2009 年度前期　データ解析序説　第１２回

分布の平均を推測する ー 区間推定

今回は，「標本調査にもとづく統計的推測」の手法のひとつとして，

> 「仮に日本人男性全員の身長を測ったとすれば，その平均は◯◯ cm 〜△△ cm の範囲にある」という推測が当たっている確率が 95%

という形式で，推測が当たっている確率まで答える「区間推定」を説明します．

区間推定

今回は，大量のデータの集まり，すなわち母集団から，いくつかのデータからなる標本を取り出し，母集団分布の平均，すなわち母平均を推測する方法を説明します．そのさい，母集団分布が正規分布で表されると仮定できる場合を考えます．

図 1(a) は，仮に何度も標本を取り出すとするときの，標本のばらつきを表したものです．母平均は，はじめからひとつに決まっています．一方，標本は無作為抽出されていますから，標本は毎回異なった値になります．このとき，標本がしたがう確率分布は母集団分布と同じですから，母平均が μ とすれば，標本の期待値も μ になります．すなわち，図 1(a) で，標本は μ を中心にしてまわりにばらついています．

さて，図 1(b) のように，標本の値の両側に幅をもたせます．そして，この幅に母平均が入っていれば，この標本は母平均に「近い」とみなすことにします．このとき，幅をある程度大きくすれば，「何度も取り出す標本のうち，95%は『母平均に近い』」ようにすることができます．すなわち，「ある幅をもたせれば，標本が『母平均に近い』確率は 95%」となります．

しかし，母集団分布の分散が大きい時は，標本の両側の幅もものすごく大きくなってしまうので，この幅の中に母平均が入っているとしても，到底「標本が『母平均に近い』」とはいえません．そこで，幅を狭める方法を考えましょう．

前回の講義の「正規分布の性質２」のところで，複数のデータを標本として取り出した時，これらの平均を標本平均とよぶことを説明しました．さらに，このとき標本平均の期待値はやはり母平均と同じであること，また標本として取り出されるデータの数，すなわち標本サイズを n とするとき，標本平均の分散は母分散の $1/n$ であることを説明しました．

この性質を使って，標本そのものではなく，標本平均の両側に幅をもたせてみたのが，図 1(c) です．標本平均 $\bar{X}$ も，標本そのものと同様に，母平均 μ を中心にしてばらついています．しかし，標本平均の分散は母分散の $1/n$ に縮んでいますから，標本平均の両側の幅を狭くしても，「ある幅をもたせれば，標本が『母平均に近い』確率は 95%」とすることができます．

この幅を**信頼区間**，信頼区間が母平均を含んでいる確率を**信頼係数**といい，この例では信頼係数が 95%なので **95%信頼区間**といいます．そして，信頼区間を，ある 1 回に標本抽出して計算した標本を用いて計算し，その信頼区間を推測結果として答えることを，**区間推定**といいます．したがって，区間推定の結論は，

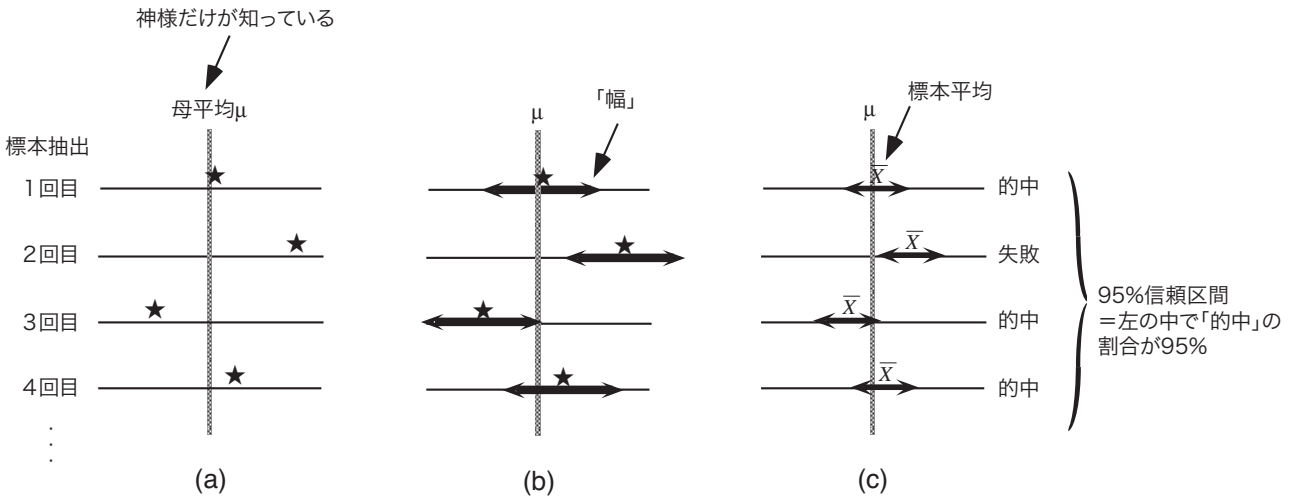

図 1: 区間推定の考え方

「母平均は，50 から 60 の間にあると推測する．この推測が当たっている確率は 95%である」

のように，母平均が入る区間を示し，さらにその推測が当たっている確率を示します．

標本サイズが大きければ大きいほど，標本平均の分散が小さくなります．図 1(c) での標本平均のばらつきが小さくなるわけですから，同じ信頼係数でも，信頼区間をより狭くすることができます．これは，「たくさんのデータを調べると，より精密な結果が出る」「たくさんくじをひけば，当たり確率はほぼ確実にわかる」という，ごくあたりまえの現象に相当しています．

台風情報では，区間推定のひとつの例を見ることができます（図 2）．テレビの画面に出ている予想進路図にある「予報円」は，区間推定によって描かれています．台風情報の「○○時に円内の範囲に達すると思われます」という予報は，「○○時に円内の範囲に達する確率が 70%である」ことを示しています．

正規分布の場合の，母平均の区間推定

では，母集団分布が正規分布モデルで表されると仮定されるときの，母平均の区間推定の方法を説明します．次の問題を考えてみましょう．

> ある試験の点数の分布は正規分布であるとします．この試験の受験者から 10 人からなる標本を無作為抽出して，この人たちの点数を平均したところ 50 点でした．この試験の受験者全体の標準偏差が 5 点であるとわかっているとき，受験者全体の平均点の 95%信頼区間を求めてください．
>
> 母集団の平均がわからないのに，母集団の標準偏差がわかっているというのはヘンな話ですが，これは説明のために用意した例です．正規分布が仮定でき，母集団の標準偏差が不明な場合については，次回の「t 分布」の項で説明します．

いまから推定する母平均を μ とし，母分散（こちらはすでにわかっているものとされています）を σ^2 とします．そうすると，母集団分布は平均 μ，分散 σ^2 の正規分布，すなわち $N(\mu, \sigma^2)$ となります．この

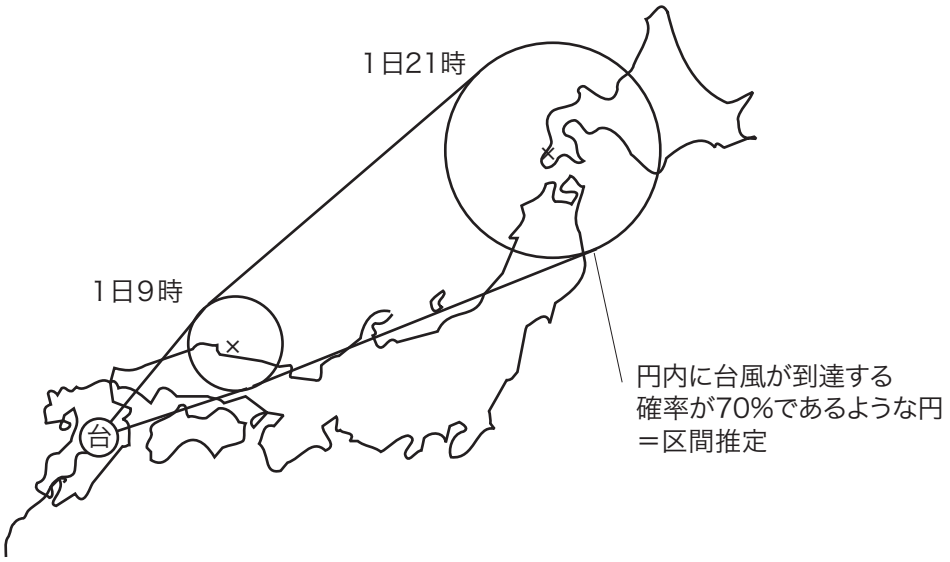

図 2: 台風情報と区間推定

とき，標本は無作為抽出されていますから，標本は確率変数で，母集団分布と同じ確率分布にしたがいます．すなわち，n 人からなる標本のそれぞれの確率分布（つまり標本分布）もまた $N(\mu, \sigma^2)$ です．

このとき，標本として取り出された n 人の点数の平均，すなわち標本平均を考え，$\bar{X}$ で表すことにします．n 人の点数を $X_1, \ldots, X_n$ で表すと，標本平均 $\bar{X}$ は $(X_1 + \cdots + X_n)/n$ で表されます．

前回説明した「正規分布の性質２」から，標本平均は $N(\mu, \sigma^2/n)$ にしたがうことがわかります．さらに，

さらに，

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \tag{1}$$

という値を計算すると，同じく前回説明した「正規分布の性質１」から，Z は標準正規分布 $N(0, 1)$ にしたがうことがわかります．

そこで，「Z が入っている確率が 95%である区間」はどういうものか考えてみましょう．第５回の講義の「連続型確率分布」のところで説明したように，Z がある区間に入る確率は，標準正規分布の確率密度関数のグラフの下の，その区間に対応する部分の面積になります．この部分の面積が全体の 95%になるように，左右対称に Z の区間をとることにし，図 3(a) のように表します．このときの Z の区間の両端を $-u$ と u とすると，Z がこの区間に入る確率すなわち $P(-u \le Z \le u) = 0.95$ となります．このとき，図 3(b) のように，$P(Z \ge u) = 0.025$ となります．$P(Z \ge u) = 0.025$ となる u は，正規分布の数表から求めることができます．数表によると，$u = 1.96$ のとき，$P(Z \ge 1.96) = 0.024998 \simeq 0.025$ であることがわかります．すなわち，$P(-1.96 \le Z \le 1.96) = 0.95$ ということがわかります．

ところで，(1) 式の関係を $P(-1.96 \le Z \le 1.96) = 0.95$ に用いると，

$$P(-1.96 \le \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \le 1.96) = 0.95 \tag{2}$$

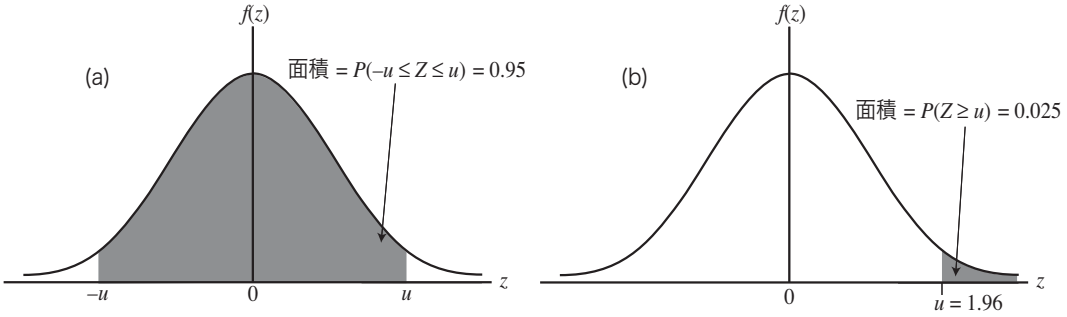

図 3: 95%信頼区間の求め方

という関係があることがわかります．ここで，今知りたいのは母集団の平均 μ の範囲ですから，(2) 式を μ の範囲に書き換えると

$$P(\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n}) = 0.95 \tag{3}$$

という関係が得られます．この範囲が，μ の 95%信頼区間となります．この問題では，標本平均 $\bar{X} = 50$，母集団の分散 $\sigma^2 = 25$ ですから，これらの数値を (3) 式に入れると，求める 95%信頼区間は「46.9 以上 53.1 以下」となります．「46.9 以上 53.1 以下」という区間を，数学では [46.9, 53.1] と書きます．

「95%信頼区間」の真の意味

前節で，母平均 μ の区間推定の結果を「求める 95%信頼区間は『46.9 以上 53.1 以下』」と書き，$P(46.9 \leqq \mu \leqq 53.1)$ とは書きませんでした．それは，**この書き方は間違い**だからです．

$P(\,)$ は，「( ) の中のことが起きる確率」という意味ですから，( ) の中にはランダムに決まる数，すなわち確率変数が入っていなければなりません．母平均 μ は，標本を調べている人が知らないだけで，実際には調べる前から 1 つの値に決まっていますから，確率変数ではありません．ですから，$P(\bar{X} - 1.96\sqrt{\sigma^2/n} \leqq \mu \leqq \bar{X} + 1.96\sqrt{\sigma^2/n})$ という式では，ランダムなのは μ ではなく $\bar{X}$ であり，不等式の上限と下限がランダムに決まることを示しています．

ところが，具体的な数値を計算して，$P(46.9 \leqq \mu \leqq 53.1)$ という式にしてしまうと，この式には確率変数がありません．したがって，この式は間違いです．具体的な数値で表された [46.9, 53.1] という信頼区間は，「いま無作為抽出された標本によって，偶然決まった標本平均 $\bar{X}$ の値を用いて，偶然そうなった値」です．母平均 μ は，[46.9, 53.1] という区間に「入っているか，入っていないかどちらかに決まっている」のであって，95%の確率で入っているわけではありません．

「母平均が入っている確率が 95%であるような区間」とは，今日の冒頭の図 1(b) で示したように，「標本を取り出して計算し信頼区間を求める」という操作を何回も行うと，

100 回あたり 95 回は，求めた信頼区間の中に確かに母平均が入っているが
残り 5 回は，求めた信頼区間の中に母平均は入っていない

となるような計算，つまり「95%の確率で当たるような，推測のやりかた」を意味しているのです．

現実には，標本を取り出して計算するのは 1 回だけです．ですから，その時にたまたま取り出された標本から計算された信頼区間，例えば [46.9, 53.1] には，母平均は入っていないかもしれません．

「1回の，信頼係数95%の推測を信じる」ことは，ある人の言っていることについて「この人が今回言っていることは本当かどうかわからないが，この人は95%の確率で本当のことを言うらしいから，今回も信じることにしよう」というのと同じです．

区間推定に関する注意

［1］ここまでの区間推定の説明では，95%信頼区間を求めました．信頼係数としては95%が一番よく使われますが，**信頼係数として95%という値を選ぶ根拠は何もありません．**「95%の確率で当っている推測」とは，「5%の確率ではずれている推測」でもありますから，信頼係数を95%とすることは，「5%くらいの確率なら，推測がはずれて失敗しても，まあいいか」と考えていることになります．また，信頼係数を例えば99%（この値も95%の次によく用いられます）にすると，図3から明らかなように，95%の場合よりも信頼区間の幅は広くなります．信頼区間の幅が広い，とは，推測のあいまいさが大きい，ということですから，場合によっては意味のある推定ができなくなってしまうこともあります．台風情報が信頼係数70%を用いているのは，台風の進路の予測は不確定の要素が多いため，信頼係数に95%や99%を使うと，予報円の範囲が広すぎて予報にならないからです．

［2］区間推定においては，**母集団の大きさは信頼区間の幅には影響しない**ことに注意してください．今回の例題でも，標本サイズが10人という条件が同じであれば，この試験の受験者全体の人数が1000人でも10万人でも，信頼区間の幅は同じです．つまり，「信頼区間の幅は，標本の**サイズそのもの**で決まり，標本サイズの母集団の大きさに対する**割合**には無関係」ということです．

「10人からなる標本」は，「1000人のうちの10人」であっても「10万人のうちの10人」であってもその価値は同じ，というのは一見不思議ですが，これは，「母集団のどの人も同じチャンスで選ばれ，しかも，ある人が選ばれるかどうかは，他の人が選ばれるかどうかには影響をうけない」という理想的な無作為抽出が，第5回の講義で説明した「復元抽出」であることに理由があります．

復元抽出の場合，「ある値のデータが標本として取り出される確率＝その値のデータが母集団中で占める**割合**」という，ここまでの講義で説明した原理が，抽出の順序によらずなりたちます．「割合」は，母集団の大きさには無関係です．したがって，その標本から計算される区間推定の結果も，母集団の大きさには無関係です．

一方，非復元抽出の場合は，標本を抽出するたびに母集団全体の人数が減ってゆきますから，「ある値のデータが標本として取り出される確率＝その値のデータが母集団中で占める割合」が，抽出の途中でだんだん変化してゆきます．この変化のしかたは，母集団のサイズに影響されます．したがって，区間推定の結果も，母集団の大きさに影響されます．この違いは，母集団が大きければさして問題になりませんが，そうでなければ，非復元抽出においては計算で補正をする必要があります．