統計研究彙報　第 69 号　2012 年 3 月（53～81）

幾何級数モデルによる標本設計の最適化と そのロバストネスに関する考察

木下　千大

A Consideration about Optimization of a Sampling Design via Geometric Series Model and its Robustness

KINOSHITA　Kazuhiro

一般に、社会調査が対象とする母集団は、有限母集団であり、間違いなく正規分布ではない。したがって、標本設計実務においては、理論的に推計精度が高いとされている手法もそのまま適用できる場合は稀である。また、母集団名簿の劣化やフィールドワークにおける非標本誤差を知ると、標本設計にどれほどの精緻さを求めるべきかという戸惑いを持つことも多いのではないだろうか。

本研究は、母集団分布が個々に異なる実データを一つのモデル関数（幾何級数）により生成し、それを基準として様々な標本抽出法の評価を数値計算によって行い、誤差精度のみならず層化変数の変動なども考慮した実務的な最適化を考察した。また、標本設計上の各変数の変化が精度にどのように影響するのか、数値計算の結果をグラフ化して考察することで、実務者にとっての設計上の勘所を示すことができたものと思う。

キーワード：幾何級数、層化抽出、比例配分、ネイマン配分、相関係数、悉皆層の設定

Generally, the population which social research deals with is a finite number, and not a normal distribution by any means. Therefore, the theoretically accurate method is not simply applicable in the actual practice of a sampling design, but only under rare circumstances. Furthermore, considering the deterioration of the population frame list, and the nonsampling errors arising from fieldworks, it is often bewildering how much accuracy should be pursued in a sampling design.

In this research, optimization is considered from a practical point of view, evaluating a range of sampling by numerical computation based on a geometric series model in various simulated data with different population distributions generated by a single functional model of geometric series, taking error and accuracy into account as well as the fluctuation of stratification variables.

Moreover, for the experts in practice, a vital point from the results of the numerical computation has been graphically presented, illustrating how the change of each variable in a sampling design affects accuracy.

Keywords: Geometric series, stratified sampling, proportional allocation, Neyman allocation, correlation coefficient, take-all strata

原稿受理日　平成 24 年 1 月 12 日　†独立行政法人統計センター情報技術部研究主幹室

はじめに

一般に、産業別売上高総額を推計するような標本調査は、大企業を悉皆調査とし、その他の企業を標本調査として設計される。これは大企業の売上高の企業間分散の大きさのため推計誤差が大きすぎることによるものであるが、統計調査を設計する実務家にとっては、しばしばこの悉皆層の区分をどのように設定すればよいのかが問題となる。また、標本層についても同様に、層をいくつに分け、その区分点をどのようにすればよいのか苦慮するところである。多くの場合は、誤差の評価の面から統計表の表章区分に基づき層の設定を行うが、単に売上総額のみ推計する場合には、層の設定そのものの最適化を図るという設計も有り得る。郵送調査のように調査コストが単純に標本数に比例的であるような場合は、標本の最適設計が特に重要性を増してくる。

標本抽出の方法について、一般式により理論が展開されているものは多いが、実務に則した最適設計への理論の適用については、定性的なこと以外、これまでほとんど論じられていない。本稿では、幾何級数モデルによる疑似データを用いて、いくつかの標準的な標本抽出方法について推計誤差量の比較を行い、実務的な標本設計の考え方及びその手法の考察を行った。

ここで本稿の構成について説明する。第 1 章では企業の売上高の分布の疑似データを幾何級数によって生成する試み及び幾何級数の基本的な特性について確認する。第 2 章では層別抽出（標本比例配分）における最適層化の区分点と推計値の誤差の算出及び層別抽出（標本ネイマン配分）における誤差の算出を行い標本抽出方法としての採用上の問題を指摘する。第 3 章では実用的な層別抽出法について比較検討する。第 4 章では標本設計における悉皆層の設定及び層別抽出法との組み合わせについて考察する。第 5 章では層化情報の影響について擬似データにより検証し、設計マージンの考察を行う。

第1章　幾何級数による疑似データの生成

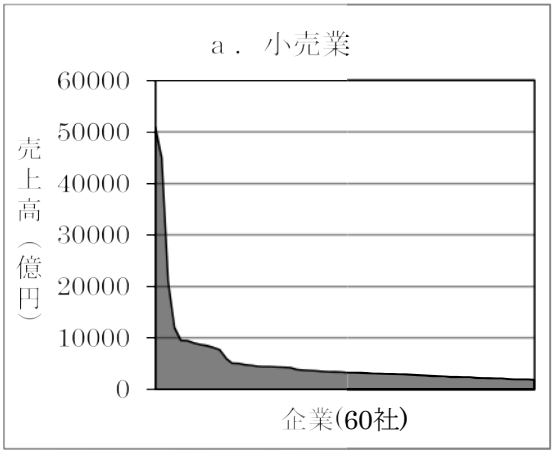
１　企業の年間売上高の分布

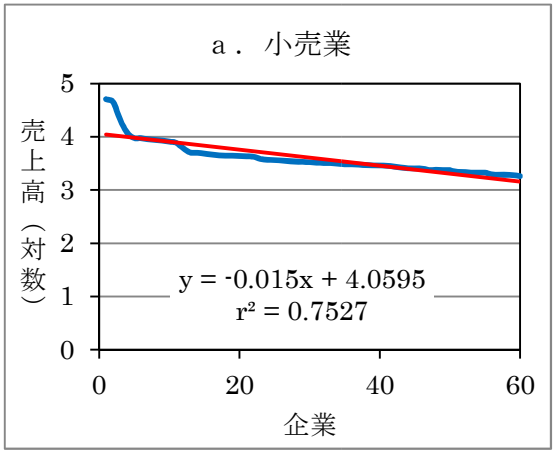
産業別の企業売上高の分布を見ると、その多くが指数分布を示すことが分かる。図 1.1.1-a～c は、小売業、卸売業、サービス業の上位 60 社の年間売上高を売上高の高い順に並べ分布をグラフにしたものである。なお、それぞれ右側のグラフは、売上高を対数変換して Y 軸にとり、売上高順位を X 軸にとって回帰線をあてはめたもので、指数関数のあてはまりのよいことが回帰式及び決定係数によって確認できる。こうした回帰式から各産業の売上高分布特性を企業の売上高順位 X の係数（R）で表現することが可能となる。各企業の売上高実数値をトップ企業の売上高に対する相対値として線形変換し、分布特性を示す R を母数とする指数関数の値として記述したグラフが図 1.1.2 である。

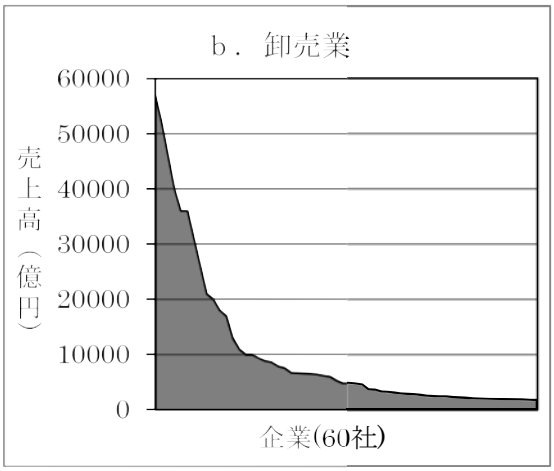
こうした指数関数は連続関数であるが、X の値が整数値のみをとる順位であることから、その関数の値 Y（売上高）は公比 R の幾何級数によって全体を表現することができる。各企業の成長が指数関数的であり、その産業界への参入時期に時間差があるとき、ある 1 時点でとらえた売上高の分布が指数関数によってモデル化されることは、あながち不自然ではないと考える。

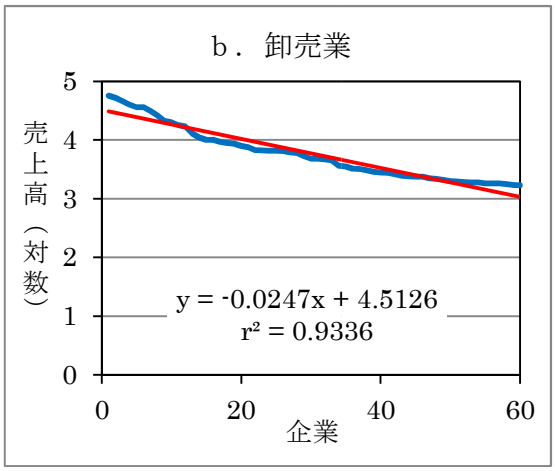
図 1.1.1　産業別企業年間売上高上位 60 社の売上高企業分布

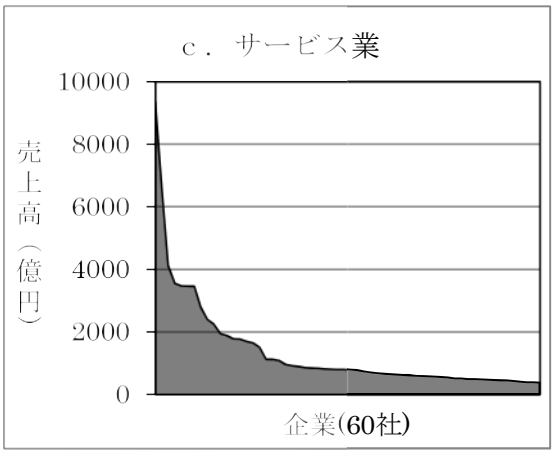
注 1：データは、Web 上に公開されている企業売上高ランキングデータを用いている。このため、産業分類は「日本標準産業分類」とは異なるものである。
注 2：左側のグラフは棒グラフで表すべきものであるが、見やすくするため面グラフとしている。

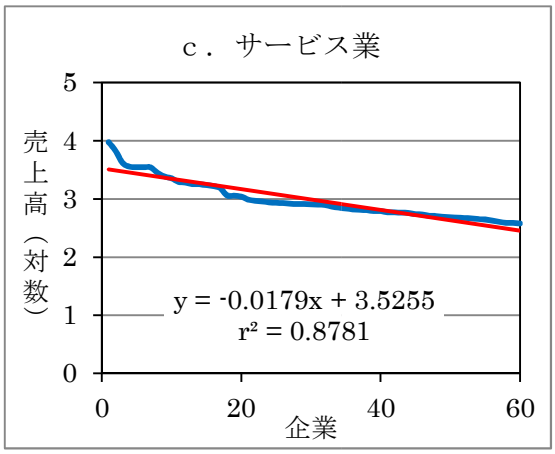






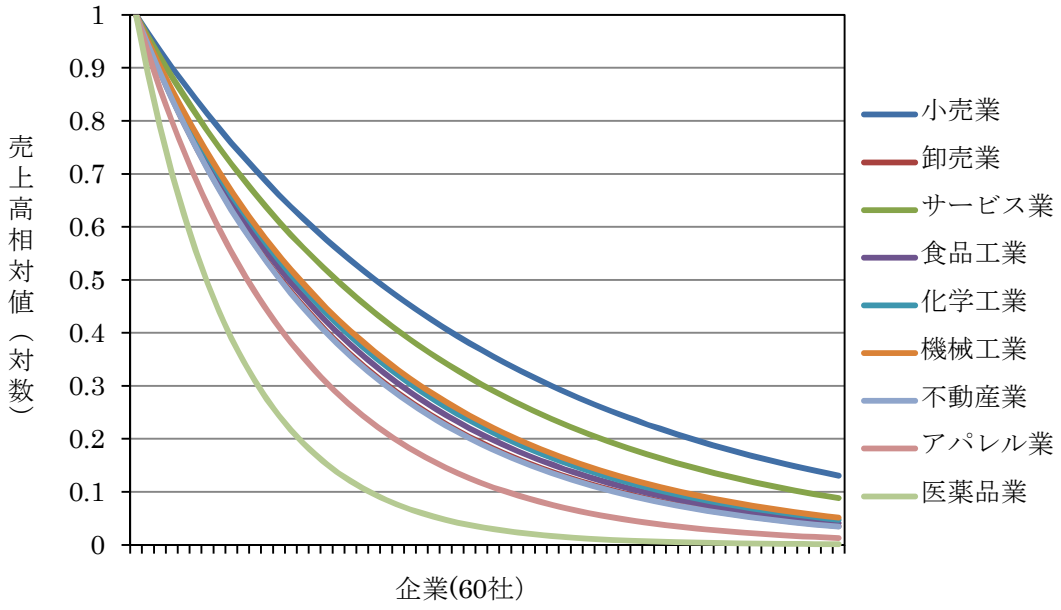






注 3：指数関数特性をとらえやすくするため、特異値であるサービス業の「株式会社　電通」を除いている。

図 1.1.2　産業別企業年間売上高の分布特性



２　幾何級数による母集団分布の表現

前項でみたように、企業の売上高分布はおおむね幾何級数で記述することが可能であることから、幾何級数を母集団分布として標本抽出法を考察することとする。その準備として幾何級数の特徴を整理しておく。

幾何級数は、次式で記述される。

$$R^0, R^1, R^2, \cdots, R^i, \cdots, R^N = 1, R, R^2, \cdots, R^i, \cdots, R^N$$

また、この数列の１番目から、（N＋１）番目までの総和 $S_{0\text{-}N}$ 及び（M＋１）番目から（N＋１）番目までの総和 $S_{M\text{-}N}$ は、

$$S_{0-N} = R^0 + R^1 + R^2 + \cdots + R^N = \sum_{i=0}^{N} R^i = \frac{R^{N+1} - 1}{R - 1}$$

$$S_{M-N} = R^M + R^{M+1} + \cdots + R^N = \sum_{i=M}^{N} R^i = \frac{R^{N+1} - R^M}{R - 1}$$

である。

幾何級数を母集団分布として標本抽出法を考察するメリットは、層区分ごとの個々の値の総和を簡単な式で表現できることであり、現実のデータのばらつきを一つの基準となる関数の上におくことで、層化や抽出法の比較についてデータの特徴を加味した一般性のある評価が可能になる点である。ここで、後に分散の計算式に用いるもう一つの級数 R_i^2 の総和についても示しておく。この数列の１番目から、（N＋１）番目までの総和 $S^2_{0\text{-}N}$ 及び（M＋１）番目から（N＋１）番目までの総和 $S^2_{M\text{-}N}$ は、

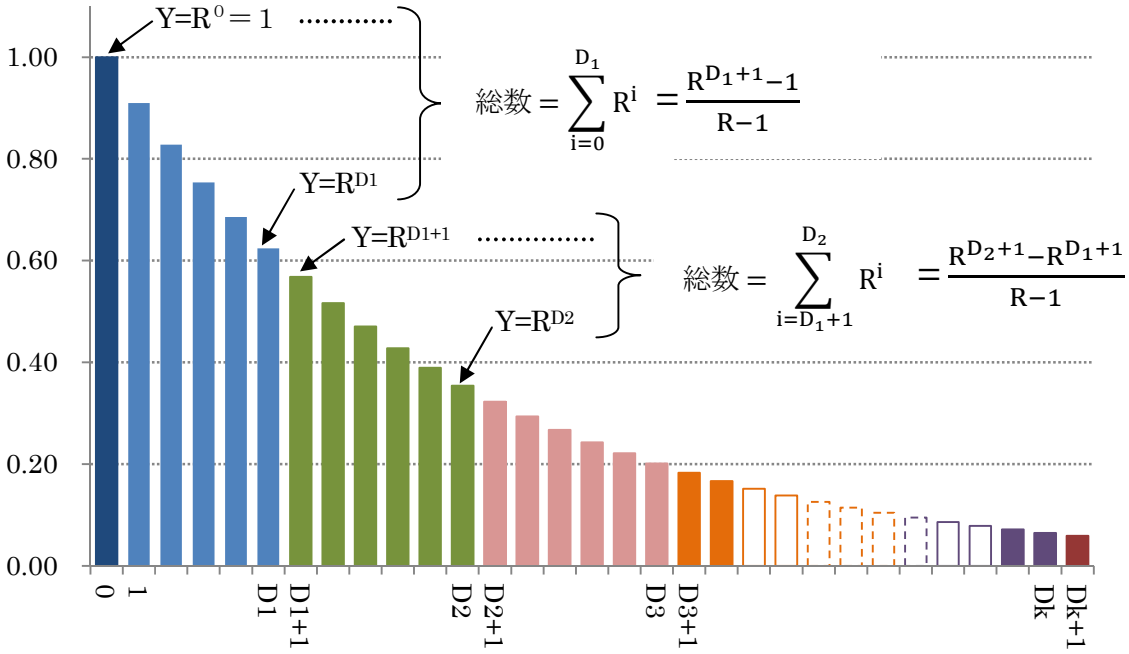
$$S_{0-N}{}^2 = R^0 + R^2 + R^4 + \cdots + R^{2N} = \sum_{i=0}^{N} R^i = \frac{R^{2N+2} - 1}{R^2 - 1}$$

$$S_{M-N}{}^2 = R^{2M} + R^{2M+2} + \cdots + R^{2N} = \sum_{i=M}^{N} (R^i)^2 = \frac{R^{2N+2} - R^{2M}}{R^2 - 1}$$

となる。

次に、この幾何級数母集団分布のグラフと区分パラメータ、区分総和の関係を図 1.2.1 のとおり設定しておくことにする。

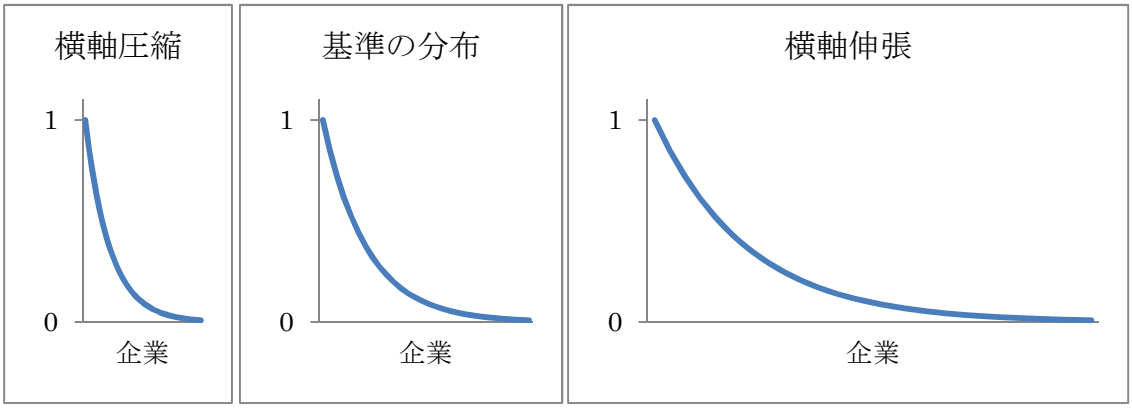
図 1.2.1　幾何級数母集団分布の区分パラメータとその総和



各産業別の企業売上高の分布特性は幾何級数の公比によって記述できることを確認したが、標本抽出法を考察する場合に重要なパラメータとして母集団数がある。幾何級数によっ

て母集団分布を記述するメリットとして、その再帰的特性から公比 R と母集団数 N を一元的に取り扱うことが可能になるということである。すなわち、母集団数 N を横軸 X にとったとき、売上高 Y の値のエンベロープを記述する関数 $Y=R^x$ のパラメータ R は、横軸を圧縮すると小さくなり伸張すると大きくなる（図 1.2.2）ことを踏まえ、基準とするモデル（本稿では後に N=1000 とおく。）による考察が有用である。

図 1.2.2　母集団数 N の線形変換と公比 R の関係



3　層化抽出

3－1　標本比例配分のための最適層区分

層化抽出法は、単純任意抽出に比べ母集団の推計精度が高いことが一般に知られているが、その層区分はいかなるときに最も精度が高くなるのであろうか。また、その区分数はいくつにするのが良いのか。標本を層の大きさに比例配分することを前提に最適な層区分について考えてみる。

ここで、最適な層区分とは次の二つの条件を満たすものとして考えることにする。

① 総売上高の推計誤差が小さい。

② 実際の標本配分に際し、その変動に対し推計結果が安定的である。

総売上高の推計誤差 ε は次式で表される。

$$\varepsilon = \sqrt{\sum_{i=1}^{L} \frac{N_i - n_i}{N_i - 1} \frac{(N_i \sigma_i)^2}{n_i}}$$

$$\sigma_i{}^2 = \frac{1}{N}\sum_{j=1}^{N} R^j{}_j{}^2 - \left(\frac{1}{N}\sum_{J=1}^{N} R^j{}_j\right)^2$$

（ただし、i=1,2,…,L の層区分、j=1,2, …,N の層内のデータ数）

$$\varepsilon = \sqrt{\frac{N_1 - n_1}{N_1 - 1}\frac{N_1}{n_1}N_1{\sigma_1}^2 + \frac{N_2 - n_2}{N_2 - 1}\frac{N_2}{n_2}N_2{\sigma_2}^2 + \cdots}$$

標本を母集団比例で配分する場合は、各層の抽出率は一定であるから、上式は G を定数として次のようになる。

$$\varepsilon = \sqrt{G\left(\frac{N_1}{N_1 - 1}N_1{\sigma_1}^2 + \frac{N_2}{N_2 - 1}N_2{\sigma_2}^2 + \cdots\right)} = \sqrt{G \cdot Z}$$

なお、図 1.2.1 のパラメータ記述で総売上高の推計誤差 ε を記述すれば

$${\sigma_1}^2 = \frac{1}{D_1 + 1}\sum_{j=0}^{D_1} R^{j2} - \left(\frac{1}{D_1 + 1}\sum_{j=0}^{D_1} R^j\right)^2$$

$${\sigma_2}^2 = \frac{1}{D_2 - D_1}\sum_{j=D_1+1}^{D_2} R^{j2} - \left(\frac{1}{D_2 - D_1}\sum_{j=D_1+1}^{D_2} R^j\right)^2$$

誤差に着目した最適な層区分は、Z の値を最小にする D_1～D_k を求めることに帰着する。しかし、母集団数 N、公比 R、区分数 k、区分点 D_1～D_k が数式上はすべて変数となってしまう（実際の場面では、母集団数 N、公比 R は決定している。）ので、いくつかの変数を固定し、他の変数を変えながら計算してみると、次に示すような結果を得ることができる。

３－２　N=1000、区分数 2～10、公比 0.95～0.99 の最適層区分

表を見ると、区分数を多くしていくにつれて売上高 Y の分散が大きい分布の左側が細分化され、分散の小さい右側の層の分割は起きないことが分かる。詳しく見ると、区分数を一つ増やした時、第１層の大きさが第２層へ、第２層の大きさが第３層へと次々にシフトしていく。すなわち、区分数を増やせば増やすほど第１層が小さくなり、やがて大きさ２の限界（計算上分母が 0 にならない最小の整数値）に達し、その後は第２層、第３層と順次大きさ２の層に分割されていく。公比 R の違いは、最初の２分割の第１層の大きさを決めることになり、R の値が小さいほど分布は急峻に減衰することから第１層が小さく、R の値が１に近いほど

減衰がなだらかで第 1 層が大きくなる。これは、R の値が 1 に近い分布ほど細分化が可能であることを示す。

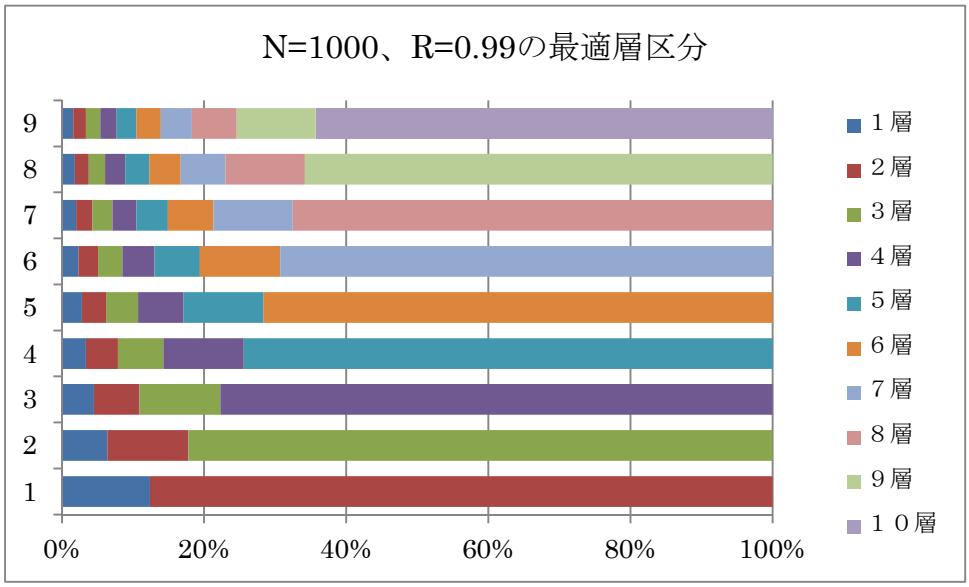
しかし、逆を言えば同じ区分数では、R の値が小さいほど層化の効果は大きく、R の値が 1 に近いほど層化の効果は小さいということでもある。

例えば、N=1000、R=0.95 の母集団を 4 層に最適分割したときの Z の値は 0.434、R=0.98 では 1.016、R=0.99 では 1.922 である。逆に Z の値を 0.434 以下にするには R=0.98 の場合は 7 層、R=0.99 の場合は 9 層以上に分割する必要がある。

表 1.3.1　R の値の違いによる最適な層区分

区分数別の最適な各層の大きさ											
R	区分数	1層	2層	3層	4層	5層	6層	7層	8層	9層	10層
0.95	2	24	976								
	3	13	24	963							
	4	9	13	24	954						
	5	7	9	13	24	947					
	6	5	7	9	13	24	942				
	7	5	5	7	9	13	24	937			
	8	4	5	5	7	9	13	24	933		
	9	3	4	5	5	7	9	13	24	930	
	10	3	3	4	5	5	7	9	13	24	927
0.98	2	62	938								
	3	33	60	907							
	4	23	33	60	884						
	5	17	23	33	60	867					
	6	14	17	23	33	60	853				
	7	12	14	17	23	33	60	841			
	8	10	12	14	17	23	33	60	831		
	9	9	10	12	14	17	23	33	60	822	
	10	8	9	10	12	14	17	23	33	60	814
0.99	2	124	876								
	3	64	114	822							
	4	45	64	114	777						
	5	34	45	64	113	744					
	6	28	34	45	64	113	716				
	7	23	28	34	45	64	113	693			
	8	20	23	28	34	44	64	112	675		
	9	18	20	23	28	34	44	63	112	658	
	10	16	18	20	23	28	34	44	63	111	643

図 1.3.1　N=1000、R=0.99 の最適層区分の様子



３－２　区分数 2～10、N=500、1000、5000 の最適層区分と公比

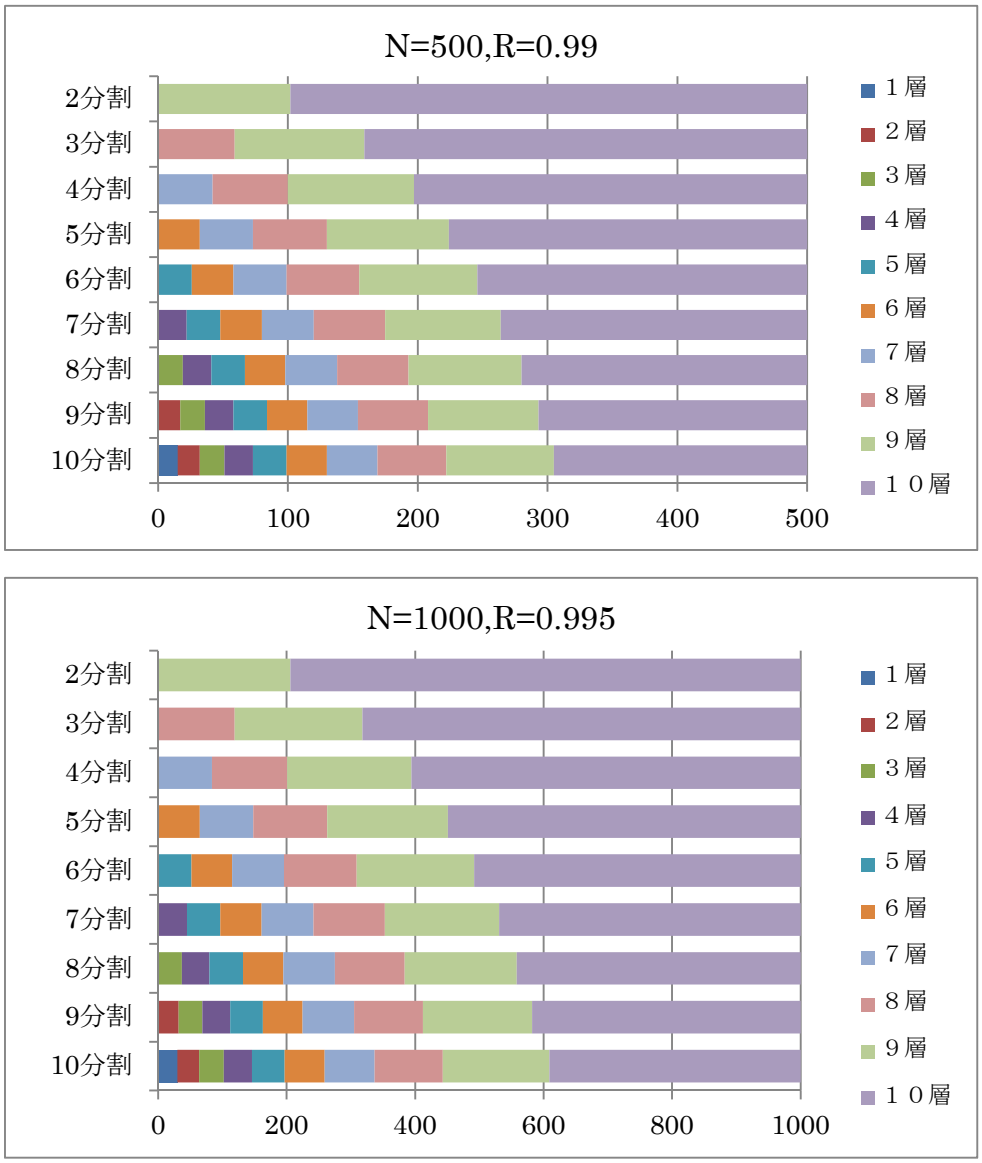
ここで、先に図 1.2.2 で示した母集団数 N と公比 R の関係を層化の際の最適区分で見てみる。

図 1.3.2 は、〔N=500、R=0.99〕と〔N=1000、R=0.995〕と〔N=5000、R=0.999〕の区分数別の最適層区分をグラフにしたものである。明らかに同じ構造を持っていることが分かり、この種の考察は N と R の組合せについて行い、実際の適用の際に N なり R なりを変換すればよいことが分かる。

また、これらのグラフをみると、4 分割は 2 分割の第 1 層及び第 2 層をそれぞれ 2 分割し、6 分割は 2 分割の第 1 層及び第 2 層をそれぞれ 3 分割、8 分割は 4 分割の第 1 層、第 2 層、第 3 層、第 4 層を各々2 分割している再帰的構造をしていることが分かる。同様に 6 分割は、3 分割の第 1 層、第 2 層、第 3 層を各々2 分割しており、2 分割点、3 分割点などの分割比が分かればその公倍数のおおよその分割点を求めることが可能である。厳密には有限母集団であるがゆえに分布の右裾の層が少しずつ小さくなっているが、それは誤差精度を向上させる方向に歪んでいるため、最適分割の点では気にする必要はない。このように 2 層分割が基本となることから、表 1.3.2 及び図 1.3.3 に N=1000 の場合の R の値別最適区分を示す。

したがって、今後は N=1000、R=0.980～0.999 の組合せによる分析を中心に行うことにする。N=1000 は我が国の産業別企業数に近く、R=0.980～0.999 も N=1000 規模の実際の売上高分布の線形変換値に近いためである。

図 1.3.2　同一の最適層区分をもつ母集団分布



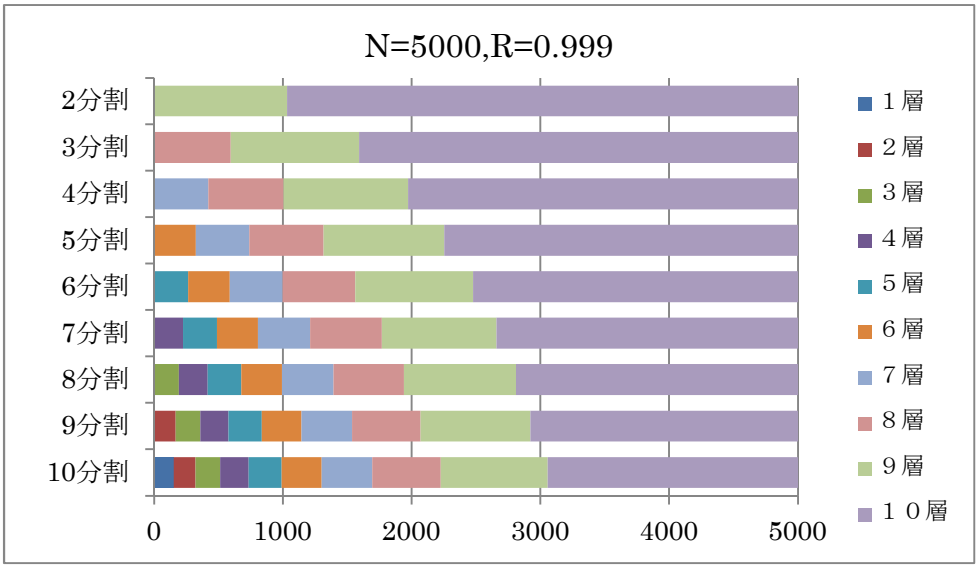
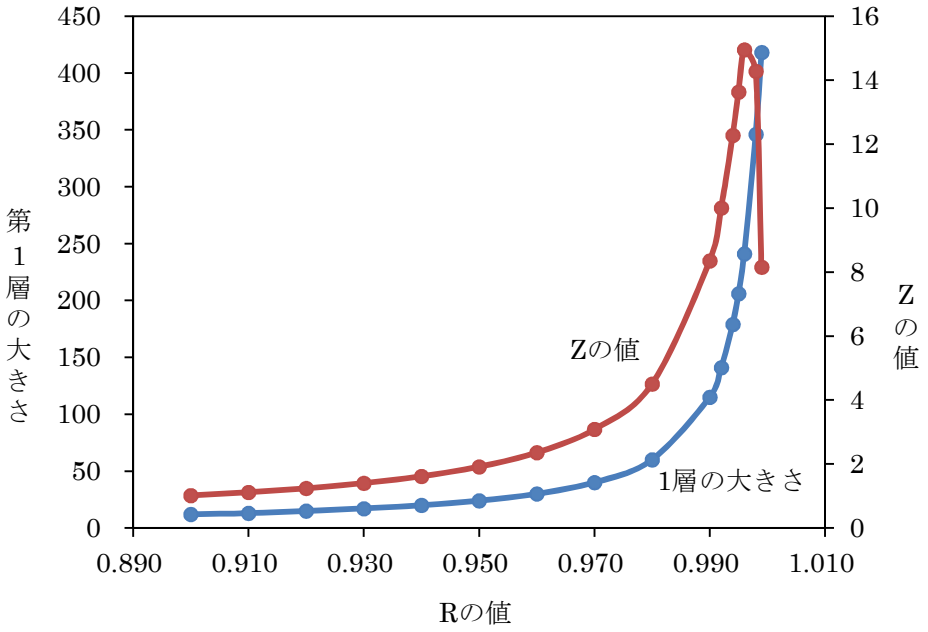


表 1.3.2　R の値別第 1 層の最適数と Z の値

R の値	1 層	Z の値	R の値	1 層	Z の値
0.90	12	1.01483	0.98	60	4.50118
0.91	13	1.11420	0.99	115	8.35019
0.92	15	1.23991	0.992	141	10.0019
0.93	17	1.40042	0.994	179	12.2710
0.94	20	1.61435	0.995	206	13.6277
0.95	24	1.91234	0.996	241	14.9453
0.96	30	2.35590	0.998	346	14.2726
0.97	40	3.08496	0.999	418	8.15274

図 1.3.3　R の値別第 1 層の最適数と Z の値



第2章　一般的層別標本抽出法の採用上の問題

1　標本配分と推計誤差

層別抽出法の精度は母集団の層化の適否のみでは決まらない。各層の抽出率が一定である標本比例配分の場合は層の設定により推計精度が決まるが、各層への標本配分のしかたによっては精度が向上する場合もあれば低下する場合もある。すなわち、売上高の分散が小さい層への標本配分を抑え、その分を分散の大きい層に配分しようという考え方であり、ネイマン配分として知られている。ネイマン配分法による各層へ配分される標本数 n_i は次式によって行われる。

$$n_i = \frac{N_i \sigma_i \sqrt{\frac{N_i}{N_i - 1}}}{\sum_{i=1}^{L} N_i \, \sigma_i \sqrt{\frac{N_i}{N_i - 1}}} n$$

ここで、層別抽出法について標本を層の規模に比例配分した場合とネイマン配分した場合の推計誤差を比較してみる。条件は、母集団数 N=1000、標本数 n=100、層の区分数 L=10 で比例配分は最適層化、ネイマン配分は等分層化である。

表 2.1.1　最適な母集団区分

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層
R=0.980	8	9	10	12	14	17	23	33	60	814
R=0.981	8	9	11	12	15	18	24	34	63	806
R=0.982	9	10	11	13	16	19	25	36	66	795
R=0.983	9	10	12	14	16	20	27	38	69	785
R=0.984	10	11	13	15	17	22	28	41	74	769
R=0.985	11	12	14	16	19	23	30	43	78	754
R=0.986	11	13	14	17	20	25	32	46	83	739
R=0.987	12	14	16	18	21	26	35	50	89	719
R=0.988	13	15	17	20	23	29	37	54	95	697
R=0.989	14	16	18	21	25	31	41	58	103	673
R=0.990	16	18	20	23	28	34	44	63	111	643
R=0.991	18	20	22	26	31	38	49	70	121	605
R=0.992	20	22	25	29	34	42	54	77	132	565
R=0.993	22	25	28	33	39	47	61	86	144	515
R=0.994	26	29	33	38	44	54	69	95	156	456
R=0.995	30	34	38	44	51	62	78	106	166	391
R=0.996	37	41	45	52	60	72	89	117	170	317

表 2.1.2　最適な母集団区分に対する標本比例配分

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層	誤差
R=0.980	1	1	1	1	1	2	2	3	6	81	1.22
R=0.981	1	1	1	1	2	2	2	3	6	81	1.22
R=0.982	1	1	1	1	2	2	3	4	7	80	1.23
R=0.983	1	1	1	1	2	2	3	4	7	79	1.30
R=0.984	1	1	1	2	2	2	3	4	7	77	1.33
R=0.985	1	1	1	2	2	2	3	4	8	75	1.41
R=0.986	1	1	1	2	2	3	3	5	8	74	1.44
R=0.987	1	1	2	2	2	3	4	5	9	72	1.46
R=0.988	1	2	2	2	2	3	4	5	10	70	1.50
R=0.989	1	2	2	2	3	3	4	6	10	67	1.57
R=0.990	2	2	2	2	3	3	4	6	11	64	1.64
R=0.991	2	2	2	3	3	4	5	7	12	61	1.70
R=0.992	2	2	3	3	3	4	5	8	13	57	1.82
R=0.993	2	3	3	3	4	5	6	9	14	52	1.90
R=0.994	3	3	3	4	4	5	7	10	16	46	2.03
R=0.995	3	3	4	4	5	6	8	11	17	39	2.18
R=0.996	4	4	5	5	6	7	9	12	17	32	2.27
R=0.997	5	5	6	6	7	8	10	12	17	25	2.34
R=0.998	6	6	7	7	8	9	11	12	15	19	2.27
R=0.999	8	8	8	9	9	10	11	12	12	14	1.71

表 2.1.3　均等な母集団区分に対する標本ネイマン配分

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層
R=0.980	87	12	2	0	0	0	0	0	0	0
R=0.981	85	13	2	0	0	0	0	0	0	0
R=0.982	84	14	2	0	0	0	0	0	0	0
R=0.983	82	15	3	0	0	0	0	0	0	0
R=0.984	80	16	3	1	0	0	0	0	0	0
R=0.985	78	17	4	1	0	0	0	0	0	0
R=0.986	76	18	5	1	0	0	0	0	0	0
R=0.987	73	20	5	1	0	0	0	0	0	0
R=0.988	70	21	6	2	1	0	0	0	0	0
R=0.989	67	22	7	2	1	0	0	0	0	0
R=0.990	63	23	8	3	1	0	0	0	0	0
R=0.991	60	24	10	4	2	1	0	0	0	0
R=0.992	55	25	11	5	2	1	0	0	0	0
R=0.993	51	25	12	6	3	2	1	0	0	0
R=0.994	45	25	14	7	4	2	1	1	0	0
R=0.995	40	24	15	9	5	3	2	1	1	0
R=0.996	34	23	15	10	7	5	3	2	1	1
R=0.997	27	20	15	11	8	6	5	3	2	2
R=0.998	21	17	14	12	9	8	6	5	4	3
R=0.999	15	14	12	11	10	9	8	7	7	6

表 2.1.4　ネイマン配分の誤差

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層	誤差
R=0.980	0.9056	0.7816	0.0919	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.45
R=0.981	1.0398	0.8504	0.1343	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.59
R=0.982	1.0863	0.9264	0.1954	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.72
R=0.983	1.2057	1.0088	0.1866	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.89
R=0.984	1.3149	1.0968	0.2683	0.0326	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.75
R=0.985	1.4111	1.1888	0.2844	0.0571	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.87
R=0.986	1.4912	1.2825	0.3189	0.0991	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.04
R=0.987	1.6337	1.2901	0.4474	0.1702	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.29
R=0.988	1.7518	1.3749	0.5119	0.1432	0.0259	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.15
R=0.989	1.8396	1.4495	0.5946	0.2400	0.0531	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.40
R=0.990	1.9738	1.5074	0.6938	0.2613	0.1072	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.78
R=0.991	1.9783	1.5407	0.7179	0.3139	0.1051	0.0348	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.64
R=0.992	2.0937	1.5400	0.8332	0.3925	0.2030	0.0823	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	3.24
R=0.993	2.0570	1.5760	0.9454	0.4956	0.2510	0.0933	0.0463	#DIV/0!	#DIV/0!	#DIV/0!	3.26
R=0.994	2.1034	1.5494	0.9521	0.6180	0.3350	0.2053	0.1245	0.0374	#DIV/0!	#DIV/0!	3.38
R=0.995	1.9638	1.5213	0.9990	0.6541	0.4510	0.2817	0.1566	0.1161	0.0426	#DIV/0!	3.35
R=0.996	1.7840	1.3802	1.0481	0.7467	0.4945	0.3173	0.2422	0.1647	0.1492	0.0670	2.53
R=0.997	1.5349	1.2451	0.9672	0.7572	0.5901	0.4408	0.2931	0.2735	0.2273	0.1246	2.54
R=0.998	1.0437	0.9076	0.7651	0.6120	0.5654	0.4309	0.3933	0.3196	0.2705	0.2442	2.36
R=0.999	0.4327	0.3840	0.3753	0.3390	0.3087	0.2839	0.2643	0.2500	0.2047	0.1976	1.74

注）標本が配分されない層については、その部分の差を誤差に加えている（赤字の部分）。

結果をみると、幾何級数モデルでは両者の差はあまりない。むしろ両者とも標本配分に関する実務上の問題が露呈している。すなわち、比例配分では売上高の高い企業層で配分する標本数が少なくなり、層の区分数 L を増やした場合や総標本数 n が少ない場合に標本が配分されない層が生じ、同様にネイマン配分では売上高の低い企業層で標本が配分されない層が生じてしまっている。それを回避しようとするならば、それはもはや「比例配分」や「ネイマン配分」とは言えないものである。

これらの結果を見ると、産業別総売上高を推計するための企業を対象とする標本調査の設計は、理論書に示されている一般的な標本抽出法を単純にそのまま用いることは適切ではないということが分かる。

２　標本配分数が少ない層の問題

標本配分数が少ない層が抱える実務上の問題について考察する。

まず、比例配分であるが、この条件設定では、表の左上の方で標本数１又は２という層が存在し、調査時点における名簿の劣化等により配分標本数のデータが取集されない恐れがあり、設定精度の確保が難しい。最適な標本設計とは、理論的な精度を高めること以上に、こうした実査上の誤差に対しても堅牢なものでなければならないと考える。

一方のネイマン配分は、R=0.980～0.995 にかけて標本が配分されない層が生じている。誤差の比較でも比例配分よりも劣る結果となっており、一般的に比例配分よりも優れていると思われているネイマン配分を現実の母集団分布に対して適用するのは得策とは思えない。

標本が配分されない層は母集団を推計していないわけであるから、誤差の算出さえもできないことになり実用には耐えない。表 2.1.4 に掲げた誤差は、推計できていない部分を加えているが、このバイアス性の誤差は母集団値が不明である実際の場面では算出が不可能である。また、少ない標本が配分された層に関しては、比例配分と異なり各層の抽出率の差、すなわち復元乗率の差が大きいことから、特異値に対して不安定な推計となるという問題を抱えている。したがって、ネイマン配分を計算通り適用した標本設計も堅牢性に欠けるものであると言える。

第3章　実用的な層別標本抽出法

前章で明らかなように、一般的な層別標本抽出法は、理論的に高い精度を求めるために層内の分散値をそのまま設計値の算出に用いており、幾何級数モデルで疑似できるような企業の売上高の母集団分布に関しては、分散値の分布の傾斜が急過ぎて使いにくいことが分かった。そこで、誤差精度的に最適な設計を目指すのではなく、標本配分誤差に対して堅牢性を持たせる二つの抽出方法を検討してみる。

１　最適層区分の母集団数の平方根に比例させる標本配分

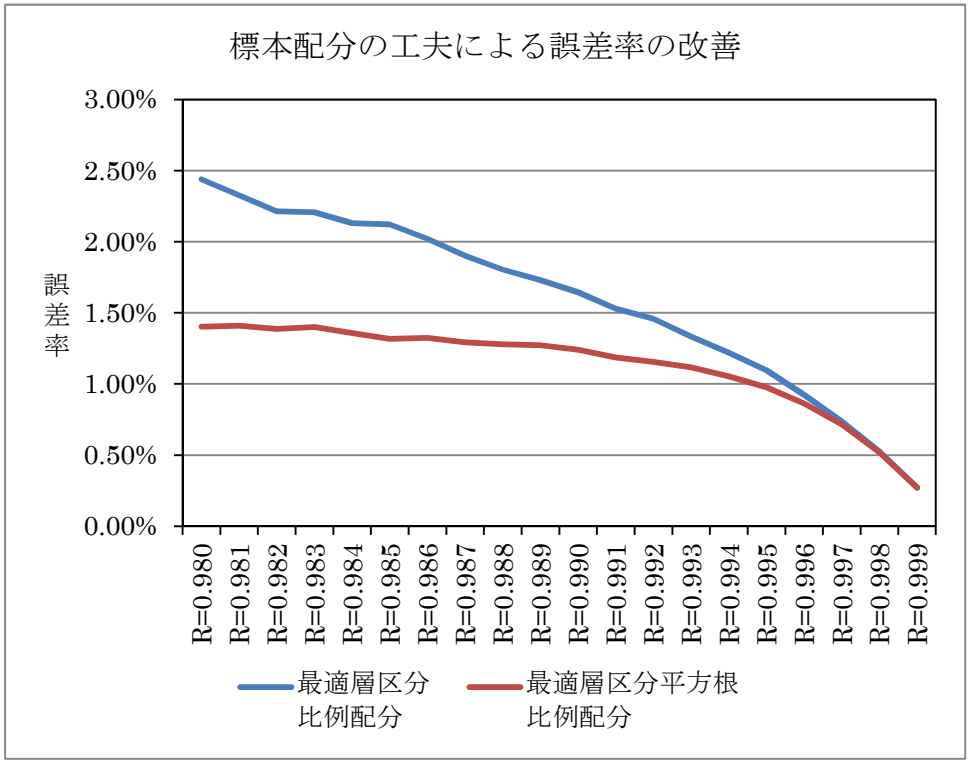
最適層区分による標本配分設計値は配分傾斜が大きいことから、配分比を各層の母集団数ではなく母集団数の平方根に比例するように配分してみる（母集団の層区分は変えていない。）。その結果は、R=0.980～0.984 の第 1 層及び R=0.980、0.981 の第 2 層の 4 と配分数が多くなり、実務上はこのあたりが最小配分数の限界である。これ以上、配分される標本数が少なくなると標本数変動に耐えられなくなる。

この方法は、各層の抽出率は一定ではなくなるが、もともと一定である必要もないことから配分数の少ない層をつくらないという考え方を優先したものである。結果的には分散の大きな層への標本配分比が高くなり、ネイマン配分の考え方を少し取り入れたものとなるため、公比 R の値の小さい分布において推計精度の向上が認められる。

表 3.1.1　最適な母集団区分の平方根に比例する標本配分

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層	誤差
R=0.980	4	4	5	5	6	6	7	9	12	42	0.70
R=0.981	4	4	5	5	6	6	7	9	12	42	0.74
R=0.982	4	5	5	5	6	6	7	9	12	41	0.77
R=0.983	4	5	5	5	6	6	7	9	12	40	0.82
R=0.984	4	5	5	5	6	7	7	9	12	39	0.85
R=0.985	5	5	5	6	6	7	8	9	12	38	0.88
R=0.986	5	5	5	6	6	7	8	9	12	37	0.94
R=0.987	5	5	5	6	6	7	8	10	13	36	0.99
R=0.988	5	5	5	6	6	7	8	10	13	35	1.06
R=0.989	5	5	5	6	6	7	8	10	13	34	1.16
R=0.990	5	5	6	6	7	7	8	10	13	32	1.24
R=0.991	5	6	6	6	7	8	9	10	14	30	1.32
R=0.992	5	6	6	6	7	8	9	11	14	29	1.44
R=0.993	5	6	6	7	7	8	9	11	14	27	1.59
R=0.994	6	6	7	7	8	8	9	11	14	24	1.75
R=0.995	6	6	7	7	8	9	10	11	14	22	1.94
R=0.996	6	7	7	8	8	9	10	12	14	19	2.12
R=0.997	7	7	8	8	9	9	10	12	13	16	2.27
R=0.998	8	8	8	9	9	10	10	11	12	14	2.23
R=0.999	9	9	9	9	10	10	10	11	11	12	1.71

図 3.1.1　母集団規模の平方根比例配分による誤差率の改善

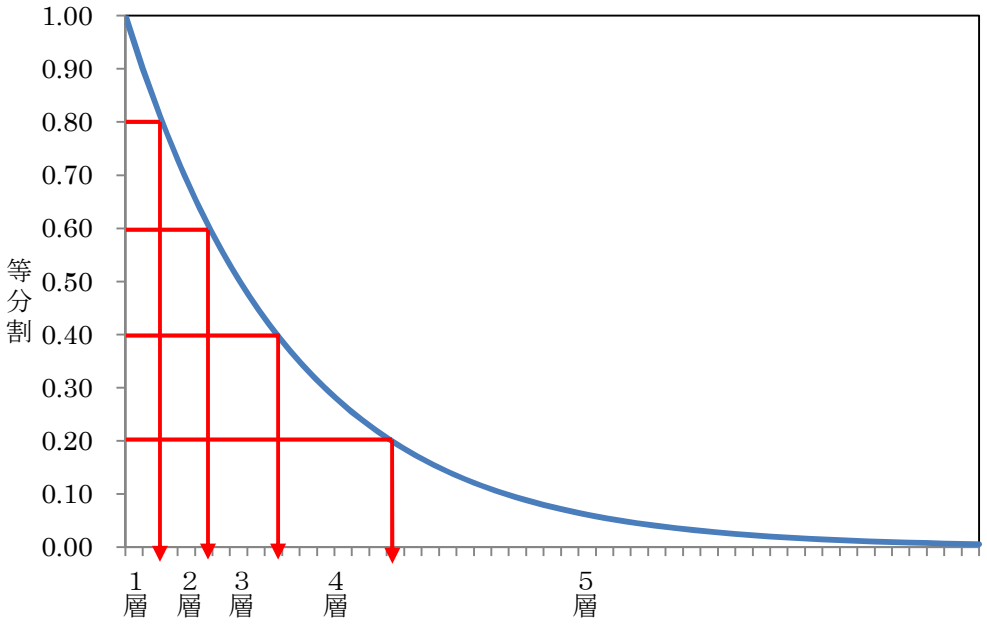


２　簡易層区分の母集団数の平方根に比例させる標本配分

抽出率一定の条件で誤差を最小にする層の分割の計算は容易ではない、そこで売上高の最大値と最小値の差を等分して層を構成する簡易な方法がどの程度実用になるのかを、幾何級数モデルで試算し比較してみる。

簡易層化とは、図 3.2.1 のように縦軸を分割数に均等に分けたときに、これに対応する横軸（母集団数）上の点を層の区分点とする方法である。この方法では、分割された各層の Y の値の分散がほぼ均等となる。

図 3.2.1　簡易層化の方法（５層への分割の例）



簡易層化による母集団の分割結果は、表 3.2.1 のとおりであり、最適な層区分にかなり近いものとなっている。この層区分を前提に、標本の配分比をこの簡易な層区分の平方根に比例させたときの誤差率は表 3.3.3 のとおりである。

表 3.2.1　簡易層化による各層の母集団数

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層
R=0.980	5	6	7	7	9	11	15	20	34	885
R=0.981	5	7	7	8	9	12	15	21	36	879
R=0.982	6	6	8	8	10	12	16	23	38	872
R=0.983	6	7	8	9	10	13	17	24	40	865
R=0.984	7	7	8	10	11	14	18	25	43	856
R=0.985	7	8	9	10	12	15	19	26	46	847
R=0.986	7	9	9	11	13	16	20	29	49	836
R=0.987	8	9	10	12	14	17	22	31	53	823
R=0.988	9	9	12	12	15	19	24	33	58	808
R=0.989	10	10	12	14	17	20	26	37	62	791
R=0.990	10	12	13	16	18	22	29	40	69	770
R=0.991	12	13	14	17	21	24	32	45	77	744
R=0.992	13	15	16	20	22	28	36	50	86	713
R=0.993	15	17	19	22	26	31	41	58	98	672
R=0.994	17	20	22	26	30	37	47	67	113	620
R=0.995	21	23	27	30	36	44	56	79	132	551
R=0.996	26	29	32	37	44	54	68	94	153	462
R=0.997	33	37	42	47	56	66	83	111	168	356
R=0.998	45	50	55	62	71	82	99	124	164	247
R=0.999	65	70	75	81	89	96	108	120	137	158

表 3.2.2　最適層化と簡易層化の各層の母集団数の差

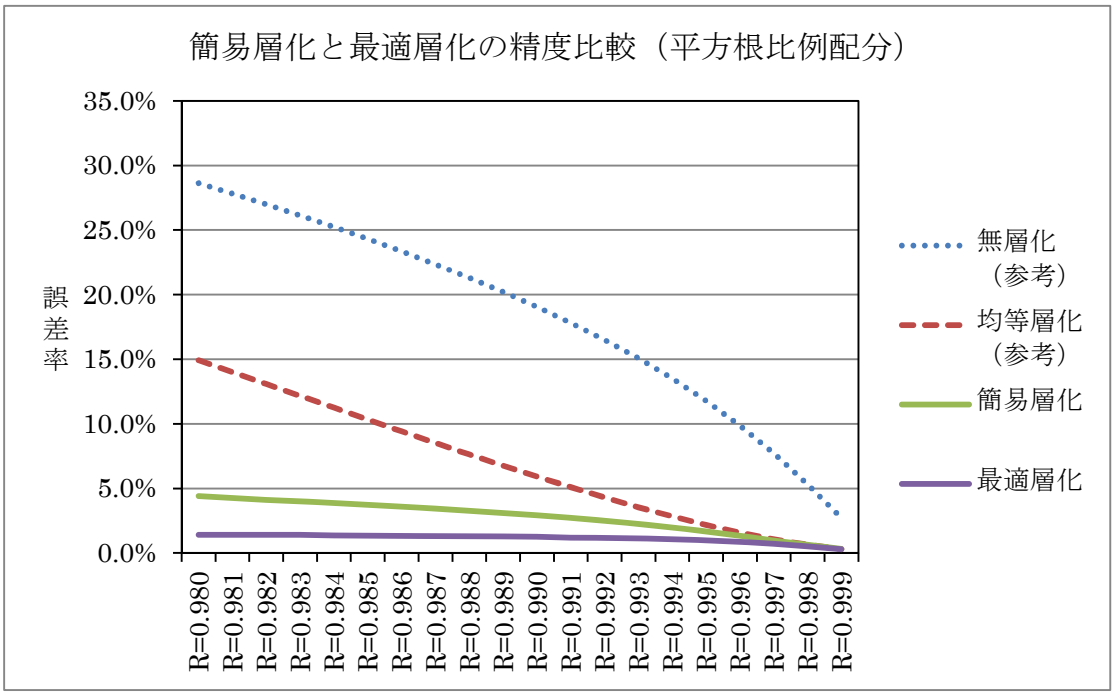
	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層
R=0.980	−3	−3	−3	−5	−5	−6	−8	−13	−26	71
R=0.981	−3	−2	−4	−4	−6	−6	−9	−13	−27	73
R=0.982	−3	−4	−3	−5	−6	−7	−9	−13	−28	77
R=0.983	−3	−3	−4	−5	−6	−7	−10	−14	−29	80
R=0.984	−3	−4	−5	−5	−6	−8	−10	−16	−31	87
R=0.985	−4	−4	−5	−6	−7	−8	−11	−17	−32	93
R=0.986	−4	−4	−5	−6	−7	−9	−12	−17	−34	97
R=0.987	−4	−5	−6	−6	−7	−9	−13	−19	−36	104
R=0.988	−4	−6	−5	−8	−8	−10	−13	−21	−37	111
R=0.989	−4	−6	−6	−7	−8	−11	−15	−21	−41	118
R=0.990	−6	−6	−7	−7	−10	−12	−15	−23	−42	127
R=0.991	−6	−7	−8	−9	−10	−14	−17	−25	−44	139
R=0.992	−7	−7	−9	−9	−12	−14	−18	−27	−46	148
R=0.993	−7	−8	−9	−11	−13	−16	−20	−28	−46	157
R=0.994	−9	−9	−11	−12	−14	−17	−22	−28	−43	164
R=0.995	−9	−11	−11	−14	−15	−18	−22	−27	−34	160
R=0.996	−11	−12	−13	−15	−16	−18	−21	−23	−17	145
R=0.997	−12	−13	−13	−15	−15	−17	−16	−13	3	110
R=0.998	−12	−12	−13	−12	−12	−11	−7	0	16	62
R=0.999	−10	−9	−8	−7	−4	−4	1	5	13	22

表 3.3.3　簡易層化した母集団数の平方根に比例した標本配分

	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層	誤差率
R=0.980	4	4	4	4	5	6	6	7	10	49	4.38%
R=0.981	4	4	4	5	5	6	6	8	10	49	4.23%
R=0.982	4	4	5	5	5	6	6	8	10	48	4.10%
R=0.983	4	4	5	5	5	6	7	8	10	47	4.01%
R=0.984	4	4	4	5	5	6	7	8	10	46	3.85%
R=0.985	4	4	5	5	5	6	7	8	11	45	3.74%
R=0.986	4	5	5	5	6	6	7	8	11	44	3.59%
R=0.987	4	5	5	5	6	6	7	8	11	43	3.42%
R=0.988	4	4	5	5	6	6	7	8	11	42	3.25%
R=0.989	5	5	5	5	6	6	7	9	11	41	3.08%
R=0.990	4	5	5	6	6	7	8	9	12	39	2.91%
R=0.991	5	5	5	6	6	7	8	9	12	37	2.71%
R=0.992	5	5	5	6	6	7	8	9	12	36	2.49%
R=0.993	5	5	6	6	7	7	8	10	13	33	2.25%
R=0.994	5	6	6	6	7	8	9	10	13	31	1.96%
R=0.995	5	6	6	7	7	8	9	11	14	28	1.64%
R=0.996	6	6	6	7	8	8	9	11	14	24	1.31%
R=0.997	6	7	7	7	8	9	10	11	14	20	0.96%
R=0.998	7	7	8	8	9	9	10	12	13	16	0.62%
R=0.999	8	8	9	9	10	10	11	11	12	13	0.32%

簡易層化では、最適層化に比べ第 1 層から第 9 層までは小さくなり第 10 層が大きくなる。一方、標本数の方は、第 10 層の標本数が相対的に少なくなるため、この層の誤差が大きくなり、公比 R の値が小さい分布で全体の精度が低下する結果となっている。

図 3.2.2　簡易層化と最適層化の精度比較



第4章　ロバストな標本設計

これまでの考察では、分布の左側での層化が層の分散に依存するため、その分散値の精度が全体の推計精度に大きく影響を与える。また、実際の場面では、この層化に用いる情報は推計する値と相関の高い他の情報を用いることになる。例えば、売上高の推計であれば従業者数などが層化の情報として用いられる。その相関係数は一般的には 0.7～0.8 程度であり、こうしたことを踏まえると、分布の左側（売上高の高い企業層）を標本によって推計するというのはいかにも不安定である。やはり、実務的に用いられている悉皆層の設定が不可欠である。

しかしながら、企業調査において、どこまでの規模を悉皆層にし、どこから標本層にすればよいかという実務上の問題に対する検討は十分ではない。標本の大半を悉皆層に割り当てる方が全体の精度は高くなるが、分布全体を正しく推計しようとするならば分布の裾の部分にも標本を割り当てる必要があることは分かる。したがって、悉皆層と標本層の最適区分はどこかに存在し、その位置は母集団の分布形によって異なるということになる。

1　最適な悉皆層の大きさ

幾何級数モデルによる悉皆層の最適な区分点は、母集団数よりも標本数に依存する。

今、母集団数を N=1000 に固定し、公比 R の値別に標本数 n を 10 から 100 まで変化したとき、誤差が最小となる区分点を求めてみる。一定の標本数 n のもとで悉皆層の区分点を変えるということは、残りの標本数で（母集団数－悉皆数）を推計することになり、この誤差が全体の誤差ということになる。

公比 R の値の違いによる悉皆区分点の変化と誤差の変化の様子を R=0.95 と R=0.98 の場合で比較すると、図のようになる。両方とも標本数が少ない場合には、悉皆数が少ないところで誤差の最小点があり、悉皆数を増やすとむしろ誤差が大きくなる。すなわち標本数が少ない場合は多くを標本として全体をバランスよく推計した方がよいと言える。一方、標本数が多い場合は、悉皆層を大きくとった方が推計精度はよくなると言えるが、一部は必ず標本層に割り当てた方がよいということが分かる。また、R=0.95 の分布では標本数が多い場合は多くを悉皆とした方がよいこと言えるが、R=0.98 の分布では悉皆数を増やしても精度の変化は小さいことが分かる。

図 4.1.1　R=0.95 の分布における標本規模別最適悉皆区分

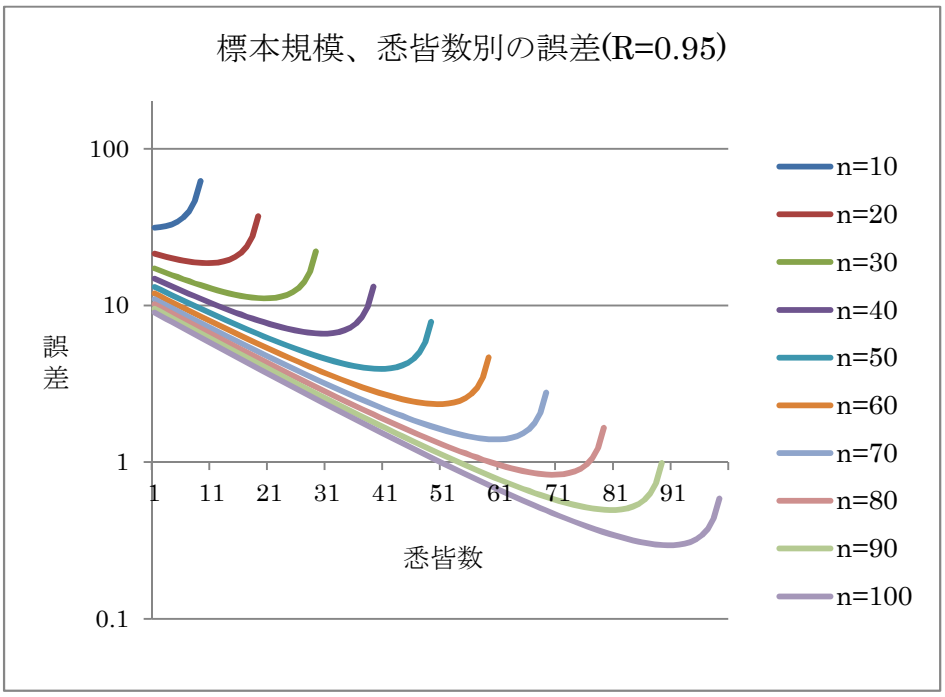


図 4.1.2　R=0.98 の分布における標本規模別最適悉皆区分

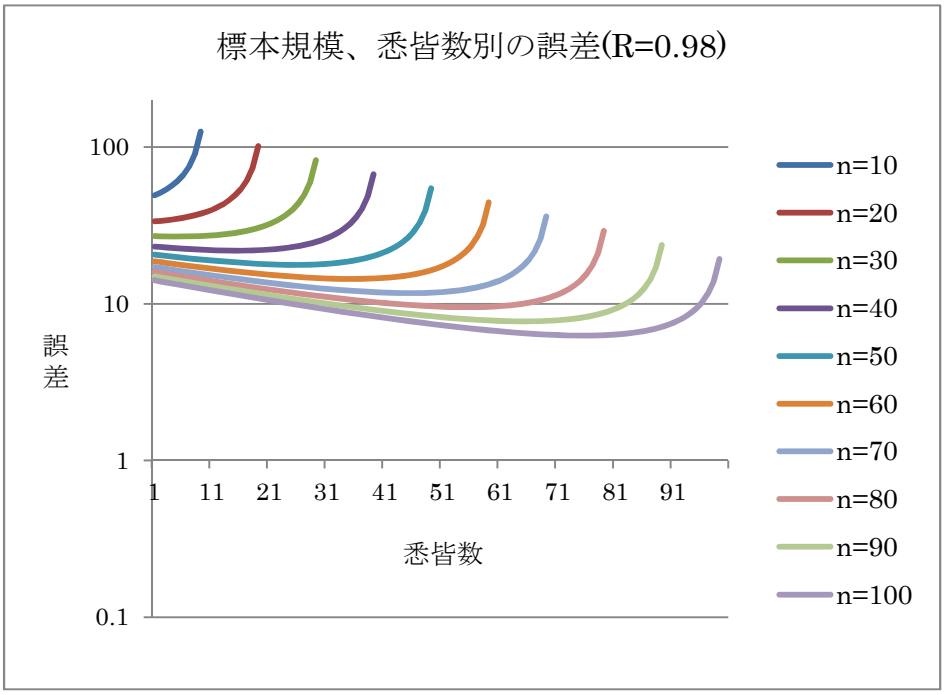
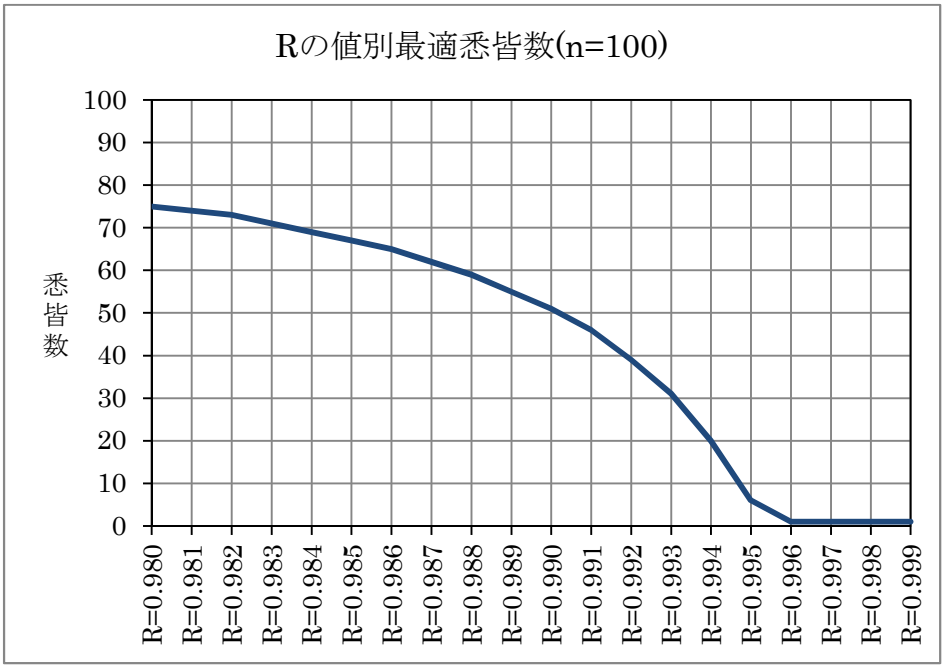


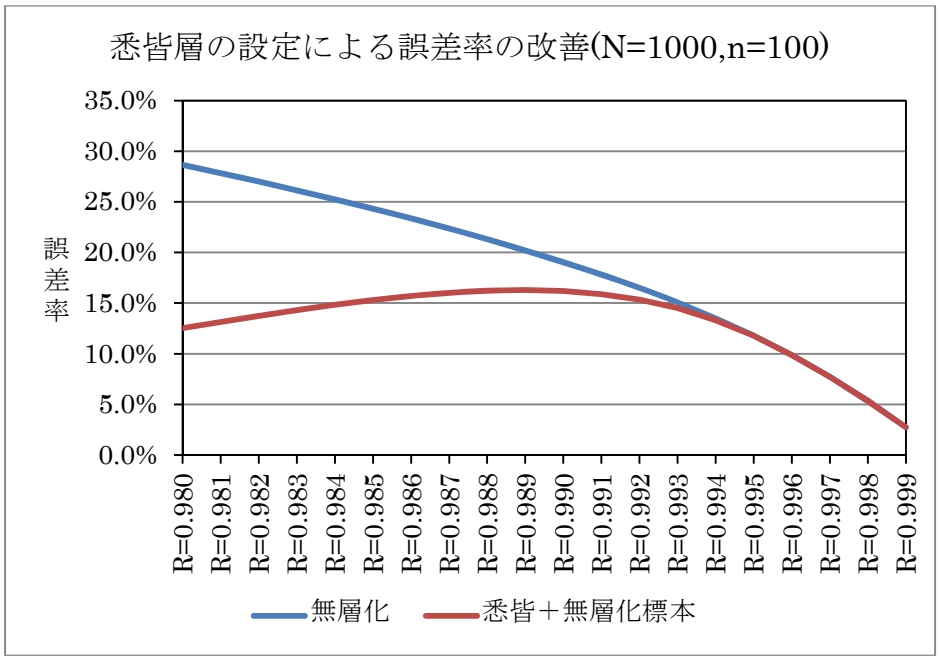
図 4.1.3　総標本数 100 の場合の R の値別最適悉皆区分



２　最適な悉皆層を設定したときの誤差

最適な悉皆層の設定により誤差率は図のように改善される。悉皆層が設定された場合の誤差は、悉皆層以外の標本層の推計誤差そのものであるから、図のようにＲの値の小さい方では大きく改善されるが Ｒ の値が大きく大部分が標本層となってしまう部分では改善されない。また、ここでは標本層は無層化の単純任意抽出であるから、悉皆層の部分が小さくなるにつれておのずと無層化単純任意抽出の誤差曲線に漸近していく。

図 4.2.1　悉皆層の設定によるＲの値の小さい分布の誤差率の改善



３　悉皆層と層化抽出標本層の組合せ

前項では、悉皆層以外の標本層が無層化単純任意抽出であったが、この部分を層化抽出にすることでより一層の誤差率の改善が見込まれる。次の表は、R=0.980 と R=0.988 の二つについて、最適悉皆層に層化抽出の標本層を組み合わせた場合の誤差を算出したものである。標本部分を単純任意抽出とした場合は、R=0.980 の場合の誤差が 6.26、R=0.988 の場合の誤差が 13.51 であるから、当然のことであるが大きく改善されている。

表 4.3.1　悉皆層を組込んだ標本設計の例

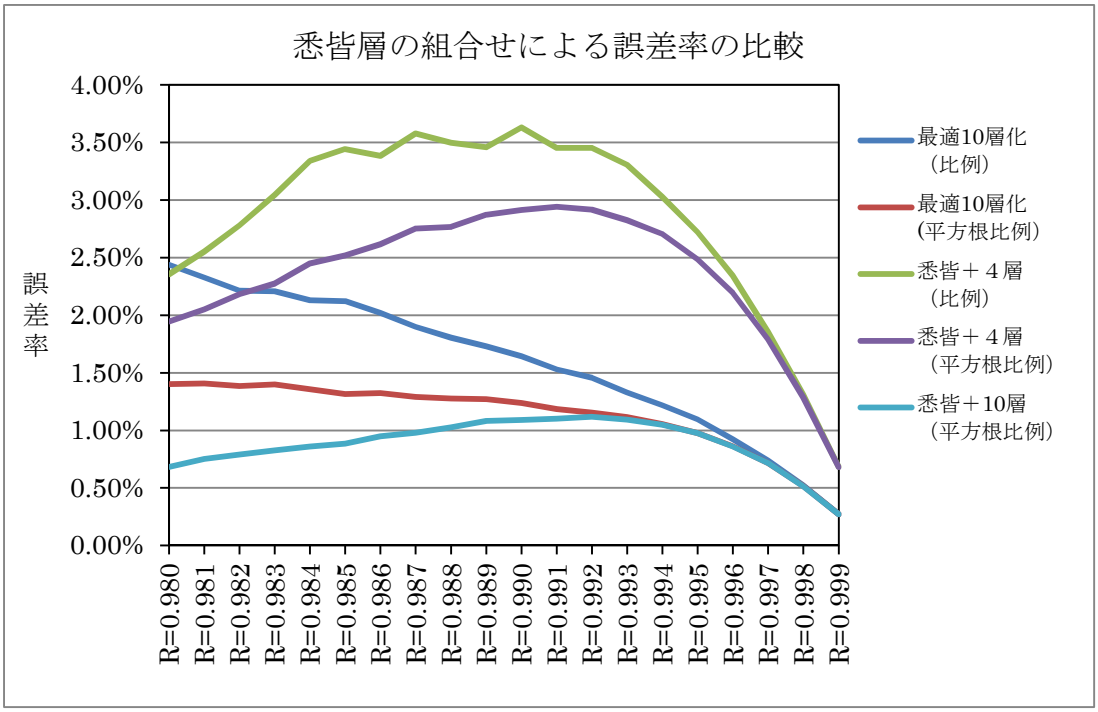
R=0.980の例											
最適層区分	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層	誤差
	8	9	10	12	14	17	23	33	60	814	
標本層比例配分	1	1	1	1	1	2	2	3	6	81	1.22
悉皆＋最適層区分	悉皆層						第1層	第2層	第3層	第4層	誤差
	76						23	33	59	809	
標本層比例配分	76						1	1	2	21	1.18
標本層平方根比例配分							2	3	4	15	0.97

R=0.988の例											
最適層区分	第1層	第2層	第3層	第4層	第5層	第6層	第7層	第8層	第9層	第10層	誤差
	13	15	17	20	23	29	37	54	95	697	
標本層比例配分	1	2	2	2	2	3	4	5	10	70	1.50
悉皆＋最適層区分	悉皆層				第1層	第2層	第3層	第4層	第5層	第6層	誤差
	60				23	29	37	54	95	702	
標本層比例配分	60				1	1	2	2	4	30	1.96
標本層平方根比例配分					3	4	4	5	7	18	1.46

例を見れば分かるように、R の値によって最適な悉皆区分が異なるため、残りをいくつに層化すべきかも異なってくる。R=0.980 の例では、最適層化 10 層の場合の第 1 層から第 6 層の合計が 70 であり、最適悉皆層は 76 であるから残りを約 4 層に分ければほぼ最適層化に見合うものとなる。誤差の程度もほぼ同じである。また、R=0.988 の例では、最適層化 10 層の場合の第 1 層から第 4 層までの合計が 65 であり、最適悉皆層は 60 であるから残りを約 6 層に分ければほぼ最適層化に見合うものとなる。誤差はわずかに大きくなるが、悉皆層の安定性を考えると、やはり悉皆層の設定は有効である。なお、どちらの例の場合も標本の配分は標本層の規模の平方根に比例させる方がよいであろう。

ここで、N=1000,n=100 の条件で、他の R 値の分布についても最適層化比例抽出、最適層化平方根比例抽出、悉皆＋4 層比例抽出、悉皆＋4 層平方根比例抽出及び悉皆＋10 層平方根比例抽出の誤差比較を行うと、下図のようになった。R の値が小さい分布では、標本抽出方法の違いによって誤差率が大きく異なる結果となっている。

図 4.3.1　悉皆層と標本層の組合せによる誤差率の変化

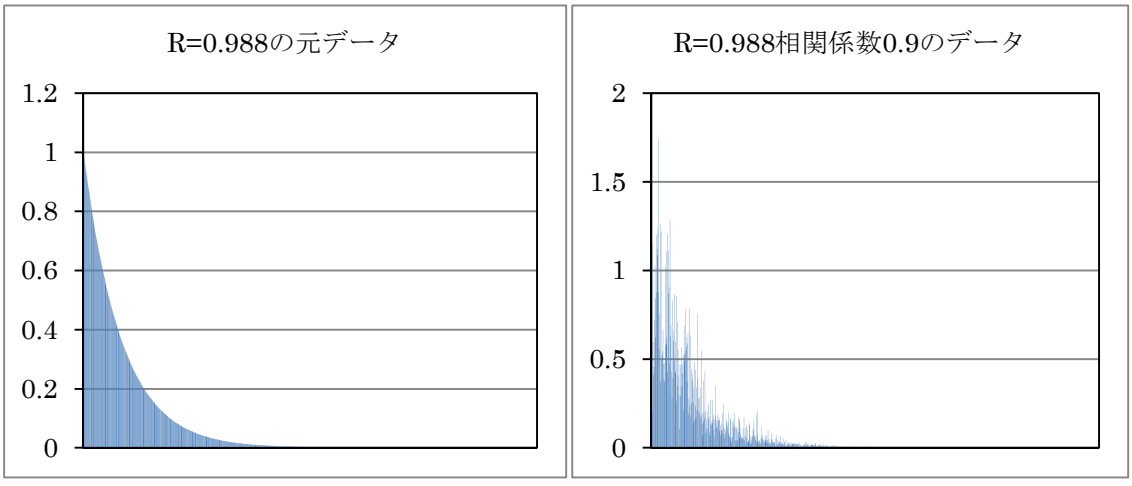


４　層化変数と推計値変数の相関係数

標本設計において直接目的変数を用いて層化できればよいが、多くの場合は目的変数と相関の高い変数を用いて行われる。そこで、当初の幾何級数モデルの値に乱数によってノイズを加え、その前後の相関係数と推計誤差の相違を分析し、各標本抽出方法がどの程度安定性を持っているのか検証した。ノイズの加え方は、正負に一様な乱数を発生させ、それをもとの値に対する比率として加えた。すなわち、分布に対して変動係数が一様な誤差である。

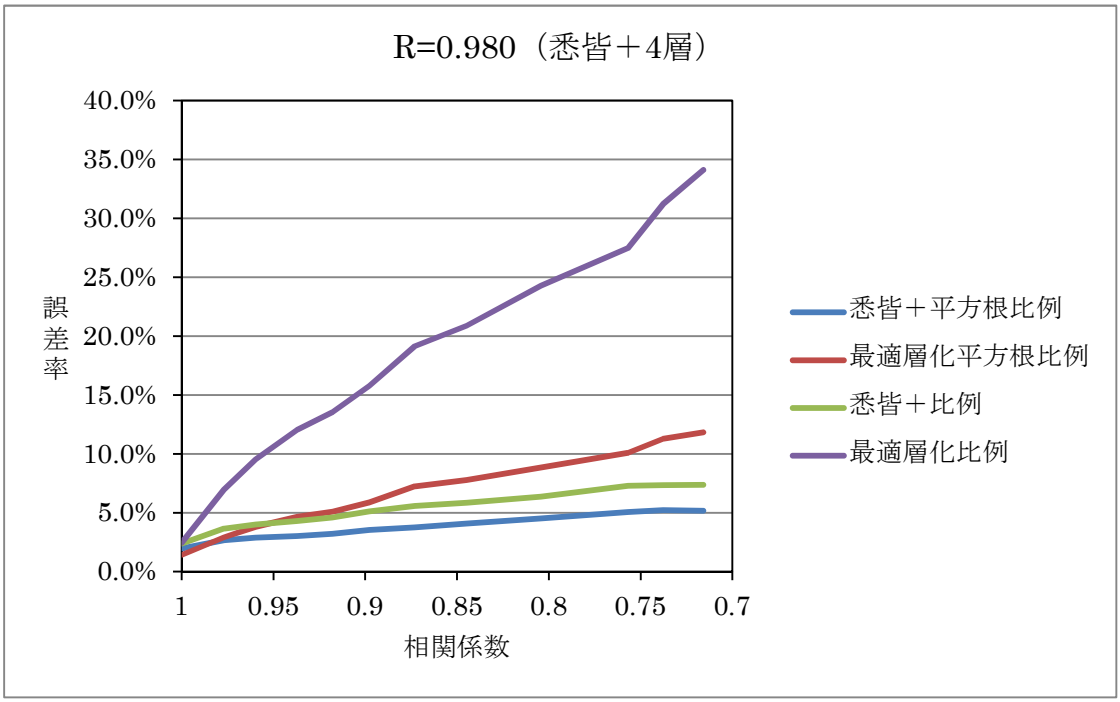
一例を図 4.4.1 に示す。

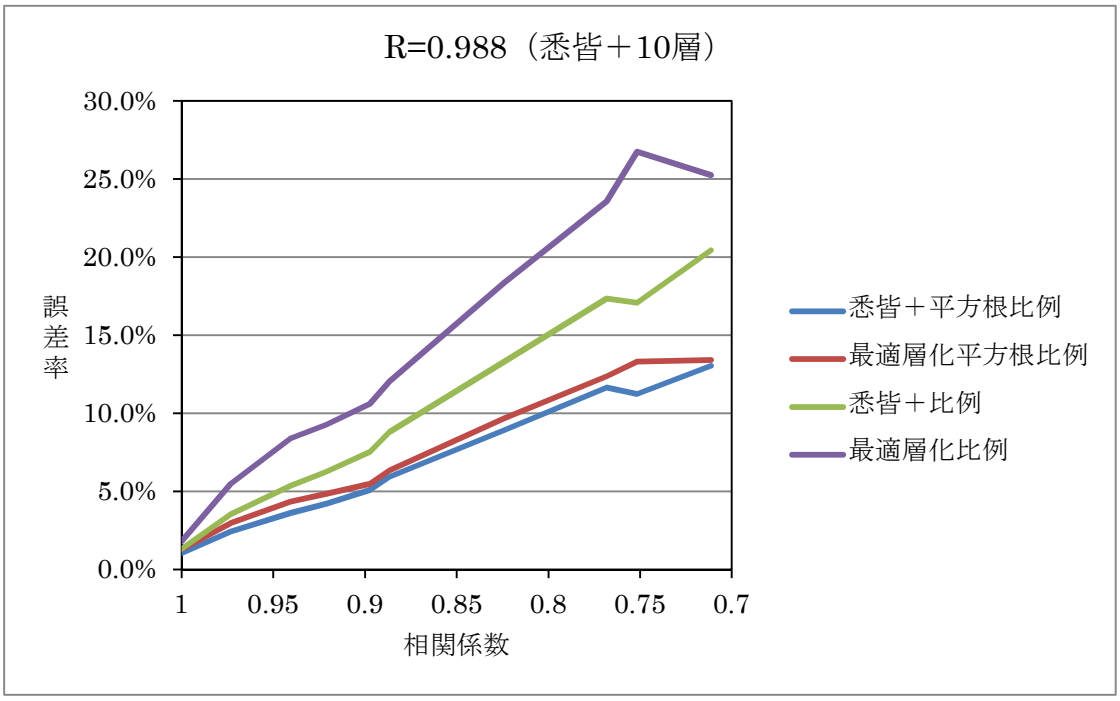
図 4.4.1　ノイズ付加による擬似的層化変数の作成の例



図を見ると、最適層化抽出法は、R=0.980、R=0.988 のどちらの分布でも目的変数と層化変数の相関係数の減少に対し誤差率の低下が大きい。一方、悉皆層を設定した抽出法では R=0.980 の分布では誤差率の低下が抑えられており、悉皆層の設定によりロバストネスが高まっていることが分かる。R=0.988 の分布については、分布の変化がなだらかに減少するため悉皆層の効果は小さいものとなっている。なお、最適層化、悉皆層＋標本のいずれも標本の抽出は平方根比例が優れていると言える。

図 4.4.2　目的変数と層化変数の相関係数の変化に伴う誤差率の変化





第5章　考察のまとめ

ロバストな標本設計として、本研究では推計の目的変数と層化に用いる変数の間の相関の強弱に対して推計誤差がどの程度影響を受けるかという観点で検討を行った。その結果、分布の歪みの大きい、公比 R が小さい幾何級数モデルに対応する母集団分布を有する場合は、最適層化後の比例抽出やネイマン配分は標本配分の傾斜が大きく実用的ではないという結論に至った。こうした問題の解決のためには、悉皆層を設定することが最も有効な手段であることを確認し、標本部分は対応する母集団部分を最適層化した上で、その層の規模の平方根に比例した標本配分とするのがよい方法であるとの結論を得た。

しかし、それでも標本数が少ない場合には、各層へ配分される標本数が少なく、母集団名簿の劣化等に対しては不安定であると推察される。さらに、標本数が少ない場合には、誤差計算が適切に行えないこともあり、標本設計としては好ましくない。これらの点を踏まえ、実務的な標本設計の考え方として、以下の点を指摘しておきたい。

（1） 歪の大きい母集団に対しては、適切な規模の悉皆層を設ける。
（2） 標本層は、標本数が少なくなるような小さな層をつくらない程度の区分とする。
（3） 各層への標本配分は抽出率（復元乗率）が一定になる比例配分が望ましい。
（4） 層の規模が小さい場合は、規模の平方根に比例させるなど一定の標本数を確保する。
（5） 標本層へ配分する標本が少ない場合は母集団を無理に層に区分しない。

その上で、最も推計精度とロバストネスのバランスがよいと思われる標本抽出法として、次のような方法を提案したい。

標本層を無層化単純任意抽出として、悉皆数可変で誤差最小点を求め、標本層とした母集団部分を総標本数から悉皆数を引いた数に分割し、その層の平均値に最も近い標本を１つ抽出するという方法である。この方法は、確率標本ではないので層化変数に対する誤差計算はできないが、明らかに層化変数に関しては不偏推定量であり、目的変数の推計誤差は層化変数と目的変数の相関誤差に依存する。層化変数を用いたすべての標本抽出法は必ずこの誤差の影響を受けるわけであるから、この誤差のみに依存するこの抽出方法は実用的なものであると考える。

今回の研究を通じて、実務的な標本設計は単に計算上の結果精度を追い求めるのではなく、適切な誤差の評価が可能となるような設計を目指すべきであることを実感した。数値計算によるグラフ化によって設計マージンと言える部分を感覚的にとらえることに成功した。ただし、今回の研究は、疑似データによったモデル分析にとどまったため、ノイズにより作成した２変数の間の関係が果たしてこのようなデータの生成方法が適切なのか更なる研究が必要である。平成 24 年に実施される経済センサス活動調査は、こうした疑似データを作成するまでもなく貴重な実データを提供してくれるはずである。また、実データを用いた標本設計の評価が可能になり、層化の情報についてもこれまで以上に適切なものが得られることから、企業を対象とした標本調査の精度は飛躍的に向上するはずである。

参考文献

[1] 森田優三、久次智雄（1993）、「新統計概論」、日本評論社
[2] 土屋隆裕、（2009）、「概説標本調査法」、朝倉書店
[3] 船津好明、（1977）、「調査統計入門－単純任意抽出法を中心として－」、共立出版
[4] 船津好明、（1986）、「続調査統計入門－1 段抽出法を中心として－」、技興社
[5] 久次智雄（1972）、標本設計の最適性についての一考察、統計局研究彙報 24 号、総理府統計局、p21－65
[6] 上田尚一（1981）、層化抽出法適用に関するノート、統計局研究彙報 36 号、総理府統計局、p95－119