統計量の計算

標本をグラフ化したりヒストグラムを描くことにより，標本の性質を視覚的に読み取ることができます．記述統計の目的は，標本の性質を視覚的・数値的に読み取ることでしたが，標本の性質を数値的に読み取る方法として，**統計量**を計算する，というものがあります．統計量とは，標本から何らかの計算によって算出された値のことで，標本の様々な性質を表す，色々な統計量が存在します．例えば，以下の性質を表すような統計量が存在します．

- 標本（データ）の中心
- 標本（データ）のばらつきの大きさ
- 2 種類の標本（データ）の関係性

ここでは，まず標本の中心を表す統計量と，標本のばらつきの大きさを表す統計量について学習していきます．

データの中心を表す統計量

データの中心を表す統計量の総称を**代表値**といいます．代表値の例としては，**平均値**，**中央値**，**最頻値**などがあります．

平均値

単に「平均」と言った場合，多くは**算術平均**を指します．算術平均は，**相加平均**ともよばれます．算術平均は，データの総和をデータ数で割ることで計算します．例として，5 人の体重の算術平均を計算してみましょう．5 人の体重がそれぞれ，

$$60.2\,\mathrm{kg},\quad 67.3\,\mathrm{kg},\quad 63.4\,\mathrm{kg},\quad 72.5\,\mathrm{kg},\quad 65.6\,\mathrm{kg}$$

のとき，この 5 人の体重の算術平均値は，

$$\frac{60.2+67.3+63.4+72.5+65.6}{5}=\frac{329}{5}=65.8(\mathrm{kg}) \tag{1}$$

で与えられます．一般的には，次のように定義されます．

算術平均

n 個のデータ $x_1, x_2, \cdots, x_n$ の算術平均 $\bar{x}$ は，以下の式で与えられる．

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

n 個のデータに対して，これを全て足しあわせて n で割ったもの，という式になっています．以降，$\bar{x}$ と書いたら，算術平均を表すこととします．

Excel による平均値の計算

ここでは，Excel で平均値を計算する方法について学習します．Excel には算術平均を計算する関数として AVERAGE 関数があり，`=AVERAGE(“データ範囲”)` のように使用します．

例題 1 「データの中心を表す統計量.xlsx」の「Sheet1」には，ある商店の 1 年間の売り上げデータがある．C15 セル（水色部分）に 1 ヶ月あたりの平均売上を計算せよ．

1 月から 12 月の売り上げの算術平均を計算すれば良いことになります．C15 セルに「`=AVERAGE(C3:C14)`」と入力すれば計算できます．

中央値

「データの中心」というとすぐに平均値をイメージしますが，平均値以外にも「データの中心」を表す統計量が存在します．なぜ，平均値以外の統計量が必要なのでしょうか？それは，平均値ではデータの中心をうまく捉えられない場合があるからです．平均値以外の代表値として中央値と最頻値があります．ここではまず中央値について学習します．

まず，以下の例題を通して，平均値でデータの中心を上手く捉えられない場合があるということを確認してみましょう．

例題 2 「データの中心を表す統計量.xlsx」の「Sheet2」には，野球選手 23 人の年俸のデータがある．この 23 人の平均年俸を，F3 セル（水色の部分）に入力せよ．

計算すると約 3669.6（万円）となります．果たして，このデータに対して，3669.6（万円）が「データの中心」を表していると言えるでしょうか．平均値を超えている選手が何人いるかを数えてみると，23 人中 4 人しかいないことがわかります．そこで，年俸が最も高い

4 億円の A さんを除いて平均値を計算してみると，その値は約 2018.2（万円）となり，1 人抜けただけで平均値が大幅に小さくなってしまうことがわかります．このように，データの中に，極端に大きなデータが存在すると，平均値はその値に引っ張られる形で大きな値をとってしまい，データ全体を代表する値とはいえなくなります．これは，極端に小さな値が存在する場合も同様です．このようにデータの集まりから極端に離れたデータの値を，**外れ値**といいます．平均値は外れ値の影響を大きく受けやすい統計量である，といえます．

外れ値が存在するデータでは，平均値より**中央値（メディアン）**の方が「データの中心」をうまく表すことができます．

中央値

中央値（メディアン）は，データを小さい順に並べて，データが奇数個なら中央に位置するデータの値，データが偶数個なら中央に位置する 2 つのデータの平均値で与えられる．

Excel による中央値の計算

Excel で中央値を計算する場合，MEDIAN 関数を利用し，`=MEDIAN(“データ範囲”)` という形で計算することができます．

例題 3　例題 2 の野球選手 23 人の年俸のデータについて，中央値を F4 セル（オレンジ色の部分）に計算せよ．

F4 セルに「`=MEDIAN(C3:C25)`」と入力すれば計算できます．結果は，1100（万円）となり，これは L さんの年俸にあたります．

外れ値の，平均値と中央値に対する影響を調べるために，以下の例題を解いてみましょう．

例題 4　例題 2 の野球選手の年俸のデータについて，以下の値を計算し比較せよ．

- C さん 〜 W さんの平均値
- C さん 〜 W さんの中央値

計算をしてみると，平均値は 1638.1（万円），中央値は 1000（万円）となります．23 人全員で計算した時の値は，平均値が 3669.6（万円），中央値が 1100（万円）でしたから，平均値に比べ，中央値は少量の外れ値の影響を受けにくいことが確認できます．

最頻値

ここまで，データの中心を表す統計量（代表値）として，平均値と中央値について学習しましたが，そのいずれも計算することができないデータがあります．次の例を考えてみましょう．

30 人の血液型を調べたら以下のようになった

A, O, A, B, O, A, A, B, O, O, B, A, O, B, A,
A, O, AB, A, AB, O, A, AB, B, A, AB, O, B, O, A

このデータの「中心」はどのように表せばよいだろうか．

すぐに分かるように，このデータに対して平均値や中央値を計算することができません．これは，血液型のデータが，順序に意味を持たない名義尺度だからです．このようなデータに対して，「データの中心」を求めるためには，**最頻値（モード）**を用います．

最頻値

最頻値は，データの集まりの中で最も度数（各データの出現回数）の多いデータの値で与えられる．出現頻度が最も多いデータが 1 つとは限らないため，最頻値は複数存在する場合がある．

上記のデータで度数を数えると，A 型：11 人，B 型：6 人，O 型：9 人，AB 型：4 人となりますから，最頻値は「A」となります．

Excel による最頻値の計算

Excel で最頻値を計算する場合，MODE 関数を利用し，`=MODE(“データ範囲”)` という形で計算することができます．ただし，Excel の MODE 関数では，関数の引数として文字をとることができず，引数は必ず数字でなくてはいけません．以下の例題を通して使い方を学習しましょう．

例題 5 「データの中心を表す統計量.xlsx」の「Sheet3」には 30 人分の血液型のデータがある.

(1) COUNTIF 関数を用いて，それぞれの血液型の人数を数えて C8〜C11（水色の部分）に出力しなさい．COUNTIF 関数は，ある範囲内に指定したデータがいくつ存在するかを数える関数で，例えば，`COUNTIF(C8:C37,F8)` とすると，「C8 から C37 の範囲で，F8 のデータが何個あるか」を求めることができます.

(2) MODE 関数でデータ範囲を C8〜C37 を指定して，G14 セル（オレンジの部分）に最頻値を求めてみなさい.

(3) MODE 関数でデータ範囲を D8〜D37 を指定して，G15 セル（オレンジの部分）に最頻値を求めてみなさい.

(1) まず, G8 セルに, A 型の人の数を出力します．G8 セルに「`=COUNTIF(C8:C37,F8)`」と入力することで，範囲 C8〜C37 の中に F8 セルの内容である “A” が幾つあるかを数えることができます．後で，他のセルにコピーをすることを考えて，「`=COUNTIF(C$8:C$37,F8)`」としておきましょう．こうすることで，このセルの内容を G9 セル〜G11 セルにコピーしたときに，データ範囲の C8〜C37 は変化しないようにすることができます（F8 の部分は，F9〜F11 と変化します）．

(2) G14 セルに「`=MODE(C8:C37)`」と入力してみましょう．すると，「#N/A」とエラーが出てしまい，計算することができません．このように，MODE 関数の引数に文字を入れてしまうとエラーとなってしまいます．その為，MODE 関数を使って最頻値を計算するためには，「A 型なら 1，B 型なら 2，0 型なら 3，AB 型なら 4」というように，予め文字データを数値データに変換する必要があります.

(3) D 列には，血液型が数値に変換されたデータが入力されています．このデータに対して MODE 関数を使って最頻値を計算してみましょう．G15 セルに「`=MODE(D8:D37)`」と入力してみましょう．すると，「1」という結果が出力されます．これは，上記の変換規則で A 型に対応する数値です.

Excel で最頻値を計算する際の注意事項として，最頻値が複数存在する場合の MODE 関数の挙動が挙げられます．最頻値が複数存在する場合でも MODE 関数はその中の 1 つしか出力しません．このような場合には，例題の（1）のように COUNTIF 関数を用いて各データの度数を調べ，その中から度数が最大のものを探すようにすると良いでしょう.

尺度水準と代表値

データの中心位置を表す統計量のことを代表値といい，代表値として平均値・中央値・最頻値を学習しました．データの尺度水準によっては，意味を持たなくなってしまう代表値があります．どのデータの尺度水準に対して，どの代表値が意味を持つかを整理しておきましょう．各尺度水準に対して，意味を持つ代表値は以下の通りです．

- 名義尺度 …… 最頻値
- 順序尺度 …… 最頻値，中央値
- 間隔尺度 …… 最頻値，中央値，平均値
- 比例尺度 …… 最頻値，中央値，平均値

既に述べたように，名義尺度については，順番や大小に本質的な意味が存在しないため，中央値や平均値を計算しても意味を持ちません[*1]．順序尺度については，大きさの大小に意味があるので，中央値は意味を持ちますが，和に意味を持たないので平均値は意味を持ちません．間隔尺度や比例尺度は，大きさの大小と共に和にも意味を持ちますから，中央値，平均値が意味を持ちます．

[*1] A, B, O, AB のように，文字になっていれば，そもそも中央値や平均値は計算のしようがありませんが，数値に変換した 1, 2, 3, 4 のデータだと計算だけはできてしまいます．