2011 年度前期　統計データ解析Ａ　第１０回

「分布」するデータを扱う (1) ー 標本調査と区間推定

第８回では，この講義の最初に述べた「くじびきの問題」に対する統計学の答えとして，検定の手法を説明しました。しかし，ここまでこの講義を聴いて，「この講義は，統計学を扱っている割には，『データの処理』がちっとも出てこない」と思っている人もいるのではないでしょうか。

実は，当たり確率に関する統計的推測の考え方は，

> 日本人男性 100 人の身長を調べた。このデータから，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何 cm くらいになるか，を推測する

といった，「標本調査にもとづく統計的推測」の手法に使えるのです。今日はこのことを説明し，さらに

> 「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。

くじびきと標本調査

第６回の講義で，「測定誤差」を例にとって，連続する数値の場合の確率分布を説明しました。そのとき，例えば「誤差が 0.0cm から 0.1cm である確率が 50%」というように，連続する数値をある幅ずつに区切って，数値がその区切りに入っている確率を表すという形で，確率分布を表現しました。

ではここで，仮に，日本人男性全体の中での身長 170 〜 175cm の人の割合が 20%だとしましょう。日本人男性全体から，あるひとりの人を「公正なくじびき」で選んだとき，その人の身長が 170 〜 175cm である確率は 20%です。

当たり前のようですが，それはなぜでしょうか？ 確率の「ラプラスの定義」（第２回）を思い出してください。さいころの１の目が出る確率が 1/6 とされているのは，どの目も出るチャンスは同じなので，１の目が出る確率は「6 通りのうちの１通り」だからです。

同様に，日本人男性から公正なくじびきでひとりを選ぶときも，日本人男性のどの人も選ばれるチャンスは同じなので，「170 〜 175cm の人が選ばれる確率」は「170 〜 175cm の人の割合」と同じになります。

さて，「165 〜 170cm」「170 〜 175cm」「175 〜 180cm」を階級と考えると，各階級に入る人が選ばれる確率を求めれば，それは冒頭で述べた確率分布と考えることができます。これらの確率は，上と同じように考えると，それぞれの階級に入る人の割合と同じになります。このような「公正なくじびき」を**無作為抽出**といい，選ばれた人を**標本**といいます。

「165 〜 170cm」「170 〜 175cm」「175 〜 180cm」といった各階級に入る人の割合を，確率分布と同じように並べたものを**度数分布**といいます。つまり，

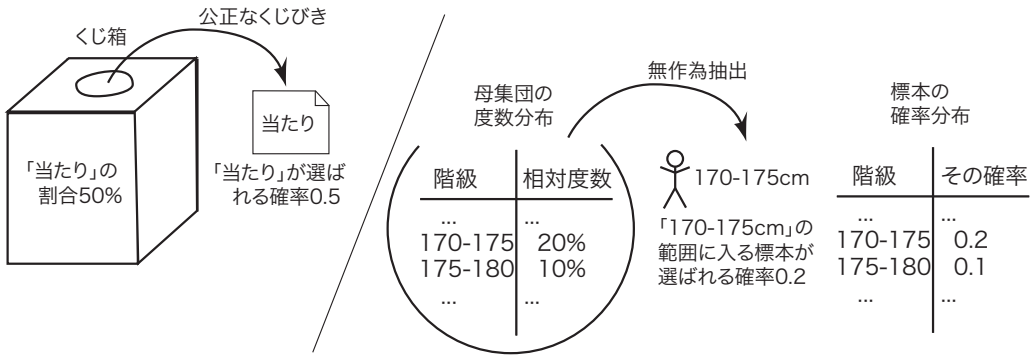

図 1: 度数分布と確率分布

あるデータの集まりの度数分布 ＝ データの集まりから標本を無作為抽出したときの，標本がしたがう確率分布

となります。

度数分布の推定

前節で述べたことは，データの集まりから標本を取り出すことによって，標本の確率分布がわかれば，データの集まりの度数分布がわかるので，それらのデータがどんなようすかがわかる，ということを示しています。

例えば，確率分布の期待値がわかれば，その確率分布を度数分布と考えることによって，「真ん中あたりのデータ」がいくらかがわかります。確率分布の期待値は，度数分布では**平均**といいます。また，確率分布の分散がわかれば，その確率分布を度数分布と考えることによって，それらのデータがどんなふうにばらついているかがわかります。確率分布の分散は，度数分布でもやはり分散といいます。

そこで，ある集団の度数分布が，正規分布モデルで表せるとしましょう。中心極限定理のところで説明したように，正規分布モデルで表せる分布は世の中にたくさんあります。そうすると，その集団から標本を無作為抽出すると，その標本がしたがう確率分布は正規分布モデルで表されます。

とはいうものの，標本としてデータをひとつだけ取り出しても，標本の確率分布を推定することはできません。しかし，データをある程度の個数取り出せば，確率分布がおおまかにわかってきます。くじを 1 本ひいても当たり確率はわかりませんが，くじを 100 本ひけば，当たり確率が大まかにわかるのと同じです。

今回は，このことを使って，大量のデータの集まり（**母集団**といいます）から，いくつかのデータからなる標本を取り出し，母集団の度数分布（**母集団分布**といいます），とくにその平均（**母平均**といいます）を推測する方法を説明します。

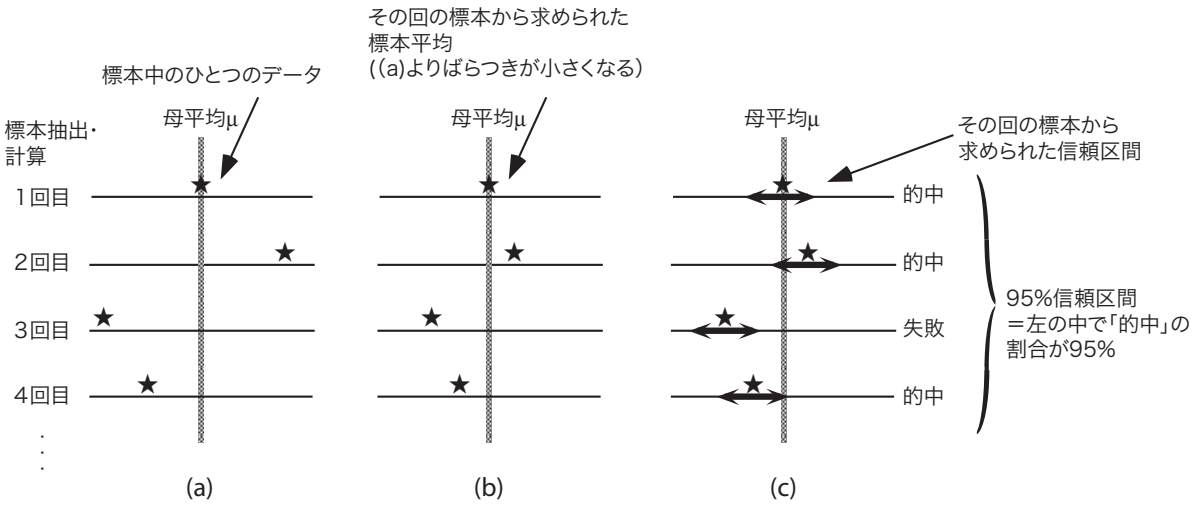

図 2: 区間推定の考え方

区間推定

今日の講義では，母集団分布が正規分布で表されると仮定できるとき，母平均を標本から推測する方法を考えます。

母集団全体のデータではなく，ほんのいくつかのデータからなる標本だけで母平均を推測するのですから，母平均の値については不確かなことしか言えません。ただ，標本として取り出したデータだけの平均，すなわち**標本平均**は，なんとなく母平均に似た値であるとは想像できます。

図 2(a) は，標本としてひとつのデータを取り出した時の，データの値を示しています。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，標本として取り出されたデータは毎回異なった値になります。図 2(b) は，標本としていくつかのデータを取り出した場合，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。標本として取り出したデータの数，すなわち**標本サイズ**[1]が大きくなると，(a) の場合に比べて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 2(c) のように，標本平均のまわりに「幅」をもたせて，例えば「母平均は，標本平均プラスマイナスいくら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。しかも，標本平均を用いた (b) の場合，ひとつのデータを用いた (a) の場合に比べてばらつきが小さくなっているので，「幅」は狭くできます。

そこで，確率分布モデルを使って，標本平均のまわりにどのくらいの「幅」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」

[1]この例のように，ひとつの母集団から n 人の標本を取り出したとき，**「標本サイズ（標本の大きさ）が n である」**といいます。なお，「標本の数」というと，統計学では「標本のセットの数」を意味します。例えば，2つの母集団からそれぞれ n 人の標本を取り出したときは，標本の数は 2 で，標本サイズはどちらも n です。

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。

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。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。

ここで，正規分布のもうひとつの重要な性質を用います。それは，

> $X_1, ..., X_n$ 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

というものです。母集団分布にくらべて，標本平均の確率分布では，分散が $1/n$ になっています。標本はランダムですから，標本となるデータをいくつか集めて計算した標本平均もまたランダムです。しかし，標本の各データにたとえ極端に母平均からかけはなれた値があっても，いくつか集めて平均すると，その極端な値は相殺されてしまいます。ですから，標本平均は，極端な値になる可能性が小さくなります（図 3)。これが，標本平均の分散が母分散より小さくなる理由です[2]。

ここでいう n，すなわち標本サイズが大きくなると，標本平均の分散が小さくなります。ということは，母集団の分散が大きくても，データをたくさん集めれば，標本平均はそうばらついた値になる可能性は少ないので，いま集めた標本から計算した標本平均も，かなり信用できる，ということになります。このことは，「くじを 1 本だけひいても当たる確率は全くわからないが，くじをたくさんひけば当たる確率はだいたいわかってくる」という，先に述べたごく当たり前の事実に対応しています。

[2] それがなぜ $1/n$ になるのかは，浅野の講義「情報統計学」(2008 年度後期) の第 4 回の講義録をネットで参照してください。

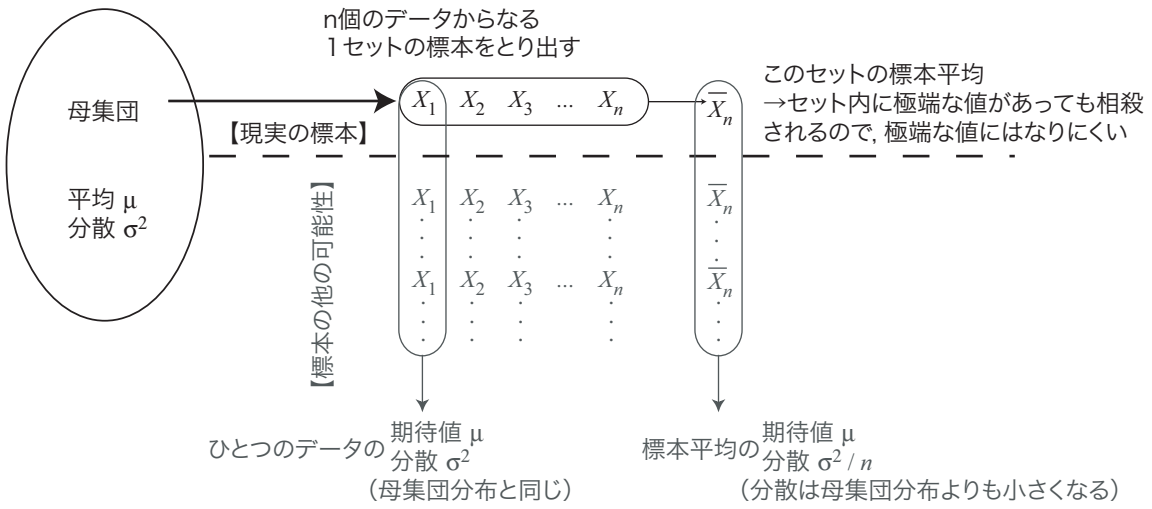

図 3: 標本平均のしたがう確率分布

この性質を，この講義では「正規分布の性質２」とよぶことにしましょう[3]。この性質により，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，

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

という値を計算すると，第６回の講義プリントにある「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 4(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 $P(-u \leqq Z \leqq u) = 0.95$ となります。このとき，図 4(b) のように，$P(Z \geq u) = 0.025$ となります。$P(Z \geq u) = 0.025$ 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，$P(Z \geq 1.96) = 0.024998 \simeq 0.025$ であることがわかります。すなわち，$P(-1.96 \leqq Z \leqq 1.96) = 0.95$ ということがわかります。

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

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

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

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

[3]標本平均の確率分布がやはり正規分布になることを**正規分布の再生性**といいます。証明は，浅野の講義「情報統計学」(2008 年度後期) 第６回の講義録（付録）をネットで参照してください。

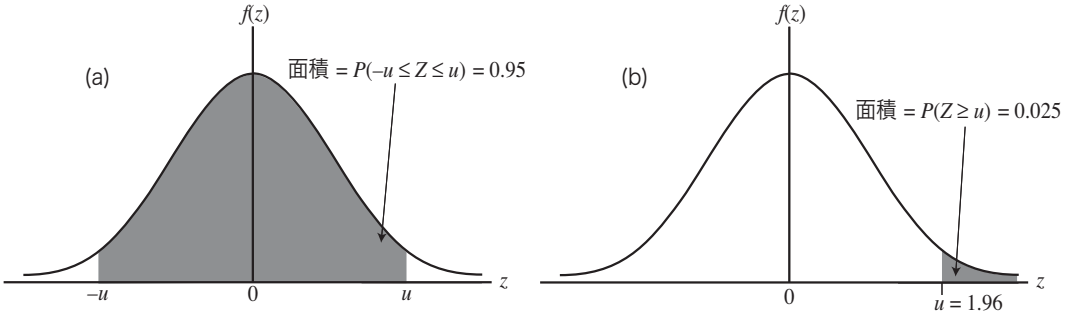

図 4: 95%信頼区間の求め方

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 2(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4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。

今日の演習

ある製品の長さを10回測定したとき，10個の測定値の平均は10.0(cm)でした。この測定では測定値の標準偏差が0.1(cm)であることがわかっており，測定値が真の長さを期待値とする正規分布にしたがうとするとき，真の長さの95%信頼区間を求めてください。また，10.0(cm)というのが20個の測定値の平均であるときは，95%信頼区間はどうなりますか。