シリーズ・技術調査報告

「統計じかけのオレンジ」
―A Statistics-work Orange―
第１回　平均、分散、標準偏差

一般社団法人 日本下水道施設業協会
技術部長 堅　田　智　洋

１．はじめに

統計学は記述統計学と推測統計学の２つに大きく分類することができる。

記述統計学は、調査や実験等で集めたデータを、整理、表やグラフ化、数値化することでイメージ的、直感的に理解しようとするものである。一方、推測統計学は、統計学の手法と確率理論を使って、「全体の把握が難しいくらい大きな対象」や「まだ起きていない､未来に起きること」といった「全体像」を、その時点で得られている「部分的な結果」から推測しようとするものである。

今回、筆者は、推測統計学のほんのさわりの部分を理解できた範囲でまとめてみた。ただ、不正確な記述も散見されると思われるため、実務への活用の際は、他の専門書で正確な情報を確認してください。

２．さまざまな統計量

２．１　統計量とは

統計量とは、記述統計学において、あるデータの特徴を１つの数値に要約して表現したものであり、よく知られているものに、平均値、分散、標準偏差、中央値、最頻値、最大値、最小値がある。ここでは、その中でも推測統計学においても重要となる平均値、分散、標準偏差を概説する。

２．２　平均値

平均値は「データを合計したものをデータ数で割った値」である。これは「算術平均」と呼ばれるもので、例えば、２つの数ｘとｙの算術平均は$\frac{x+y}{2}$で求まる。

平均値の意味合い、性質として重要なポイントを以下に挙げる。

・データをヒストグラムで表すと、平均値はヒストグラムをやじろべえとみなした時のつり合いの支点となる。
・データは平均値の周辺に分布している。
・多く現れるデータの平均値への影響力が大きい。

※ヒストグラム：データの分布状況を視覚的に表現するため、測定値の存在する範囲をいくつかの区分に分け、各区分を底辺としてその区間に属する測定値の出現（相対）度数に比例する面積をもつ長方形を並べた図。

実は、平均の求め方はこの「算術平均」のほかに「相乗平均（あるいは幾何平均)」、「二乗平均」、「調和平均」等がある。「二乗平均」は各データを２乗して合計し個数で割ってその後ルートにしたもので、２つの数ｘとｙの二乗平均は$\sqrt{\frac{x^2+y^2}{2}}$である。この「二乗平均」の手順は後で標準偏差のところでも出てくるので、算術平均と対比して示しておく（図２－１)。

二乗平均は、各データを最初に二乗してから算術平均を行い、最後に、最初の二乗操作を戻すためにルートを行うのだ、と解釈するとわかりやすいかもしれない。

他の平均の説明は割愛するが、いずれも「ｘとｙの間にある１つの数」を代表値として選び出していることは共通している。「ｘとｙを１つの数で代表するのにどの平均がふさわしいのか」は「そのデータ全体に関して何を知りたいのか」に依存して決まるため、用途に従って使い分けることになる。

本稿では、以降、特に断りがない限り「平均」は算術平均を指すこととする。

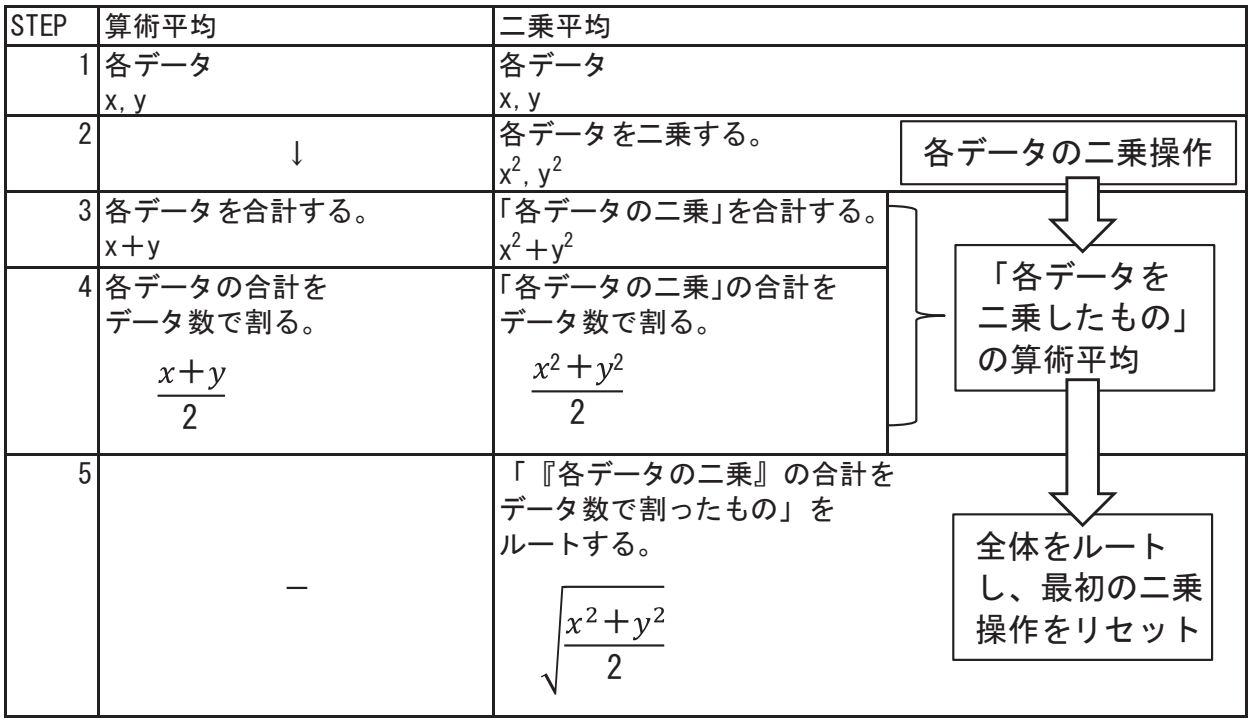

図２－１　算術平均と二乗平均の手順

２．３　分散、標準偏差

２．３．１　データのバラツキの重要性

私たちは、日常生活においては何かと平均値だけで物事を判断することも少なくないが、実際には平均だけで事象の様子がわかったというわけにはいかないことも多い。例えば､「ある歌手のファン層の平均年齢は25歳である。」という話を聞いたとする。それだけで、この歌手のファンには20歳代の若者が多いのだろうと早とちりしてはいけない。子供から高齢者まで幅広い年齢層から支持されていることが、平均年齢という統計量においては25歳という数値で表されたに過ぎない可能性もある。ファンの年齢のばらつきの情報は与えられていないのだから、年齢構成を推し測ることはできないのである。

上記は極端な例だとしても、データの特徴を把握するためには平均値だけでなく散らばりやバラツキを知ることが非常に重要である。小島は自身の著書[1)]で、そのことを「バスの運行状態」を例にとってわかりやすく説明しているので、以下に要約して紹介する。

ある目的地へ行くために、バスAとバスBのどちらかを利用しようとしている。バスAは、これから乗車しようとするバス停に、予定到着時刻に対して等確率で２分遅れたり２分早く来たりする。同じく、バスBは、等確率で10分遅れたり10分早く来たりする。この場合、バス停への到着時刻を平均値だけで見る分には、どちらのバスも時刻表通りに運行しているバスと見なすことができ、甲乙つけられない。しかし、実際の到着時刻が予定時刻からどの程度前後するか（ばらつくか）を考慮すると、実際には、バスAを選択するのではないだろうか。上記のバスAにおける「２分」とバスBにおける「10分」がダイヤの乱れ、つまりバス停到着時刻のバラツキ、散らばり具合を表している統計量だと考えることができる。そして、利用するバスの選択には、平均値よりこの散らばり具合を知ることの方が重要だということになる。

２．３．２　分散、標準偏差の算出方法

前述したデータの散らばり具合を示す統計量が、分散と標準偏差である。その意味と算出方法を、A高校１年の５人の女子の体重（kg）を例[2)]に以下に示す。

その前にまず､「偏差」という用語がある。偏差とは各データから平均値を引いて得られる値で、

番号	体重(kg)
1	51
2	49
3	50
4	57
5	43

平均値の算出

$$\frac{51 + 49 + 50 + 57 + 43}{5} = 50\ (\mathrm{kg})$$

図２－２　平均値の算出

番号	体重(kg)	偏差(kg)
1	51	51 － 50 ＝ 1
2	49	49 － 50 ＝ −1
3	50	50 － 50 ＝ 0
4	57	57 － 50 ＝ 7
5	43	43 － 50 ＝ −7

図２－３　各データの偏差の算出

各データが平均値からどれだけ離れているかを表す（図２－２、図２－３）。

偏差＝データの数値－平均値　　…（式２－１）

私たちは、調査から得た複数のデータを分析する場合、個々のデータの偏差がどのようなものかということよりも、データ全体の傾向としての散らばり具合を知りたいことの方が多いだろう。そこで、図２－３で求めた全データの偏差を平均すればいいと考えるのだが、図２－４で示すように各データの偏差を単純に算術平均するとゼロになってしまう。そもそも、平均値は、平均値より大きかったり小さかったりする各データを均（なら）したものなので、プラス、マイナスどちらの値も存在する偏差を算術平均すると打ち消しあってゼロになるのは当然である。

そこで、「『偏差を二乗したもの』の平均値」をとってみる。偏差を二乗してから平均を取る（二乗したものを合計し、データ数で割る）ことでプラス・マイナスが打ち消しあわないようにできるので、資料のバラツキの指標となりうる。この値が「分散」である（図２－４）。

番号	体重(kg)	偏差(kg)
1	51	51 － 50 ＝ 1
2	49	49 － 50 ＝ −1
3	50	50 － 50 ＝ 0
4	57	57 － 50 ＝ 7
5	43	43 － 50 ＝ −7

偏差の単純な算術平均値は0になってしまう。

$$\frac{1 + (-1) + 0 + 7 + (-7)}{5} = 0\ (\mathrm{kg})$$

⇩

そこで
偏差を二乗してから算術平均を行う。

$$\frac{1^2 + (-1)^2 + 0^2 + 7^2 (-7)^2}{5} = 20\ (\mathrm{kg}^2)$$

これが「分散」である。

図２－４　分散の算出

$$分散=\frac{(偏差の２乗)の合計}{データ数}\qquad\cdots(式２-２)$$

しかし、この「分散」には２つの問題がある。ひとつはバラツキを表す数値として大きすぎること、もうひとつは単位が変わってしまっている（kg→kg^2）ことである。この２つの問題はいずれも、偏差を二乗してから平均を取るために生ずる。そこで、それを解消するために、最後にこの分散のルートをとる。これが「標準偏差」である。

先述した「二乗平均」の定義から、「偏差の二乗」の平均値のルートをとると、「偏差の二乗平均」になるから、次式が得られる。

$$標準偏差=\sqrt{分散}=偏差の二乗平均\cdots(式２-３)$$

標準偏差は、「各データの偏差の平均値」であり、すなわち、データ全体の「散らばり具合」を示すものである。「各データの偏差（平均値からの離れ方）の二乗平均（二乗し、合計し、データ数で割り、ルートにしたもの）」であるから、「各データの偏差（平均値からの離れ方）の平均」を表しているのである。

例題では、標準偏差＝$\sqrt{20}$＝4.472≒4.5（kg）となり、各人の体重の平均体重（50kg）からのバラツキの標準的な幅が4.5kgであるとわかる。

＜次号に続く＞

【本稿の全体構成】

【参考文献】

１）完全独習　統計学入門　小島　寛之　2006年９月　ダイヤモンド社
２）まずはこの一冊から　意味がわかる統計解析　涌井　貞美　2013年２月　ベレ出版

シリーズ・技術調査報告

「統計じかけのオレンジ」―A Statistics-work Orange―　第２回　標準偏差、正規分布

一般社団法人 日本下水道施設業協会
技術部長 堅田 智洋

2.3.3　標準偏差の意味

標準偏差を用いることでデータの中のいったい何がわかるのだろう。これについては、２つの見方が考えられる。まず第一に、「１セットのデータの中のある１個のデータが持つ意味（特殊性）がわかる」ということ、そして第二に、「複数のデータのセットを比較した場合に出てくる違いがわかる」ということである。

第一の「１セットのデータの中のある１個のデータが持つ意味（特殊性）がわかる」事例を以下に示す[1)]。数学のテストで自分の結果が75点で平均点が60点だったとする。そのとき、自分はその結果をどれくらい喜んでいいものだろう。このとき知っておく必要があるのは「標準偏差が何点か」ということである。ここでの標準偏差は、テストを受けた人の点数と平均点との差（偏差）を全員分、二乗平均したものである。標準偏差が12点だったとする。そうなると、自分の点数は平均点より「おおよそ標準偏差1個分程度高い点数」だとわかる。標準偏差は「各データの平均値からの離れ方を平均化した値」であるから、自分の点数は「平均点から良い方に『普通程度に』離れた値」だということができる。そのような人は大勢いるだろうから、それほど大喜びすることではないといえるかもしれない。

標準偏差がもっと低く、８点だったとする。このとき、自分の点数は「平均的な離れ方である８点」の２倍近くも良い方向に離れていることから、その価値は標準偏差が12点の場合より大きいと評価できるだろう。

以上のように、「１セットのデータ中のある１個のデータが持つ意味（特殊性）」は、平均値との比較、つまり平均値からの離れ方（偏差）の数値自体だけでは評価できず、標準偏差を基準に相対的に見る必要があるということである。つまり、「そのデータの偏差が標準偏差いくつ分であるか」という視点が重要となってくる。

統計学の世界では、次のようなおおまかな基準が広く了解されているようである。すなわち、１セットのデータの中の「あるデータ」の偏差が、標準偏差±１個分であれば、そのデータは「月並みなデータ」、そして±２個分から外側のデータである場合は「特殊なデータ」だということである。

続いて、第二の「複数のデータのセットを比較した場合に出てくる違いがわかる」事例を以下に示す[1)]。例えば、X君は模擬テストを10回受けて平均点が60点で標準偏差が10点、Y君は同じ模擬テストを10回受けて平均点が50点、標準偏差が30点だったとする。このことから何が言えるだろう。

今度は、二人が持つ複数（10回）のデータ（模擬テストの点数）のセットを比較してみる。平均点をみれば、X君の点数の方がY君の点数より優秀である。しかし、これだけでこの２人の本番の受験の結果を見通すことはできない。「標準偏差１個分のバラツキは普通に起こりうる」ということから考えて、X君はおおよそ60±10＝50～70点の範囲の点数を取る生徒だと見積もることができる。同様に、Y君は50±30＝20～80点の範囲の点数を取る生徒だと推測できる。つまり、X君は「安定した成績」を取る人で、Y君は「ムラのある成績」を取るタイプだと言える。平均点だけでは評価しきれない側面である。例えば、X君は50点で

入れる学校ならきっと不合格になることはないだろうが、80点取らないと合格しない学校にはなかなか入れないだろう。それに対してY君は、40点で入れる学校にも不合格になる可能性がある反面、80点を要する学校にも合格のチャンスがある、ということもいえるのである。

以上のように、複数のデータセットを比較する際にも、平均値だけでなく標準偏差を考慮すれば別の見方の評価ができることがわかる。

ここまで、データの特徴を表現する「平均値」、「分散」、「標準偏差」という代表的な統計量を説明してきた。ところで、少し抽象的な表現になるが、データというものは世の中で起こった（あるいは将来起こる）さまざまな現象を観測（あるいは予測）することで得られるが、その現象を発現させるシステムの大部分は何かしらの「不確実性」を持っている。そのため、そこから生み出されるデータは判で押したようにぴったり同じ数値になることはほとんどなく、まちまちな値になるのが一般的である。「データがまちまちな数値をとる」ことを専門的には「分布する」という。分布が生じるのは、その数値が決まる背後に何らかの「不確実性」が働いているからであるが、「不確実性」と一口に言っても、そこには固有の「特徴」、「癖」がある。その固有の特徴、癖を専門的には「分布の特性」と呼ぶ。その「分布の特性」を、平均値と標準偏差を用いて「おおよそこの辺を中心にだいたいこれくらいの散らばり具合で存在（分布）する」と表現するわけである。

３．正規分布

３．１　正規分布の特徴

データの分布について触れたところで、データの分布でもっとも代表的なものと言われる「正規分布」を紹介する。平均値μ、分散をσ^2の正規分布を図３－１に示す。

このグラフは横軸を連続的な確率変数、縦軸を確率密度関数として正規分布をとったものである。確率変数とはいわばそのデータが取りうる数値のことである。それは、サイコロの目（1,2,3,4,5,6）のように飛び飛びの値のこともあるが、人の身長や流入下水のBOD値等、切れ目なくどんな

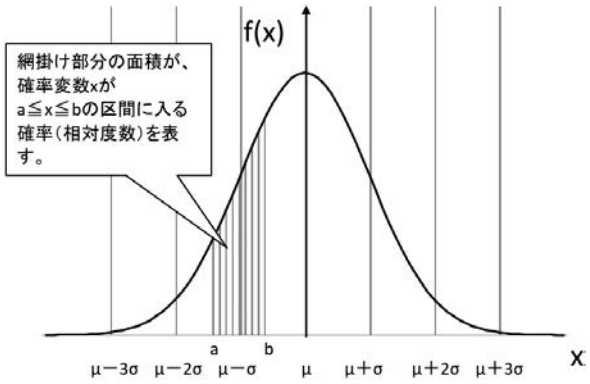

図３－１　平均値μ、分散をσ^2の正規分布

値でも取りうるものは「連続的な確率変数」という。「連続的な確率変数（xとする）に対する確率分布を関数（f(x)とする）で表現したもの」が確率密度関数である。この確率分布を表す確率密度関数のグラフでは、横軸とグラフとで囲まれる部分の面積が確率を表す。例えば図４－１で確率変数xが区間a≦x≦bに入る確率は図４－１に示す「網掛け部分の面積」となる。xの値から算出されるf(x)の値自体が確率を示すのではない。したがって、xがこのグラフの左端から右端までのすべての範囲の確率（横軸とグラフとで囲まれる部分の面積）は１（100％）となる。

正規分布はこのような確率密度関数の中でも最も有名なものである。その特徴は左右対称な釣り鐘型の分布となっていることである。まず平均値が一番高く、その近辺にもデータが集中して存在し（曲線が盛り上がり）、そこから左右に急激に減少していき平均値μから左右±２σを超えた辺りではf(x)の値は非常に小さくなる、すべり台のような曲線は正規分布の特徴を示す重要なポイントである。

正規分布は、自然現象や社会現象として観測されるデータに非常に頻繁に現れるものとされている。例えば、人間も含め生物の身長（体長）のデータは正規分布となることが知られている。また、株の収益率のデータも正規分布だと考える研究者は多いということである。

正規分布の姿を示す確率密度関数f(x)は数式で表すことができる。したがって、先述の区間a≦x≦bをどの範囲で指定しても、そこにデータが入る確率（相対度数）ははっきり決まる（算出

される)。例えば、その中でよく使われるものとして以下のものがある。

・平均から標準偏差±１個分の範囲（$\mu - \sigma \leqq x \leqq \mu + \sigma$）のデータの相対度数は0.6826（＝70％弱)。

・平均から標準偏差±２個分の範囲（$\mu - 2\sigma \leqq x \leqq \mu + 2\sigma$）のデータの相対度数は0.9544（＝95％強）（図３－２)。

このことは、前項で説明した「１セットのデータ中のある１個のデータが持つ意味（特殊性)」は、「そのデータの偏差が標準偏差いくつ分であるか」という観点から評価する必要があり、それが±１個分であればそのデータは月並みなデータで、±２個分から外側のデータであれば特殊なデータだ、ということを裏付けている。

３．２　標準正規分布

標準正規分布とは、平均＝０、標準偏差＝１の正規分布のことで、正規分布の中で最も基礎になるものである（図３－３)。あるデータxが平均値＝μ、標準偏差＝σの一般正規分布のデータであるとき、

$$z = (x - \mu) \div \sigma \qquad \cdots(式３－１)$$

という加工をすると、データzは標準正規分布のデータに変換される。

＜次号へ続く＞

【参考文献】

１）完全独習　統計学入門　小島　寛之　2006年９月　ダイヤモンド社

２）まずはこの一冊から　意味がわかる統計解析　涌井　貞美　2013年２月　ベレ出版

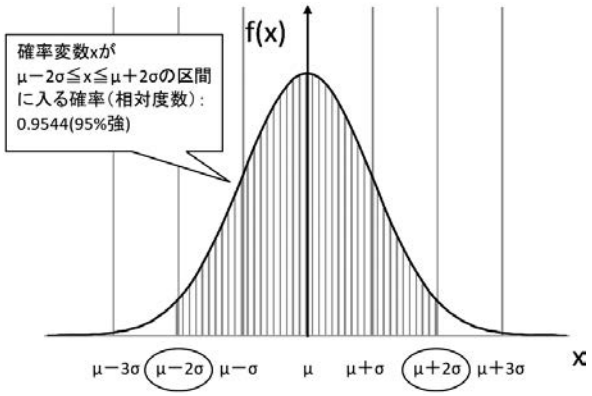

図３－２　平均値μ、分散をσ^2の正規分布

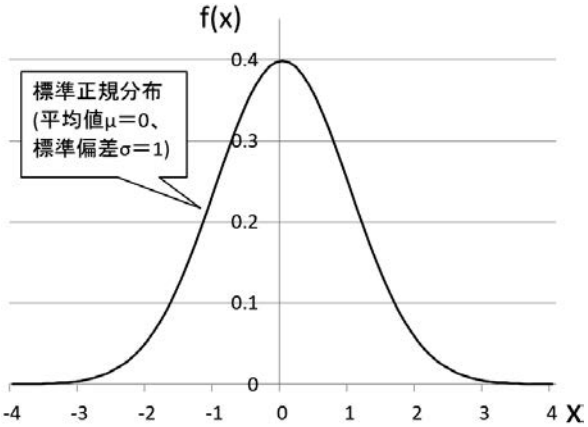

図３－３　標準正規分布

シリーズ・技術調査報告

「統計じかけのオレンジ」
―A Statistics-work Orange―
第３回　推定（１）

一般社団法人 日本下水道施設業協会
技術部長 堅　田　智　洋

４．推定

４.１　統計的推定とは

ある集団に対して何かしらの調査を行うとする。その調査対象となる集団のことを母集団と呼ぶ。その調査は、正確を期すなら対象全部に対して調査を行う「全数調査」が望ましいが、現実には、母集団の規模や費用等の事情から、母集団全部を調査し尽くすことは現実的に不可能なことが多い。そこで、その母集団から一部を選び出し（抽出という）調査することになる。このとき、調査対象として選び出された母集団の一部を「標本」と呼び、標本に対して行う調査を「標本調査」という。

標本調査では、母集団から標本を抽出し、その標本から母数（母集団の分布の特徴を示す、平均値や分散、標準偏差といった数値のこと）に関連する情報を得る。しかし、その調査結果は抽出する標本ごとに異なる。例えば、A市の20歳男子を母集団として、その平均身長を調べるためにその中から10人の男子を抽出したとする。その10の標本の変量（調査項目のこと。ここでは身長）をx_1、x_2、…x_{10}とすると、それらは母集団に従った確率変数（そのデータが取りうる数値）である。この10人の平均身長を$\bar{x}$とすると、

$$\bar{x}=\frac{x_1+x_2+\cdots+x_{10}}{10} \qquad \cdots（式４－１）$$

となる。このように、「標本から算出された平均値」を標本平均と呼ぶ。重要なことは、この標本平均（ここでは抽出した10人の平均身長）自体も、抽出する標本ごとに値を変える確率変数であるということである。つまり、この10人の標本の平均身長は、抽出する10の標本のセットごとにバラツキが出る。それは、無作為に抽出される男子10人の顔触れがその都度変わるからである。ちなみに、この標本平均のように、母数の推定に利用される統計量をその母数の推定量と呼ぶ。標本平均は母平均の推定量である。そして、実際の標本から算出された推定量の値を推定値という。

以上のように、バラツキのある標本調査の結果から、元の母集団に関する真の値（母数）を推定しようというのが「推定（統計的推定）」である。つまり、部分的に明らかになった結果から全体を推測するのである。では、私たちはどのような母集団について、どのような仕組みで統計的推定を行うことになるのだろうか。

例えば、１個のデータxが現実に観測されたとして、そのことからxが属する母集団について何が言えるだろう。まず、「母平均μはこのxの近くだろう」程度の推量は可能であろう（あくまで推量だが）。なぜなら平均値とは分布の中から選ばれた大小いずれにも偏っていない代表的な点だからである。さらに、ここでもし仮に、母集団の標準偏差σが何らかの理由でわかっているならどうだろう。先程、１セットのデータの中のあるデータの偏差が標準偏差±１個分なら、そのデータは「月並みなデータ」で、±２個分から外側のデータである場合は「特殊なデータ」だと述べた。これは、たいがいのデータは平均値を中心にその前後標準偏差２個以内の範囲におさまることを意味する。つまり、「データxは母平均μから$\sigma\times２$程度以内の離れ方だろう」と考えることができる。逆に言えば「xから$\sigma\times２$程度以内の隔たりでμが存在するだろう」、つまり、観測されたデータは、ある程度母平均に近いものだとみてよい。このことは、母集団が正規分布の場合には強く支持される。

次に、観測されたデータが１個ではなく、複数個の場合を考える。もちろん、数個のデータでも母集団の分布を再現するほどの情報にはならない。しかし、母平均μの推量なら、１個の観測データのときよりもずっと精度を上げることができる。さきほど、複数個の観測データの平均値を母平均

と区別するために標本平均と呼んだが、標本平均については、次の定理が成立する。「１つの母集団から、n個のデータを観測しその標本平均$\bar{x}$を作る。このとき、nが大きければ大きいほど、標本平均$\bar{x}$は母平均μに近い数値をとる可能性が高くなる。」これを「大数の法則」という。「同じ標本平均をとるにしても、より多くのデータでとった方が正確である」という意味のイメージしやすい定理だといえる。

つまり、私たちがこのように標本平均をとるのは、偶然に起きるデータの散らばりをできるだけ打ち消して、実際の値に近い値を得たいからである。この標本平均の発想が統計的推定でも大きな効力をもつ。

さて、統計的推定として最も理想的なものは、もちろん、「母集団の分布がまるでわからないときの推定」だろう。しかし、これは、母集団に関する情報が何もないのだから、さすがに「原理的には無理だ」と直感的に感じさせる。実際には、いくつかのアプローチの方法があるようだが、本稿では取り扱わない。また、正規分布をとらない母集団は、たとえ母集団自体の分布がわかっていても、標本平均の分布はそれとは変わってしまうため、推定の対象とするには不都合である。

他方、正規母集団（正規分布している母集団のこと）は、「標本平均を作っても、その分布は正規分布のまま」という都合のいい性質をもっている。そして、私たちが、ある特定の不確実現象に関して何か知りたい場合の多くは、その母集団の母平均や母分散（あるいは標準偏差）を知ればよいことが多い。それは、前述したように、自然現象や社会現象において正規分布が非常に頻繁に現れるからである。

したがって、本稿では、統計的推定の対象を正規母集団とする。そして、正規母集団の母平均、母分散をばらつきのある標本調査の結果から推定する方法を下記の４パターンに分類してみていく。

・正規母集団における母分散σ^2がわかっているときの母平均μの推定
・正規母集団における母平均μがわかっているときの母分散σ^2の推定
・正規母集団における母平均μがわからないときの母分散σ^2の推定
・正規母集団における母分散σ^2がわからないときの母平均μの推定

４.２　統計的推定のパターン別アプローチ

４.２.１　統計的推定のファーストアプローチ

まず、「区間推定」の重要なポイントは、100%を望まずにピンポイントでなく、例えば「95%の確率で」と幅をもって母集団の平均値や分散値を言い当てようとすることである。本題の「正規母集団とわかっていて、母分散もわかっているときの母平均の推定」に入る前に、その区間推定の考え方を、標本数（観測データ）が１個の場合で確認する。たった１個の標本数で区間推定を行うことは実際には現実的ではないが、区間推定のアプローチを学ぶことができる最もシンプルな事例となりうるのでここでみておくことにする。

【例題１】[2)]

お金が入っている箱がたくさん並んでいる。各々の箱の額は不明で、その平均金額を知るために、標本として１箱だけ無作為に抽出して調べたところ、箱の中には500円が入っていた。すべての箱の中の平均金額を区間推定せよ。ただし、箱の中の金額Xは分散30^2（標準偏差30）の正規分布に従うとする。

推定するといっても確率現象の話である。「100%こうなる」と断言することは不可能である。そこで、区間推定では、「ここからここまでの幅（区間）に母平均μが含まれる確率は95%」というように、推定した値に「当たる確率は何%」と精度を与える。この95%のことを「信頼度」と呼ぶ。得られる推定区間がどれくらい信頼できるかを確率的に表現したものである。区間推定においてよく用いられる信頼度は95%と99%である。以降、本稿では、実際に最もよく使われる信頼度95%でみていくこととする。

推定したい平均金額すなわち母平均をμとおく。推定量となる金額Xは平均値μ（未知）、分散30^2の正規分布に従うので、抽出した標本ごとに異なる金額で散らばりながら図４－１のように分布する。

前述したとおり、確率分布を表す確率密度関数のグラフでは、横軸とグラフとで囲まれる部分の面積が確率を表す。この図の正規分布でも同じであるので、図４－１で平均値μを対称にして左右同じ幅だけ面積（すなわち確率）が信頼度0.95（95%）となるように網掛けをする（図４－２）。

箱の中の金額Xは95%の確率でこの網掛けした区間に現れるはず、と考えるのである。むろん、この網掛けした範囲の右側と左側に、左右合わせて５

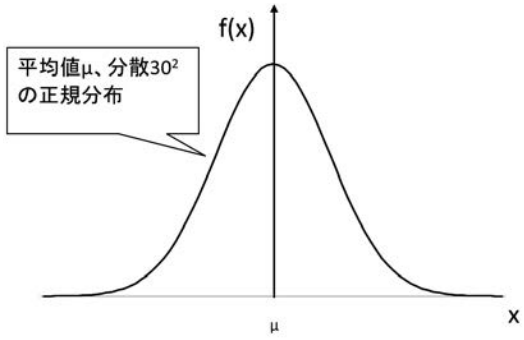

図４－１　平均値μ、分散30²の正規分布

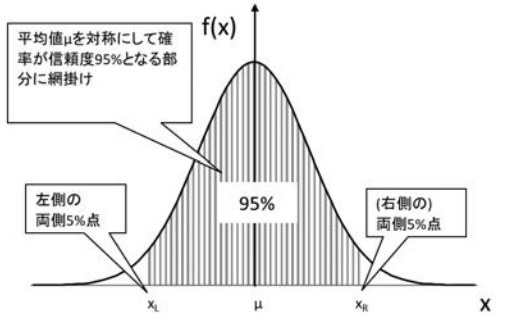

図４－２　平均値μ、分散30²の正規分布（μを対称に信頼度95%となる部分に網掛け）

%の確率でXが現れる可能性が残っているが、そこも含めた100%の確率で現れるXの区間は－∞以上＋∞以下になってしまうので、それでは推定する意味がない。そこで、信頼度95%を設定して残り５%の確率で起こり得る事象をカバーすることは諦めることで「有限の区間を推定する」のである。

図４－２の正規分布のように、確率密度変数が左右対称の場合、信頼度95%で網を掛けたときの右端の方の点を「両側５%点」と呼ぶ（x_R）。グラフが左右対称なので、「両側」と言いながら右の方の点のみを指定すれば情報としては足りるのである。ただ、平均値μを中心として右側の「両側５%点」x_Rと対称となるx_Lを説明したいときには、x_Rを「右側の両側５%点」、x_Lを「左側の両側５%点」と別々の呼称で呼ぶこともある。

平均値μ、分散σ^2の正規分布の「右側の両側５%点」、「左側の両側５%点」は次のように与えられる（図４－３）。

右側の両側５%点：$\mu+1.96\times\sigma$…（式４－２）

左側の両側５%点：$\mu-1.96\times\sigma$…（式４－３）

よって、網を掛けたXの範囲は次のように表される。

$\mu-1.96\times\sigma\leqq X\leqq\mu+1.96\times\sigma$…（式４－４）

これは、「正規分布では、Xが95%の確率で平均値±1.96×標準偏差の区間に出現する」ことを意味する。前節で、正規分布で平均から標準偏差±２個分の範囲にデータが入る確率は0.9544（95.44%）

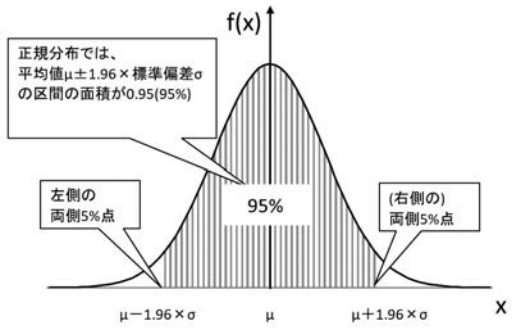

図４－３　平均値μ、分散σ^2の正規分布における両側５%点

であると述べた。そこから95%ぴったりにするために標準偏差±２個分からほんの少しだけ範囲を狭めて±1.96個分とした、と考えてよい。

この例題では標準偏差$\sigma=30$であるから、

$\mu-1.96\times30\leqq X\leqq\mu+1.96\times30$…（式４－５）

とし、金額X（500円だった）が95%の確率でこの区間に現れるものとして、μを逆算するのである。

すなわち、現実に観測したX（500円）がこの範囲に入らないようなμは現実の母集団の母平均としてありえないとして棄却する（採用しない）、というふうに考える。

$\mu-1.96\times30\leqq500\leqq\mu+1.96\times30$…（式４－６）

から

$441.2\leqq\mu\leqq558.8$　　…（式４－７）

この範囲に入る母平均μは棄却せずに残ることになるので、これが母平均μの区間推定の結果、すなわち「母平均μの95%信頼区間」となる。

ここで、相対度数が95%になる区間は、（式４－４）の「$\mu-1.96\times\sigma\leqq X\leqq\mu+1.96\times\sigma$」以外にもいろいろあるのではないかという疑問が浮かぶ方もいると思われる。その通りである。例えば、（式４－４）の区間を少しずらして「$\mu-2.1\times\sigma\leqq X\leqq\mu+1.86\times\sigma$」としても、相対度数は同じ95%になる。しかし、相対度数95%となる区間は、前者の3.92に対して後者は3.96と長くなってしまう。同じ相対度数で推定する区間は短いほど精度が高いだけ望ましいので、正規分布においては、ヒストグラムが左右対称で、対称軸に近いほど頻度が高いことを考慮すれば、同じ確率で推定される区間を選び出すには、「左右対称の区間」を選択すべきということがわかるだろう。

〈次号に続く〉

【参考文献】

１）完全独習　統計学入門　小島　寛之　2006年９月　ダイヤモンド社

２）まずはこの一冊から　意味がわかる統計解析　涌井　貞美　2013年２月　ベレ出版