第9章 ウェイトの利用

ここまでの章におけるデータ分析では，データが集められた過程については注意を払う必要がないと仮定してきた。わが国においては社会制度上の特殊な条件（選挙人名簿や住民基本台帳が国によって整備され調査研究者に広く解放されてきた）がかつて存在し，理想的な単純無作為抽出（シンプル・ランダム・サンプリング）に近いかたちで調査が可能であったために，データの収集過程を考慮した分析手法というものが必要なく，紹介もされてこなかった。

しかし，個人情報保護への関心の高まりから，サンプリングの困難は大きくなっている。今後は電話調査や，名簿ではなくフィールドにおいて直接サンプリングを行なうエリア・サンプリングといった手法が広まるだろう。このことを考えれば★1，抽出の確率がサンプルで一定でない状況に応じた分析方法を紹介しておくことに意義があると思われる。そのため，この章ではやや理論的な説明を行ない，その後にStataを用いた具体的な分析手順を紹介することにする。

★1：一般に電話調査やエリア・サンプリングでは，電話・訪問先の家庭で1人の回答者をランダムに選ぶが，その際に家庭の構成員の人数は一定でないために，世帯人数の小さい家庭の構成員ほど選ばれる確率が高くなり，世帯人数が多い家庭の構成員ほど選ばれる確率が小さくなるという現象が起きる。このような状況に対応するためには，サンプリング・ウェイトをつくりそれを用いた分析をすることで，偏りのない分析が可能になる。

1節 サンプリングの理論について

抽出（サンプリング）を行なう理由は何であろうか。サンプリングは，まず第一に経済的理由・利用可能な資源の限界を出発点に，サンプルを抽出しようとする動機から生まれた。ジニ係数に名を残しているイタリアの経済学者コッラド・ジニが20世紀の初頭にイタリア国勢調査の調査票からサンプルを抽出しようと考えた動機は，紙の調査票を保存しておく部屋が足りないという現実的な問題の前に，いかにして後の再分析に耐えるような「代表的な」サンプルを保存しておくか，というものであった。

ジニの取った解決法は現在の視点からみて最適な方法ではなかったが，その後のイェジ・ネイマン（1934），ウィリアム・コクラン（1977）らによる理論的発展およびレスリー・キッシュ（1965）らによる調査の実施の体系化を通じて標本調査法は20世紀を通じて各国政府の公的機関をはじめとして社会科学一般において広く利用されることになった。

サンプリング理論の基本は，①母集団名簿に含まれるすべての個体にゼロでない抽出確率が与えられている情況で，サンプリングを繰り返せば理論的には必ずすべての

個体をサンプルに含めることができ，②抽出された個体を観測（通常は調査員による面接や自記式の郵送調査など）して，個体のデータを収集し，③既知の抽出確率とサンプル・データを基に母集団の属性について偏りのない（unbiased）な推定を行なうことができるということにある。サンプリングが受け入れられる以前に行なわれていた全数調査に比べて，サンプルだけを調べればよいので費用と時間を大幅に減らすことができるという点でこれは革命的な発見であった。

また，サンプリング理論のもう1つの強みは，④得られたデータの「信頼性(reliability)」を数学的に明解な方法で一意に計算できるという点である。サンプル・サイズ（社会調査では回答者数）と抽出比率およびデータのもつ分散の3つの変数から得られた推定値のもつ変動を計算することができる（信頼区間）。これらはネイマンによる理論的貢献の主たるものであり，大幅なコストの削減と，コストを減らしたことに伴う信頼性の低下というトレードオフを管理できることが，標本調査法が世の中に広く受け入れられた理由の1つである。

2節　調査の非標本誤差に対処するためのウェイトの考え方

実際には，いくつかの理由によって①～②の前提は満たされない。まず母集団の構成員を完全に網羅した理想的な名簿はまず存在しない。たとえば選挙人名簿は，日本であれ米国であれ，名簿に住民の移動が反映されるまでに多少の時間がかかる。また転居者が住民票を移さなかったり，米国であれば新しい住所で選挙人登録をしないこともある。役所が名前や住所をまちがえるということもある。

選挙人名簿からサンプリングを行なって調査をすれば，そのような世帯は調査対象から欠落する。やや古い研究であるが，NHKによる研究（1971）では3月から4月にかけて多くの人口移動が発生し，名簿からの欠落が大量にこの時期に発生していた(補足漏れによる誤差)。入学，卒業だけでなく，多くの人事異動が年度をまたぐかたちで行なわれるからである。米国の選挙人名簿に載っている名前と住所から電話番号が判明するのはせいぜい6割であり，電話調査に頼る米国の選挙調査ではこれらの世帯は調査対象から漏れている。

このために，無作為なサンプリングを行なっても，常に特定の対象は調査対象から欠落するのであり，サンプルと母集団の間に乖離が生じるほうが一般的である。この乖離に系統的な特徴があれば（たとえば低所得者ほど電話を引いていないし，若い世代ほど移動の機会が多い），調査結果にも系統的な影響を与える。そのため，この補足漏れによる誤差は重要である。一方で，補足漏れがランダムに生じている場合は，それはサンプリング・エラーであり分析結果には影響を与えない。

近年では，②の抽出された全個体を観察するという条件も達成できない。ネイマン

の理論は調査における欠測（non observation）はいっさい考慮されていないので，調査への系統的な未回答（たとえば忙しい人ほど家を空けることが多く，調査員が対象を補足しにくい）による影響を直接に被るのである。また特定の質問項目への無回答（たとえば収入を尋ねる質問などは，無回答が多くなりがちである）も基本的には欠測と同じ効果をもたらす。

得られたデータに何らかの修正を加えて補足漏れ誤差や欠測の問題に対処しようという試みは，じつは標本調査の黎明期から行なわれていた。サンプルを集計した後に，既知の母集団の集計と比較する事が可能である場合に，その乖離を修正しようという初期の例としてはデミングとステファン（1940）による米国の国勢調査のサンプルを用いた例がある。国勢調査のデータは膨大で，高速なコンピューターが存在しなかった当時，結果の集計は多大な労力と時間を要する作業であった。そのために国勢調査の結果からサンプルを抽出し，そのサンプルを分析することで効率化を図ったのである。しかし，サンプルにはサンプリング・エラーが含まれるため，標本の集計結果と母集団の集計結果に乖離が生じる。この問題を克服するためにデミングらが取ったのが，サンプルの集計結果が母集団の集計結果に一致するような比率（ウェイト）を計算し，そのウェイトを用いた推定をすることでサンプルの周辺分布と母集団の周辺分布を一致させるという方法であった。この方法は後にレイキングとよばれるようになり，今日でも広く使われている方法である。

デミングらのアプローチはサンプリング・エラーによる乖離の最小化を主たる目的としたものであったがその後の理論の発展により上述の名簿の不備による誤差や欠測による誤差についても，ウェイトを用いて誤差を減らす試みが広く行なわれるようになった。

3節　調査の抽出確率の違いに対処するためのウェイトの考え方

ここでウェイトと各種誤差の関係をまとめておこう。調査の世界では誤差には主に以下のものがある。

1．サンプリング・エラー：母集団の構成員からサンプルを抽出する過程で生じる誤差。ランダムな抽出をしている限り，この偏りは一定の方向性をもたずに個別の誤差がお互いを打ち消しあうために問題にならない。

2．非標本誤差

a．補足漏れによる誤差：調査対象の母集団と利用する名簿の間の対応関係の不備から生ずる誤差。

b．欠測による誤差：対象者のうち特定の属性をもった人間だけが回答者となる，あるいはならないことによる誤差。

3．サンプリング・バイアス：この本に付属のデータセットのように，都市部と農村部においてサンプリングの確率を変えている場合や，電話調査，エリア・サンプリングのように，世帯内抽出を必要とする場合に，世帯人数によってサンプリングの確率が変わってしまうことによる誤差。

通常1はサンプル数を大きくすることで任意に小さくすることが可能であり，また偏りとしては特定の方向性をもたないために問題とならない。次節以降では3の例を扱う。通常，サンプリング・ウェイト（sampling weight）を用いることで，ほぼ完全に3による偏りは取り除くことができる。

2の補足漏れによる誤差や欠測による誤差もまた，3とほぼ同じ方法で対応することが可能である。その場合，作成されたウェイトはノンレスポンス・ウェイト（non-response weight）と通常呼ばれサンプリング・ウェイトとは区別されるが，ノンレスポンス・ウェイトはサンプリング・ウェイトを内包していることが多い★2。注意点としては，サンプリング・ウェイトと異なり，欠側による誤差はノンレスポンス・ウェイトを用いたところで，必ずしもなくなるわけではないということである。

★2：これはすなわち，ノンレスポンス・ウェイトを用いて分析を行なえば2と3の両方の誤差に一度に対応できることを意味する。

欧米ではノンレスポンス・ウェイトおよびサンプリング・ウェイトはたいていの社会調査・疫学調査の公開データに付属しており，またその利用についても政治学・公衆衛生学などの分野では使うべきであるという一定のコンセンサスが存在するが，日本ではあまり受け入れられていないようである。

4節 ウェイト作製の実際

（1）事後層化

ウェイトを作るにあたっては母集団に関する正確な情報が必要である。母集団の名簿の集計が手に入れば，それを用いる。官庁による統計など，公表されている，比較的質が高い情報が利用できるなら，それらを母集団の母数としてみなす。通常よく用いられる変数は，階級別年齢，性別，学歴などである。

ここで紹介するウェイトの作成法は，事後層化（cell weighting）による方法である。事後層化によってウェイトを作る場合，まず2変量もしくは多変量の分布から得られる母集団の集計を用いる。典型的には，国勢調査などのデータのクロス表の数値や母集団名簿の数値を利用する。クロス表のような複数の変数の組み合わせに対応する母集団の数値が手に入らない場合は，周辺度数のみを用いたレイキングという方法が用いられることも多々あるが，基本的には事後層化を単純化したものとみなすことで同

じコマンドが利用できるので省略する。

(2) svrコマンドのインストール

初期状態では，Stataにはウェイトそのものを作るコマンドはインストールされていない。spostと同様に，ウェイトを作成するsvrコマンドをインストールする必要がある。spostと同じく，このコマンドをインストールするためには，インターネットへの接続が必須である。

インターネットに接続された状態で，

```
net search svr
```

と入力すると，以下の表示がされるので（出力例9－1），□で囲んだエリアをクリックする。インストールされるファイルについての情報を示した別ウィンドウが開くので，Click here to installをクリックすれば，インストールできる。

■出力例9－1

```
. net search svr
(contacting http://www.stata.com)

1 package found (Stata Journal and STB listed first)
-----------------------------------------------------

svr from http://fmwww.bc.edu/RePEc/bocode/s
    'SVR': module to compute estimates with survey replication (SVR) based
    standard errors / The prefix svr stands for SurVey Replication, and refers
    to / commands that analyze complex survey data using replication /
    methods.  The available methods are balanced repeated replication / (BRR),
```

(3) Survwgtによるウェイトの作成

表9－1は，教科書付属の調査のサンプル計画である。この調査では，対象者が地方と都市規模で層化されて抽出されていた。その分配は純粋なランダムではなく，むしろ大きく偏っている。

表9－1aでは対象者1人あたりが何人の母集団を代表しているか（N/n）を計算しているが，たとえば関東・甲信越では母集団での人口31298578人からサンプルとして684人を抽出しているから，1人の対象者が45,758人を代表していることになる。一方で，九州では1人の対象者が代表するのはずっと少ない29,703人になる。一方，都市規模の表に目を向けてみれば，14大都市では39,243人に1人が抽出されている一方で，5000人未満の小さな郡部では2,541人に1人が抽出されている。

これは，この調査が都市部と郡部との比較を目的としており，その目的のためには単純に人口規模で比例配分をしては郡部の対象者が少なくなりすぎると判断されたためである。回答者が少なければ十分な推定の精度を保てないために，意図的に郡部を

多めに抽出（over sampling）したのである。しかし，そのことが郡部と都市部を合わせて推定した際に，郡部のデータがより強く反映されてしまうという問題を引き起こしている。適切なウェイトをかけて分析すれば，この問題を調整可能である。以下に，その手順を示そう。

まず表9－2に示したデータを表9－1bの層化表を元に準備する。regionは地域であり1が北海道・東北地域を示し，2が関東・甲信越。3が東海・北陸，4が近畿，5が中国・四国，6が九州である。またcitysizeは5が5000人未満の都市規模，1が14大都市で，人口希望が小さいほど，値が大きくなるようになっている。つまり，サンプルデータには6つの地域に5つの都市規模が存在し，全部で30の層が存在する。

まずおのおのの層を識別する変数を作る。これは都市規模と地域の組み合わせを識別できればなんでもよい。ここでは，表9－2のように30の層に1～30の数値を振ることにしよう。StratumPopの列は各層における母集団の人口である。

表9－1　サンプル・データの調査計画と回収数

a. 調査計画と回収数

地域	北海道・東北	関東・甲信越	東海・北陸	近畿	中国・四国	九州	合計
母集団数（N）	10,059,764	31,298,578	11,772,961	13,766,472	7,633,324	9,386,293	83,917,392
計画サンプル数（n）	408	684	247	294	251	316	2200
抽出率の逆数（N/n）	24,656	45,758	47,664	46,825	30,412	29,703	38,144
回収数	268	386	174	197	182	238	1,445

区分	市部		郡部			
都市規模	14大都市	その他の市	10000以上	5000人以上	5000人未満	合計
母集団数（N）	19,621,591	49,858,837	10,518,590	2,901,916	1,016,458	83,917,392
計画サンプル数（n）	500	500	400	400	400	2200
抽出率の逆数（N/n）	39,243	99,718	26,296	7,255	2,541	38,144
回収数	284	269	287	302	303	1445

b. 層化表

地域	市部		郡部			合計
	14大都市	その他の市	10000以上	5000人以上	5000人未満	
北海道・東北	1,910,605 49(3)	5,190,967 52(4)	1,806,037 69(5)	841,226 116(7)	310,929 122(8)	10,059,764 408(27)
関東・中信越	10,356,420 264(16)	17,190,217 173(10)	3,114,404 118(7)	474,498 65(4)	163,039 64(4)	31,298,578 684(41)
東海・北陸	1,436,947 36(3)	8,194,987 82(5)	1,833,729 70(5)	244,534 34(2)	62,764 25(2)	11,772,961 247(17)
近畿	3,639,376 93(6)	8,793,660 88(6)	940,016 36(3)	303,996 42(3)	89,424 35(3)	13,766,472 294(21)
中国・四国	744,837 19(2)	5,199,204 52(4)	1,065,317 40(3)	414,430 57(4)	209,536 83(5)	7,633,324 251(18)
九州	1,533,406 39(3)	5,289,802 53(4)	1,759,087 67(4)	623,232 86(6)	180,766 71(5)	9,386,293 316(22)
母集団数 サンプル (地点数)	19,621,591 500(33)	49,858,837 500(33)	10,518,590 400(27)	2,901,916 400(26)	1,016,458 400(27)	83,917,392 2200(146)

〈層変数を指定〉

```
gen strataID = 0
replace strataID=1 if region==1 & citysize==1
replace strataID=2 if region==1 & citysize==2
replace strataID=3 if region==1 & citysize==3
replace strataID=4 if region==1 & citysize==4
replace strataID=5 if region==1 & citysize==5
replace strataID=6 if region==2 & citysize==1
(途中省略)
replace strataID=28 if region==6 & citysize==3
replace strataID=29 if region==6 & citysize==4
replace strataID=30 if region==6 & citysize==5
```

地域と都市規模に対応する層変数（strataID）を指定した後は，下のように，各層の人口を指定する。

〈各層の人口を指定する〉

```
gen StratumPop= 0
```

表9－2　事後層化を行なうために準備するデータ

region	citysize	StrataID	StratumPop
1	1	1	1,910,605
1	2	2	5,190,967
1	3	3	1,806,037
1	4	4	841,226
1	5	5	310,929
2	1	6	10,356,420
2	2	7	17,190,217
2	3	8	3,114,404
2	4	9	474,498
2	5	10	163,039
3	1	11	1,436,947
3	2	12	8,194,987
3	3	13	1,833,729
3	4	14	244,534
3	5	15	62,764
4	1	16	3,639,376
4	2	17	8,793,660
4	3	18	940,016
4	4	19	303,996
4	5	20	89,424
5	1	21	744,837
5	2	22	5,199,204
5	3	23	1,065,317
5	4	24	414,430
5	5	25	209,536
6	1	26	1,533,406
6	2	27	5,289,802
6	3	28	1,759,087
6	4	29	623,232
6	5	30	180,766

```
replace StratumPop=1910605 if strataID==1
replace StratumPop=5190967 if strataID==2
replace StratumPop=1806037 if strataID==3
(途中省略)
replace StratumPop=1759087 if strataID==28
replace StratumPop=623232 if strataID==29
replace StratumPop=180766 if strataID==30
```

tabulate を用いて層別の人口が正しく追加された事を確認してみよう（出力例9－2）。

■出力例9－2

StratumPop	Freq.	Percent	Cum.
62764	17	1.18	1.18
89424	24	1.66	2.84
163039	40	2.77	5.61
180766	79	5.47	11.07
209536	56	3.88	14.95
244534	26	1.80	16.75
303996	27	1.87	18.62
310929	68	4.71	23.32
414430	43	2.98	26.30
474498	31	2.15	28.44
623232	59	4.08	32.53
744837	16	1.11	33.63
841226	83	5.74	39.38
940016	23	1.59	40.97
1065317	32	2.21	43.18
1436947	24	1.66	44.84
1533406	26	1.80	46.64
1759087	40	2.77	49.41
1806037	49	3.39	52.80
1833729	52	3.60	56.40
1910605	30	2.08	58.48
3114404	91	6.30	64.78
3639376	64	4.43	69.20
5190967	38	2.63	71.83
5199204	35	2.42	74.26
5289802	34	2.35	76.61
8194987	55	3.81	80.42
8793660	59	4.08	84.50
1.04e+07	143	9.90	94.39
1.72e+07	81	5.61	100.00
Total	1,445	100.00	

上記の例に拠れば1から30までの値をもったstrataという変数が作成される。続いて，ウェイトを作成しよう。

次のページから

・事後層化法のためのウェイト変数を作る

survwgt poststratify *varname*, by(*strata ID*) totvar(*strata size*) replace
survwgt poststratify ウェイト変数, by(層の ID) totvar(層のサイズ) replace

survwgt は，各層のサイズを元に，自動でウェイトを作成するコマンドである。ただ，まず最初に基になる変数が存在しないといけないので，回答者全員に同じウェイトとして定数1を与える★3。その後，survwgt を実行する。

```
gen weight=1
survwgt poststratify weight, by(strataID) totvar(StratumPop) replace
```

★3：「作成する」と書いたが，generate コマンドより replace コマンドに近い。今ある変数をウェイト変数に置き換えるといったほうがよいだろう。

上記のコマンドを実行し終わったら tab weight　として確認する（出力例9－3）。

■出力例9－3

(post-stratified)	Freq.	Percent	Cum.
2288.177	79	5.47	5.47
3692	17	1.18	6.64
3726	24	1.66	8.30
3741.714	56	3.88	12.18
4075.975	40	2.77	14.95
4572.485	68	4.71	19.65
9405.154	26	1.80	21.45
9637.907	43	2.98	24.43
10135.25	83	5.74	30.17
10563.25	59	4.08	34.26
11259.11	27	1.87	36.12
15306.39	31	2.15	38.27
33291.16	32	2.21	40.48
34224.22	91	6.30	46.78
35264.02	52	3.60	50.38
36857.9	49	3.39	53.77
40870.26	23	1.59	55.36
43977.17	40	2.77	58.13
46552.31	16	1.11	59.24
56865.25	64	4.43	63.67
58977.15	26	1.80	65.47
59872.79	24	1.66	67.13
63686.83	30	2.08	69.20
72422.52	143	9.90	79.10
136604.4	38	2.63	81.73
148548.7	35	2.42	84.15
148999.8	55	3.81	87.96
149045.1	59	4.08	92.04
155582.4	34	2.35	94.39
212224.9	81	5.61	100.00
Total	1,445	100.00	

・Stata にサンプリング・ウェイトの情報を伝える
svyset psu [pweight=*weight*], strata(*strata ID*)
svyset 抽出地点の ID [pweight=ウェイト変数], strata(層の ID)

では，実際にウェイトを用いた推定を行ない，予定通りの結果になるかを確認しよう。Stataでは調査のデザインを反映した分析を行なう際には，最初にsvysetコマンドを用いてサンプリング計画にかかわる設定をStataに伝える必要がある。基本的に設定するべき項目は3つあり，（1）抽出地点のIDおよび（2）ウェイト変数，そして（3）層のIDである。

今回の例では，以下のように入力する。

```
svyset cityid [pweight=weight], strata(stratalD)
```

cityidは抽出地点（町村）を区別する変数である。単純無作為抽出の場合は省略して構わない。同じ町や村から抽出された標本は似通った属性をもっていることが多く，サンプリングの効率を下げてしまう。svysetコマンドに，抽出地点IDを渡すことで，この抽出地点内の相関による効率の低下を考慮したより正確な分散を計算することができる★4。

★4：抽出地点内の相関が非常に高い場合（極端な例として郵便番号の分布を調べた場合）町村内の郵便番号にはあまり変動がないので実質的な標本数は1445ではなく，町村部の147に近くなり，分散の推定値もn=147として計算した場合にほぼ等しくなる。一方で町村内における相関がほとんどない場合，もしくは負の相関である場合（近隣に住んでいるほど属性や回答が似通っていない傾向がある場合）は，実質的な標本数は単純無作為抽出と同じになる場合やそれを上回る効率をもつ場合も理論的には存在するが，通常はあまりそういうことはない。詳しくはKishによるdesign effectの議論を参照の事。

設定がすんだらうまくいったかどうかを確認する。確認には**svydescribe**コマンド(svydesと省略できる）を用いる。上段はウェイト，層，および抽出地点の指定結果を表示し，下段の表では各層ごとの抽出地点数が表示される（出力例9－4)。svyコマンドを実行する場合最低でも1つの層に付き2つ以上の抽出地点がなくてはいけない。もし特定の層の抽出地点が1つしかない場合は，できるだけ形質の近い層と併合して構わない（たとえば，10番目の層の抽出地点が1しかなければ，9番目か11番目と併合する)。そうしなければ分散が推定できない。また併合の結果は点推定（平均値や，回帰係数）には影響を与えない★5。単に

★5：Stata 10以降ではsingleunitオプションを指定することで，層の中に1つしか抽出地点がない場合の処理を自動化できるようになった。デフォルトではそのような抽出地点がある場合にsurveyコマンドは分散を計算せずに，分散の欄が空の出力を返す。そのような事態が生じた場合は，svy describeコマンドで問題となっている層を調べて層を併合するか，オプションを指定することでその手間を省くことができる。ただし，どのオプションにも一長一短があり特に1つのオプションが勧められるわけではない。

```
svydes
```

と入力すれば，出力例9－4を得る。

svydescribeコマンドによる確認がすんだら以下のようなコマンドを入力して，svyコマンドによるtabulateコマンドの結果と，通常のtabulateコマンドの結果を比較し

■出力例9－4

```
Survey: Describing stage 1 sampling units

      pweight: weight
          VCE: linearized
  Single unit: missing
     Strata 1: strataID
         SU 1: cityid
        FPC 1: <zero>
```

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	3	30	9	10.0	11
2	5	38	5	7.6	10
3	5	49	9	9.8	11
4	8	83	8	10.4	12
5	6	68	9	11.3	14
6	15	143	4	9.5	13
7	9	81	5	9.0	12
8	9	91	6	10.1	14
(途中省略)					
26	3	26	7	8.7	10
27	4	34	6	8.5	10
28	4	40	7	10.0	14
29	6	59	9	9.8	11
30	7	79	10	11.3	13
30	147	1445	4	9.8	14

てみる。通常の度数分布はもちろんtabulateコマンドで出力できる（出力例9－5）。

■出力例9－5

①

1= 14th largest cities	Freq.	Percent	Cum.
1	303	20.97	20.97
2	302	20.90	41.87
3	287	19.86	61.73
4	269	18.62	80.35
5	284	19.65	100.00
Total	1,445	100.00	

svyコマンドを用いたtabulateは，svy：を頭につけて実行する（ここでは比較のため，percentオプションも指定している）。対応したどのコマンドでも，同じように頭にsvy：とつければウェイトを考慮した分析が行なわれる（svyに対応するコマンドについては，章末のまとめを参照）。

結果は出力例9－6のようになる。

svy：tabulate citysize, percent

■出力例9－6

```
②
 (running tabulate on estimation sample)

Number of strata   =        30          Number of obs      =       1445
Number of PSUs     =       147          Population size    =   83917391
                                        Design df          =        117

----------------------------
1= 14th     |
largest     |
cities      | percentages
          1 |       23.38
------------+---------------
          2 |       59.41
          3 |       12.53
          4 |       3.458
          5 |       1.211
            |
      Total |         100
----------------------------
  Key:  percentages  =  cell percentages
```

最初の結果（①の領域）はウェイトを用いていない回答者データそのままの結果である。カテゴリー5の郡部が全体の約2割を占めている。次の結果（②の領域）はウェイトを考慮しており，層の数（30）および抽出地点数（147）が左肩に表示されて，右肩にはサンプル数（1445）および母集団数（83917391）が表示され，その下には自由度（抽出地点数－層の数）が表示されている。層化表における母集団数(83,917,392）とがほぼ一致しているのを確認してほしい（完全に一致しないのは，丸め誤差のため)。都市規模の比率において5000人未満の郡部が本来の日本の人口比に従って1.2％になっているのがわかる。

次にtabulateコマンドで都市規模と地域でクロス表を作ってみよう。度数分布表と同じく，単純にtabとすれば通常のクロス表が作られ（出力例9－7；cellオプションと，nofreqオプションで，セルの比率だけを表示してある)，svy：tabとすることで，

小文字のq です

■出力例9－7

region	Citysize 1	2	3	4	5	Total
1	2.08	2.63	3.39	5.74	4.71	18.55
2	9.90	5.61	6.30	2.15	2.77	26.71
3	1.66	3.81	3.60	1.80	1.18	12.04
4	4.43	4.08	1.59	1.87	1.66	13.63
5	1.11	2.42	2.21	2.98	3.88	12.60
6	1.80	2.35	2.77	4.08	5.47	16.47
Total	20.97	20.90	19.86	18.62	19.65	100.00

ウェイトや層の構造を考慮したクロス表を作成することができる（出力例 9－8)。両者の比較のために

svy：tab region citysize，percent format(%3.2f)

percent オプションを用いてセルの比率を表示している。また format（%3.2f）は表示する比率の桁を指定するオプションである。

■出力例 9－8

```
(running tabulate on estimation sample)

Number of strata   =        30                 Number of obs      =       1445
Number of PSUs     =       147                 Population size    =   83917391
                                               Design df          =        117

-----------------------------------------------------
          |                 Citysize
   region |      1       2       3       4       5  Total
----------+------------------------------------------
        1 |   2.28    6.19    2.15    1.00    0.37  11.99
        2 |  12.34   20.48    3.71    0.57    0.19  37.30
        3 |   1.71    9.77    2.19    0.29    0.07  14.03
        4 |   4.34   10.48    1.12    0.36    0.11  16.40
        5 |   0.89    6.20    1.27    0.49    0.25   9.10
        6 |   1.83    6.30    2.10    0.74    0.22  11.19
          |
    Total |  23.38   59.41   12.53    3.46    1.21 100.00
-----------------------------------------------------
  Key:  cell percentages

  Pearson:
    Uncorrected   chi2(20)         =  116.5435
    Design-based  F(8.25, 965.30)  =   25.9284     P = 0.0000
```

svy：tabulate コマンドは標準で独立性の検定結果がついてくるが，上段の uncorrected は調査のデザインを無視した通常のカイ2乗検定による統計量が参考に表示されており（tabulate コマンドでカイ2乗検定を行なった場合と同じ結果である），下段の design-based は調査のデザインを考慮した Rao-Scott の方法による検定結果である。この方法の詳細は本書の範囲を超えるので説明は省くが，常にこちらの結果を参考すればよい。

5節　ウェイトを用いた分析例

例として，q4 の現在の町内に何年住んでいるかという質問を分析してみよう。以下のようにして，同じ町に生まれてからずっと住んでいるというダミー変数を作り，また都市規模についても都市部と1万人以上の郡部を Urban とみなし，それ以外を suburban とすることにする。以下のようにしよう。

gen SamePlace＝q4

recode SamePlace 2/7＝0 9＝.★6

```
gen Urban=citysize
recode Urban 1/3=1 4/5=0
```

★6：recodeの際に2/7=0とすることは 1=0 2=0 3=0 4=0 5=0 6=0 7=0 とすることに等しい。2行下の1/3=1，4/5=0も同様である。

単純なクロス表からも，都市部と郊外でこの変数の分布には大きな違いがあることがわかる（出力例9－9）。生まれてから同じ町に住み続けていると答えた回答者は人口1万人以上の都市部で8.6％で人口1万人以下の都市では22％である，回答者全体の平均を見ると13.7％である。はたしてこれは日本人全体の状況を正確に反映しているといえるだろうか？（means SamePlaceとして表示した出力例9－10も参照）

■出力例9－9

Key
frequency
row percentage

Urban	SamePlace 0	1	Total
0	429	121	550
	78.00	22.00	100.00
1	814	77	891
	91.36	8.64	100.00
Total	1,243	198	1,441
	86.26	13.74	100.00

■出力例9－10

Mean estimation　　Number of obs = 1441

	Mean	Std. Err.	[95% Conf.	Interval]
SamePlace	.1374046	.0090724	.119608	.1552011

同じ分析を，都市規模による抽出率の違いを考慮したウェイトを用いて行なってみよう。先ほどと同じようにsvysetコマンドで，調査のデザイン変数を指定し（出力例9－11），先ほど同様に，tabulate（tab）コマンドにsvy：をつけることで，ウェイトを考慮した計算を指示する（出力例9－11）。meansコマンドにもsvy：をつけて，同じ指示を行なう（出力例9－12）。

母集団における同じ住居に住み続けている回答者は8.0％であり，先ほどのウェイトなし推定量の13.7％からみて大幅にその値が小さくなっていることが見て取れる。こちらの数字が，母集団全体の推定値としてはより適切である。

一方で，クロス表で見た限りでは，特に都市と地方，それぞれの内部での分布につ

■出力例 9－11

```
(running tabulate on estimation sample)

Number of strata   =        30                  Number of obs      =      1441
Number of PSUs     =       147                  Population size    =  83864888
                                                Design df          =       117

           |      SamePlace
     Urban |         0         1     Total
-----------+----------------------------------
         0 |    .8084     .1916         1
         1 |    .9257     .0743         1
           |
     Total |    .9203     .0797         1
----------------------------------------------

Key:  row proportions

Pearson:
  Uncorrected   chi2(1)         =   11.9865
  Design-based  F(1, 117)       =   20.4439     P = 0.0000
```

■出力例 9－12

```
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =        30                  Number of obs      =      1441
Number of PSUs   =       147                  Population size    =  8.4e+07
                                              Design df          =       117

             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
   SamePlace |   .0797438   .012699      .0545941    .1048936
-----------------------------------------------------------
```

いては，ウェイトの有無による推定値の違いはあまり見られない。母集団全体の平均値を推定すると大きな差が生じるのは，同じ程度の都市規模の回答者の間ではウェイトに大きな差がないのに対して，都市規模が異なる回答者間では大きな差が存在するからである。これは都市規模と同じ町内の居住期間の間に相関があるともいえる。逆に言えば，都市規模や地域変数と相関がない変数であれば，ウェイトを用いた推定も用いない推定もほとんど違いは生じない。

6節　まとめ

この章で紹介したsvy：mean，svy：tabulate以外にも，以下のコマンドが基本的な統計量を推定する際に利用する事ができる。それぞれ，比率（%）の推定，比推定および合計を求める。

svy：proportion

svy：ratio

svy：total

また回帰分析系のコマンドもほぼすべて使うことができる。主なものとしては，順に線形回帰，ロジスティックモデル，順序ロジスティックモデル，多項ロジットモデル，およびポワソン回帰である。

svy：regress

svy：logit および svy：logistic

svy：ologit および svy：oprobit

svy：mlogit

svy：poisson,

svy 系統のコマンドは速いペースで update されており，これら以外にも多くの STATA のコマンドが対応している。最初に述べたように歴史的な経緯からも日本語で読める書籍（実践的な本）にはこの分野について解説した本はあまりない。英語で書かれた理論書ではない実践書としては Lehtonen & Pahkinen（2004）や Levy & Lemeshow（1999）がある。読むには初等の統計の知識があったほうがよいであろう。

●引用文献

Deming, W. & Stephan, F. 1940 On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, 11(4), 427-444.

Fischer, C. 1984 *To dwell among friends: personal networks in town and city.* The University of Chicago Press. 松田康・前田尚子（訳） 2003 友人のあいだで暮らす：北カリフォルニアのパーソナル・ネットワーク 未来社

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. 2004 *Survey Methodology.* New York: John Wiley & Sons.

放送世論調査所 サンプリング研究会 1971 サンプリングをめぐる諸問題6 文研月報10, NHK.

Kish, L. 1965 *Survey Sampling.* New York: John Wiley & Sons.

Levy, P. & Lemeshow, S. 1999 *Sampling of Populations.* New York: John Wiley & Sons.

Lehtonen, R. & Pahkinen, E. 1994 *Practical Methods for Design and Analysis of Complex Surveys.* New York: John Wiley & Sons.

Long, J. S. 1997 *Regression models for categorical and limited dependent variables.* Sage Publications.

Neyman, J. 1934 On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97**(4), 558-625.

Royston, P. & Altman D. G. 1994 Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics*, **43**, 429-467.

William, C. 1977 *Sampling techniques.* New York: John Wiley & Sons.

●参考文献

足立浩平 2006 多変量データ解析法―心理・教育・社会系のための入門 ナカニシヤ出版

Romesburg, H. C. 1990 *Cluster analysis for researchers(Reprint ed.).* Malabar, Fla.: Robert E. Krieger Pub. Co. 西田英郎・佐藤嗣二（訳） 1992 実例クラスター分析 内田老鶴圃

齋藤堯幸・宿久洋 2006 関連性データの解析法―多次元尺度構成法とクラスター分析法 共立出版

上田尚一 2003 クラスター分析（講座・情報をよむ統計学） 朝倉書店