてはわずかに異なっているが，実質的な違いとは言えない．これは一段目の抽出率 m_h/M_h が非常に小さいためである (p.159)．

演習 9.1 解答例 -7

```
##### サイズとの比の推定値 #####
> svymean(~height + weight, si)
          mean     SE
height 163.311 0.6248
weight  57.813 0.5302
```

演習 9.2 解答例

一段目は学校を層化確率比例抽出であり，二段目は児童を層化無作為抽出なので抽出ウェイトは $w_i = N_h/(m_h N_a) \times N_{ag}/n_{ag}$ となる．

演習 9.2 解答例 -1

```
##### 抽出ウェイトの作成 #####
> otona$w <- otona$N.h / (otona$m.h * otona$N.a) * otona$N.ag / otona$n.ag
```

一段目は復元抽出なので引数 fpc は不要だが，二段目は非復元抽出である．そこで $M_h \to \infty$ とすることで $1 - m_h/M_h \to 1$ とする．以下の例ではそのための変数 inf を追加している．

演習 9.2 解答例 -2

```
##### 一段目のfpc #####
> otona$inf <- Inf
```

関数 svydesign() の第一の引数 ids には，一段目の抽出単位 school と二段目の抽出単位 obs を指定する．引数 strata には一段目の層化変数 city と二段目の層化変数 gender を指定する．引数 fpc には，上記で作成した変数 inf を一段目とし，二段目として変数 N.ag を指定する．

演習 9.2 解答例 -3

```
##### 層化二段抽出法の指定 #####
> pps.si <- svydesign(ids=~school + obs, strata=~city + gender, fpc=~inf + N.ag,
                                                          weights=~w, data=otona)
```

変数 Q1 と Q3 をカテゴリとして扱うために factor() を用いて関数 svymean() を適用する．デザイン効果を推定するために引数 deff=TRUE も指定する．deff は 1 よりも大きく，同じサイズの単純無作為抽出標本よりも推定値の精度は劣ることが分かる．

演習 9.2 解答例 -4

```
##### 母集団割合の推定 #####
> svymean(~factor(Q1) + factor(Q3), pps.si, deff=TRUE)
                  mean        SE   DEff
factor(Q1)1 0.8138770 0.0177347 1.7720
factor(Q1)2 0.1861230 0.0177347 1.7720
factor(Q3)1 0.1873693 0.0187879 1.9785
factor(Q3)2 0.5465327 0.0201991 1.4050
factor(Q3)3 0.2079296 0.0157456 1.2847
factor(Q3)4 0.0581684 0.0077522 0.9362
```

以下の例では関数 svydesign() において二段目を指定せず，引数 fpc も指定していない．関数 svydesign() による指定結果は pps に代入している．

演習 9.2 解答例 -5

```
##### fpcと二段目を無指定 #####
> pps <- svydesign(ids=~school, strata=~city, weights=~w, data=otona)
```

pps を用いて質問項目 Q1 と Q3 の各カテゴリの母集団割合を推定すると以下のとおりとなる．この結果は pps.si を用いた結果に一致する．これは本書の (9.32) 式から分かるように，一段目が復元抽出であれば，推定量の分散を推定するときに二段目の分散成分の推定量は用いられないからである．

演習 9.2 解答例 -6

```
##### 母集団割合の推定 #####
> svymean(~factor(Q1) + factor(Q3), pps, deff=TRUE)
                 mean         SE   DEff
factor(Q1)1 0.8138770 0.0177347 1.7720
factor(Q1)2 0.1861230 0.0177347 1.7720
factor(Q3)1 0.1873693 0.0187879 1.9785
factor(Q3)2 0.5465327 0.0201991 1.4050
factor(Q3)3 0.2079296 0.0157456 1.2847
factor(Q3)4 0.0581684 0.0077522 0.9362
```

演習 9.3 解答例

抽出ウェイトは各段の抽出ウェイトをかけ合わせればよい.

演習 9.3 解答例 -1

```
##### 抽出ウェイトの作成 #####
> otona$w <- otona$M.h / otona$m.h * otona$N.class * otona$N.a / otona$n.a
```

学級を表す変数はデータフレーム otona に含まれていないので，以下の例では変数 one を追加している.

演習 9.3 解答例 -2

```
##### SSUを表す変数の追加 #####
> otona$one <- 1
```

以下の例では，関数 svydesign() の引数 ids と fpc に三段目まで全て指定している．その結果は si.si.si としている.

演習 9.3 解答例 -3

```
##### 層化三段抽出法の指定 #####
> si.si.si <- svydesign(ids=~school + one + obs, strata=~city, fpc=~M.h + N.class + N.a,
                                                              weights=~w, data=otona)
```

si.si.si を用いて推定を行おうとするとエラーとなる．これは二段目で一学級しか抽出しておらず，二段目の分散成分を推定できないからである.

演習 9.3 解答例 -4

```
##### 母集団割合の推定 #####
> svymean(~factor(Q1), si.si.si)
 以下にエラー switch(lonely.psu, certainty = crossprod(x * sqrt(scale)), remove =
  crossprod(x *  :
  Stratum (1.1) has only one PSU at stage 2
```

以下の例では関数 svydesign() の引数 ids と fpc で一段目しか指定していない.

演習 9.3 解答例 -5

```
##### 二段目以降を無指定 #####
> si <- svydesign(ids=~school, strata=~city, fpc=~M.h, weights=~w, data=otona)
```

si を用いると推定値が得られる.

演習 9.3 解答例 -6

```
##### 母集団割合の推定 #####
> svymean(~factor(Q1), si)
                mean     SE
factor(Q1)1 0.82871 0.0182
factor(Q1)2 0.17129 0.0182
```

一段目の抽出率は最大でも 0.004132 であり，推定量の分散を推定するに当たって二段目以降を省略しても問題はない．

演習 9.3 解答例 -7

```
##### 一段目の抽出率の分布 #####
> summary(otona$m.h / otona$M.h)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001584 0.001584 0.001584 0.002809 0.004132 0.004132
```

第12章　クロス集計

12.a　クロス表の推定

クロス表の推定

```
TAB <- svytable(~ROW+COL, design=DES, Ntotal=N)
```

クロス表の推定を行うには関数 svytable() を用いる．母集団において各セルに該当する要素の数の推定値 $\hat{N}_{rc} = \sum_s w_i y_{i,rc}$ が求められる．ただし $y_{i,rc}$ は (12.1) 式のとおりである．ROW にはクロス表の行の変数，COL には列の変数を指定する．必ずしも+COL は必要ではなく，指定しなかった場合には変数 ROW の周辺分布が求められる．引数 Ntotal を指定すると，セルの値の合計が N になるよう調整した結果を表示する．

割合への変換

```
PTAB <- prop.table(TAB, margin=M)
```

クロス表の各セルの割合 p_{rc} を推定するには，svytable() の結果求められた TAB を第一の引数として関数 prop.table() を用いればよい．関数 prop.table() は base パッケージに含まれる．引数 margin を指定しないと全てのセルの合計が 1 となる．margin=1 を指定すると行計 (横に足した値) が 1 となり，margin=2 を指定すると列計 (縦に足した値) が 1 となるよう変換する．

12.a.1　演習問題

演習 12.1　層化二段集落抽出法とクロス表の推定

データフレーム otona が層化二段集落抽出されたものとして，質問項目 Q1 と Q2 とのクロス表および Q3 と Q4 とのクロス表を推定してみよう．一段目は市郡 city で層化し，学校 school を非復元単純無作為抽出したものとし，二段目は一学級を単純無作為抽出したものとする．

ヒント：層ごとの母集団学校数は変数 M.h であり，標本学校数は変数 m.h である．各学校の学級数は変数 N.class にある．

12.b　独立性の検定

クロス表の独立性の検定

```
TEST <- svychisq(~ROW+COL, design=DES, statistic=S)
```

関数 svychisq() はクロス表における変数間の独立性の検定を行う．ROW にはクロス表の行変数，COL には列変数を指定する．引数 statistic には，"Wald" (Wald F 検定)，"adjWald" (修正 Wald F 検定)，"Chisq" (Rao-Scott の一次修正による χ^2 検定)[1]，"F" (Rao-Scott の二次修正による F 検定) のいずれかを指定する．デフォルトは"F"である．

svychisq() を用いる代わりに summary(TAB, statistic=S) としてもよい．第一の引数に指定する TAB は svytable() の結果のオブジェクトである．デフォルトは statistic="F"である．

自由度の取り出し

```
DF <- degf(design=DES)
```

標本抽出デザインに基づく分散の自由度を調べるには関数 degf() を用いる．degf() は本書の (11.23) 式による自由度を返す．引数 design の DES には，svydesign() による指定結果である survey.design オブジェクトを指定する．

[1]表示される検定統計量の値は Rao-Scott 修正前の検定統計量の値である．

12.b.1　演習問題

演習 12.2　層化二段集落抽出法とクロス表の独立性の検定

データフレーム otona が層化二段集落抽出されたものとして，性別 gender と質問項目 Q3 とが独立かどうかを検定してみよう．標本は，一段目は市郡 city で層化し，学校 school を非復元単純無作為抽出したものとし，二段目は一学級を単純無作為抽出したものとする．分散の自由度 ν を求め，検定方法としては，2 つの Wald 検定と 2 つの Rao-Scott 修正の全てを試みること．

ヒント：層ごとの母集団学校数は変数 M.h であり，標本学校数は変数 m.h である．各学校の学級数は変数 N.class にある．

12.c　演習問題解答例

演習 12.1 解答例

以下の例ではまず抽出ウェイトを変数 w としている．一段目の抽出率が非常に小さいので，関数 svydesign() では一段目のみを指定している．

演習 12.1 解答例 -1

```
##### 抽出ウェイトの作成 #####
> otona$w <- otona$M.h / otona$m.h * otona$N.class

##### 標本抽出デザインの指定 #####
> des <- svydesign(ids=~school, strata=~city, fpc=~M.h, weights=~w, data=otona)
```

以下の例では関数 svytable() を用いて，母集団において各セルに該当する人数を推定している．例えば Q1 が 1 で，Q2 が 1 の人数の推定値は $\hat{N}_{1,1} = 629111.13$ である．

演習 12.1 解答例 -2

```
##### Q1とQ2とのクロス表の推定 #####
> svytable(~Q1 + Q2, des)
   Q2
Q1          1         2
  1 629111.13 103718.93
  2  94333.47  68335.27

##### Q3とQ4とのクロス表の推定 #####
> svytable(~Q3 + Q4, des)
   Q4
Q3           1          2          3          4
  1   9164.133  42001.200  66031.533  60812.600
  2  12267.933 131569.600 247660.667  90924.800
  3   2083.267  38423.733  96185.067  49187.133
  4   4608.333   4650.533  11015.733  28912.533
```

以下の例では関数 prop.table() の第二の引数に 1 を指定することで，行計 (横計) が 100% となるようにしたクロス表を求めている．

演習 12.1 解答例 -3

```
##### 行計が100となるクロス表の推定 #####
> 100 * prop.table(svytable(~Q1 + Q2, des), 1)
   Q2
Q1         1        2
  1 85.84680 14.15320
  2 57.99115 42.00885
```

以下の例では関数 prop.table() の第二の引数に 2 を指定することで，列計 (縦計) が 100% となるようにしている．

演習 12.1 解答例 -4

```
##### 列計が100となるクロス表の推定 #####
> 100 * prop.table(svytable(~Q1 + Q2, des), 2)
   Q2
Q1         1        2
  1 86.96051 60.28271
  2 13.03949 39.71729
```

以下の例では関数 svytable() の第三の引数に 100 を指定し，4 つのセルの合計が 100 となるようにしている．

演習 12.1 解答例 -5

```
##### 合計が100となるクロス表の推定 #####
> svytable(~Q1 + Q2, des, 100)
   Q2
Q1          1         2
  1 70.252594 11.582253
  2 10.534181  7.630972
```

演習 12.2 解答例

演習 12.1 解答例 -1 と同様に標本抽出デザインを指定する．

演習 12.2 解答例 -1

```
##### 抽出ウェイトの作成 #####
> otona$w <- otona$M.h / otona$m.h * otona$N.class

##### 標本抽出デザインの指定 #####
> des <- svydesign(ids=~school, strata=~city, fpc=~M.h, weights=~w, data=otona)
```

関数 `degf()` を用いて分散の自由度を求めると $\nu = 26$ となる．これは標本 PSU の数が 28 であり，層の数が 2 であることから求まる．

演習 12.2 解答例 -2

```
##### 分散の自由度 #####
> degf(des)
[1] 26
```

以下の例はまず Wald 検定の結果である．本書の (12.28) 式に基づく検定統計量の値は `F = 2.2446` である．自由度は分子が $P = (R-1)(C-1) = (2-1) \times (4-1) =$3，分母が $\nu =$26 である．p 値は `0.1069` となって有意ではない．

演習 12.2 解答例 -3

```
##### Wald検定 #####
> svychisq(~gender + Q3, des, statistic="Wald")

    Design-based Wald test of association

data:  svychisq(~gender + Q3, des, statistic = "Wald")
F = 2.2446, ndf = 3, ddf = 26, p-value = 0.1069
```

以下の例は修正 Wald 検定の結果である．本書の (12.29) 式に基づく検定統計量の値は `F = 2.0719` である．自由度は分子が $P = (R-1)(C-1) = (2-1) \times (4-1) =$3，分母が $\nu - P + 1 = 26 - 3 + 1 =$24 である．$p$ 値は `0.1306` となっている．

演習 12.2 解答例 -4

```
##### 修正Wald検定 #####
> svychisq(~gender + Q3, des, statistic="adjWald")

    Design-based Wald test of association

data:  svychisq(~gender + Q3, des, statistic = "adjWald")
F = 2.0719, ndf = 3, ddf = 24, p-value = 0.1306
```

以下の例は，Rao-Scott の一次修正を用いた検定結果である．

演習 12.2 解答例 -5

```
##### Rao-Scottの一次修正 #####
> svychisq(~gender + Q3, des, statistic="Chisq")

    Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~gender + Q3, des, statistic = "Chisq")
X-squared = 10.8684, df = 3, p-value = 0.05737
```

Rao-Scott の二次修正は以下のとおりとなる．

演習 12.2 解答例 -6

```
##### Rao-Scottの二次修正 #####
> svychisq(~gender + Q3, des, statistic="F")

    Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~gender + Q3, des, statistic = "F")
F = 2.5024, ndf = 2.650, ddf = 68.888, p-value = 0.07363
```

第13章　回帰分析

13.a　重回帰分析

重回帰分析

```
GLM.RES <- svyglm(formula=Y~X1+X2+...+XK, design=DES, family=FML)
```

回帰分析を行うには関数 `svyglm()` を用いる．Y には基準変数，X1 から XK には説明変数を指定する．説明変数 XK がカテゴリカルな変数であれば `factor(XK)` などとすればよい．定数項の指定は不要である．また重回帰分析では引数 `family` の指定は不要である．

回帰分析の結果の要約

```
summary(GLM.RES, df.resid=NULL)
```

`svyglm()` では各回帰係数だけが表示される．回帰係数がそれぞれ 0 であるか否かの (無修正) Wald 検定の結果は `summary()` を用いて表示させる．回帰係数ごとの検定であるので，検定統計量の分子の自由度は $P = 1$ である．そのため F 検定ではなく t 検定となっている．

`df.resid=Inf` と指定すると z 検定が行われる．これは本書の (13.18) 式の χ^2 検定に相当する (表示は `t` のままである)．`df.resid=degf(DES)` と指定すると，自由度を $\nu =$ PSU の数 $-$ 層の数 とした t 検定 となる．これは分子の自由度を 1，分母の自由度を ν とした (13.19) 式の F 検定に相当する．`df.resid=NULL` と指定すると，自由度を $\nu =$ PSU の数 $-$ 層の数 $-$ 説明変数の数 $+ 1$ とした t 検定となる．デフォルトは `NULL` である．

回帰係数の検定

```
regTermTest(GLM.RES, test.terms=X, df=Inf)
```

P (≥ 2) 個の回帰係数が全て同時に 0，つまり $H_0 : \boldsymbol{\beta}_{P\times 1} = \mathbf{0}$ という帰無仮説を (無修正) Wald 検定するには `regTermTest()` を用いる．引数 `test.terms` には検定の対象となる説明変数 X を指定する．引数 `df` を指定しなければ (無修正) Wald χ^2 検定，`df=degf(DES)` とすれば分母の自由度を $\nu =$ PSU の数 $-$ 層の数 とした (無修正) Wald F 検定，`df=NULL` とすれば分母の自由度を $\nu =$ PSU の数 $-$ 層の数 $-$ 説明変数の数 $+ 1$ とした (無修正) Wald F 検定となる．

例題 13.1　回帰分析

表 9.1 (p.154) の標本を用いて身長を基準変数とし，体重を説明変数とした回帰分析を行ってみよう[1]．以下の例ではまずデータフレーム data を作成している．

例題 13.1-1

```
##### データの作成 #####
> (data <- data.frame(a=c(rep(2,12), rep(3,6), rep(5,10)), id=c(1:28),
     y=c(162, 172, 170, 160, 172, 168, 172, 154, 161, 161, 155, 159,
         168, 173, 154, 167, 156, 159,
         171, 168, 168, 168, 169, 160, 152, 160, 154, 161),
     x=c(73, 60, 68, 64, 66, 53, 55, 47, 49, 55, 70, 55,
         57, 65, 53, 47, 55, 69,
         59, 61, 48, 59, 55, 52, 52, 46, 59, 76),
     gen=c(rep('男',7), rep('女',5), rep('男',2), rep('女',4), rep('男',6), rep('女',4)),
     M=336, m=3, N.a=c(rep(33,12), rep(14,6), rep(30,10)),
     n.a=c(rep(12,12), rep(6,6), rep(10,10))))
   a id   y  x gen   M m N.a n.a
1  2  1 162 73  男 336 3  33  12
2  2  2 172 60  男 336 3  33  12
...
```

標本は二段抽出で，一段目・二段目ともに非復元単純無作為抽出とする．以下の例では抽出ウェイトの変数 w を作成し，標本抽出デザインを指定した結果は des としている．

例題 13.1-2

```
##### 抽出ウェイトの作成 #####
> data$w <- data$M / data$m * data$N.a / data$n.a

##### 標本抽出デザインの指定 #####
> des <- svydesign(ids=~a + id, fpc=~M + N.a, weights=~w, data=data)
```

以下の例では基準変数を変数 y とし，説明変数を変数 x とした単回帰分析を行っている．切片は $\hat{\beta}_0 = 157.3925$，回帰係数は $\hat{\beta}_{\mathrm{x}} = 0.1028$ となる．

例題 13.1-3

```
##### 単回帰分析 #####
> (glm.res <- svyglm(y ~ x, des))
2 - level Cluster Sampling design
With (3, 28) clusters.
svydesign(ids = ~a + id, fpc = ~M + N.a, weights = ~w, data = data)

Call:  svyglm(y ~ x, des)

Coefficients:
(Intercept)            x
   157.3925       0.1028

Degrees of Freedom: 27 Total (i.e. Null);  1 Residual
Null Deviance:      1171
Residual Deviance: 1152         AIC: 189.6
```

[1] 本書の表に示した結果はいずれも SUDAAN を用いて計算したものである．したがって検定統計量の値などが異なる．

以下の例では関数 svyglm() の結果 glm.res に対して関数 summary() を適用し，回帰係数の標準誤差を求めるとともに，それが 0 かどうかの検定を行っている．回帰係数 $\hat{\beta}_{\mathrm{x}}$ の標準誤差は $\widehat{SE}(\hat{\beta}_{\mathrm{x}}) = 0.0280$ となる．検定統計量は $t = 3.671$ となり，p 値は $p = 0.0668$ である．これらの結果は本書の表 13.2 (p.228) に対応する[2]．

例題 13.1-4

```
##### 単回帰分析の結果の要約 #####
> summary(glm.res, df.resid=degf(des))

Call:
svyglm(y ~ x, des)

Survey design:
svydesign(ids = ~a + id, fpc = ~M + N.a, weights = ~w, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.3925     1.5346 102.565  9.5e-05 ***
x             0.1028     0.0280   3.671   0.0668 .
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

(Dispersion parameter for gaussian family taken to be 42.65798)

Number of Fisher Scoring iterations: 2
```

なお決定係数 R^2 の値は以下で求まる．

例題 13.1-5

```
##### 決定係数 #####
> 1 - glm.res$deviance / glm.res$null.deviance
[1] 0.01646031
```

以下の例では説明変数に性 gen を追加した結果である．カテゴリカルな変数として扱うために factor() を用いている．

例題 13.1-6

```
##### 重回帰分析 #####
> (glm.res <- svyglm(y ~ x + factor(gen, levels=c('男', '女')), des))
2 - level Cluster Sampling design
With (3, 28) clusters.
svydesign(ids = ~a + id, fpc = ~M + N.a, weights = ~w, data = data)

Call:  svyglm(y ~ x + factor(gen, levels = c("男", "女")), des)

Coefficients:
 (Intercept)           x  factor(gen, levels = c("男", "女"))女
   168.90165    -0.01529                             -10.20006

Degrees of Freedom: 27 Total (i.e. Null);  0 Residual
Null Deviance:      1171
Residual Deviance: 454.6        AIC: 165.6
```

[2]本書では検定統計量が $F_{1,2} = 13.4765$ となっているが，これは $t^2 = 3.671^2 = 13.47624$ として得られる．

以下の例では各回帰係数の検定結果を示している．これらの結果は本書の表 13.3 (p.228) に対応する[3].

例題 13.1-7

```
##### 重回帰分析の結果の要約 #####
> summary(glm.res, df.resid=degf(des))

Call:
svyglm(y ~ x + factor(gen, levels = c("男", "女")), des)

Survey design:
svydesign(ids = ~a + id, fpc = ~M + N.a, weights = ~w, data = data)

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          168.90165    7.50314  22.511  0.00197 **
x                                     -0.01529    0.12015  -0.127  0.91039
factor(gen, levels = c("男", "女"))女 -10.20006    0.31126 -32.771  0.00093 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

(Dispersion parameter for gaussian family taken to be 16.83884)

Number of Fisher Scoring iterations: 2
```

[3]本書では体重の回帰係数の検定統計量が $F_{1,2} = 0.0162$ となっているが，これは $t^2 = (-0.127)^2 = 0.016129$ として得られる．また性別 (女) の回帰係数の検定統計量が $F_{1,2} = 1073.9139$ となっているが，これは $t^2 = (-32.771)^2 = 1073.9384$ として得られる．

例題 13.2　部分母集団平均の差の検定

二段抽出された表 9.1 (p.154) の標本を使って，平均身長の男女差の有無を検定してみよう．データと標本抽出デザインは**例題 13.1** と同じなので，以下では同じ survey.design オブジェクト des を用いることにする．以下の例では身長の変数 y を基準変数とし，性別 gen を説明変数とした回帰分析を行っている．回帰係数 $\hat{\beta}_{\text{性別 (女)}} = -10.1536$ が身長の部分母集団平均の差の推定値となり，その標準誤差は 0.1819 である．検定統計量は $t = -55.83$ となり，明らかに有意な差が認められる．これらの結果は本書の表 13.4 (p.229) に対応する[4]．

例題 13.2-1

```
##### 部分母集団平均の差の検定 #####
> summary(svyglm(y ~ factor(gen, levels=c('男', '女')), des), df.resid=degf(des))

Call:
svyglm(y ~ factor(gen, levels = c("男", "女")), des)

Survey design:
svydesign(ids = ~a + id, fpc = ~M + N.a, weights = ~w, data = data)

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         167.9920     0.4917  341.65 8.57e-06 ***
factor(gen, levels = c("男", "女"))女 -10.1536     0.1819  -55.83 0.000321 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

(Dispersion parameter for gaussian family taken to be 16.85408)

Number of Fisher Scoring iterations: 2
```

[4] 本書では検定統計量が $F_{1,2} = 3116.5547$ となっているが，これは $t^2 = (-55.84)^2 = 3118.1056$ として得られる．

13.a.1　演習問題

演習 13.1　層化二段抽出法と回帰分析

データフレーム kenko が層化二段抽出されたものとして，身長 height を基準変数とし，体重 weight と所在地 area と性 gender を説明変数とした重回帰分析を行ってみよう．ただし所在地 area と性 gender はカテゴリカルな変数として扱うこと．層化変数は所在地 area である．PSU は学校 school であり層内で PSU を非復元単純無作為抽出，SSU は生徒 obs であり PSU 内で非復元単純無作為抽出とする．

ヒント：一段目は各層において変数 M.h にある PSU 数から変数 m.h にある PSU 数を非復元単純無作為抽出し，二段目は各 PSU において変数 N.a にある SSU 数から変数 n.a にある SSU 数を非復元単純無作為抽出している．

13.b　ロジスティック回帰分析

ロジスティック回帰分析

```
GLM.RES <- svyglm(formula=Y~X1+X2+...+XK, design=DES,
                                                family=binomial())
```

ロジスティック回帰分析を行うには，基準変数 Y に 0/1 の二値変数を指定し，引数 family には binomial() を指定する．抽出ウェイト w_i が整数でない場合には警告メッセージが表示されるが問題ない．これを避けるためには引数 family を quasibinomial() とすればよい．

13.b.1　演習問題

演習 13.2　層化二段確率比例抽出法とロジスティック回帰分析

データフレーム otona が層化二段抽出されたものとして，質問項目 Q2 が 1 であるか否かを基準変数とし，市郡 city と性 gender と質問項目 Q3 と Q4 を説明変数としたロジスティック回帰分析を行ってみよう．ただし一段目は市郡 city で層化し，学校 school を在籍児童数 N.a で復元確率比例抽出したものとする．二段目は性 gender で層化し，児童 obs を非復元単純無作為抽出したものとする．各説明変数は全てカテゴリカルな変数として扱うものとし，説明変数ごとに回帰係数が全て 0 かどうかの検定を行うこと．

ヒント：在籍児童数の層総計は変数 N.h にあり，層ごとの標本学校数は変数 m.h にある．各学校の性別在籍数は変数 N.ag とし，標本となった児童数は変数 n.ag にある．

13.c 演習問題解答例

演習 13.1 解答例

一段目・二段目ともに非復元単純無作為抽出なので，抽出ウェイトは $w_i = M_h/m_h \times N_a/n_a$ となる．以下の例では標本抽出デザインを指定した結果を des としている．

演習 13.1 解答例 -1

```
##### 抽出ウェイトの作成 #####
> kenko$w <- kenko$M.h / kenko$m.h * kenko$N.a / kenko$n.a

##### 標本抽出デザインの指定 #####
> des <- svydesign(ids=~school + obs, strata=~area, fpc=~M.h + N.a, weights=~w, data=kenko)
```

以下の例ではまず関数 svyglm() を用いて重回帰分析を行った後に，関数 summary() を利用して回帰係数を求めている．回帰係数が 0 かどうかの検定を行うと，体重 weight の回帰係数は有意であり，性 gender 差もあることが分かる．

演習 13.1 解答例 -2

```
##### 重回帰分析 #####
> glm.res <- svyglm(height ~ weight + factor(area) + factor(gender), des)

##### 重回帰分析の結果の要約 #####
> summary(glm.res)

Call:
svyglm(height ~ weight + factor(area) + factor(gender), des)

Survey design:
svydesign(ids = ~school + obs, strata = ~area, fpc = ~M.h + N.a,
    weights = ~w, data = kenko)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     162.31750    1.13032 143.603 2.09e-13 ***
weight            0.10516    0.01955   5.378  0.00103 **
factor(area)2     0.17532    0.29173   0.601  0.56680
factor(gender)2 -10.39747    0.12924 -80.449 1.21e-11 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

(Dispersion parameter for gaussian family taken to be 8.16356)

Number of Fisher Scoring iterations: 2
```

ちなみに決定係数の値は $R^2 = 0.8037848$ となる．

演習 13.1 解答例 -3

```
##### 決定係数 #####
> 1 - glm.res$deviance / glm.res$null.deviance
[1] 0.8037848
```

演習 13.2 解答例

関数 svydesign() を用いて標本抽出デザインを指定するときには，一段目が復元抽出なので一段目だけを指定すればよい．

演習 13.2 解答例 -1

```
##### 抽出ウェイトの作成 #####
> otona$w <- otona$N.h / (otona$m.h * otona$N.a) * otona$N.ag / otona$n.ag

##### 標本抽出デザインの指定 #####
> des <- svydesign(ids=~school, strata=~city, weights=~w, data=otona)
```

ロジスティック回帰分析を行うには，関数 svyglm() において引数 family=binomial() を指定する．Warning message:が表示されるが特に問題ない．

演習 13.2 解答例 -2

```
#####  ロジスティック回帰分析 #####
> glm.res <- svyglm((Q2==1) ~ factor(city) + factor(gender) + factor(Q3) + factor(Q4), des,
     family=binomial())

Warning message:
In eval(expr, envir, enclos) : non-integer #successes in a binomial glm!
```

関数 summary() を用いると，各回帰係数の推定値とともにそれぞれが 0 かどうかの検定結果が表示される．市郡 city と性 gender はいずれもカテゴリが 2 つであり，それぞれの変数でカテゴリ間に 5% 水準で有意な差があることが分かる．

演習 13.2 解答例 -3

```
#####  ロジスティック回帰分析の結果の要約 #####
> summary(glm.res, df.resid=degf(des))

Call:
svyglm((Q2 == 1) ~ factor(city) + factor(gender) + factor(Q3) +
    factor(Q4), des, family = binomial())

Survey design:
svydesign(ids = ~school, strata = ~city, weights = ~w, data = otona)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.9371     0.4550   4.257 0.000238 ***
factor(city)2    -0.5568     0.2445  -2.277 0.031246 *
factor(gender)2   0.4437     0.1291   3.437 0.001991 **
factor(Q3)2      -1.0829     0.3892  -2.782 0.009914 **
factor(Q3)3      -1.6466     0.3772  -4.365 0.000179 ***
factor(Q3)4      -2.8058     0.4924  -5.699 5.38e-06 ***
factor(Q4)2       0.7750     0.4192   1.849 0.075881 .
factor(Q4)3       0.7375     0.3550   2.078 0.047733 *
factor(Q4)4       0.8833     0.3838   2.301 0.029647 *
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

(Dispersion parameter for binomial family taken to be 1.004718)

Number of Fisher Scoring iterations: 5
```

以下の例では関数 regTermTest() を用いて説明変数 Q3 の全てのカテゴリの回帰係数が 0 かどうかの (無修正) Wald 検定を行っている．p 値は $p = 1.1139e-05$ であり，全てのカテゴリの回帰係数が 0 という帰無仮説は棄却される．

演習 13.2 解答例 -4

```
#####  説明変数Q3の検定 #####
> regTermTest(glm.res, ~factor(Q3), df=degf(des))
Wald test for factor(Q3)
 in svyglm((Q2 == 1) ~ factor(city) + factor(gender) + factor(Q3) +
    factor(Q4), des, family = binomial())
F =  14.21227  on  3  and  26  df: p= 1.1139e-05
```

同様に以下の例では説明変数 Q4 の全てのカテゴリの回帰係数が 0 かどうかの検定を行っている．p 値は $p = 0.17312$ であり有意ではない．

演習 13.2 解答例 -5

```
#####  説明変数Q4の検定 #####
> regTermTest(glm.res, ~factor(Q4), df=degf(des))
Wald test for factor(Q4)
 in svyglm((Q2 == 1) ~ factor(city) + factor(gender) + factor(Q3) +
    factor(Q4), des, family = binomial())
F =  1.793721  on  3  and  26  df: p= 0.17312
```

第14章　演習問題で用いるデータ

14.a　kigyo

企業を対象に，売上高を調査した仮想データである $(n = 2,000)$.

- ◇ area は企業の所在地であり，1 から 3 の 3 つの値をとる.
- ◇ gyoshu は企業の業種であり，1 から 5 の 5 つの値をとる.
- ◇ shihon は企業の資本金である.
- ◇ uriage は企業の売上高である.
- ◇ uriage.na は企業の売上高であり，275 社は値が欠測 NA となっている.
- ◇ N は母集団サイズであり，全ての企業の値が 10000 である.
- ◇ n は標本サイズであり，全ての企業の値が 2000 である.
- ◇ N.h は所在地・業種別の母集団サイズであり，各所在地と業種の組み合わせ内では，全ての企業の値が同一である.
- ◇ n.h は所在地・業種別の標本サイズであり，各所在地と業種の組み合わせ内では，全ての企業の値が同一である.
- ◇ total.shihon は資本金の母集団総計であり，全ての企業の値が 1725000 である.
- ◇ total.shihon.h は所在地・業種別の資本金の母集団総計であり，各所在地と業種の組み合わせ内では，全ての企業の値が同一である.

例題 14.1-1

```
> kigyo[seq(1,2000,250), ]
      obs area gyoshu shihon uriage uriage.na     N    n  N.h n.h total.shihon
1       1    1      1     15     21        21 10000 2000  400 100       1725000
251   251    1      3    189     65        65 10000 2000  300 100       1725000
501   501    2      1     21     15        15 10000 2000  600 140       1725000
751   751    2      2    150    118       118 10000 2000  900 140       1725000
1001 1001    2      4     21     20        20 10000 2000 1100 140       1725000
1251 1251    3      1     95     60        60 10000 2000  900 160       1725000
1501 1501    3      2     11     16        16 10000 2000  800 160       1725000
1751 1751    3      4    161     85        85 10000 2000  900 160       1725000

     total.shihon.h
1             78000
251           71000
501          127000
751          101000
1001         117000
1251         149000
1501         168000
1751         144000
```

14.b kenko

生徒を対象に，身長と体重を調査した仮想健康診断データである ($n = 884$).

- ◇ area は生徒が在籍する学校の所在地であり，1 と 2 の 2 つの値をとる．
- ◇ school は生徒が在籍する学校であり，1 から 12 の 12 の値をとる．
- ◇ gender は生徒の性別であり，1 (男子) と 2 (女子) の 2 つの値をとる．
- ◇ height は生徒の身長であり，weight は生徒の体重である．
- ◇ N は母集団サイズであり，全ての生徒の値が 749569 である．
- ◇ n は標本サイズであり，全ての生徒の値が 884 である．
- ◇ N.g は性別の母集団サイズであり，男子は 375574，女子は 373995 である．
- ◇ n.g は性別の標本サイズであり，男子は 465，女子は 419 である．
- ◇ M は母集団における学校数であり，全ての生徒の値が 3576 である．
- ◇ M.h は所在地別母集団学校数であり，所在地が 1 は 1999，所在地が 2 は 1577 である．
- ◇ m.h は所在地別標本学校数であり，所在地が 1 は 7，所在地が 2 は 5 である．
- ◇ N.h は所在地別母集団サイズであり，所在地が 1 は 439996，所在地が 2 は 309573 である．
- ◇ N.a は学校別在籍生徒数であり，各学校内では全ての生徒の値が同一である．
- ◇ n.a は学校別標本生徒数であり，各学校内では全ての生徒の値が同一である．
- ◇ N.ag は学校・性別在籍生徒数であり，各学校と性別の組み合わせ内では全ての生徒の値が同一である．
- ◇ n.ag は学校・性別の標本サイズであり，各学校と性別の組み合わせ内では全ての生徒の値が同一である．

例題 14.2-1

```
> kenko[seq(1,884,100),]
    obs area school gender height weight      N   n    N.g n.g    M  M.h m.h    N.h N.a
1     1    1      1      1  168.0   59.5 749569 884 375574 465 3576 1999   7 439996 520
101 101    1      1      2  159.1   54.5 749569 884 373995 419 3576 1999   7 439996 520
201 201    1      3      2  154.7   56.9 749569 884 373995 419 3576 1999   7 439996 120
301 301    1      5      1  170.9   62.3 749569 884 375574 465 3576 1999   7 439996 100
401 401    1      6      2  154.7   50.0 749569 884 373995 419 3576 1999   7 439996 450
501 501    2      8      1  167.4   67.3 749569 884 375574 465 3576 1577   5 309573 240
601 601    2      9      1  165.5   65.9 749569 884 375574 465 3576 1577   5 309573  90
701 701    2     10      2  164.7   48.9 749569 884 373995 419 3576 1577   5 309573  80
801 801    2     11      2  158.8   60.7 749569 884 373995 419 3576 1577   5 309573 480
    n.a N.ag n.ag
1   106  275   60
101 106  245   46
201  55   75   38
301  92   58   57
401  79  305   53
501 105  138   63
601  65   45   43
701  56   42   37
801  81  251   51
```

14.c otona

小学 6 年生を対象に実際に行われた調査データの一部を加工したものである ($n = 853$).

- ◇ school は児童が在籍する学校であり，1 から 28 まで 28 の値をとる.
- ◇ gender は児童の性別であり，1 (男子) と 2 (女子) の 2 つの値をとる.
- ◇ Q1 は『あなたの身の回りには「あのようになりたくない」と思う大人の人がいますか.』という質問項目への回答であり，1 (いる) と 2 (いない) の 2 つの値をとる.
- ◇ Q2 は『それでは「あのようになりたい」と思う大人の人がいますか.』という質問項目への回答であり，1 (いる) と 2 (いない) の 2 つの値をとる.
- ◇ Q3 は『あなたは，近所の大人の人からほめられたことがありますか.』という質問項目への回答であり，1 (よくある)，2 (時々ある)，3 (あまりない)，4 (全くない) の 4 つの値をとる.
- ◇ Q4 は『それでは，近所の大人の人から注意されたりしかられたことがありますか.』という質問項目への回答であり，1 (よくある)，2 (時々ある)，3 (あまりない)，4 (全くない) の 4 つの値をとる.
- ◇ N は母集団サイズであり，全ての児童の値が 505252 である.
- ◇ city は学校の市郡別であり，1 (市部) と 2 (郡部) の 2 つの値をとる.
- ◇ N.h は市郡別母集団サイズであり，市郡が 1 は 225634，市郡が 2 は 279618 である.
- ◇ M.h は市郡別母集団学校数であり，市郡が 1 は 3146，市郡が 2 は 9469 である.
- ◇ m.h は市郡別標本学校数であり，市郡が 1 は 13，市郡が 2 は 15 である.
- ◇ N.a は学校別在籍児童数であり，各学校内では全ての児童が同一の値である.
- ◇ N.ag は学校・性別在籍児童数であり，各学校と性別の組み合わせ内では全ての児童が同一の値である.
- ◇ N.class は学校別学級数であり，各学校内では全ての児童が同一の値である.
- ◇ n.a は学校別標本児童数であり，各学校内では全ての児童が同一の値である.
- ◇ n.ag は学校・性別標本児童数であり，各学校と性別の組み合わせ内では全ての児童が同一の値である.

例題 14.3-1

```
> otona[seq(1,nrow(otona), 80), ]
    obs school gender Q1 Q2 Q3 Q4      N city    N.h  M.h m.h N.a N.ag N.class n.a n.ag
1     1      1      2  1  1  2  4 505252    1 225634 3146  13  60   32       2  37   19
81   81      3      1  1  1  2  3 505252    1 225634 3146  13  76   40       2  29   14
161 161      6      1  2  2  2  4 505252    1 225634 3146  13  54   27       2  25   13
241 241      9      1  1  1  1  3 505252    1 225634 3146  13 121   59       4  34   19
321 321     11      2  1  1  2  2 505252    1 225634 3146  13  95   46       3  26   15
401 401     13      1  1  1  2  3 505252    1 225634 3146  13  75   40       2  35   20
481 481     16      1  2  1  4  4 505252    2 279618 9469  15 100   49       3  27   17
561 561     19      2  1  1  3  1 505252    2 279618 9469  15  30   16       1  30   16
641 641     21      1  1  1  2  3 505252    2 279618 9469  15  31   15       1  31   15
721 721     23      2  1  1  3  4 505252    2 279618 9469  15  40   18       1  40   18
801 801     26      2  1  1  3  3 505252    2 279618 9469  15  95   47       3  37   17
```

関数索引

改訂履歴

2009.08.25 第 1.0 版を発行しました。