# ５章 基本統計量

3.5 節で量的データの集計方法について簡単に触れ、前章でデータの分布について学びましたが、データの特徴を１つの数値で示すこともよく行なわれます。これは統計量と呼ばれ、主に分布の中心や拡がりなどを表わします。この章ではよく利用される分布の統計量を特徴で分類して説明します。数式表示を統一的に行なうために、データの個数を $n$ 個とし、それらを $x_1, x_2, \cdots, x_n$ と表わすことにします。ここで学ぶ統計量は統計分析の基礎となっており、基本統計量とも呼ばれています。

## 5.1 分布の中心を表わす基本統計量

分布の特徴を表わすには、まず分布の中心がどこにあるのかを示さなければなりません。この分布の中心を表わす統計量には重要なものが３つあります。

### 1. 平均値（mean, average）

これは最もよく使われている中心を表わす統計量で、特に統計を学んでいなくても、知っていると思います。平均値はデータから以下のような式で与えられます。

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

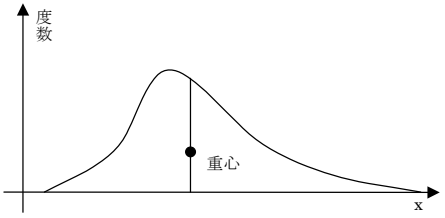


図 5.1a 平均値

この定義は図 5.1a のように、ヒストグラムの重心を通り $x$ 軸に下ろされた垂線の $x$ 座標を表わしています。

Excel にもこの統計量を求める以下の関数があります。

平均値 = average(範囲)

### 2. 中央値（median）

これは中間値またはメジアンとも呼ばれ、データを小さい方から大きい方に並べた、真中の値です。度数分布を用いると、面積が度数を表わしますので、図 5.1b のように左右の面積の等しい位置が中央値となります。例として、以下のデータの中央値を求めてみましょう。

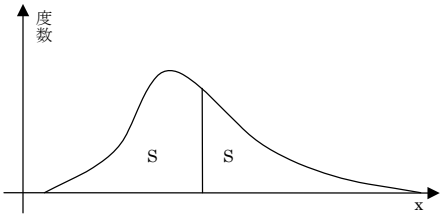


図 5.1b 中央値

1, 2, 3, 5, 7, 8, 9

1, 2, 3, 4, 5, 7, 8, 9

最初のデータは奇数個ですから、中央値は 5 です。2 番目のデータは偶数個で中央値は、$n/2$ 番目と $n/2+1$ 番目の値の平均を取ります。この場合 $(4+5)\div 2 = 4.5$ となります。

Excel にも中央値を求める以下の関数があります。

中央値 = median(範囲)

### 3. 最頻値（mode）

度数分布で最も頻度の高い値を最頻値またはモードといいます。Excel にも最頻値を求める関数があります。

最頻値 = mode(範囲)

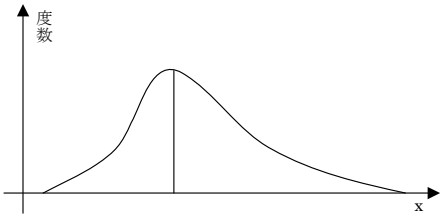


図 5.1c 最頻値

しかし、この関数を利用するときには注意が必要です。例えば 1, 2, 4, 5, 6, 6 というデータで最頻値を求めてみます。このデータだと 6 が 2 つ、他は 1 つですから、最頻値は 6 ということになってしまいます。データ数が少ない場合や、データが多くても殆ど同じ値を持たないとき、利用には注意が必要です。度数分布表を作って、最も頻度の高い値を最頻値とするのが無難なようです。

## 5.2 分布の拡がりを表わす基本統計量

分布の中心と同様、拡がりも分布の特徴を表わす大切な指標です。ここでは、分布の拡がりを与える統計量を見てみましょう。

### 1. レンジ（range）

最も単純な分布の拡がりを表わす統計量は、データの最大値と最小値の差です。これをレンジまたは範囲といいます。

$$R = \max(X) - \min(X) \qquad \text{ここに、} X = \{x_1, x_2, \cdots, x_n\}$$

これは単純な定義で分かり易いのですが、飛び離れたデータがある場合には、レンジがそのデータによって拡がりすぎて、必ずしも現実の拡がりを表わしていると言えなくなります。

Excel では最大値として =max(範囲)、最小値として =min(範囲) という関数があり、

その差がレンジを表わしています。

レンジ = max(範囲) – min(範囲)

## 2. 分散（variance）

レンジは飛び離れた 1 つのデータに大きく左右されるのが欠点でした。この欠点を除いて、現在最もよく利用されている統計量は、ここで述べる分散（または不偏分散）と分散から得られる標準偏差です。

分散は各データの平均からのずれの 2 乗を合計して、データ数で割ったもので、以下の式によって与えられます。

$$s^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

我々は分散を$s^2$という表式で表わします。分散はデータのばらつきが平均からずれているほど大きな値となります。また、1 つのデータの寄与は$(x_i - \bar{x})^2 / n$ですので、全体に対してレンジのように大きな影響はありません。

また、分散は以下のようにも変形できます。

$$s^2 = \frac{1}{n}({x_1}^2 + {x_2}^2 + \cdots + {x_n}^2) - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

これはコンピュータでプログラムする際にデータを読みながら、平均と分散が同時に計算できる便利な公式です。

Excel では分散を求める以下のような関数が用意されています。

分散 = varp(範囲)

名前は variance から取られています。

## 3. 不偏分散（unbiased estimator of variance）

分散にはもう 1 つの定義があり、不偏分散と呼ばれています。場合によってはこちらの定義の方がよく利用されているかも知れません。Excel で分散というとこの不偏分散を示しています。我々は不偏分散を分散と区別して表わすために$u^2$という記号を用います。不偏分散の定義と通常の分散との関係は以下のように与えられます。

$$u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}s^2$$

分散と不偏分散はどのように使い分けるのでしょうか。通常母集団の分散は通常の分散を、標本から母集団の分散を推測する場合は不偏分散を使います。Excel には不偏

分散を表わす以下のような関数があります。

不偏分散 = var(範囲)

### 4. 標準偏差（standard deviation）

分散はデータと平均との差の 2 乗を取ったせいで、データの単位の 2 乗の単位を持っていますので（例えばデータが cm なら分散は $cm^2$）、これから直接データの広がりを見ることはできません。そこで、データの単位に合わせるために、分散の平方根を取って標準偏差と呼びます。これにより分布の拡がりという意味がはっきりとします。また標準偏差には分散から求められるものと不偏分散から求められるものがあります。我々はそれらを区別するために、それぞれ $s$ と $u$ の記号を用いて表わします。

$s = \sqrt{分散}$　　または　$u = \sqrt{不偏分散}$

Excel にもこれらを表わす関数が以下のように与えられています。名前は standard deviation から取られています。

標準偏差 = stdevp(範囲)　または　=stdev(範囲)

Excel では通常、標準偏差というと不偏分散から得られるものを指しており、後者がそれに当ります。

## 5.3 分布の形を表わす基本統計量

分布の中心と広がりは分かりましたが、分布の形についてはこれらの統計量からは推測できません。そこである程度分布の形が分かるような統計量も考案されていますが、頻繁に利用されているかというとそうでもないようです。

### 1. 歪度（skewness）

分布の歪みを表わす統計量には、歪度と呼ばれるものがあります。これは以下のような定義で与えられます。

$$a_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

この値は、裾が右に伸びている場合に正、左に伸びている場合に負になります。Excel の関数は以下の尖度ともに定義が少し異なっていますので、ここでは省略します。

### 2. 尖度（kurtosis）

次に分布の尖り方を示す統計量を紹介します。尖度と呼ばれる値で、以下の式によって表わされます。

$$a_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

これは、これから学ぶ標準的な分布（正規分布）より裾が伸びている場合に３以上の値になることが分かっています。

**問題**

分散の以下の 2 つの表式が同等であることを示せ。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

**解答**

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2\bar{x}}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

**問題**

1) 以下のデータで、分散 $s^2$ を定義に従って求めよ。

2) Excel の関数を使って、以下のデータの平均値、中央値、レンジ、分散、不偏分散、それぞれの標準偏差を求めよ。

身長(cm)　　171, 181, 172, 166, 172, 175, 168, 174, 171, 170

**解答**

1)

| No. | $x$ | $x$-$a$ | $(x$-$a)^2$ |
|---|---|---|---|
| 1 | 171 | -1 | 1 |
| 2 | 181 | 9 | 81 |
| 3 | 172 | 0 | 0 |
| 4 | 166 | -6 | 36 |
| 5 | 172 | 0 | 0 |
| 6 | 175 | 3 | 9 |
| 7 | 168 | -4 | 16 |
| 8 | 174 | 2 | 4 |
| 9 | 171 | -1 | 1 |
| 10 | 170 | -2 | 4 |
| 平均 | 172 | 分散 | 15.2 |

まず、データ $x$ を入力し、平均を求める。それを $a$ として $x-a$ を計算する。さらに

$(x-a)^2$ を求め、それを平均して、分散の値を求める。

2)

| | | | |
|---|---|---|---|
| 平均値 | 172 | 分散 | 15.2 |
| 中央値 | 171.5 | 標準偏差 | 3.898718 |
| レンジ | 15 | 不偏分散 | 16.88889 |
| | | 標準偏差 | 4.109609 |

## 5.4 ２変量の関係を表わす統計量

3.4 節で２つの量的データの関係を表わす量として相関係数を紹介しましたが、ここではこの相関係数について少し詳しく説明したいと思います。今、以下のような対になった２つの変数を考えます。

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

この 2 変数の間の相関係数は以下のように与えられます。

$$r = \frac{s_{xy}}{s_x s_y}$$

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

ここに、$s_x$ と $s_y$ は変数 $x$ と $y$ の標準偏差で、$s_{xy}$ は $x$ と $y$ の共分散と呼ばれる量です。

さて、相関係数はどのような値を取るのでしょうか。図 5.2 を見て下さい。

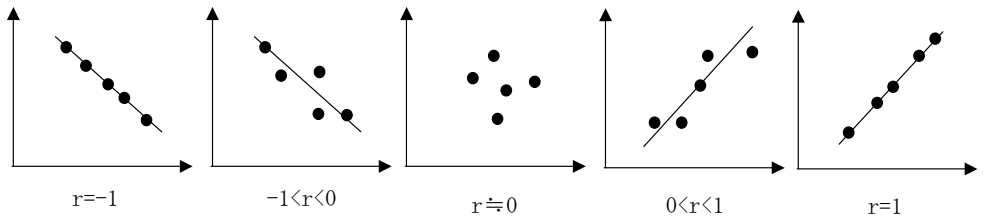


図 5.2 相関係数と散布図

これは、変数 $x$ を横軸に $y$ を縦軸にして、各データを点で表示した散布図です。相関係数は、２つの変数間に完全な $y = ax + b$ の線形関係があるとき、$a$ の正負に応じて $r = \pm 1$ となるように作られています。そしてそれから外れるごとに 0 に近づいて行き、軸のスケールを適当にとることによりデータが球状に分布するときほとんど 0 になります。

相関係数は上の定義から、単位が分子と分母で打ち消されており、どんな単位を使っても（例えば m か cm, kg か g 等）その値は変化しません。

**問題**

1) 以下の対になった身長と体重のデータで、相関係数の定義に従ってその値を Excel で計算せよ。

2) Excel の関数を利用して、これらのデータの基本統計量及び相関係数を求めよ。

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 身長(cm) | 171 | 181 | 172 | 166 | 172 | 175 | 168 | 174 | 171 | 170 |
| 体重(kg) | 71 | 74 | 65 | 58 | 66 | 70 | 60 | 63 | 72 | 61 |

**解答**

1)

| No. | *x* | *x*-*a* | (*x*-*a*)^2 | *y* | *y*-*b* | (*y*-*b*)^2 | (*x*-*a*)(*y*-*b*) |
|---|---|---|---|---|---|---|---|
| 1 | 171 | -1 | 1 | 71 | 5 | 25 | -5 |
| 2 | 181 | 9 | 81 | 74 | 8 | 64 | 72 |
| 3 | 172 | 0 | 0 | 65 | -1 | 1 | 0 |
| 4 | 166 | -6 | 36 | 58 | -8 | 64 | 48 |
| 5 | 172 | 0 | 0 | 66 | 0 | 0 | 0 |
| 6 | 175 | 3 | 9 | 70 | 4 | 16 | 12 |
| 7 | 168 | -4 | 16 | 60 | -6 | 36 | 24 |
| 8 | 174 | 2 | 4 | 63 | -3 | 9 | -6 |
| 9 | 171 | -1 | 1 | 72 | 6 | 36 | -6 |
| 10 | 170 | -2 | 4 | 61 | -5 | 25 | 10 |
| 平均 | 172 | | 15.2 | 66 | | 27.6 | 14.9 |

まず、$x$ と $y$ のデータを入力し、それぞれの平均を求める。それらの平均を $a$, $b$ として、$x-a$ と $y-b$ を求める。それらの 2 乗 $(x-a)^2$, $(y-b)^2$ とそれらの積 $(x-a)(y-b)$ を求めて、それぞれ平均を計算する。その値を使って相関係数を求める。

$$r = 14.9/\sqrt{15.2\times 27.6} = 0.727461$$

2)

特によく利用されるものだけ結果を示す。最頻値はこのようなデータでは意味を持たない。

| | | | | | |
|---|---|---|---|---|---|
| 平均値 | 172, 66 | 分散 | 15.20, 27.6 | 相関係数 | 0.727 |
| 中央値 | 171.5, 65.5 | 標準偏差 | 3.90, 5.25 | | |
| レンジ | 15, 16 | 不偏分散 | 16.89, 30.67 | | |
| | | 標準偏差 | 4.11, 5.54 | | |

**興味ある人に　[Skip OK]**

相関係数は$-1 \le r \le 1$の値を取ると言いましたが、これを証明してみましょう。まず、$c$を何らかの数として、以下の式を考えます。

$$\frac{1}{n}\sum_{i=1}^{n}\{c(x_i-\bar{x})-(y_i-\bar{y})\}^2 = c^2 s_x^2 - 2cs_{xy} + s_y^2 \ge 0$$

ここで、$c = s_{xy} \big/ s_x^2$とすると、この式は以下のように変形できます。

$$\frac{s_{xy}^2}{s_x^2} - 2\frac{s_{xy}^2}{s_x^2} + s_y^2 = -\frac{s_{xy}^2}{s_x^2} + s_y^2 \ge 0$$

これから、$r^2 = \dfrac{s_{xy}^2}{s_x^2 s_y^2} \le 1$となり、$-1 \le r \le 1$が示されます。

**問題**

分布図のデータが完全に直線$y = ax + b$の上に並ぶとき、相関係数$r$の値は$a$の正負により$r = \pm 1$となることを示せ。

**解答**

$y_i = ax_i + b$（$i = 1, \cdots, n$）とする。

$s_y^2 = \dfrac{1}{n}\sum_{i=1}^{n}[(ax_i+b)-(a\bar{x}+b)]^2 = \dfrac{a^2}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = a^2 s_x^2$ より、$s_y = |a| s_x$

$s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})[(ax_i+b)-(a\bar{x}+b)] = \dfrac{a}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = as_x^2$

$r = \dfrac{s_{xy}}{s_x s_y} = \dfrac{as_x^2}{s_x \cdot |a| s_x} = \dfrac{a}{|a|}$ となり、$a$の正負により$r = \pm 1$となる。