2024 年度春学期　統計学　第５回

# 分布をまとめる ー 記述統計量（平均・分散など）

## 代表値

度数分布のヒストグラムによる表現は，視覚的にはよくわかる表現です。しかし，度数分布から取り出される情報を今後の処理に用いたり，比較したりするには，分布を 1 つの数字で表現する必要があります。これを，**代表値**といいます。ここでは，もっともよく使われる代表値である算術平均と，分布を表現するもうひとつの重要な指標である分散について説明します。

### 算術平均

データが $n$ 個の数値 $x_1, x_2, \ldots, x_n$ からなるとき，**算術平均（相加平均）**は次の式で定義されます。

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

つまり，**算術平均 ＝（データの合計）／（データサイズ）**です。データサイズとは，データを構成する数値の個数をいいます。ふつう，単に「平均」といえば算術平均のことをさします[1]。

### 度数分布から算術平均を求める

データの度数分布がわかっているときに，その平均を求めるにはどうすればよいでしょうか？ 平均とはデータを構成する数値の合計をデータサイズで割ったものです。一方，ある階級の度数は「その階級値をとる数値が，何個あるか」を表しています。そこで，

**平均＝（データの合計）／（データサイズ）**
**＝（[階級値×度数］の合計）／（データサイズ）**
**＝［階級値×（度数／データサイズ)］の合計**
**＝［階級値×相対度数］の合計**

ですから，**「平均 ＝［階級値×相対度数］の合計」**ということになります。

表 1 の度数分布で，この計算を見てみましょう。1 行目を見ると，0〜9 点の階級は，階級値 5 点で，相対度数は 0.04 です。ですから，［階級値×相対度数］は $5 \times 0.04 = 0.2$ となります。各行について［階級値×相対度数］を計算して，これを全階級にわたって合計すると，一番下の行にあるように，合計は 49.8 となり，これがこの度数分布の平均になります。

---

## 分散と標準偏差

### 「ばらつき」を表すには

分布をもっとも簡単にひとつの数字で表したのが代表値ですが，代表値だけでは，その分布が「どのくらいばらついているか」は表現できません。その例を見てみましょう。つぎのようなデータの組 A, B, C があるとします。

[1] (1) 式にある「Σ 記号」については，「数学の補足説明」プリントを参照してください。

表 1: 度数分布から分散を求める

| 階級 | 階級値 | 相対度数 | 階級値×相対度数 |
|---|---|---|---|
| 0 〜 9（点） | 5 | 0.04 | $5 \times 0.04 = 0.2$ |
| 10 〜 19 | 15 | 0.16 | 2.4 |
| 20 〜 29 | 25 | 0.08 | 2.0 |
| 30 〜 39 | 35 | 0.12 | 4.2 |
| 40 〜 49 | 45 | 0.10 | 4.5 |
| 50 〜 59 | 55 | 0.10 | 5.5 |
| 60 〜 69 | 65 | 0.12 | 7.8 |
| 70 〜 79 | 75 | 0.08 | 6.0 |
| 80 〜 89 | 85 | 0.18 | 15.3 |
| 90 〜 100 | 95 | 0.02 | 1.9 |
| 合計 | | 1.0 | 49.8<br>(=平均 ) |

A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10
B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10
C: 3, 4, 4, 5, 5, 5, 5, 6, 6, 7

これらの平均はいずれも 5 で，平均値ではこれらの分布を区別して表現することはできません。これらの分布の違いは，ばらつきにあります。

A と B は分布の幅（レンジ）は違いませんが，分布の平均値への集まり具合がちがいます。レンジは分布の両端の値しか使っていないので分布の平均値への集まり具合を表現することはできませんが，次に述べる**分散**や**標準偏差**は，分布内のすべてのデータを使うので，集まり具合を表現できます。

### 偏差と分散

各データと平均との差を**偏差**といい，各データが平均からどのくらい離れているかを表します。例えばデータ組Aについて，表 2 のように偏差が求まります。「偏差の平均」を求めれば，このデータ組の「データの平均（=5）からの散らばり具合」がわかりそうですが，平均値はデータ組のちょうど真ん中の値ですから，「偏差の平均」は 0 になってしまいます。

そこで，「偏差の平均」のかわりに「(偏差)$^2$ の平均」を用います。(偏差)$^2$ はすべて正ですから，「(偏差)$^2$ の平均」，すなわち「各データについての偏差の 2 乗の合計をデータの個数で割ったもの」でばらつきの程度を表現できます。これが**分散 (variance)** です。式で書くと，各数値を $x_1, x_2, \ldots, x_n$，データサイズを $n$，平均を $\bar{x}$ とするとき，分散 $\sigma^2$ はつぎのようになります。

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\left\{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{aligned} \tag{2}$$

| データ | 偏差 | $(\text{偏差})^2$ |
|---|---|---|
| 0 | $0-5=-5$ | 25 |
| 3 | $3-5=-2$ | 4 |
| 3 | $-2$ | 4 |
| 5 | 0 | 0 |
| 5 | 0 | 0 |
| 5 | 0 | 0 |
| 5 | 0 | 0 |
| 7 | 2 | 4 |
| 7 | 2 | 4 |
| 10 | 5 | 25 |
| 平均=5 | 平均=0/10=0 | 平均=66/10=6.6（分散） |

表 2: 分散

#### 標準偏差

また，分散の平方根を**標準偏差 (standard deviation, SD)** といいます。データの単位が m（メートル）のとき，分散の単位は $\mathrm{m}^2$，すなわち平方メートルになってしまいますが，標準偏差の単位は同じ m です。

#### 分散を求めるとき，なぜ偏差の絶対値をとらずに偏差を 2 乗するのか？

確かに，偏差の絶対値を使って計算しても，「偏差を全部正の値にしてから平均する」という目的は達せられます。しかし，絶対値の計算は 2 乗よりも簡単そうですが，実はそうではありません。2 乗の計算は，どんな数に対しても同じ手続きでできますが，絶対値の計算は，正の数と負の数とで別の手続きが必要です。みなさんも，高校の数学の時間に，「$y=2x+3$ のグラフを描け」といった問題で，ややこしい場合分けをやった記憶があると思います。こういう事情で，偏差の絶対値の平均は用いられず，偏差の 2 乗の平均である分散が用いられているのです。さらに，2 乗を考えると，これを 3 乗，4 乗，…に発展させることができます。これらの「拡張された分散」のことをモーメントといいます（この講義では取り扱いません）

### 度数分布から分散を求める

上で，度数分布から平均を求める方法として「平均 ＝［階級値×相対度数］の合計」となることを示しました。分散は**「(偏差)$^2$ の平均」**ですから，上の計算を利用すると，**「分散＝［(偏差)$^2$× 相対度数］の合計」**すなわち**「分散＝［(階級値 − 平均)$^2$× 相対度数］の合計」**という計算で求められます。

表 3 の度数分布で，この計算を見てみましょう。1 行目を見ると，0〜9 点の階級は，階級値 5 点，相対度数は 0.04 で，表 1 で計算したように，［階級値×相対度数］は $5\times0.04=0.2$ です。

一方，この度数分布の平均は，やはり表 1 で計算したように，49.8 です。したがって，この行について，偏差は $5-49.8=-44.8$ で，$(\text{偏差})^2$ は $(-44.8)^2=2007.04$ となります。よって，$[(\text{偏差})^2\times$ 相対度数］は $2007.04\times0.04=80.28$ となります。

各行について［$(\text{偏差})^2\times$ 相対度数］を計算して，これを全階級にわたって合計すると，一番下の行にあ

るように，合計は 696.96 となり，これがこの度数分布の分散です。さらに，標準偏差は，$\sqrt{696.96} = 26.4$ となります。

表 3: 度数分布から分散を求める

| 階級 | 階級値 | 相対度数 | 階級値×相対度数 | 偏差 | (偏差)$^2$ | (偏差)$^2$ × 相対度数 |
|---|---|---|---|---|---|---|
| 0 〜 9（点） | 5 | 0.04 | $5 \times 0.04$ $= 0.2$ | $5 - 49.8$ $= -44.8$ | $(-44.8)^2$ $= 2007.04$ | $2007.04 \times 0.04$ $= 80.28$ |
| 10 〜 19 | 15 | 0.16 | 2.4 | −34.8 | 1211.04 | 193.77 |
| 20 〜 29 | 25 | 0.08 | 2.0 | −24.8 | 615.04 | 49.20 |
| 30 〜 39 | 35 | 0.12 | 4.2 | −14.8 | 219.04 | 26.28 |
| 40 〜 49 | 45 | 0.10 | 4.5 | −4.8 | 23.04 | 2.304 |
| 50 〜 59 | 55 | 0.10 | 5.5 | 5.2 | 27.04 | 2.704 |
| 60 〜 69 | 65 | 0.12 | 7.8 | 15.2 | 231.04 | 27.72 |
| 70 〜 79 | 75 | 0.08 | 6.0 | 25.2 | 635.04 | 50.80 |
| 80 〜 89 | 85 | 0.18 | 15.3 | 35.2 | 1239.04 | 223.03 |
| 90 〜 100 | 95 | 0.02 | 1.9 | 45.2 | 2043.04 | 40.86 |
| 合計 | | 1.0 | 49.8 (=平均 ) | | | 696.96（=分散） $\sqrt{696.96} = 26.4$ (=標準偏差) |

---

## 標準得点

### 同じ「70 点」でも，周りと比べたときの「地位」はいろいろ

ある人が，数学の試験で 100 点満点で 70 点をとったとします。70 点という点数そのものには，問題全体の 70%に正答したので，その試験についてはまあまあの出来，という意味はもちろんあります。しかし，大学などの受験においては，同じ試験を受けたすべての受験生の中で上位に入らなければ合格はできませんから，その 70 点という点数が「同じ試験を受けた他の受験生に比べて，上位なのか下位なのか」を知ることが重要です。同じ 70 点でも，他の受験生が皆 50 点そこそこなら，70 点をとった人は上位に位置するでしょうし，他の受験生が皆 90 点以上なら，他の人より大幅に劣っていることになって，意味合いは全然違います。

#### 「地位」を表す標準得点

このように，分布中のある数値が，分布の中でどのぐらいの位置にいるかを表現するために，「その数値が，分布の平均に比べて，標準偏差の何倍上回っているか（あるいは下回っているか)」を求めます。この値を**標準得点**（Z スコア，Z 値）といいます。たとえば，「ある数値を標準得点に換算すると 1.0 である」ということは，その数値が平均にくらべて標準偏差の 1.0 倍上回っていることを意味しています。また，標準得点が −1.5 なら，平均にくらべて標準偏差の 1.5 倍下回っていることを意味しています。

### 分布の変換

標準得点を求めるために，分布を平均 0・標準偏差が 1 になるように「変換」することを考えます。「分布を変換する」とは，分布に含まれる各々の数値について，一斉に同じ計算を施して，別の分布を作ることです。このような計算をしたとき，ある数値が変換の結果 1.0 になったとすれば，それは平均 0・標準偏差 1 である分布においての 1.0 という値ですから，これは平均よりも 1.0 倍上回っていることを示しており，標準得点に変換されていることになります。

#### 平均が 0 になるように変換する

どういう計算をすれば，分布を平均 0・標準偏差が 1 になるように「変換」することができるでしょうか？ 例えば，分布に含まれる各数値について一斉に「10 を引く」という計算をすると，どの数値も 10 小さくなるわけですから，平均も 10 小さくなることは容易に理解できます。そこで，10 を引くかわりに「元の平均を引く」という計算をすると，平均は「(元の平均) — (元の平均)」，すなわち 0 になります。

#### 標準偏差が 1 になるように，さらに変換する

標準偏差のほうはどうでしょうか。いま述べた「平均を 0 にする変換」をしたあと，各数値を一斉に「2 倍する」という計算をするとしましょう。各数値が 2 倍になったとき，平均は「数値の合計／データサイズ」ですから，平均も 2 倍になりますが，いまは平均は 0 なので，2 倍してもやはり 0 です。そうすると，各数値と平均との差である偏差が 2 倍になります。したがって，分散は「偏差の 2 乗の平均」ですから，分散は 2 の 2 乗，すなわち 4 倍になります。標準偏差は分散の平方根ですから，4 倍になったものの平方根で，やはり 2 倍になります。そこで，2 倍するかわりに「(1/元の標準偏差) 倍する」という計算をすると，標準偏差は「(元の標準偏差) の (1/元の標準偏差) 倍」で，すなわち 1 になります。

以上のことから，どんな分布でも「平均を引いて，標準偏差で割る」という計算をすれば，平均 0・標準偏差 1 である分布に変換されます。したがって，ある数値に「その数値が含まれる分布の平均を引いて，標準偏差で割る」という計算をすると，標準得点に変換できます。

### 分布の変換をヒストグラムでみると

この変換を，図 1 でヒストグラムを使って説明しています。ここでは，元の平均を $\mu$，元の標準偏差を $\sigma$ で表しています。ヒストグラムにおいて，度数を表しているのは柱の高さではなく面積であるため，「$(1/\sigma)$ 倍する」という変換でヒストグラムの横方向の広がりが変わると，それに応じて高さも変わることに注意してください。

### 分布の変換をより一般的に

より一般的に，分布中の数値にすべて同じ定数 $a$ をかけて，さらにすべて同じ定数 $b$ を加えることを考えます。この計算を式で書くと，定数 $a,b$ をもってきて，もとの分布の各数値 $x_i$ に対して $z_i = ax_i + b$ という計算をして，別の数値 $z_i$ を作ることになります。これを，「各数値 $x_i$ を，$z_i = ax_i + b$ という 1 次式で $z_i$ に変換する」といいます。また，分布そのものを代表して $X$ であらわして，「分布 $X$ を，$Z = aX + b$ という 1 次式で分布 $Z$ に変換する」という言い方をすることもあります。

このとき，変換前の分布の平均を $\mu_x$・分散を $\sigma_x^2$・標準偏差を $\sigma_x$ とし，変換後の分布の平均を $\mu_z$・分散を $\sigma_z^2$・標準偏差を $\sigma_z$ とすると，

$$\mu_z = a\mu_x + b, \quad \sigma_z^2 = a^2\sigma_x^2, \quad \sigma_z = |a|\sigma_x \tag{3}$$

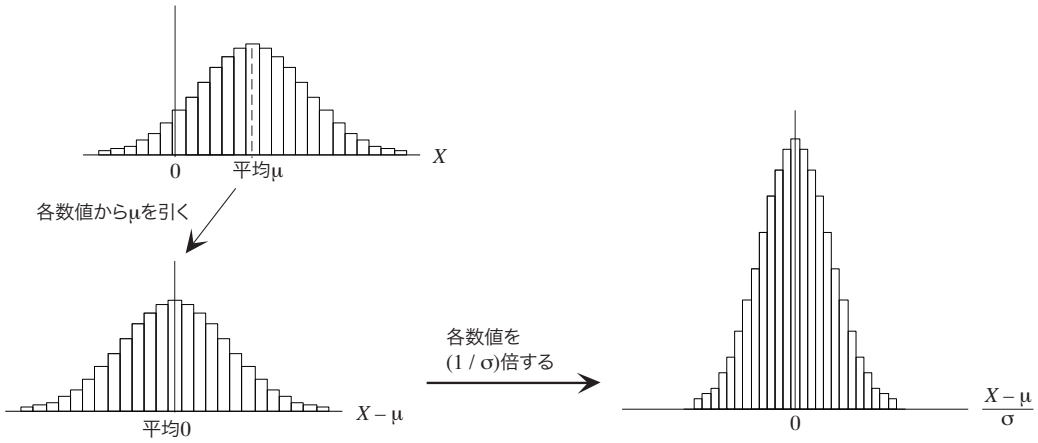


図 1: 度数分布の変換

となります。計算は，付録に載せています。標準得点に変換する計算は，$a = \frac{1}{\sigma_x}, b = -\frac{\mu_x}{\sigma_x}$ とおいた場合に相当し，このとき

$$\mu_z = \frac{1}{\sigma_x}\mu_x + \left(-\frac{\mu_x}{\sigma_x}\right) = 0, \quad \sigma_z = |\frac{1}{\sigma_x}|\sigma_x = 1 \tag{4}$$

となるので，この計算で新しい分布をつくると，その平均は 0，標準偏差は 1 となります。

**受験産業でいう「偏差値」**

標準得点（平均 0，標準偏差 1）に対して，さらに $a = 10$, $b = 50$ とおいて各データをもう一度変換してみます。すると，分布を変換する (3) 式に $\mu_x = 0, \sigma_x = 1, a = 10$, $b = 50$ を代入すると分かるように，変換後の分布は平均 50 点，標準偏差 10 点となります。このように各数値を変換して得られる得点が，受験でおなじみの偏差値です。例えば，偏差値 70 点とは，その試験の平均点よりも標準偏差の 2 倍だけ高い点数であることを表しています。これは，学力テストは 100 点満点で行われることが多いため，30 点〜 70 点あたりのなじみのある値を使って分布中の位置を表現するために考案されたものです。

---

**演習問題**

ある 10 人のクラスで試験を行ったところ，その得点は 35, 45, 50, 55, 60, 60, 60, 70, 75, 90（点）でした．このとき，

1. 平均と標準偏差を求めてください．
2. 得点が 75 点の人の偏差値は何点ですか．
3. 偏差値 65 点は，試験の得点では何点に相当しますか．

---

## 付録：(3) 式の導出

算術平均および分散の定義から，

$$\begin{aligned}\mu_z &= \frac{z_1 + z_2 + \cdots + z_n}{n} \\ &= \frac{(ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b)}{n} \\ &= \frac{a(x_1 + x_2 + \cdots + x_n) + nb}{n} = a\mu_x + b\end{aligned} \tag{A1}$$

となります。また，

$$\begin{aligned}\sigma_z^2 &= \frac{1}{n}\left\{(z_1 - \mu_z)^2 + (z_2 - \mu_z)^2 + \cdots + (z_n - \mu_z)^2\right\} \\ &= \frac{1}{n}\left\{((ax_1 + b) - (a\mu_x + b))^2 + ((ax_2 + b) - (a\mu_x + b))^2 + \cdots + ((ax_n + b) - (a\mu_x + b))^2\right\} \\ &= \frac{1}{n}\left\{a^2(x_1 - \mu_z)^2 + a^2(x_2 - \mu_z)^2 + \cdots + a^2(x_n - \mu_z)^2\right\} \\ &= a^2\frac{1}{n}\left\{(x_1 - \mu_z)^2 + (x_2 - \mu_z)^2 + \cdots + (x_n - \mu_z)^2\right\} = a^2\sigma_x^2\end{aligned} \tag{A2}$$

となりますから，$\sigma_z = |a|\sigma_x$ となります。