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。

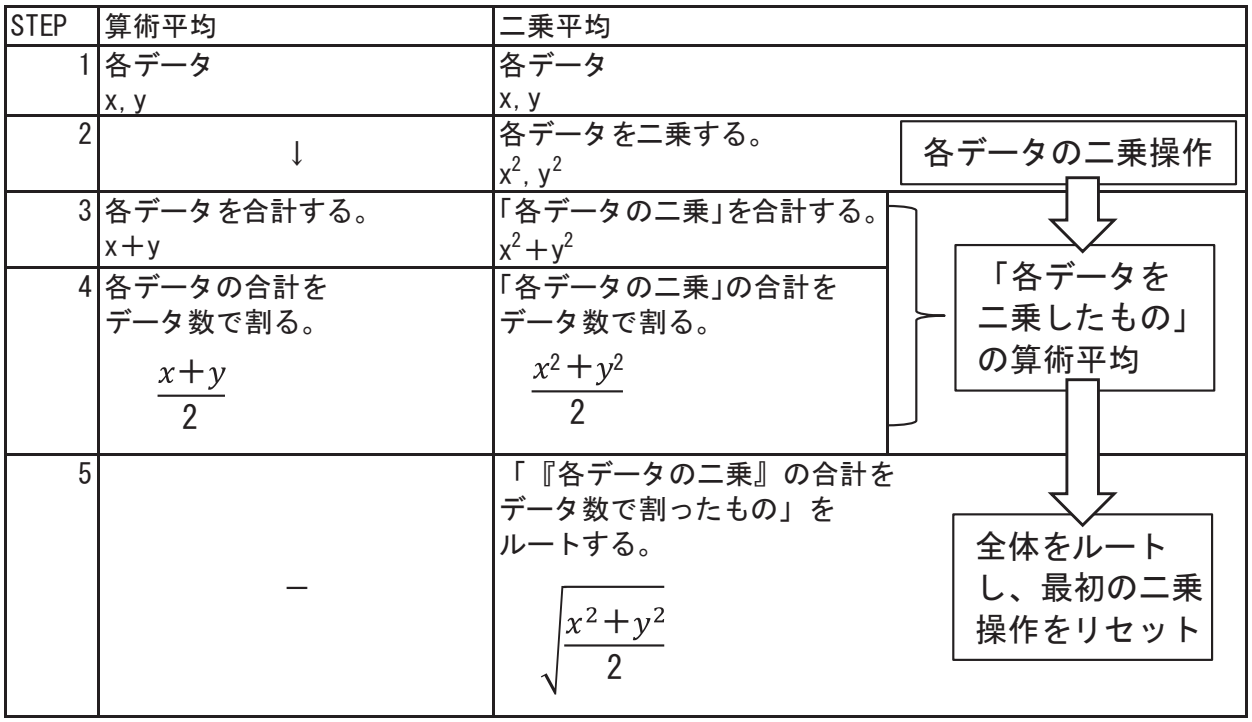

図２－１　算術平均と二乗平均の手順

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