気象庁97年間（明治14～昭和52年）の統計によれば，梅雨のはしりは５月23日，梅雨の入りは６月11日になるといいます。

食べ物が湿気のために腐りやすい季節です。胃腸の弱い人はもちろんのこと，何を食べても消化してしまうという人も，食中毒にはくれぐれも気をつけて，仕事にはげんで下さい。

今月のおもな行事

○１日　　農家意識調査調査日
○１～３日　漁業センサスブロック会議（神奈川県）
○15日　　事業所統計調査調査日
○23～24日　毎月勤労統計調査ブロック会議（千葉県）
○26～27日　住宅統計ブロック会議（栃木県）
○30日　　消費者動向調査調査日

標本数をどう決めるか（下）　標本調査のために……

標本抽出法（サンプリング）を使って調査を行うとき，単純任意抽出法のやり方だけで調査が実施されることはあまり見かけません。多くの場合には調査の性格，客体の性格や散らばり具合，調査費用などの面から，より能率的・効果的でありしかも精度の落ちないような方法が考えられ，使われています。特に国や県が行うような広範な地域を対象とした調査では，単純任意抽出法はおよそ適しません。そこでは，層別抽出法や2段抽出法，さらにはそれらを組みあわせたいくつもの方法が使われます。しかし，その中で単純任意抽出法の理論もまた生きているわけです。

層別抽出法

層別抽出法というのは，母集団の抽出単位をいくつかのグループに分け，各グループごとに抽出単位の標本を選ぶ方法で，この分けられたグループを**階層**といいます。このグループというのは，たとえばある市の世帯の消費支出額の平均値を調べる際に，調査区を繁華街地区と農村地区というように分けて，それぞれから世帯を選んで調査する場合の調査区をさします。この階層については，階層間の抽出単位は異質に，階層内の抽出単位は等質にするのがよい方法です。この階層作りは調査項目に関する情報を利用して，上手にやらなければなりません。調査項目がひとつであれば，その項目と**相関関係**が高い情報を基準に階層を作ります。調査項目がたくさんある場合には，そのうち最も重要な項目について相関の高い情報を基準にする方法や母集団の分布が偏っているかどうかを基準にする方法，また，**分散**が大きいものを基準にする，といった方法がとられます。

階層の作り方とともに大切なこととして，各階層からいくつの抽出単位を選ぶか，という問題があります。100世帯を調査するとしても，各階層から何世帯ずつ抽出して100世帯を調査するのかという問題ですが，この各階層への抽出単位の配分については代表的な配分法が2つあります。ひとつは**比例配分**と呼ばれるもので，実務上最もよく使われます。この方法をとった場合には，階層の作り方が多少悪かったとしても単純任意抽出法よりも精度が悪くなることはないという利点があります。そのやり方は名称のとおりで，階層の大きさに比例して配分します。すなわち，母集団Nで，第i番目の階層にN_iの単位が含まれているとき，全体の標本数nのうち第i番目の階層に配分される標本数n_iは，

$$n_i = n \times \frac{N_i}{N}$$

で求められます。

もうひとつの配分法は**最適配分**というもので，調査費用を考えた配分法です。費用が一定のとき標準誤差が最小になるよう各階層に標本数n_iを配分する方法を，**ネイマンの最適配分**といいます。この最適配分は，各階層内の分散σ_i^2がわかる場合には比例配分よりも精度のよい標本設計ができるという利点があるのですが，反対に，分散の推定を誤った場合には，最適どころか最不適にもなりかねない危険があります。うまくやるといいのですが，まちがったらヒドイのです。

さて，標本数をいくつとったらよいか，という本題にはいりましょう。標本誤差の**許容限界**をd，決められた**信頼水準**をα_0とし，α_0に対応する値をZ_0とすれば，

$$Z_0 \cdot \sigma(\bar{x}) = d$$

でした。比例配分の場合，標本数nは，

$$n = \frac{\frac{Z_0^2 \sigma_w^2}{d^2}}{\frac{1 + Z_0^2 \sigma_w^2}{Nd^2}} \qquad \left(\sigma_w^2 = \frac{1}{N}\Sigma N_i \sigma_i^2 : \text{層内分散}\right)$$

によって求められます。

抽出率が十分小さいとき上の式は，次の式で代用されます。

$$n = \frac{Z_0^2 \sigma_w^2}{d^2}$$

ネイマンの最適配分の場合には下の式

$$n \fallingdotseq \frac{Z_0^2 N^2 \bar{\sigma}_w^2}{d^2} \qquad \left(\begin{array}{l}\bar{\sigma}_w \text{は層内標準偏差平均で，} \\ \bar{\sigma}_w = \dfrac{1}{N}\Sigma N_i \sigma_i\end{array}\right)$$

で求められます。

〔例題1〕

ある県で，製品の出荷額を知るため標本抽出により調査を行うことにした。工場一覧表と前回の調査結果により下表のことがわかっている。出荷高の合計を95%の信頼水準で，標本誤差を8,000千万円以内に押さえるには標本としていくつの工場を調べればよいか。

冨永重己

〔表1〕

従業員規模	県内の工場数	標準偏差
1～ 9人	5,500	5 千万円
10～ 99	1,800	15
100～999	660	80
1,000～	40	120
—	8,000	—

従業員規模別に出荷額の標準偏差が前回の調査でおおよそわかっており，また現在の母集団の従業員規模別の事業所数もわかっているので，層別抽出法をとり，最適配分を行う。最適配分による場合の標本数の決め方は先にあげた式，

$$n \fallingdotseq \frac{Z_0^2 N^2 \bar{\sigma}_w^2}{d^2}$$

を使います。95%の信頼水準で標本からの誤差を8,000千万円以内におさえることから，

$Z_0 = 2$，$d = 8{,}000$

$$\bar{\sigma}_w = \frac{1}{N}\Sigma N_i \sigma_i = \frac{1}{8{,}000}(5{,}500\times 5 + 1{,}800\times 15 + 660\times 80 + 40\times 120)$$

$$= \frac{1}{8{,}000}(27{,}500+27{,}000+52{,}800+4{,}800)$$

$$= \frac{1}{8{,}000}\times 112{,}100$$

$$= 14.01$$

公式に代入すると

$$n \fallingdotseq \frac{Z_0^2 N^2 \bar{\sigma}_w^2}{d^2} = \frac{2^2\times 8{,}000^2\times 14.01^2}{8{,}000^2} \fallingdotseq 785$$

785の工場を調べればよいことになります。さらに，層別抽出法ですから各層にこの785の標本数を配分しなければなりません。そこまでやらなければこの問題は The End の字幕が出ません。

ネイマンの最適配分の配分された標本 n_i の求め方は，

$$n_i = n\frac{N_i \sigma_i}{N\bar{\sigma}_w}$$

で求められますから，

$$n_1 = 785\times\frac{5{,}500\times 5}{8{,}000\times 14.01}$$

$$n_2 = 785\times\frac{1{,}800\times 15}{8{,}000\times 14.01}$$

$$n_3 = 785\times\frac{660\times 80}{8{,}000\times 14.01}$$

$$n_4 = 785\times\frac{40\times 120}{8{,}000\times 14.01}$$

これらを解いて，

$n_1 = 193$，$n_2 = 189$，$n_3 = 370$，$n_4 = 34$

こうして従業者規模1～9人の工場を193，同じく10～99人の工場を189，100～999人の工場を370，1,000人以上の工場を34，あわせて786調査すればよいことがわかりました。

〔例題2〕

ある中学校（生徒数：1年生400名，2年生500名，3年生600名）で生徒の平均身長を知るため各学年からそれぞれ10名を抽出して測定し，下表の結果を得た。下表から，

①各学年及び全校生徒の平均身長を推定せよ。

②同じく標準偏差を推定せよ。

③全校生徒の平均身長は信頼水準95%でどの範囲にあると考えられるか。

④次回，標本数を50人にして調査するとすれば，標本を各学年にどのように配分すればよいか。

1 年	2 年	3 年
135 cm	138 cm	145 cm
140	143	148
140	146	149
143	150	155
144	152	157
146	153	159
147	154	163
148	156	166
152	163	168
155	165	170
合計 1,450	合計 1,520	合計 1,580

① （1年） $\bar{x}_i = \frac{\Sigma x_i}{n_i} = \frac{1{,}450}{10} = 145$ cm

（2年） $\bar{x}_i = \frac{1{,}520}{10} = 152$ cm

（3年） $\bar{x}_i = \frac{1{,}580}{10} = 158$ cm

（全校） これは $\bar{x} = \frac{\Sigma \bar{x}_i \cdot N_i}{n}$ で求められるから，

$$\bar{x} = \frac{(145\times 400)+(152\times 500)+(158\times 600)}{1{,}500}$$

$$= 152.5$$

② まず各学年について，標準偏差 $\sigma(\bar{x})$ は，

$$\sigma(\bar{x}) = \sqrt{\frac{N_i - n}{N_i - 1}} \times \frac{\sigma}{\sqrt{n}} \fallingdotseq \frac{\sigma}{\sqrt{n}}$$

σ_i^2（母集団分散）はわからないのでその階層の標本から，

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

によって推定すれば，

1　年			2　年			3　年		
x	$x-\bar{x}$	$(x-\bar{x})^2$	x	$x-\bar{x}$	$(x-\bar{x})^2$	x	$x-\bar{x}$	$(x-\bar{x})^2$
135	−10	100	138	−14	196	145	−13	169
140	− 5	25	143	− 9	81	148	−10	100
140	− 5	25	146	− 6	36	149	− 9	81
143	− 2	4	150	− 2	4	155	− 3	9
144	− 1	1	152	0	0	157	− 1	1
146	1	1	153	1	1	159	1	1
147	2	4	154	2	4	163	5	25
148	3	9	156	4	16	166	8	64
152	7	49	163	11	121	168	10	100
155	10	100	165	13	169	170	12	144
1,450	—	318	1,520	—	628	1,580	—	694

（1年）$s_i^2 = \frac{1}{10-1} \times 318 = 35.3$　　$s_i = 5.94$

（2年）$s_i^2 = \frac{1}{10-1} \times 628 = 69.8$　　$s_i = 8.35$

（3年）$s_i^2 = \frac{1}{10-1} \times 694 = 77.1$　　$s_i = 8.78$

これを先の式に代入すれば，

（1年）$\frac{5.94}{\sqrt{10}} = 1.88$，　同様にして，2年・3年生が出ます。

全校生徒の標準誤差は，

$$\sigma(\bar{x}) \fallingdotseq \frac{1}{N} \times \sqrt{\Sigma N_i^2 \cdot \frac{\sigma_i^2}{n_i}}$$

または，

$$\sigma(\bar{x}) \fallingdotseq \sqrt{\Sigma W_i^2 \cdot \frac{\sigma_i^2}{n_i}} \qquad (W_i = \frac{N_i}{N})$$

どちらの式を使っても答は同じですが，いま，上の式を使えば，

$$\sigma(\bar{x}) \fallingdotseq \frac{1}{1{,}500} \times \sqrt{(400^2 \cdot \frac{35.3}{10}) + (500^2 \cdot \frac{69.8}{10}) + (600^2 \cdot \frac{77.1}{10})}$$
$$\fallingdotseq 1.5$$

③ 95%の場合 $Z_0 = 2$

$\bar{x} \pm Z_0 \times \sigma(\bar{x})$ により，問2の値を使って，

$152.5 \pm 2 \times 1.5$

$\therefore 149.5 \sim 155.5$

よって149.5cmから155.5cmの範囲にあると考えられる。

④ 〔比例配分〕

$n_i = n \frac{N_i}{N}$ により，

（1年）$n_i = 50 \times \frac{400}{1{,}500} = 13$

（2年）$n_i = 50 \times \frac{500}{1{,}500} = 17$

（3年）$n_i = 50 \times \frac{600}{1{,}500} = 20$

〔ネイマンの最適配分〕

$$n_i = n \frac{w_i \sigma_i}{\Sigma w_i \sigma_i}$$

または，

$$n_i = n \frac{N_i \sigma_i}{N \bar{\sigma}_w} \qquad (\bar{\sigma}_w = \frac{1}{N} \Sigma N_i \sigma_i)$$

下の式を使って計算すると，

$$\bar{\sigma}_w = \frac{1}{1{,}500}(400 \times 5.94 + 500 \times 8.35 + 600 \times 8.78) \fallingdotseq 7.88$$

（1年）$n_i = 50 \times \frac{400 \times 5.94}{1{,}500 \times 7.88} \fallingdotseq 10.05$

（2年）$n_i = 50 \times \frac{500 \times 8.35}{1{,}500 \times 7.88} \fallingdotseq 17.66$

（3年）$n_i = 50 \times \frac{600 \times 8.78}{1{,}500 \times 7.88} \fallingdotseq 22.28$

それぞれ，10人，18人，22人を調べればよいことになります。

2段抽出法

たとえば，ある市で市内に住む世帯の平均実収入を標本をとって調査する場合，まず市内をいくつもの調査区に分けてその中から一部の調査区を抽出し，抽出された調査区内に含まれる世帯からさらに，実際に調査する世帯を抽出する方法を2段抽出法といいます。なおこの場合，はじめに抽出した調査区に含まれる世帯についてはすべて調査する方法をとる場合には，これを**集落抽出法**といいます。2段抽出法がさらに発展して，3段，4段に分けて抽出する調査もありますが，これらを総称して**多段抽出法**と呼んでいます。精度の点からみると，調査単位数がほぼ同数の場合には一般に，集落抽出法よりも2段抽出法の方が精度がすぐれています。単純任意抽出法に比べると精度は落ちるのですが，しかし何と言っても費用の点では優れた方法です。また，実査上の労力の点でも優れていますし応用もきくので，実務上非常によく使われています。総理府統計局で行っている**家計調査**は層別3段抽出法という方法になっています。

さて，2段抽出法における標本数の決定と，その配分についてですが，配分は層別抽出法で出てきた，調査にかかる費用の範囲内で標準誤差を最小にするように各第1次抽出単位内の調査単位数 n_i を決定する最適配分を考えなけれ

ばなりません。費用には，標本数とは関係のない固定費用，第1次抽出単位の標本数にかかる単位あたりの費用，第2次抽出単位の標本数にかかる単位あたりの費用とに区別し，それらを各々，C_0, C_1, C_2，であらわせば全費用Cについては，

$$C = C_0 + C_1 m + C_2 n$$

$$\begin{pmatrix} m: \text{第1次抽出単位の数} \\ n: \text{第2次抽出単位の数} \end{pmatrix}$$

であらわされます。抽出されるmとnの数によって費用が変わるわけですが，実際の場合は費用のワクが決められていることが多いので，ここでは全費用Cが一定のとき，標準誤差を最小にするmと$\bar{n}$（mが一定の大きさのときmから抽出される調査単位数）を求める方法を考えます。mから抽出されるn_iが一定のとき（$n_i = \bar{n} = \frac{n}{m}$）第1次抽出単位$m$は，

$$m = \frac{C - C_0}{C_1 + C_2 \bar{n}}$$

であらわされます。$\bar{n}$については，mの大きさが均一の場合，

$$\bar{n} = \sqrt{\frac{C_1}{C_2}} \times \frac{\sigma_w}{\sqrt{\sigma_b^2 - \frac{\sigma_w^2}{\bar{N}}}}$$

$$\begin{pmatrix} \bar{N}: \text{第1次抽出単位に含まれる} \bar{n} \text{の平均} (\frac{N}{m}) \\ \sigma_b: \text{第1次抽出単位間分散} \\ \sigma_w: \text{第1次抽出単位内分散} \end{pmatrix}$$

なお，抽出率$\frac{\bar{n}}{N_i}$，$\frac{m}{M}$が小さくて無視できる場合には，

$$\bar{n} = \sqrt{\frac{C_1}{C_2}} \times \frac{\sigma_w}{\sigma_b}$$

でもかまいません。

なお，このとき標準誤差は，

$$\sigma(\bar{x}) \fallingdotseq \sqrt{\frac{1}{m}(\sigma_b^2 + \sigma_b \sigma_w \sqrt{\frac{C_2}{C_1}})}$$

で計算されます。簡単な設例で実際にやってみましょう。

〔例題3〕

ある市で1世帯あたりの平均収入を知るため，調査区～調査世帯の2段抽出法による調査を行うことになった。この市には400調査区があり，おのおのの調査区にはおおむね50世帯が含まれている。調査費用は1調査区あたり10,000円，1世帯あたり100円である。過去の資料によると，調査区間分散は10,000円，調査区内分散は20,000円である。

信頼水準95%で，標準誤差を40,000円以下に押さえるように標本設計をせよ。また，その場合，費用はどの位かかるか。

いま，わかっていることは，

調査区数 M：400

調査区内平均世帯数 $\bar{N} = 50$

全世帯数 $N = 400 \times 50 = 20,000$

費用 $C_1 = 10,000$円

費用 $C_2 = 100$円

第1次抽出単位間分散 σ_b：10,000円

第1次抽出単位内分散 σ_w：20,000円

許容限界dについて，$Z\sigma(\bar{x}) = d$より，95%信頼水準だから$Z_0 \fallingdotseq 2$で，$2\sigma(\bar{x}) \leqq 4,000$，よって$\sigma(\bar{x}) \leqq 2,000$円。

以上の値を先にあげた公式に代入すれば，

$$\bar{n} \fallingdotseq \sqrt{\frac{C_1}{C_2}} \times \frac{\sigma_w}{\sqrt{\sigma_b^2 - \frac{\sigma_w^2}{\bar{N}}}} = \sqrt{\frac{10,000}{100}} \times \frac{20,000}{\sqrt{10,000^2 - \frac{20,000^2}{50}}}$$

$$\fallingdotseq 21$$

抽出率を無視すれば，

$$\bar{n} \fallingdotseq \sqrt{\frac{C_1}{C_2}} \times \frac{\sigma_w}{\sigma_b} = 10 \times \frac{20,000}{10,000} = 20$$

となって，各調査区からそれぞれ20世帯を調査することになります。次に調査区の数については，前の標準誤差の公式を使って，

$$\sigma(\bar{x}) \fallingdotseq \sqrt{\frac{1}{m}(\sigma_b^2 + \sigma_b \sigma_w \sqrt{\frac{C_2}{C_1}})}$$

$$2,000 \fallingdotseq \sqrt{\frac{1}{m}(10,000^2 + 10,000 \times 20,000 \times \sqrt{\frac{100}{10,000}})}$$

$$= \sqrt{\frac{1}{m} \times 120,000,000}$$

これをmについて解けば，$m = \frac{120,000,000}{2,000^2} = 60$

抽出する調査区の数は60調査区に決まります。

この調査にかかる総費用については，

$$m = \frac{C - C_0}{C_1 + C_2 \bar{n}}$$

より，C_0を費用0としてCについて変形すると，

$$C = m(C_1 + C_2 \bar{n})$$

$$= 60(10,000 + 100 \times 20) = 720,000$$

により72万円かかることになります。

以上，単純任意抽出法と層別抽出法，2段抽出法について簡単な例を使ってみてみました。実務的には，まだまだ多くの情報（または情報の不足）やさまざまな条件が加わって，標本抽出法は困難をきわめます。

今回は，これだけでもう頭がすっかり疲れてしまいましたので，これでおしまい。

（前県統計課・消費統計係）