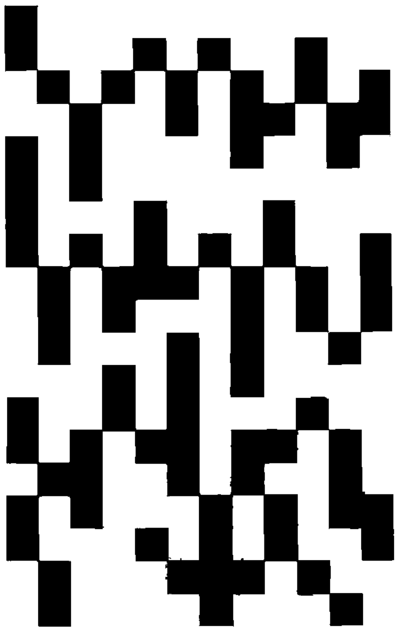
あたらしい統計学

羽鳥 裕久　著

培風館

あたらしい統計学

羽鳥 裕久　著



培風館

羽 鳥 裕 久 略歴
は とり ひろ ひさ

1960年 東京工業大学大学院修了
1960年 理学博士
現 在 東京理科大学教授

主 要 著 書

確率統計演習 1 (培風館)
初等確率論 (共著, 培風館)
ダイナミックプログラミングと
マルコフ過程 (共訳, 培風館)
有限マルコフ連鎖 (共著, 培風館)
確率論とその応用 II (上, 下)
(共訳, 紀伊國屋書店)
確率論の基礎 (コロナ社)
現代教養統計学 (共著, サイエンス社)
数理統計演習 (共著, サイエンス社)

昭和59年11月15日 初 版 発 行

あたらしい統計学

著 者 羽鳥裕久
発行者 山本健二
発行所 株式会社 培 風 館
東京都千代田区九段南 4-3-12・郵便番号 102
電話(03)262-5256(代表)・振替東京 4-44725

定 価 ¥ 1300.

前田印刷・坂本製本

ISBN4-563-00842-7
C3033 ¥1300E

まえがき

‘やさしい推測統計の書物’を書くことがここ数年来脳裏にあったが，内容について構想がなかなかまとまらず，他の仕事の多忙さを自分自身に対する言い訳にしてそのままにしており，私としては半ば断念していたのであった．ところが最近になって，ある機会があって再びこれに挑戦してみる気持ちとなり，新たに構想を練って作った素案をもとにこのような書物を編んでみた．基本的な方針としては，“どうしてそれを考えるのか”，“どうしてそうなるのか”ということをできるだけ丁寧に解説することを試みようとした．‘やさしい’ということは“なるべく数式を使わない”ということにも通じようが，上の方針を貫くために，その程度を高校教科書の延長という気持ちで推測理論の考え方を平易に解説するよう努力した．ただ高校数学をマスターしていることが前提ではなく，その一部分，すなわち総和記号 Σ，簡単な不等式，区分求積と定積分の概念——‘区分求積としての定積分 = 面積’ぐらいの基礎概念であって，実際に積分を計算することはない——などの予備知識があれば十分であり，必要があれば高校教科書などを参照して頂ければよいであろう．しかし，私自身これで満足しているわけではなく，不備な点についてはできるだけの修正をしていこうと考えているので，読者諸賢の御批判をお願いしたい．なお，a, b を端点とする区間を $a \sim b$ で表わした† ことは，常用の記号ではないので注意して頂きたい．

森俊夫氏は原稿を校閲して下さり，また三上敏夫氏は原稿を通読していくつかの注意を与えて下さった．また，出版については牧野末喜氏の御尽力によるところが多く，校正については石田恵理子さんが適切な処置をして下さった．これらの方々に謝意を表する．

昭和59年7月7日

著　　者

† 開区間 (a, b) や閉区間 $[a, b]$，半開区間 $(a, b]$，$[a, b)$ などを区別せずに $a \sim b$ で表わした．要するに a と b の間の範囲である．

目　次

I　序論として——推測の対象と手がかり

1.　何を知りたいのか——母平均と母集団分布

▶母集団と母平均　例 1　鋼球をつくっている工場で

（a）　男子従業員の扶養家族数　（例 2 に続く）

（b）　できあがった鋼球の直径　（例 3 に続く）

について知りたいものとする.

一般に，調査（あるいは検査，実験，観測など）を行なうとき，その対象となる '個々のもの' を**資料**†ということにし，資料全体の集まりを**母集団**††，資料の総数をその**母集団の大きさ**という．(a) では男子従業員 1 人 1 人が資料であり，(b) では 'できあがった鋼球' 1 つ 1 つが資料である．また，'扶養家族数' や '鋼球の直径' のように，母集団の個々の資料についての '数値的に表わせる性質' を**変量**という.

大きさ N の母集団の各資料についての 'ある変量' の値を

$$\xi_1,\ \xi_2,\ \cdots,\ \xi_N \qquad (1)$$

とするとき，これらの平均

$$\mu=\frac{1}{N}(\xi_1+\xi_2+\cdots+\xi_N)=\frac{1}{N}\sum_{i=1}^{N}\xi_i \qquad (2)$$

を**母平均**†††という.

---

†　'個体'，'要素'，'単位' などということもある．'資料' という言葉は後出の '標本(値)' とまぎらわしいので適切ではないが，他に考えつかなかった．しかし，この両者ははっきり区別しなければならない.

††　'もとになる集団' の意.

†††　'母集団平均' の略.

例 2　この工場で，実際に全男子従業員 $N=20$ 名についてその扶養家族数を調べたら

$$\begin{matrix} 3 & 5 & 3 & 0 & 1 & 2 & 2 & 3 & 3 & 0 \\ 2 & 1 & 4 & 2 & 1 & 3 & 0 & 4 & 1 & 2 \end{matrix}$$

であった．この 20 名の集まりが母集団であり，上の 20 個の数値が変量‘扶養家族数’の値で(1)の $\xi_1, \xi_2, \cdots, \xi_N$ に相当する．母平均すなわち平均扶養家族数は

$$\mu=\frac{1}{20}\times(3+5+3+\cdots+1+2)=\frac{1}{20}\times 42=2.1\ (\text{人})$$

である．これは，この工場の男子従業員についての‘1 世帯あたりの扶養家族数’の大よその目安になるものといえる．これを記録しておけば，世帯数 $N=20$ を掛けることにより

$$2.1\times 20=42$$

のように‘扶養家族数の総和’を復元することもできる．

例 3　できあがっている $N=20$ 個の鋼球の直径を 1/100 cm の単位まで測定したら，

$$\begin{matrix} 3.19 & 3.07 & 3.16 & 3.21 & 3.17 & 3.13 & 3.20 & 3.08 & 3.11 & 3.13 \\ 3.16 & 3.21 & 3.15 & 3.23 & 3.25 & 3.12 & 3.10 & 3.13 & 3.17 & 3.18 \end{matrix}$$

であった．この 20 個の鋼球の集まりを母集団とみなす†ならば，上の数値が変量‘鋼球の直径’の値で，(1)の $\xi_1, \xi_2, \cdots, \xi_N$ に相当している．母平均すなわち平均直径は

$$\mu=\frac{1}{20}\times(3.19+3.07+\cdots+3.18)=3.1575\ (\text{cm})$$

であり，これらの鋼球の直径の大よその目安となるものである．

一般に，母平均 μ を知ることは

（イ）　μ は変量の値 $\xi_1, \xi_2, \cdots, \xi_N$ を代表する値である††．

（ロ）　母集団の大きさ N が（大略にもせよ）わかっているときには，変量の値の総和 $\xi_1+\cdots+\xi_N$ が $\mu\times N$ のように簡単に求められる．

などの理由†††で重要である．

---

†　II 章以降で標本とみなすこともある．

††　代表値としては，平均の他に**メディアン**，**モード**などがある．高校の教科書などをみよ．

†††　この他の理由も後に説明する．

問 1 50 名の生徒に計算テストをした結果，次の得点が得られた．

6	4	2	7	3	7	5	6	6	5
6	5	7	5	6	6	6	7	5	3
5	8	5	6	8	7	7	4	3	7
5	2	6	8	5	1	5	6	5	6
3	7	4	5	6	4	5	4	6	6

この $N=50$ 名の集まりを母集団とみなして，変量 '得点' についての母平均すなわち平均点を求めよ．

問 2 上の 50 名の身長を測定して次の結果を得た (cm 未満 4 捨 5 入)．

121	127	130	138	133	127	121	130	127	123
120	125	128	127	132	122	137	127	125	125
131	129	120	128	122	123	132	126	118	126
126	123	120	121	126	128	118	132	126	127
127	127	119	127	126	126	121	132	112	132

この変量 '身長' の場合に，母平均すなわち平均身長を求めよ．

変量 '扶養家族数' の可能な値は 0, 1, 2, … である．一般に，変量の可能な値が

$$a_1,\ a_2,\ \cdots \tag{3}$$

のように 'バラバラな値' ——一般的な用語ではないが，以下これを**変量値**ということにする——である† とき，この変量を**離散変量**という．これに対し，'鋼球の直径' の可能な値は 'ある連続した範囲†† 内のすべての値' である．このような変量を**連続変量**といっている．

▶**離散変量の場合**に，(1) の変量の値 $\xi_1,\ \xi_2,\ \cdots,\ \xi_N$ は (3) の変量値 $a_1,\ a_2,\ \cdots$ のどれかと一致するが，(1) の中に各 a_ν ($\nu=1,\ 2,\ \cdots$) が出てくる度数を勘定し，a_1 が r_1 回，a_2 が r_2 回，… あったものとする．もちろん

$$r_1+r_2+\cdots=\sum_\nu r_\nu=N$$

となっている．これを表 1.1 のような形にまとめ，**度数分布**という．各 ν に対し r_ν/N は '系列 (1) の中に出てくる a_ν の度数の全度数 N に対する割合' を表わしており，これを**相対度数**††† という．これに対して

$$\frac{r_1}{N}+\frac{r_2}{N}+\cdots=\sum_\nu \frac{r_\nu}{N}=1 \tag{4}$$

---

† これらの値は小さいほうから順に書くのが普通である．なお，'変量値' と '変量の値' を混同しないように．

†† この範囲を前もって正確に知るのは困難なことが多い．

††† 100 倍してパーセントで表わすこともある．

表 1.1

変量値	a_1	a_2	$\cdots$	計
度　数	r_1	r_2	$\cdots$	N

表 1.2

変量値	a_1	a_2	$\cdots$	計
相対度数	r_1/N	r_2/N	$\cdots$	1

が成り立つことは容易にわかる．これも表 1.2 のようにまとめて，**相対度数分布**あるいは**母集団分布**†という．ξ_1, ξ_2, $\cdots$, ξ_N の中には a_1 が r_1 個，a_2 が r_2 個，$\cdots$ あるから

$$\xi_1+\xi_2+\cdots+\xi_N=a_1\times r_1+a_2\times r_2+\cdots=\sum_\nu a_\nu r_\nu$$

と書くことができ，母平均 μ は

$$\mu=\frac{1}{N}(a_1r_1+a_2r_2+\cdots)=a_1\frac{r_1}{N}+a_2\frac{r_2}{N}+\cdots$$

すなわち

$$\mu=\frac{1}{N}\sum_\nu a_\nu r_\nu=\sum_\nu a_\nu\frac{r_\nu}{N} \tag{5}$$

と表わすことができる．この最後の式からわかるように母平均は母集団分布だけで定まるので，これを**分布の平均**ともいう．

例 4　例 2 の 20 個の数値の中に出てくる 0, 1, 2, $\cdots$ の個数を‘正の字’を書くようにして勘定すると，表 1.3 の度数分布が得られる．

表 1.3

変量値 a_ν	0	1	2	3	4	5	計
チェック	下	𠀋	正	正	丅	一	
度　数 r_ν	3	4	5	5	2	1	20
$a_\nu r_\nu$	0	4	10	15	8	5	42

これに $\sum_\nu a_\nu r_\nu$ の値を計算するための欄をつけ加えて (5) 式を用いると

$$\text{母平均 } \mu=\frac{1}{20}\times 42=2.1$$

を得る．N が大きいほど，(2) 式による単調な和の計算より，この (5) 式による方法のほうが速くて間違いが少ないであろう．これからさらに表 1.4 のように母集団分布をつくることができる．この表から，20 世帯の中の半数は扶養家族が2～3人であり，扶養家族数が 0～1人のものは1/3強である，といったこと

† a_1, a_2, $\cdots$ を明示しないでも了解できるときは，単に $\{r_1/N, r_2/N, \cdots\}$ で表わすこともある．

が読みとれよう.

表 1.4

変量値 a_ν	0	1	2	3	4	5	計
相対度数 r_ν/N	0.15	0.20	0.25	0.25	0.10	0.05	1.00

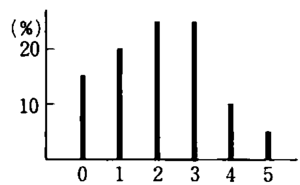


図 1.1　母集団分布のグラフ[†]

問 3　問 1 の場合に, 度数分布をつくって (5) 式を用いて母平均を計算せよ. また相対度数分布をつくり, そのグラフを描け.

母集団分布を知ることは,

(ハ)　これは, 変量に着目したときの母集団の構成状況を表わしている.

(ニ)　母集団分布がわかれば, 母平均 (や後に述べる母分散など) を計算することができる.

などの理由で重要である.

▶**連続変量の場合**, その可能な値の範囲を幅が一定である[††]いくつかの小区間 C_1, C_2, … に分け, これらを**階級**という. (1) の変量の値 ξ_1, ξ_2, …, ξ_N について, これらの階級 C_1, C_2, … に属するものの度数 r_1, r_2, … をそれぞれ勘定して表 1.5 のような形にまとめたものが**度数分布**である. 2 行目に書き入れた a_1, a_2, … はそれぞれ階級 C_1, C_2, … の中央の値であって, **階級値**といわれる. この度数分布から母平均 μ (の近似値) を計算することを考えてみよう. N が大きいときにはそのほうが速くて大きな間違いも少なく, また実際の

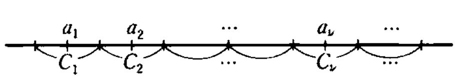


図 1.2

表 1.5

階　級	C_1	C_2	…	計
階級値	a_1	a_2	…	
度　数	r_1	r_2	…	N

† 以下, グラフの縦・横の目盛は異なることが多い.

†† ときには幅を一定にしないこともある.

場合に資料の変量の値が‘なま’の形でなく度数分布の形にまとめて与えられていることもあるからである．ところで，ξ_1, ξ_2, …，ξ_N の中で階級 C_ν に属する r_ν 個のものの実際の値はわからないので，これらはすべて階級値 a_ν に等しいものとみなしてしまう．こうすると ξ_1, ξ_2, …，ξ_N の中には a_1 が r_1 個，a_2 が r_2 個，… あるとみなされるので

$$\xi_1+\xi_2+\cdots+\xi_N \fallingdotseq a_1r_1+a_2r_2+\cdots=\sum_\nu a_\nu r_\nu$$

したがって，近似的に

$$\text{母平均 } \mu=\frac{1}{N}(a_1r_1+a_2r_2+\cdots)=\frac{1}{N}\sum_\nu a_\nu r_\nu \tag{6}$$

が成り立つことになる．

問4　幅が d の階級をつくったとき，(6)式で母平均を計算することによる誤差の絶対値は $d/2$ 以下である†．その理由を考えよ．[ヒント：各 ξ_i を階級値でおきかえたとき，その誤差の絶対値は $d/2$ 以下である.]

r_ν/N は‘系列(1)の中で階級 C_ν に属するものの度数の全度数 N に対する割合’を表わしており，これを**相対度数**という．表1.6のような形にまとめたものが**相対度数分布**である．もちろん

$$\frac{r_1}{N}+\frac{r_2}{N}+\cdots=\sum_\nu \frac{r_\nu}{N}=1 \tag{7}$$

が成り立っている．相対度数分布を図示するには，横軸上に階級 C_1, C_2, … を設けてその上に面積がそれぞれ r_1/N, r_2/N, … である(あるいは，r_1/N, r_2/N, … に比例する)長方形を図のように立てる．これを**ヒストグラム**という．(7)式はこれらの面積の和が1であることを示している．

表1.6

階　級	C_1	C_2	…	計
相対度数	r_1/N	r_2/N	…	1

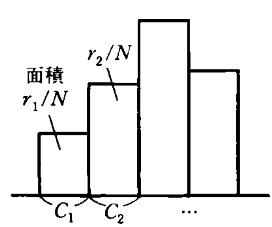


図1.3

† 実際の場合には，階級値でおきかえたときの誤差には正のものと負のものとがあって，それらの和は打ち消し合うから，(6)式で母平均を計算したときの誤差は $d/2$ よりもずっと小さいことが多い．

例 5 例 3 の 20 個の鋼球について，その直径の最小は 3.07 cm，最大は 3.25 cm であるが，表 1.7 のように度数分布をつくってみても各変量値ごとの度数が 0 ないし 3 では，はっきりした全体的傾向は読みとりにくい．元来，直径が 3.13 cm のものが 3 個あるといっても，実はこの 3 個の鋼球の直径は 3.125 cm 以上 3.135 cm 未満であると考えるべきであろう．もう少しまとめて，たとえば 3.065 cm からはじめて幅が 0.04 cm の階級をつぎつぎにつくり度数を勘定して度数分布 (表 1.8) を得る．

表 1.7

直径	3.07	3.08	3.09	3.10	3.11	3.12	3.13	3.14	3.15	3.16	3.17	…
度数	1	1	0	1	1	1	3	0	1	1	2	…

表 1.8

階 級 C_ν	3.065 以上 ～3.105 未満	3.105 ～3.145	3.145 ～3.185	3.185 ～3.225	3.225 ～3.265	計
階級値 a_ν	3.085	3.125	3.165	3.205	3.245	
チェック	下	正	正 一	下	丅	
度 数 r_ν	3	5	6	4	2	20
$a_\nu r_\nu$	9.255	15.625	18.990	12.820	6.490	63.180

これに $\sum_\nu a_\nu r_\nu$ を計算するための欄をつけ加えて† (6) 式を用いると

$$母平均\ \mu=\frac{1}{20}\times 63.180=3.159$$

を得る．例 3 で計算した μ の真の値 3.1575 との違いは 0.0015 で階級の幅 $d=0.04$ の半分 $d/2=0.02$ の 1/10 程度である．相対度数分布を表 1.9 のようにつくってヒストグラムを描くと，20 個の数値が母平均 μ を中心に分布している状況が直観的に把握できる (図 1.4)．区間 3.105～3.225 に対する相対度数は

$$0.25+0.30+0.20=0.75$$

すなわち直径がこの範囲にある鋼球は全体の 75% あることが読みとれる．

表 1.9

階 級 C_ν	3.065 ～3.105	3.105 ～3.145	3.145 ～3.185	3.185 ～3.225	3.225 ～3.265	計
相対度数 r_ν/N	0.15	0.25	0.30	0.20	0.10	1.00

† メモリーつきの電卓などを使えば，この欄をつくる必要はない．

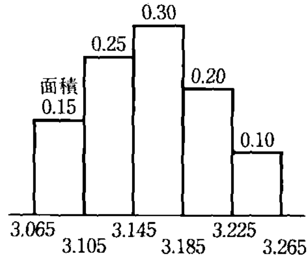


図 1.4

問 5 問 2 の場合に，111.5 cm からはじめて幅が 4 cm の階級をつぎつぎにつくり，度数分布を求めよ．また，これから (6) 式を用いて母平均 μ の近似値を計算し，問 2 の結果と比較せよ．さらに，相対度数分布をつくりヒストグラムを描け．

離散変量のときと同様に，連続変量のときにも

（ホ） 相対度数分布は，変量に着目したときの母集団の構成状況を表わしている．

が，階級のつくり方により若干の相違がみられる．階級のつくり方に無関係な相対度数分布（母集団分布）の概念は，3 節で解説する．

▶**母平均の簡便計算**について述べておこう．3 つの数 1030，1040，1050 の平均を求めようと思えば，われわれは無意識の中につぎのようにするであろう．

$$\left.\begin{matrix}1030\\1040\\1050\end{matrix}\right\}\xrightarrow{\text{1000 を引く}}\left.\begin{matrix}30\\40\\50\end{matrix}\right\}\xrightarrow{\text{10 で割る}}\left.\begin{matrix}3\\4\\5\end{matrix}\right\}\xrightarrow{\text{平均を計算する}}4$$

$$\xrightarrow{\text{10 倍する}}40\xrightarrow{\text{1000 を加える}}1040\ \text{(求める平均)}$$

この考えを (6) 式の計算の場合に適用しよう．階級値 a_1，a_2，… の中で（度数分布も考慮して）ほぼ中央にある値を a として，新しい階級値

$$b_1=\frac{a_1-a}{d},\ b_2=\frac{a_2-a}{d},\ \cdots \tag{8}$$

をつくる．ここに d は階級の幅であり，a のことを**仮平均**という†．b_1，b_2，… は連続した整数で，途中に 0 が入っているので桁数が小さく，この新しい階級値に基づく平均

$$\mu'=\frac{1}{N}(b_1r_1+b_2r_2+\cdots)=\frac{1}{N}\sum_\nu b_\nu r_\nu \tag{9}$$

† はじめの説明の 1000，10 に相当するのがそれぞれ a，d である．

表 1.10

もとの階級値	a_1 a_2 $\cdots$	計
新しい階級値	b_1 b_2 $\cdots$	
度　数	r_1 r_2 $\cdots$	N

を計算することは，(6) 式の μ の計算よりは簡単である．μ' から μ を計算するための公式を与えよう．(9) 式に (8) 式を代入して

$$\mu'=\frac{1}{N}\left\{\frac{a_1-a}{d}r_1+\frac{a_2-a}{d}r_2+\cdots\right\}$$

$$=\frac{1}{d}\left\{\frac{1}{N}(a_1r_1+a_2r_2+\cdots)-a\cdot\frac{1}{N}(r_1+r_2+\cdots)\right\}$$

$$=\frac{1}{d}(\mu-a)$$

$$\therefore\quad \mu=\mu'd+a \tag{10}$$

これが求める母平均である．離散変量の場合でも．変量値 a_1, a_2, $\cdots$ が等間隔に並んでいれば上の方法が適用できる†.

例 6　例 5 の場合に，階級値の中央の値 $a=3.165$ を仮平均にとろう．$d=0.04$ であるから，'新' 階級値は

$$b_3=\frac{a_3-a}{d}=0 \qquad \therefore\quad b_4=1,\ b_5=2;\quad b_2=-1,\ b_1=-2$$

となる．表 1.11 をつくって μ' を計算すると

表 1.11

もとの階級値 a_ν	3.085	3.125	3.165	3.205	3.245	計
新しい階級値 b_ν	−2	−1	0	1	2	
度　数 r_ν	3	5	6	4	2	20
$b_\nu r_\nu$	−6	−5	0	4	4	−3

$$\mu'=\frac{1}{20}\times(-3)=-0.15$$

$$\therefore\quad \mu=-0.15\times0.04+3.165=3.159$$

のように μ の値が求められ，例 5 の計算よりは楽であるといえる．たとえ電卓で計算しても，キーを押す回数が減り，したがってキーの押し間違いに注意が

† 度数分布を用いずに直接 ξ_1, ξ_2, $\cdots$, ξ_N から母平均 μ を計算するときにも，この方法が適用できる．8 節を参照せよ．

行き届きやすい†.

問 6 問5の場合に，簡便計算を用いて母平均(の近似値)を求めよ．

▶**推測の必要性** 変量および母集団が与えられたとき，**母集団分布**(**相対度数分布**)**を知ることが重要である**ことは上でみてきた．これがむずかしいときでも，**母平均**(や次節で述べる**母分散**など)**を知ることが望まれる**ことが多い．上の例 2, 例 3 の場合には $N(=20)$ が小さかったから資料全部についての調査が容易にできたが，母集団の大きさ N が大きいときにはこのような**全数調査**は

(ヘ) 調査の実施や分類，集計などに多くの時間や費用を必要とする．

(ト) 調査の際の間違いや集計の誤りに対し注意することが困難となる．

(チ) 製品の寿命試験のような破壊検査では，全数調査をするとすべての製品が使いものにならなくなる．

のような理由††で好ましくないことが多い．このため，母集団の中からいくつかの資料——これを**標本**という——を選び，それらの標本の変量の値——これを以下では**標本値**ということにする——x_1, x_2, …, x_n をもとにして母平均や母集団分布などの母集団の特性についての推測を行ないたい．このための数学的理論が数理統計学であって，この理論に基づいて推測を行なう方法を**標本調査**ということがある†††．実際の推測方法は次章以降で述べることにするが，**標本数**——これを以下では**標本の大きさ**ということもある——n がさほど大きくなければ，(ヘ), (チ) の問題点は緩和ないし解消されるし，(ト) についても間違いを起こさないよう十分注意することが可能になろう．

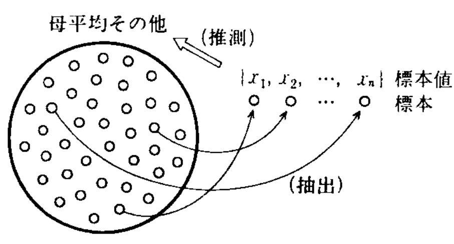


図 1.5

† 大型計算機やパソコンなどを用いるときは，(後述のものを含めて) 簡便計算そのものが不用であるかも知れない．ソロバン，電卓による手計算には便利なこともあろう．

†† 5節で説明する '仮想的母集団' では，もちろん全数調査はできない．

††† この言葉は有限母集団の場合に使うことが多い．