第 7 章

推定

標本抽出によって集められたデータは母集団の状況を反映はするものの，そのままの真実を伝えてはくれない．私たちは確率分布を考えることで，データから母集団の真の統計量を推定することができる．その手法を学ぼう．

7.1 点推定と区間推定

世論調査で発表される内閣支持率などの数字は標本平均であり，人々はその数字を無意識のうちに母平均として認識している．しかし，前章で見たように，抽出された標本の平均は抽出のたびごとに異なった値をとる確率変数であり，その期待値（平均）は母平均に一致するが，ある分散をもつ．

つまり $\overline{X}$ を復元抽出によって得られた標本平均とすると，その期待値と分散は

$$E[\overline{X}] = \mu \tag{7.1}$$

$$V[\overline{X}] = \frac{\sigma^2}{n} \tag{7.2}$$

で表されることになる．たとえば内閣支持率が 35% であると発表された場合には，その数値は一定の幅をもっているのである[*1]．

そこで，抽出によって標本平均が得られたとして，それがどの程度信用できるかという問題を考える必要が出てくる．標本平均 $\overline{X}$ は確率変数であるから，母平均とちょうど一致する確率はゼロかきわめて小さいからである．そこで $\overline{X} \pm \alpha$ のようにある範囲をとってやれば，その中に入る確率を指定できることになる．その場合に幅 α を大きくとれば

[*1] 世論調査では 1000 人から 3000 人ほどの無作為抽出標本をもとに統計量が算出されるのが普通だが，「無回答」という層も少なからず存在する．もし，これらの回答をしない人々が一定の傾向を持っているならば，得られた結果はそれを反映した偏りをもつであろう．その種の結果の偏りは，数学的な取り扱いだけでは取り除くことが難しい．

正しい確率は大きくなるが，情報としては意味がなくなってくるし，幅を小さくとれば情報としてシャープになるが，間違う危険は大きくなる．

すなわち，

> **この調査結果が 95% 正しいというためには，どれくらいの幅を持たせておかなければならないか？**

といったことを検討しなければ正しい統計にはならない．

すなわち，「母平均はこれこれの値であると考えられる」というふうに誤差を許容しつつ一点の値で答えるような推定を**点推定 (point estimation)** といい，このようにして一点で示された値を**点推定量 (point estimator)** という．

一方,「母平均は $x_1 \sim x_2$ の範囲にある確率が 95% である」というふうに「幅」をもって推定することを**区間推定 (interval estimation)** といい，なされた推定結果を**区間推定量 (interval estimator)** という．言い換えると，区間推定量は**信頼区間 (confidence interval)** を伴った推定量である．

7.1.1　パーセント点と信頼区間

ここで正規分布に関わって，統計的推定でよく用いられる量を定義しておこう．図 7.1 の標準正規分布 $N[0,1]$ で，右端から面積が α となるようにとった点を z_α と書く．この点は通常，**$100(1-\alpha)$ パーセント点**と呼ばれることが多い．たとえば，$\alpha = 0.05$ であれば，この点は $z_{0.05}$，すなわち 95 パーセント点であり，正規分布表を参照して，$\Phi(z) = 0.95$ となる z を見つけて，**95 パーセント点は 1.645** であるということになる．

なおパーセント点は，第 1 章で登場したパーセンタイル（→ p.12）と同じ内容である．

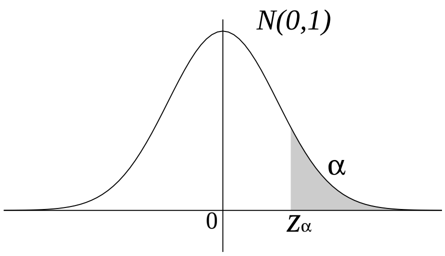


図 7.1　パーセント点の定義．z_α の点を $100(1-\alpha)$ パーセント点という．

いくつかの重要なパーセント点を表 7.1 に示した．

ここで注意しておかなければならないことは，一般的な正規分布の利用法では，平均値を中心として対称な面積を考えることが多いということである．つまりたとえば 95 パーセント点と 5 パーセント点とが対になって，その内側の 90% の領域を作っていると捉え

るのである．

そこで，z_α ではなくて，$z_{\alpha/2}$ もよく用いられる．これは，その右側の面積が $\alpha/2$ であるような点であるから，$-z_{\alpha/2}$ の左側にも対称に領域を作ったとすると，合わせて α の面積が両端にできることになり，その間には $1-\alpha$ の面積が残ることになる．図 7.2 を参照のこと．

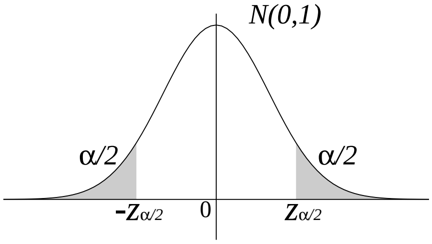


図 7.2　$z_{\alpha/2}$ の意味

▼パーセント点と正規分布表　表 7.1 を使わずに，巻末の正規分布表から 90% 点などを求めよ．

90% 点について求めてみる．正規分布表をみると，$z=1.28$ のときに $\Phi(z)=0.899727$, $z=1.29$ のときに $\Phi(z)=0.901475$ となっている．短い区間であるから直線関係を仮定して，$\Phi(z)=0.9$ となる z を比例配分で求めればよい (比例配分の計算方法については巻末付録参照)．

$$z = 1.28 + \frac{0.9-0.899727}{0.901475-0.899727} \times (1.29-1.28) = 1.28156 \approx 1.282$$

▼z_α と $z_{\alpha/2}$　標準正規分布 $N[0,1]$ がある．平均値を中心に 90% および 95% の面積を含む区間の両端の値を表 7.1 から読み取れ．

90% を含む区間は，**両側の端**に 5% ずつを含むようなとり方をするのであるから，95 パーセント点を使えばよい．したがって，$z_{0.05}$ を選んで，区間の右端は 1.645, 左端は

表 7.1　正規分布の各パーセント点

パーセント	90	95	97.5	99	99.5
α	0.10	0.05	0.025	0.01	0.005
z_α	1.282	1.645	1.960	2.326	2.576

-1.645 となる．同様に 95% の場合には，両端は ± 1.960 となる．これらはよく使われる値であるので気に止めておくとよい．

7.2　不偏推定量

μ や σ^2 のような母集団の統計量 θ が，$\overline{X}$ や s^2 のような標本の統計量 Θ の期待値として推定されるとき，すなわち，

$$\theta = E[\Theta] \tag{7.3}$$

として推定されるとき，Θ を母数 θ の**不偏推定量 (unbiased estimation)** [*2]という．つまり，Θ ($\overline{X}$ や s^2) を多数回標本抽出して得て，それらの平均値を知ることができるならば，θ になるであろうということである．

すでにこの形の式の例は，

$$E[\overline{X}] = \mu \tag{7.1}$$

が与えられていて，これから，**母平均 $\boldsymbol{\mu}$ の不偏推定量は標本平均 $\overline{\boldsymbol{X}}$ である**．また母分散 σ^2 の不偏推定量については，式 (6.5)，すなわち，

$$E[s^2] = \frac{n-1}{n}\sigma^2$$

から，

$$\sigma^2 = E\left[\frac{n}{n-1}s^2\right]$$

と変形でき，**$\frac{n}{(n-1)}s^2$ は母分散の不偏推定量と見なせる**ことが分かる．これは更に，

$$s^2 = \frac{1}{n}\left((X_1-\overline{X})^2 + (X_2-\overline{X})^2 + ... + (X_n-\overline{X})^2\right)$$

という定義を用いれば，次のようになる．

$$\frac{1}{n-1}\left((X_1-\overline{X})^2 + (X_2-\overline{X})^2 + ... + (X_n-\overline{X})^2\right) \tag{7.4}$$

これが σ^2 の不偏推定量であり，式 (7.4) の値を**標本不偏分散**，その平方根を**標本不偏標準偏差**という[*3]．

結局，もしもたった 1 回の標本抽出で母分散を知りたいときには，式 (7.4) を使って計算した値をもって最も確からしい推定値にするのである．このことを理解しておいてほしい．

[*2] unbiased というのはバイアス，つまり偏りが入っていないという意味である．

[*3] Excel で分散，標準偏差を求める関数として使われている VAR,STDEV　は本来の分散や標準偏差ではなく，この標準不偏分散および標本不偏標準偏差を指している．そのため，誤解して使っているケースもかなりあるはずである．

7.3　母平均の推定

一般に行われる統計調査で最もよく利用されるのは，平均である．政党や内閣の支持率，番組の視聴率，人の寿命，身体測定のデータ，その他数えきれないほどの統計で平均が登場する．この場合，標本平均 $\overline{X}$ を真の平均，つまり**母平均** μ と考えることを，私たちは暗黙のうちに認めている．そのことを保証しているのは，**標本平均**が母平均の**不偏推定量**であるという事実だ．このことは，母平均の**点推定量**として標本平均を使っていると言い換えてもよい．

一方，知りたい母平均が確率的に見てどんな範囲に広がっているかを知ることも大切だ．そのためには**区間推定**を行うことが必要になる．

そのための理論的な枠組みはすでに第 6 章で取り上げた (p.86, 正規母集団)．それに従って，次の 3 つの**モデル**に分けて考えてこう[*4]．

最初は，母分散 σ^2 が知られているケースである．この場合には，式 (7.1)，(7.2) の関係をそのまま利用して，標本平均 $\overline{X}$ の広がり，つまり分布を知ることができる．最も単純なモデルである．

しかし，母分散が分かっているのに母平均が分からないなどという幸運な，あるいは虫のよい状況はそうそうあるものではない．そんなときでも，標本のサイズが大きければ，$\overline{X}$ が正規分布していることが**中心極限定理**から期待できる．また，84 ページの式 (6.5) を使って標本分散から母分散を推定することができるので，それを使って標本平均の広がりを知ることが可能である．

しかし，標本のサイズが小さい場合には，$\overline{X}$ が正規分布をするという想定はできなくなるので，上記の方法はもはや適用できなくなる．その場合にも一定の条件の下で Student の t-分布を仮定することで，$\overline{X}$ の広がりを考えることができることになる．

7.3.1　母分散が既知の場合

このケースは 6.4.2 節 (p.87) で検討してある．すなわち，$\overline{X}$ を次のように標準化変換したとすると，

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \tag{7.5}$$

式 (7.5) で定義される Z は標準正規分布に従うことになる．

[*4] ここで言っている「モデル」というのは，考えている状況をどのように数学的な表現で記述できるかという事を指す言葉である．

仮に 90% の信頼区間を求めるとしよう．このときには 95% 点を考えればよいことになるが，表 7.1 から，それが 1.645 であることがわかる．結局，上の Z の値が ±1.645 になるような 2 つの $\overline{X}$ の間が，求める信頼区間だということになる．

そこで，式 (7.5) で $Z = \pm1.645$ と置いて $\overline{X}$ を求めると，信頼区間の両端が得られ，次のように信頼区間が決まる．

$$\left[\overline{X} - 1.645 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right] \tag{7.6}$$

一般化して，考えるべきパーセント点を λ として信頼区間を書き直そう．

$$\left[\overline{X} - \lambda \times \frac{\sigma}{\sqrt{n}}, \overline{X} + \lambda \times \frac{\sigma}{\sqrt{n}}\right] \tag{7.7}$$

この式から次のことが言える．

母分散 $\boldsymbol{\sigma^2}$ が既知の場合の母平均の信頼区間は，標本平均 $\overline{\boldsymbol{X}}$ を中心として，前後に標準誤差 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ の $\boldsymbol{\lambda}$ 倍の幅をもつ．

7.3.2　母分散が未知の場合—大標本

母分散が未知で，標本のサイズが十分に大きい場合については，88 ページで扱った．結果を再度書くと，次の式で与えられる Z は標準正規分布 $N[0,1]$ に従う．

$$Z = \frac{\sqrt{n-1}(\overline{X} - \mu)}{s}$$

考えるべきパーセント点を λ とすると，ここでも次のようにして信頼区間が表される．

$$\left[\overline{X} - \lambda \times \frac{s}{\sqrt{n-1}},\ \overline{X} + \lambda \times \frac{s}{\sqrt{n-1}}\right] \tag{7.8}$$

もしも信頼度を 90% に取るならば，95% 点 1.645 を λ として信頼区間を計算すればよい．

7.3.3　母分散が未知な小標本　— Student の t-分布

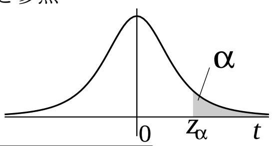
■Student の t-分布を使う

小さな標本の場合には，母分散を標本分散から直接に推定して，それから母集団の平均を区間推定する．これについては，89 ページ以降で扱った．それによれば，次の式で定義される T は，自由度 $n-1$ の t-分布に従う．

$$T = \frac{\sqrt{n-1}(\overline{X}-\mu)}{s} \tag{7.9}$$

注意しておくと,，t-分布は標本のサイズによって分布の形が異なるので，パーセント点も正規分布とは異なった表にまとめられることになる (表 7.2).

表 7.2　t-分布の形とパーセント点：α の意味は図を参照



パーセント	90	95	97.5	99	99.5	99.75
α	0.10	0.05	0.025	0.01	0.005	0.0025
$\nu=1$	3.078	6.314	12.706	31.821	63.657	
$\nu=2$	1.886	2.920	4.303	6.965	9.925	14.089
$\nu=5$	1.476	2.015	2.571	3.365	4.032	4.773
$\nu=6$	1.440	1.943	2.447	3.143	3.707	4.317
$\nu=7$	1.415	1.895	2.365	2.998	3.499	4.029
$\nu=8$	1.397	1.860	2.306	2.896	3.355	3.832
$\nu=9$	1.383	1.833	2.262	2.821	3.250	3.690
$\nu=10$	1.372	1.812	2.228	2.764	3.169	3.581

さて，式 (7.10) の形を見ると，式 (7.7) や式 (7.8) と全く同じである，つまり信頼度 90% で区間推定を行うのであれば，ここでも 95% 点を使って (正規分布ではなく t-分布なのでその値は異なることに注意) 区間の両端を求める．これまでと同様に，与えられた信頼度に相当するパーセント点を λ とすると，次のように信頼区間が得られる．λ の値は t-分布のパーセント点の表から探せばよい．

$$\left[\overline{X} - \lambda\frac{s}{\sqrt{n-1}},\ \overline{X} + \lambda\frac{s}{\sqrt{n-1}}\right] \tag{7.10}$$

いくつかの n の値に対する t-分布のパーセント点が表 7.2 に掲げてある．詳しい表は巻末付録に掲載してある．これを用いて，実際に推定を行うことができる．なお，表によっては異なる面積の取り方をしているものもあるが，その場合でも意味を考えれば，それほど混乱することはない．

7.3.4　平均値の区間推定の例

例題 7–1　平均値の区間推定

等級 L の卵の大きな箱から 10 個を抽出して，その質量 (g) を測定したところ，以下のようになった．

65.1, 67.5, 71.5, 68.4, 70.1, 72.2, 68.7, 69.3, 70.6, 67.1

この等級の卵の質量はほぼ正規分布しており，母分散は 4.0 g^2 であることが知られているものとする．この箱の卵の質量の平均値を 90% と 95% の信頼区間で推定せよ．

まず，これらから標本平均を計算すると 69.05，また題意より $n = 10$, $\sigma^2 = 4.0$ である．そこで 90% の信頼区間をとるときには，$\alpha = 0.05$ に相当するから，表 7.1 より，$z_{\alpha/2} = z_{0.05} = 1.645$ を読み取る．したがって式 (7.6) から，$69.05 - 1.645 \times \sqrt{4.0}/\sqrt{10} = 68.0$ と $69.05 + 1.645 \times \sqrt{4.0}/\sqrt{10} = 70.1$ が信頼区間の両端である．

すなわち，この卵の母平均 μ の信頼区間は，90% の信頼水準で $68.0 < \mu < 70.1$ と推定される．また信頼水準を 95% にとったときには，同様の計算により，信頼区間は $67.8 < \mu < 70.3$ となる．

例題 7–2　大標本の平均値の区間推定

ある県の高校生から 40 人を無作為に選んで，体重を測定したところ，標本平均 $\overline{X}$ と標本分散 s^2 はそれぞれ，53.8 kg, 18.81 kg^2 であった．これから母集団の平均値を 90% の信頼区間で推定せよ．

上の議論に従えば，信頼区間は

$$53.8 - 1.645 \times \frac{\sqrt{18.81}}{\sqrt{40-1}} = 52.6575..., \ 53.8 + 1.645 \times \frac{\sqrt{18.81}}{\sqrt{40-1}} = 54.9424...$$

となる．

答え：平均値は 90% の信頼水準で $52.7 < \mu < 54.9$ と推定される．

例題 7–3　小標本の平均値の区間推定

等級 L の卵の大きな箱から 10 個を抽出して，その質量 (g) を測定したところ，

65.1, 67.5, 71.5, 68.4, 70.1, 72.2, 68.7, 69.3, 70.6, 67.1

であった．箱全体の卵の質量は正規分布しているとして，この箱の卵の質量の平均値を 90% と 99% の信頼区間で推定せよ．

この問題は，105 ページの例題と同様に小標本のサンプルに関するものであるが，母分散は知られていない．母分散が分かっているケースというのはむしろまれであるから，この問題のほうが実際的である．

まず標本平均 $\overline{X}$ と標本標準偏差 s を求める．

$$\overline{X} = (65.1 + 67.5 + 71.5 + ... + 67.1)/10 = 69.05$$

$$s^2 = (65.1^2 + 67.5^2 + 71.5^2 + ... + 67.1^2)/10 - \overline{X}^2 = 4.1845$$

$$s = \sqrt{4.1845} = 2.045$$

また，自由度は $\nu = n - 1 = 9$ である．

90% の区間に含まれるためには $\nu = 9$ の時の t-分布の 95 パーセント点 $z_{0.05}$ を表から読みとって，1.833 を得る．したがって求める区間は，

$$\left[69.05 - 1.833 \times \frac{2.045}{\sqrt{9}}, 69.05 + 1.833 \times \frac{2.045}{\sqrt{9}}\right]$$

となって，平均値 μ は $67.8 < \mu < 70.3$ と推定される．

同様にして，信頼水準を 99% にとると，平均値は $66.8 < \mu < 71.3$ と推定される．確かめていただきたい．

7.3.5　平均値の推定のまとめ

■なぜ t-分布を使うのか

母分散 σ^2 が知られていない場合，標本平均 $\overline{X}$ の分散 $V[\overline{X}]$ が σ^2/n で与えられるというおなじみの関係式 (式 (6.4) 参照) は使えない．

そこで母分散の不偏推定量である不偏分散 $\frac{ns^2}{n-1}$ を利用して，$\overline{X}$ の分布の広がりを正規分布で扱おうというのが，7.3.2 節での大標本の取り扱いの基本的な考え方であった．

しかし，標本の大きさが小さくなると，中心極限定理によって標本平均 $\overline{X}$ が正規分布するとみなしてもよいという仮定はもはや成立しなくなる．その場合でも，**もし母集団が正規分布しているのであれば，**標本平均は母平均 μ にピークを持つ山形の分布をするであろう．

この場合，標本が小さいほど分布のすそは正規分布よりも広がるであろうことが直観的に予想される．また，標本が十分大きい極限では，その分布は正規分布に一致するようになるはずである．t-分布はそのような条件を満たすように作られている．このことは図 6.6 を見るとよく分かる．

7.3.6　区間推定におけるモデルの使い分け

ここまでの 3 つの場合の平均値の信頼区間の推定の手法について，表 7.3 にまとめた．ここで表の λ のパーセント点は，正規分布なら区間の幅だけで決まるが，t-分布の場合には標本の大きさ n に対して自由度として $n-1$ を選ぶ必要があることに注意しよう．

それでは，これらのモデルの使い分けについては，どういう基準を設ければよいのだろうか？

まず，大標本と小標本の区別は，標本のサイズ n がどのへんのところを境目にすればよいのだろうか．これについては，大体 $n = 20$ 程度を境にするということで実用的には問題ない．つまり Student の t-分布を使う目安としては抽出するデータの数が 20 程度以下としておけばよい．

また，十分な大きさの大標本の場合，標本不偏分散の値 $\frac{n}{n-1}s^2$ と標本分散 s^2 との比は，ほぼ 1 に等しいとしてもよい．この近似は n が 100 程度でも十分に成り立つ．

$$\frac{n}{n-1} \approx 1 \tag{7.11}$$

この時には，s^2 を母分散と同一してしまうこともできるので，数式の上では母分散が既知の場合と同じ扱いになる．

表 7.3　平均値推定のモデルの使い分け: λ はそれぞれの分布におけるパーセント点

	事前の知識	区間の幅	確率分布関数
大小標本	母分散既知	$\pm\lambda \times \frac{\sigma}{\sqrt{n}}$	正規分布
大標本	母分散未知	$\pm\lambda \times \frac{s}{\sqrt{n-1}}$	正規分布
小標本	母分散未知	$\pm\lambda \times \frac{s}{\sqrt{n-1}}$	t-分布

■3 つのモデルのつながり

ここまで述べてきた 3 通りの扱いについて，その関連を考えておきたい．

母集団から無作為抽出するデータの数が少ない場合，そこから得られる情報の信頼性が低いことはだれでも分かる．「クイズ 5 人に聞きました」という番組タイトルでは，ぜんぜん真実味を感じられないだろう．しかし，その場合でもデータから一定の情報は得たいということがしばしばある．標本平均が正規分布することが保証できないようなケースで，どうやったら定量的な信頼性のある結果を得ることができるのだろうか．

ビール会社で生産管理をしていたゴセットが突き当たったのはその問題だった．そこで彼は，母集団が正規分布している場合には，小標本から式 (6.8) で与えられる T を作ってやると，正規分布よりも信頼区間が広くなるような確率分布である t-分布が成立すること

を示して，小標本の統計的推定に道を開いた．

誤解してはいけないのだが，ゴセットがやったことは「少ないデータからでも信頼性の高い推定をできるようにした」というわけではない．むしろ「少ないデータから得られる結果は信頼性が低いけれど，低いなりの信頼性をきちんと明らかにした」と理解したほうがよい．

今，n 個，ただし n は数個程度のデータを，母集団からランダムサンプリングで得たとしよう．そこから計算して得られる標本分散

$$s^2 = \frac{1}{n}\left((x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2\right)$$

は，とても不安定な量であることに注意してほしい．つまり，標本平均 $\overline{X}$ は抽出のたびに大きく変動してしまう．ということは，標本平均を使って導かれる標本分散も，大きく変動することになる．

それでは，標本から得られる分散を真の平均である母平均 μ を使って次の式で定義される

$$\frac{1}{n}\left((x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2\right)$$

こういう「なんとか分散」を計算して使えばよいかというと，そもそも未知の μ を知りたくて推定の手続きを進めようとしているのだから，それが最初から使えれば苦労はない．まったくナンセンスなことになってしまう．

ただし，仮に母平均はわからないのに母分散だけがわかっているといううまい話があったとすると（すごいレアなケースだ！），そこを手がかりにして母平均を推定できる．

しかし普通はどちらの量も分かっていないだろう．その時でも，もしも標本のサイズが大きければ，標本平均は正規分布することが期待でき，かつ標本不偏分散をもって母分散として計算することができる．

そしてゴセットは，標本のサイズが小さい時に，ただし母集団が正規分布している場合に，標本不偏分散と，t-分布という特別な分布を持ち込むことで，問題を解決した．

【章末問題】

問題 7–1　100 人の有権者を無作為に選び，今の内閣を支持しているかどうかを尋ねたところ，支持率は 30% であった．この時の真の内閣支持率を 95% の信頼区間で推定しなさい．

問題 7–2　母集団が正規分布しているとして，そこから 12 個のデータを無作為抽出したところ標本平均が 16, 標本標準偏差が 4 であった．このとき，母平均 μ について 99% 信頼区間を推定しなさい．

問題 7–3　問題 7–2 で，データの数が 80 であったとし，他の量は同一であったとする．このとき，母平均 μ について 99% 信頼区間を推定しなさい．