2008 年度後期　情報統計学　第９回

区間推定

区間推定

今日の講義では，母集団分布が正規分布であると仮定できるとき，標本平均を使って母平均を推測する方法を考えます．

図 1(a) は，仮に何度も標本平均を計算したとするときの，標本平均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，標本平均は毎回異なった値になります．第４回の講義で説明したように，標本平均の期待値は母平均と同じですから，図のように，標本平均は母平均のまわりにばらついていることになります．標本サイズが大きくなると，やはり第４回の講義で説明したように，ばらつきは小さくなりますが，標本平均が母平均そのものでないことは変わりません．

そこで，図 1(b) のように，標本平均のまわりに余裕をもたせて，例えば「母平均は，標本平均 ±10 の範囲に入っているだろう」というように推測します．このようにすると，標本平均はばらついていますが，母平均がその範囲の中に入っている確率は大きくなります．

そこで，標本平均のまわりにどのくらいの余裕をもたせれば，その範囲の中に母平均が入っている確率がいくらになるかを，母集団分布をあらわす確率分布モデルを仮定したうえで計算します．この方法で，

「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」

というように，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．この方法を**区間推定**といい，「当たっている確率が 95%である」ような母平均の値の範囲（ここでは 50 〜 60）を**95%信頼区間**といいます．またこの「当っている確率」（ここでは 95%）を**信頼係数**といいます．

正規分布の場合の，母平均の区間推定

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，第１５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しています）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．標本は無作為抽出されていますから，たとえ今はある値のデータが抽出されているとしても，他の値になる可能性もあったわけです．すなわち，標本は「ランダム」な数値（確率変数）です．しかし，今日の前半で述べたように，標本は母集団分布と同じ確率分布にしたがいますから，それぞれの標本の確率分布もまた $N(\mu, \sigma^2)$ です．

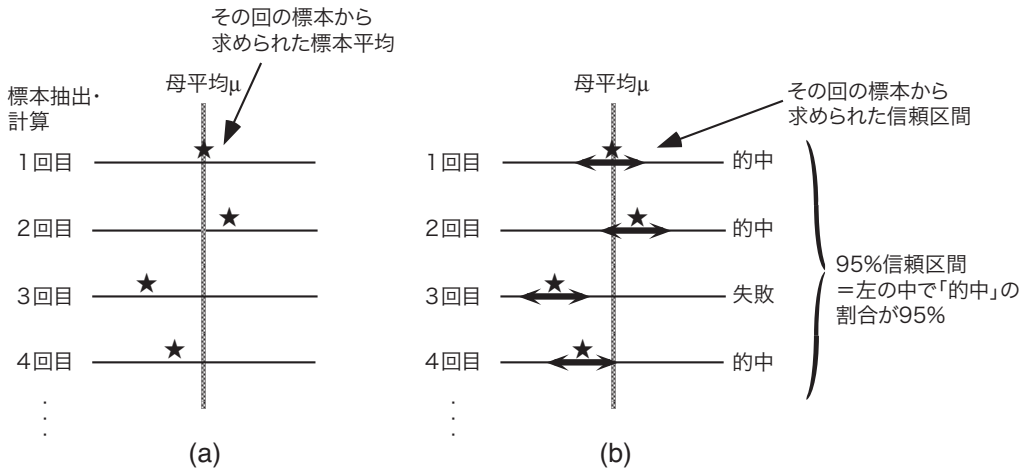

図 1: 区間推定の考え方

このとき，n 人の標本の平均（標本平均といい，$\bar{X}$ で表します）を考えてみましょう．n 人の標本を $X_1, \ldots, X_n$ で表すと，標本平均 $\bar{X}$ は $(X_1 + \cdots + X_n)/n$ で表されます．

ここで，第 6 回の講義で述べた「正規分布の再生性」，すなわち

> $X_1, ..., X_n$ が独立で，いずれも正規分布 $N(\mu, \sigma^2)$ にしたがうならば，それらの平均 $(X_1 + ... + X_n)/n$ は $N(\mu, \sigma^2/n)$ にしたがう

を用います．標本平均の分散が母分散の $1/n$ になる理由は，第 4 回の講義で述べました．「正規分布の再生性」は，さらに，標本平均もまた正規分布にしたがうことを述べています．

さらに，

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

という値を計算すると，第 6 回の講義で説明した正規分布の性質のひとつから，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 6 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 2(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 2(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，

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

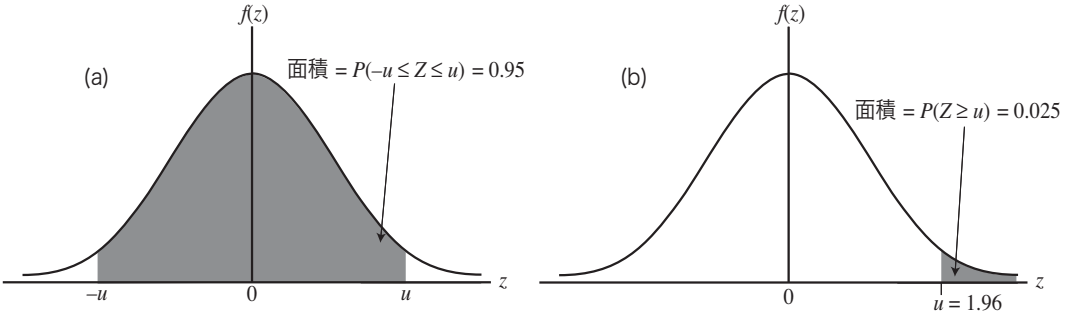

図 2: 95%信頼区間の求め方

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2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．

［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．しかし，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，**復元抽出**であることに理由があります．復元抽出とは，標本としていくつかのデータを取り出すときは，まずひとつのデータを取り出した後に，そのデータを母集団に戻してから，あらためて次のデータを選ぶ，という方法です．

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，一度取り出したデータは母集団にはもどさないという**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．

なお，復元抽出をしている場合は，この理屈でいえば，母集団の大きさが「無限」であっても，本質的違いはありません．これは，「600本中100本の当たりを含むくじを作り，そこから1本ひく」というくじびきと，「さいころを1回ふって，1の目が出たら当たり」というくじびきが，理想的なくじびきであればまったく同じ意味である，というのと同じです．前者は，母集団に含まれるくじの本数は600本です．後者は，さいころはいくらでもふることができますから，母集団の大きさが無限の場合に相当します．