統計的な推測【母集団と標本】p.88～92

青チャートB: 例題 83

【態度目標】しゃべる、質問する、説明する、動く、協力する、貢献する

【内容目標】統計調査に関する基本的な用語について理解できる。

また，標本調査の考え方について理解が深まる。

□全数調査と標本調査

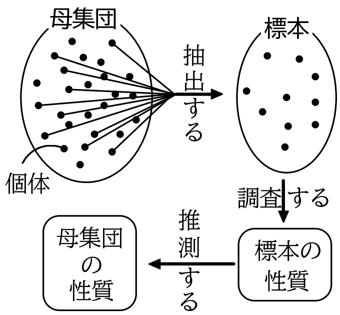
標本調査については中学のおさらいです

統計的な調査には，調べたい対象全体のデータを集める **全数調査** と，対象全体から一部を抜き出して調べ，その結果から，全体の状況を推測する **標本調査** がある。たとえば，5 年に 1 度行われる国勢調査などは全数調査である。しかし，全数調査は普通多くの労力，時間，費用を必要とする。また，工場で多数の製品を検査する場合や，検査によって製品が傷つくときなどは，全数調査は好ましくない。このような調査の対象全体が多数のときには，少ない費用と労力で全体の傾向をつかむことが有効な方法として標本調査が広く採用されている。たとえば，テレビ番組の視聴率を調べる方法は標本調査である。

出口調査で選挙結果が分かるからつまらない。
開票率５％で当確なんておかしいと
数学者・秋山仁と話したら。
「それが統計学ですよ」
「まだ開票率５％なのに？」
「あなたね、味噌汁作って味見するのに
丼鉢でグーッと飲む？」
「・・・小皿ですよね」
「それが５％よ」

左の話は立川志の輔さんの落語にある話です。
味噌汁全飲みが全数検査ということになり、鍋の味噌汁全体に均等に味噌が行き渡っていると仮定して、鍋一杯の全量に対してほんの少しだけ小皿にとって味見して、鍋の味噌汁の味噌の濃度を考えるのが標本調査となります。ただし、「ちゃんと混ざってるのか」とか「どれぐらいのちょっとならよいのか」という課題があります。そのような事をこの後考えていきます。

標本調査では，調べたい対象全体の集合を **母集団**，調査のために母集団から抜き出された要素の集合を **標 本** といい，母集団から標本を抜き出すことを，標本の **抽 出** という。また，母集団，標本の要素の個数を，それぞれ，**母集団の大きさ**，**標本の大きさ** という。



標本調査では，抽出された標本が母集団のなるべく正しい縮図になるようにしなければならない。すなわち，標本に属する要素が**特別なものに偏らないように選ぶこと，標本が偏りなく公平に抽出されることが必要**である。

母集団の各要素を等しい確率で抽出する方法を **無作為抽出** といい，無作為抽出で選ばれた標本を **無作為標本** という。無作為抽出では，**乱数さい** や **乱数表** を用いたり，コンピュータによって発生させた乱数を利用したりする。

□無作為抽出の方法

無作為抽出では，**乱数さい** や **乱数表** などが使われる。

乱数さいは，正二十面体のさいころで，0 から 9 までの数字が 2 回ずつ刻まれている。たとえば，青白赤の 3 個の乱数さいを投げて，それぞれに出た目の数を，3 桁の数の百，十，一の各位の数とすることで，0 から 999 までの数の 1 つを無作為抽出することができる。

	10	20	
	07 11 51 83 35	12 35 37 59 13	64
	77 75 52 21 20	82 26 52 29 06	47
	44 25 57 16 26	73 48 47 15 19	32
	86 79 39 57 33	99 69 31 32 42	36
	82 66 49 35 24	65 79 11 37 74	06
	62 66 89 51 70	44 20 34 68 35	55
	98 06 44 45 67	76 92 33 82 72	49
	71 28 23 55 52	67 72 00 05 18	07
	67 72 64 67 06	84 71 08 74 47	76
10	12 73 22 89 84	99 87 32 57 09	40
	42 25 31 79 54	98 70 23 91 40	41

世界最初の乱数表はティペット（L.H.C Tippett）が1927年頃に作ったもの。乱数表はどこから始めても、どのような規則に従って選んでも０から９までの数字がほぼ等しい規則で出現する。

乱数の発生

また，非常に大きい母集団から標本を抽出するときは，コンピュータを利用すると便利である。右の画像は，コンピュータを利用して，1 から 100000 までの 10 万個の数から 40 個の数字を無作為に抽出したものである。

A1 　=RANDBETWEEN(1,100000)

	A	B	C	D	E	F	G	H
1	70120	92561	14289	11126	45989	8359	96016	84718
2	51038	57182	93874	76599	77620	60457	61400	24480
3	56201	57559	82828	89808	3698	60134	11024	74483
4	70638	14696	99537	571	21772	91843	51397	3732
5	38653	82897	62425	54345	86044	65549	78386	17034
6								

コンピュータが発生するのは真の意味の乱数ではなく疑似乱数と呼ばれる非常に大きな周期をもつ周期列である。

コラム

標本の抽出方法

標本調査では，標本が母集団の縮図になるように，要素を適切に抽出する必要があるが，すべての要素を母集団全体から無作為抽出することは簡単ではない。実際には，標本に属する要素が偏らないように抽出するいろいろな方法がある。

世論調査などを行うとき，調査事項によっては，性別，職業別，年代別などに区分することがある。このように，調査結果に影響をもつと思われる性質によって，あらかじめ母集団を意識的にいくつかの組に分け，その各組から，標本を無作為抽出する方法がある。

このとき，分けたおのおのの組を層，層に分けることを層別化または層化するといい，このような標本の抽出法を **層化抽出法**(stratified sampling) という。例えば，下の表は，ある高等学校の学年別の通学手段とその人数をまとめたものである。この母集団から，層化抽出法によって，50 人の標本を抽出する方法を考えてみよう。例えば，次の ① や ② のような方法が考えられる。

学年	1 年	2 年	3 年	計
公共交通機関	257	233	271	761
自転車，徒歩	62	55	71	188

① 通学手段の 2 つの層を考えると，その比はほぼ 8 : 2 である。50 人の生徒を 8 : 2 の比に分けると，40 人と 10 人になるから，無作為抽出法によって，公共交通機関を使用している生徒から 40 人，自転車，徒歩通学の生徒から 10 人を選び，これを標本とする。

② 1 年，2 年，3 年を通学手段別にして 6 つの層を考え，各層から，その人数にほぼ比例するように計 50 人を選び，標本とする。

層化抽出法は，少ない要素で精度の高い標本を得ることができるが，母集団についての情報が事前に必要となる。層化抽出法以外にも，クラスター抽出法(下記参照)，多段抽出法など，いろいろな方法があり，それぞれの方法に長所，短所がある。

■クラスター抽出法

母集団を地域など複数の部分集団(クラスター)に分割し，いくつかの部分集団を抽出して，その集団に対しては全数調査を行う方法を，クラスター抽出法という。あらかじめ部分集団ごとの名簿があれば時間と費用を軽減することができる。

□母集団分布

統計調査の対象には，身長，血液型，不良品の個数などのように，特定の性質がある。これを**特性**といい，ある特性を表すものを **変量** という。統計調査で実際に調べるのは，母集団に属する要素についてのデータであるが，それはある変量の値の集合として表現できることが多い。 以下，特に断らない限り，データをある変量の値の集合と考える。

大きさ N の母集団において，変量 x のとりうる異なる値を x_1，x_2，……，x_r とし，それぞれの値をとる個体の個数を f_1，f_2，……，f_r とすると，この母集団における変量 x の度数分布表は，右の表のようになる。

x の値	度数
x_1	f_1
x_2	f_2
⋮	⋮
x_r	f_r
計	N

大きさ N で割ることで相対度数になる
⇒和が1
⇒大きさも1
⇒確率分布の話に

また，母集団における変量 x の平均値 m，標準偏差 σ は，次の式で与えられる。

$$m=\frac{1}{N}(x_1f_1+x_2f_2+\cdots\cdots+x_rf_r)=\frac{1}{N}\sum_{k=1}^{r}x_kf_k=\sum_{k=1}^{r}x_k\frac{f_k}{N}$$

$$\sigma=\sqrt{\frac{1}{N}\{(x_1-m)^2f_1+(x_2-m)^2f_2+\cdots\cdots+(x_r-m)^2f_r\}}$$

いま，この母集団から 1 個の要素を無作為に抽出するとき，その要素における変量 x の値 X は偶然に支配されるが，$X=x_k$ となる確率 p_k は

$$p_k=\frac{f_k}{N},\quad k=1,\ 2,\ 3,\ \cdots\cdots,\ r$$

X	x_1	x_2	…	x_r	計
P	$\frac{f_1}{N}$	$\frac{f_2}{N}$	…	$\frac{f_r}{N}$	1

である。したがって，X は右の表のような確率分布をもつ確率変数と考えられる。**この確率分布は，母集団の相対度数の分布と一致する**。したがって，母集団における変量 x の平均値 m，標準偏差 σ と，この確率変数 X の期待値 $E(X)$，標準偏差 $\sigma(X)$ について，次のことが成り立つ。

$$\boxed{E(X)=m,\quad \sigma(X)=\sigma}$$

一般に，母集団における変量 x の分布を **母集団分布**，その平均値を **母平均**，標準偏差を **母標準偏差** という。

したがって，大きさ 1 の無作為標本における変量 x の値 X は，母集団分布に従う確率変数で，その**期待値，標準偏差は，それぞれ，母平均，母標準偏差と一致する。**

1回しか取り出していなくても「どれぐらい出るか」は母集団のルールに従う。

1回しか取り出していないからばらつかない訳ではなく，まだ取り出していないだけで，繰り返せば母集団と同じだけばらつくと想定される。

母集団分布が正規分布のとき、この母集団を正規母集団という

例１１） 0 と書かれたカードが 5 枚，1 と書かれたカードが 3 枚，2 と書かれたカードが 2 枚ある。

この 10 枚のカードを母集団とみて，カードの数字を変量とすると，**母集団分布は，大きさ 1 の無作為標本の値 X の確率分布と一致する**から，右の表のようになる。よって，母平均 m と母標準偏差 σ は

X	0	1	2	計
P	$\frac{5}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	1

標本の期待値と標準偏差が適用される

$$m=0\cdot\frac{5}{10}+1\cdot\frac{3}{10}+2\cdot\frac{2}{10}=\frac{7}{10}$$

$$\sigma=\sqrt{\left(0^2\cdot\frac{5}{10}+1^2\cdot\frac{3}{10}+2^2\cdot\frac{2}{10}\right)-\left(\frac{7}{10}\right)^2}=\frac{\sqrt{61}}{10}$$ 終

実際の統計においては，母集団の大きさは**非常に大きいのが普通**であり，母集団分布は**連続型確率変数の分布で近似されることが多い**。とりわけ，母集団分布として**正規分布**をとると具合のよい現象が多い。

統計的な推測【母集団と標本】p.88～92

青チャートB: 例題 83

□復元抽出・非復元抽出

母集団の中から標本を抽出するのに，

独立な確率変数（独立な試行や反復試行のときをイメージしよう）

毎回もとに戻しながら次のものを 1 個ずつ取り出すことを **復元抽出** という。

これに対して，取り出したものをもとに戻さずに続けて抽出することを **非復元抽出** という。

標本調査のときは非復元抽出であることが多い

互いに独立ではない（条件付き確率のようなもの）

問6） 4 枚のカードに，それぞれ 1，2，3，4 の数字が書いてある。

この 4 枚のカードからなる母集団から，次のような方法で大きさ 2 の無作為標本を抽出し，

そのカードの数字を順に X_1，X_2 とする。X_1，X_2 の同時分布を，次のそれぞれの場合に求めよ。

(1) 復元抽出　　(2) 非復元抽出

解答 $p_{ij}=P(X_1=i,\ X_2=j)$ $(i=1,\ 2,\ 3,\ 4\ ;\ j=1,\ 2,\ 3,\ 4)$ とすると

(1) 復元抽出であるから

$p_{ij}=\frac{1}{4}\cdot\frac{1}{4}=\frac{1}{16}$

よって，同時分布は

左の表のようになる。

(2) 非復元抽出であるから，

$i\neq j$ で

$p_{ij}=\frac{1}{4}\cdot\frac{1}{3}=\frac{1}{12}$

よって，同時分布は

右の表のようになる。

(1)

X_1 \ X_2	1	2	3	4	計
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
計	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1

(2)

X_1 \ X_2	1	2	3	4	計
1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$
2	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$
3	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{1}{4}$
4	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{4}$
計	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1

母集団から大きさ n の標本を無作為に抽出し，その n 個の要素における変量 x の値を X_1，X_2，……，X_n とする。この抽出が復元抽出ならば，それは母集団から大きさ 1 の標本を無作為に抽出するという試行を n 回繰り返す反復試行であるから，X_1，X_2，……，X_n は，それぞれが母集団分布に従う**互いに独立な確率変数となる。**

非復元抽出の場合でも，母集団の大きさが標本の大きさ n に比べて十分大きい場合には，非復元抽出と復元抽出の違いは小さくなる。

したがって，**非復元抽出による標本も近似的に復元抽出による標本とみなす**ことができ，X_1，X_2，……，X_n は，それぞれが母集団分布に従う互いに独立な確率変数と考えてよい。

工場の製品の個数は無作為標本より十分に大きいと考えられる

例12） ある工場で生産された製品について，どの程度の割合で不良品が混ざっているかを，抜き取り検査によって調べる。この場合，母集団である全製品のおのおのについて，不良品のとき 1，そうでないとき 0 と表すと，母集団から抽出された大きさ n の標本について，0 と 1 の値をとる変数 X_1，X_2，……，X_n が得られ，これらは同じ母集団分布に従う**互いに独立な確率変数となる。** 終