MAT220 第3講：中心傾向とばらつき

Contents

第1講では、データ、変数、母集団、標本、標本抽出法など、統計学の基本的な言葉を導入しました。第2講では、ヒストグラム、棒グラフ、円グラフ、時系列グラフ、散布図を用いてデータを視覚的に表す方法を学びました。この講義では、記述統計の数値的な側面に入ります。主な目標は、少数の慎重に選ばれた統計量を用いてデータセットを要約する方法を学ぶことです。今回は二つの大きな考え方に焦点を当てます。一つはデータの中心がどこにあるかを表す中心傾向であり、もう一つはデータがどれだけ散らばっているかを表すばらつきです。

1　中心傾向

この講義では、量的データを扱います。このデータについて構造的に話す必要があるので、まず重要な記法を導入します。この講義で使うすべての公式では、データの値を文字 x で表し、添字を付けて次のように書きます。

$$x_1,\ x_2,\ x_3,\ x_4,\ \ldots,\ x_n.$$

添字は、後で各値を参照するための単なるラベルです。データのラベルは 1 から始まり、n に到達するまで数えていきます。この最後の数 n は、その値が何であれ、データセット全体の大きさを表します。順序が重要な場合には、データセットを小さいものから大きいものへ並べて、次のように書きます。

$$x_1 \le x_2 \le x_3 \le x_4 \le \cdots \le x_n.$$

1.1　中心傾向の意味

多くのデータを集めたとき、私たちはしばしば、それを一つの代表的な数で要約したいと考えます。その一つの方法は、データの「中心」がどこにあるかを記述することです。このようにすると、データ点がどこに集まりやすいかについて述べることができます。これがまさに中心傾向の考え方です。

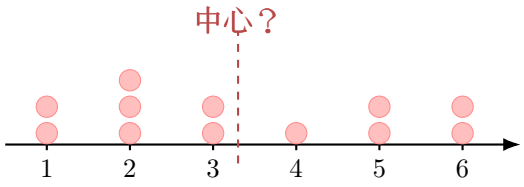
定義：中心傾向

中心傾向とは、データセットの中心を記述しようとすることです。

例えば、次のデータセットを考えます。

$$(1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6).$$

各観測値を小さな球として表すと、このデータを数直線上に描くことができます。



図の中で、データの中心はこの破線であるべきでしょうか。それとも、何か別の場所でしょうか。中心傾向の課題は、中心の位置を見つけるための意味のある方法を考えることです。直感は明確ですが、数学的な問いはそれほど単純ではありません。なぜなら、「中心」という言葉にはいくつもの解釈があるからです。このため、中心傾向には複数の尺度があります。

1.2　総和記号

これから導入する公式では、データに対していくつかの算術操作を行う必要があります。そのため、効率よく書き表す方法も必要になります。これを説明するために、データセットの大き

さが $n = 1000$ で、すべてのデータ値を足し合わせたい場合を考えます。この状況では、毎回手で次のように書き続けるのは避けたいところです。

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + \dots$$

この問題を解決するために、総和記号を使います。

総和記号

$$\sum_{i=1}^{n} x_i \quad \text{は} \quad x_1 + x_2 + x_3 + \cdots + x_n \quad \text{を意味する。}$$

文字 i は**添字**です。これは、現在どの項を足しているかを表します。下の数はどこから始めるかを表し、上の数はどこで止めるかを表します。

上から分かるように、総和記号は非常に大きな和を効率よく書くための方法です。この記法は、単に x_i を足す場合だけでなく、他の場合にも使えます。例えば、次のように書くことができます。

$$\sum_{i=1}^{n} 2x_i = 2x_1 + 2x_2 + 2x_3 + \cdots + 2x_n.$$

ここでは、個々の x_i の値を足し合わせるのではなく、すべての項 $2x_i$ を足し合わせています。

例。

$$(x_1, x_2, x_3) = (4, 10, 12)$$

とします。このとき、

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 4 + 10 + 12 = 26,$$

$$\sum_{i=1}^{2} x_i = x_1 + x_2 = 4 + 10 = 14,$$

また、

$$\sum_{i=2}^{3} x_i^2 = x_2^2 + x_3^2 = 10^2 + 12^2 = 100 + 144 = 244.$$

練習

$$(x_1, x_2, x_3, x_4, x_5) = (2, 4, 6, 8, 10)$$

とします。次を計算しなさい。

(a) $\displaystyle\sum_{i=1}^{4} x_i$

(b) $\displaystyle\sum_{i=3}^{5} x_i$

(c) $\displaystyle\sum_{i=1}^{5} x_i$

(d) $\displaystyle\sum_{i=2}^{3} x_i$

(e) $\displaystyle\sum_{i=1}^{3} \frac{x_i}{2}$

解答

(a) $\displaystyle\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 2 + 4 + 6 + 8 = 20.$

(b) $\displaystyle\sum_{i=3}^{5} x_i = x_3 + x_4 + x_5 = 6 + 8 + 10 = 24.$

(c) $\displaystyle\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 2 + 4 + 6 + 8 + 10 = 30.$

(d) $\displaystyle\sum_{i=2}^{3} x_i = x_2 + x_3 = 4 + 6 = 10.$

(e) $\displaystyle\sum_{i=1}^{3} \frac{x_i}{2} = \frac{x_1}{2} + \frac{x_2}{2} + \frac{x_3}{2} = \frac{2}{2} + \frac{4}{2} + \frac{6}{2} = 1 + 2 + 3 = 6.$

1.3 三つの一般的な平均：平均値、中央値、最頻値

日常語では、「平均」という言葉は曖昧です。これは複数の異なる統計量を指すことがあります。ここでは、最も一般的な三つの平均である**平均値**、**中央値**、**最頻値**を確認します。これらはすべて中心傾向の尺度ですが、それぞれ少し異なる問いに答えます。

統計量	基本的な考え方	主な弱点
平均値	データの釣り合いの中心。	外れ値の影響を受けやすい。
中央値	並べ替えた後の中央の値。	すべての値の数値的な大きさを利用しない。
最頻値	最も頻繁に現れる値。	一意に決まらない場合があり、数値データではあまり多くを語らないことがある。

これから、中央値から始めて、一つずつ確認していきます。

1.3.1 中央値

中央値とは、データセット全体を半分に分ける値です。中央値が見つかると、少なくとも半分のデータは中央値以下にあり、少なくとも半分のデータは中央値以上にあります。

定義：中央値

中央値とは、順序付けられたデータセットの中央の値です。

中央値を見つけるには、まず n の値に注意する必要があります。n が奇数か偶数かによって、少し異なる操作を行います。手順は次の通りです。

Step 1: すべての値を小さいものから大きいものへ並べる。
Step 2: n が奇数なら、中央の項を選ぶ。これは $\frac{n+1}{2}$ 番目の項である。
Step 3: n が偶数なら、二つの中央の項を見て、そのちょうど中間の数を計算する。

例えば、次のデータセットを考えます。

$$(1, 2, 3, 4, 5).$$

ここには 5 個のデータ点があります。つまり $n = 5$ です。したがって、中央値を見つけるには、n に $+1$ を足してから半分にします。つまり $\dfrac{5+1}{2} = 3$ です。したがって、データセットの 3^{rd} 番目の項へ行き、その値を読み取ります。この場合、中央値は 3 です。

次に、次のデータセットを考えます。

$$(2, 3, 4, 5, 6, 7).$$

ここには 6 個の数があります。つまり $n = 6$ であり、これは偶数です。中央値の位置は $\dfrac{6+1}{2} = 3.5$ です。これは、中央値がデータセットの 3^{rd} 番目と 4^{th} 番目の項のちょうど中間にある数になることを意味します。見ると、この二つの項はデータ点 4 と 5 に対応しています。したがって、中央値はこの二つの数の中間、つまり $\dfrac{4+5}{2} = 4.5$ です。

1.3.2 平均値

平均値とは、ふつう私たちがデータセットの「平均」と呼ぶものです。標本

$$x_1, x_2, \ldots, x_n$$

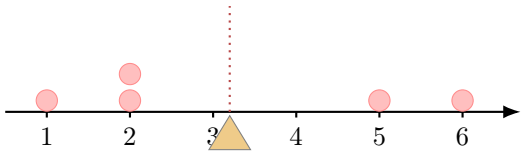
に対して、平均値は $\bar{x}$ と書き、次のように定義されます。

定義：標本平均

標本 $x_1, x_2, \ldots, x_n$ に対して、**標本平均**は

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

言い換えると、平均値 $\bar{x}$ は、データセット内のすべての値を足し合わせ、それを値の個数で割ることによって計算されます。幾何学的には、平均値をデータの釣り合いの点として想像することができます。各データ点を数直線上に置かれた重りだと考えると、平均値は全体の系が釣り合う点です。例えば、データセット $(1, 2, 2, 5, 6)$ では、平均値を中央にある釣り合いの点として考えることができます。



この例では、数値的には次のように計算します。

$$\bar{x} = \frac{1+2+2+5+6}{5} = \frac{16}{5} = 3.2.$$

平均値の重要な性質の一つは、平均値からの正の偏差と負の偏差が常に打ち消し合うことです。

平均値についての便利な事実

任意のデータセットに対して、

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

言い換えると、データから平均値までの符号付き距離の合計は常にゼロです。

上のデータセット $(1, 2, 2, 5, 6)$ を使って、この事実を手で確認できます。平均値 $\bar{x} = 3.2$ の左側には三つの値、すなわち $1, 2, 2$ があり、平均値より大きい値は二つ、すなわち 5 と 6 です。$\bar{x}$ の左側にある値は 3.2 より小さいので、$x_i - \bar{x}$ の値は負になります。絶対値で見ると、実際の距離は

$$(3.2 - 1) + (3.2 - 2) + (3.2 - 2) = 2.2 + 1.2 + 1.2 = 4.6,$$

一方、$\bar{x}$ の右側のデータ点については、

$$(5 - 3.2) + (6 - 3.2) = 1.8 + 2.8 = 4.6.$$

これらの値は、平均値の左側への距離の合計が、平均値の右側への距離の合計と等しいことを示しています。これが、平均値を釣り合いの中心と呼ぶのが自然である理由です。

別の方法として、この平均値に関する事実は代数的にも確認できます。

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x}) &= (x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})\\ &= x_1+x_2+\cdots+x_n-\bar{x}-\bar{x}-\cdots-\bar{x}\\ &= x_1+x_2+\cdots+x_n-n\cdot\bar{x}\\ &= \sum_{i=1}^{n}x_i-n\cdot\frac{\sum_{i=1}^{n}x_i}{n}\\ &= \sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i\\ &= 0.\end{aligned}$$

1.3.3 最頻値

簡単に言えば、最頻値とはデータセットの中で最も頻繁に現れる値です。

定義：最頻値

最頻値とは、データセットの中で最も多く現れる値です。

同じ最高頻度で現れるデータ点が複数ある場合、そのデータセットには一意な最頻値がないと言います。

例。 データセット

$$(1,1,1,3,4)$$

の最頻値は 1 です。なぜなら、1 が三回現れるからです。
データセット

$$(1,1,2,2,3)$$

には一意な最頻値がありません。なぜなら、1 と 2 がどちらも二回ずつ現れるからです。

1.4 例題

次のデータセットを考えます。

$$(1,1,2,2,3,4,5).$$

平均値は

$$\bar{x}=\frac{1+1+2+2+3+4+5}{7}=\frac{18}{7}\approx 2.57.$$

データ点は 7 個あるので、中央の項は

$$\frac{7+1}{2}=4$$

番目です。したがって、中央値は 4 番目の項です。4 番目の項は 2 なので、中央値は 2 です。

値 1 と 2 はどちらも二回現れ、これより多く現れる値はありません。したがって、一意な最頻値はありません。

練習

次のデータセットを考えます。

$$(1, 1, 2, 2).$$

次を求めなさい。

(a) 平均値
(b) 中央値
(c) 最頻値

解答

(a) 平均値は

$$\bar{x} = \frac{1+1+2+2}{4} = \frac{6}{4} = 1.5.$$

(b) データ点は 4 個あるので、二つの中央の値は 2 番目と 3 番目の項です。これらは 1 と 2 なので、中央値は

$$\frac{1+2}{2} = 1.5.$$

(c) 値 1 と 2 はどちらも二回現れます。したがって、一意な最頻値はありません。

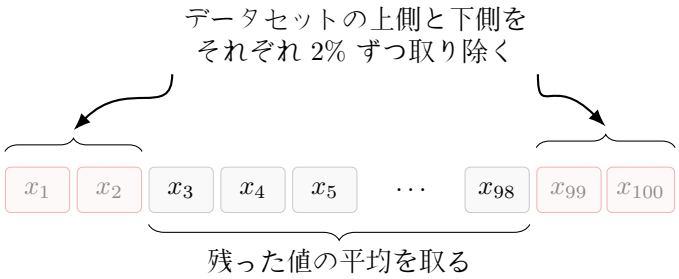
1.5 トリム平均

平均値はデータセット内のすべての値を使います。これは多くの場合に有用ですが、極端な外れ値によって平均値が影響を受けることも意味します。もし一つのデータ値が片側に大きく離れているなら、平均値は釣り合いを保つために大きく補正されることになります。ある状況では、この平均値の性質は許容できるかもしれません。しかし、別の状況では、極端な外れ値の影響を意図的に取り除きたい場合があります。これがトリム平均の考え方を動機付けます。

定義：トリム平均

$x\%$ **トリム平均**とは、順序付けられたデータセットから最小側の $x\%$ と最大側の $x\%$ を取り除き、残った値の平均を計算することで得られる平均です。

言い換えると、$x\%$ トリム平均は、順序付けられたデータセットの下位 $x\%$ と上位 $x\%$ を切り落とし、残りの平均を取ることで作られます。例えば、順序付けられた大きさ 100 のデータセットがある場合、2% トリム平均は二つの最小値と二つの最大値を取り除きます。その後、残った 96 個の値の平均を取ります。



トリム平均は、人間の判断の集まりを平均化するときに特に役立ちます。例えば、フィギュアスケートでは、審査員団が一つの演技に点数を付けます。一人の審査員が非常に高い点を付け、別の審査員が偏りや人為的な誤りによって非常に低い点を付けた場合、それらの極端な点数が平均値に不公平な影響を与えるかもしれません。この効果を抑えるために、演技の得点をトリム平均で計算し、極端な判断の影響を小さくすることがあります。

練習：トリム平均

次の順序付けられたデータセットを考えます。

$$1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24.$$

ここで $n = 20$ です。

(a) 5% トリム平均を計算しなさい。
(b) 20% トリム平均を計算しなさい。
(c) ここで 50% トリム平均が意味を持たない理由を説明しなさい。

解答

(a) 20 の 5% は 1 なので、最小値と最大値を取り除きます。残ったデータは

$$1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23.$$

これらの値の合計は 149 なので、5% トリム平均は

$$\frac{149}{18} \approx 8.28.$$

(b) 20 の 20% は 4 なので、下から四つの値と上から四つの値を取り除きます。残ったデータは

$$4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11.$$

これらの値の合計は 84 なので、20% トリム平均は

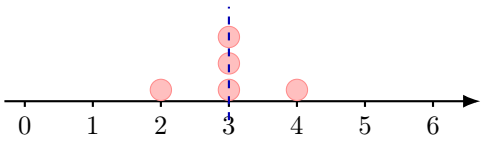
$$\frac{84}{12} = 7.$$

(c) 下から 50%、上から 50% を取り除くと、データセット全体が取り除かれます。平均を取る値が残らないので、50% トリム平均は意味を持ちません。

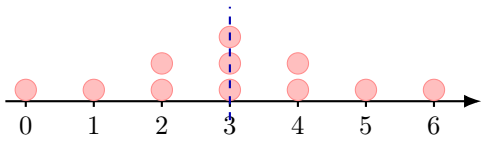
2　ばらつき

中心傾向の尺度を使うと、データが中心との関係でどこに集まりやすいかを特定できます。しかし、中心傾向だけでは、データがどれだけ散らばっているかは分かりません。例えば、下に描かれた二つのデータセットを考えます。

$A = (2, 3, 3, 3, 4)$



$B = (0, 1, 2, 2, 3, 3, 3, 4, 4, 5, 6)$



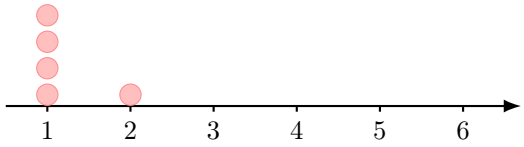
この二つのデータセットは同じ平均値を持ちます。どちらも 3 です。しかし、データセット B はデータセット A よりも明らかに大きく散らばっています。データの中心を一つだけ記述しても、散らばりという特徴を捉えることはできません。そこで、新しい概念であるばらつきを導入する必要があります。

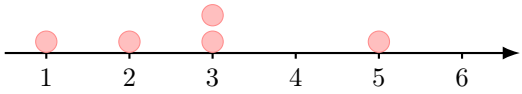
定義：ばらつき

ばらつきとは、データがどれだけ散らばっているかを表すものです。

2.1　範囲

下に描かれた二つのデータセット $A = (1, 1, 1, 1, 2)$ と $B = (1, 2, 3, 3, 5)$ を考えます。





視覚的には、B の方がより散らばっており、A の方がより密集していることは明らかです。これを数量化する一つの方法は、データセットの最大値と最小値の間の距離を見ることです。A ではデータセット全体が 1 と 2 の間に含まれていますが、B ではデータセット全体が 1 から 5 まで広がっています。

この単純な観察が、散らばりに関する最初の定義であるデータセットの範囲を動機付けます。

定義：範囲

データセットの**範囲**は

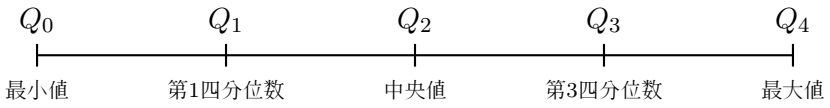
$$\text{範囲} = \max(\text{データ}) - \min(\text{データ})$$

です。

例えば、データセット $A = (1, 1, 1, 1, 2)$ の範囲は $2-1 = 1$ であり、データセット $B = (1, 2, 3, 3, 5)$ の範囲は $5 - 1 = 4$ です。したがって、範囲は視覚的に見たことを確認しています。つまり、B は A よりも散らばっています。

2.2 四分位数

「四分の一」という言葉は、何かの四分の一を意味します。これに対して、四分位数とは、順序付けられたデータセットを四つの部分に分けるのに役立つ値を意味します。全体には四つの部分があるので、それらを分けるには五本の線が必要です。中央に三本、端点に二本です。



五つの値

$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4$$

は、最小値、第1四分位数、中央値、第3四分位数、最大値を示します。

四分位数

順序付けられたデータセットに対して、

- Q_0 は最小値です。
- Q_1 はデータの下半分の中央値です。
- Q_2 はデータセット全体の中央値です。
- Q_3 はデータの上半分の中央値です。
- Q_4 は最大値です。

教科書によって、四分位数には少し異なる慣習が使われることがあります。特に、データセットの値の個数が奇数の場合に違いが出ます。この講義では、次の実用的な規則を使います。まず中央値を見つけ、その後で下半分の中央値と上半分の中央値を見つけます。n が奇数の場合、Q_1 と Q_3 を求めるときには、中央値をどちらの半分にも含めません。

2.3 四分位範囲

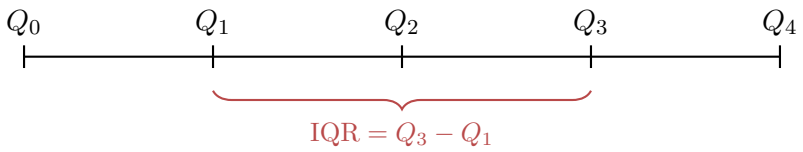
定義：四分位範囲

四分位範囲、略して **IQR** は

$$\mathrm{IQR} = Q_3 - Q_1$$

です。これはデータの中央半分の散らばりを測ります。

視覚的には、四分位範囲はこの距離を測ります。



IQR は、範囲よりも外れ値に対して頑健であることが多いです。なぜなら、データの中央半分に焦点を当て、下位四分の一と上位四分の一を無視するからです。

2.3.1 例題：四分位数と IQR

次の順序付けられたデータセットを考えます。

$$(1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24).$$

四分位数を計算するために、まず中央値を見つけます。データ点は全部で 20 個あるので、中央値は 10 番目と 11 番目の項の中間にあります。これらの項は 6 と 7 なので、中央値は $Q_2 = \frac{6+7}{2} = 6.5$ です。Q_1 の値を求めるには、データの下半分の中央値を計算する必要があります。書き出すと、次のデータの中央を探していることになります。

$$(1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6).$$

この下半分の中央値は 4 と 5 の中間にあるので、

$$Q_1 = \frac{4+5}{2} = 4.5.$$

同様に、データの上半分は

$$(7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24)$$

です。この上半分の中央値は 10 と 11 の中間にあるので、$Q_3 = \frac{10+11}{2} = 10.5$ です。

これらをまとめると、五数要約は

$$Q_0 = 1, \qquad Q_1 = 4.5, \qquad Q_2 = 6.5, \qquad Q_3 = 10.5, \qquad Q_4 = 24,$$

そして四分位範囲は次のように計算できます。

$$\mathrm{IQR} = Q_3 - Q_1 = 10.5 - 4.5 = 6.$$

2.4 分散

範囲は二つの数、つまり最小値と最大値だけを使います。このため計算は速いですが、データセットの大部分には目を向けていません。また、範囲は外れ値に非常に敏感です。例えば、次の二つのデータセットを考えます。

$$A = (1, 1, 1, 2, 2, 2, 3, 6) \qquad B = (1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6)$$

どちらのデータセットも同じ範囲を持ちます。つまり $6-1=5$ です。しかし、視覚的には、データセット A はデータセット B よりも散らばりが小さいという感覚がまだあります。データセット A では、ほとんどのデータが $1,2,3$ の周辺に集まっており、範囲を 5 にしているのは一つの外れ値です。対照的に、データセット B ではデータがより一様に広がっており、外れ値はありません。

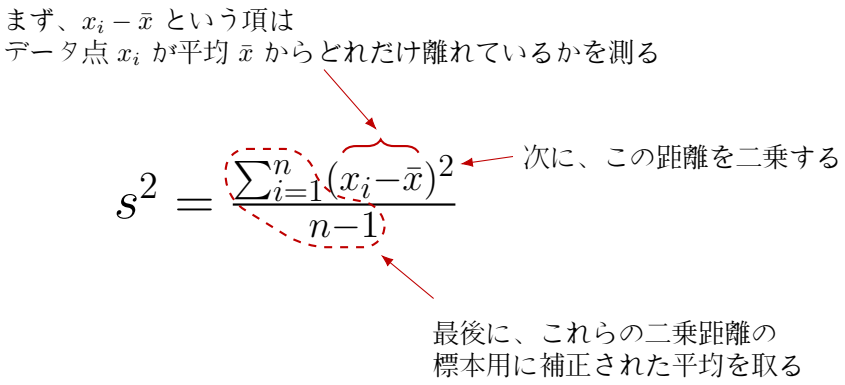
ここで、データセットの範囲に関する根本的な弱点が明らかになりました。範囲はばらつきの初期的な尺度としては有用ですが、外れ値に非常に敏感であり、そのためデータセットの内部構造を見逃すことがあります。これを補うために、より慎重な散らばりの尺度である分散を導入します。

定義：標本分散

標本 $x_1, x_2, \ldots, x_n$ と標本平均 $\bar{x}$ に対して、**標本分散**は

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

見て分かるように、この公式はこれまで見てきたものよりかなり複雑です。そこで、s^2 が実際に何を表しているのかを説明するために、これを段階的に分解します。下は公式の注釈です。



要するに、分散 s^2 は、各データ値と平均値の間の二乗距離について、標本用に補正された平均を測るものです。

1. 式 $x_i - \bar{x}$ は、点 x_i が中心 $\bar{x}$ からどれだけ離れているかを測ります。これは負の値を取ることがあるので、厳密には距離ではなく、偏差と呼ばれます。
2. 式 $(x_i - \bar{x})^2$ は、これらすべての偏差を二乗します。
3. 次に、総和記号 $\sum_{i=1}^{n}(x_i - \bar{x})^2$ は、これらすべての二乗偏差を足し合わせることを意味します。つまり、
$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2.$$
4. 最後に、$n-1$ で割ることで標本分散を得ます。

これから、なぜ特にこの公式になるのかを詳しく説明します。

2.4.1　なぜ偏差を二乗するのか

分散の公式がなぜ偏差を二乗するのかを尋ねるのは自然なことです。代わりに、平均値からの生の偏差をただ足し合わせてはいけないのでしょうか。実は、その問題は以前見た平均値の重要な性質の中に隠れています。

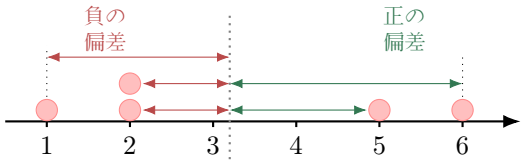
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

これは、すべての生の偏差の和が常にゼロになることを述べています。もし偏差を二乗せずにそのまま使おうとすると、常に次のようになります。

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{0}{n-1} = 0,$$

どのようなデータから始めても同じです。例えば、次のデータセットを考えます。

データ：$(1, 2, 2, 5, 6)$



この平均値は $\bar{x} = 3.2$ です。 生の偏差を足し合わせようとすると、

$$(1 - 3.2) + (2 - 3.2) + (2 - 3.2) + (5 - 3.2) + (6 - 3.2) = 0.$$

となります。

二乗することで、この問題は一般に解決されます。偏差を二乗すると、負の偏差も正の偏差も非負の二乗になります。

$$(-2)^2 = 4, \qquad 2^2 = 4.$$

これにより、和は非負になるので、上で述べた問題を避けられます。[1]

2.4.2　なぜ $n-1$ で割るのか

通常、平均を取るときには、値を足し合わせて値の個数で割ります。したがって、標本分散が n ではなく $n-1$ で割る理由を尋ねるのは自然です。

短く言うと、s^2 は**標本統計量**であり、**母集団パラメータ**ではないからです。通常、私たちは標本を使って、より大きな母集団の散らばりを推定します。公式の中で標本平均 $\bar{x}$ を使うとき、私たちはすでに標本から一つの量を推定しています。このため、二乗偏差は平均的に少し小さくなります。$n-1$ で割ることで、この効果を補正します。

[1] また、二乗することで、絶対値の大きい偏差に特に大きな重みが与えられます。

ベッセルの補正

標本分散の公式で n ではなく $n-1$ を使うことを**ベッセルの補正**と呼びます。この講義での主なポイントは実用的なものです。

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

が標本分散の標準的な公式です。

この講義では、この補正を証明しません。今のところ、どの公式を使うべきか、そしてそれが何を測っているのかを知っていれば十分です。

2.4.3 例題：分散

例。 次のデータセットを考えます。

$$(1, 1, 2, 3).$$

まず平均値を計算します。

$$\bar{x} = \frac{1+1+2+3}{4} = \frac{7}{4} = 1.75.$$

次に、偏差と二乗偏差を計算します。

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

二乗偏差の和は

$$0.5625 + 0.5625 + 0.0625 + 1.5625 = 2.75.$$

$n = 4$ なので、$n - 1 = 3$ で割ります。

$$s^2 = \frac{2.75}{3} \approx 0.917.$$

2.5 標準偏差

分散は有用ですが、一つ扱いにくい特徴があります。それは、二乗された単位で測られることです。例えば、データが分単位で測られている場合、分散は平方分で測られます。この量は、元のデータと直接結び付けて解釈しにくいことがあります。この問題は、分散の平方根を取ることで解決できます。そうすれば、結果は元のデータと同じ単位になります。これがまさに標準偏差の定義です。

定義：標本標準偏差

標本標準偏差は

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

です。

標準偏差はデータと同じ単位を持ちます。そのため、平均値からの距離として解釈しやすくなります。

前の例では、

$$s^2 \approx 0.917$$

であることを求めました。したがって、

$$s = \sqrt{0.917} \approx 0.958.$$

よって、標準偏差は約 0.958 です。

2.6 変動係数

時には、非常に異なる尺度を持つデータセットの散らばりを比較したいことがあります。例えば、標準偏差 2 は、平均値が 4 なら大きいですが、平均値が 2000 なら非常に小さいです。データセットの大きさに対する相対的な散らばりを表すために、変動係数を使うことができます。

定義：変動係数

正の比率尺度データに対して、変動係数は相対的な散らばりを表す単位のない尺度です。これは次の比として定義されます。

$$\mathrm{CV} = \frac{s}{\bar{x}}.$$

変動係数が大きいほど、典型的な値の大きさと比較して、より大きなばらつきがあることを意味します。

この統計量は、平均値が正であり、ゼロに近すぎないときに最も有用です。平均値がゼロに非常に近い場合、平均値で割ることによって、変動係数は誤解を招いたり不安定になったりすることがあります。

2.7 標本統計量と母集団パラメータ

第1講で述べたように、推測統計では、標本統計量を用いて母集団パラメータを推定することがよくあります。このため、標本と母集団には異なる記法を使います。

量	標本統計量	母集団パラメータ
平均	$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$	$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$
分散	$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$	$\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
標準偏差	$s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

ここで、n は標本サイズ、N は母集団サイズです。標本分散と標本標準偏差にはベッセルの補正を使っていることに注意してください。しかし、母分散や母標準偏差では、その補正は必要ありません。なぜなら、もはや推定をしているわけではないからです。

3 箱ひげ図

3.1 五数要約

箱ひげ図は、五数要約を視覚的に表したものです。

五数要約

データセットの**五数要約**は、

$$\text{最小値}, \quad Q_1, \quad \text{中央値}, \quad Q_3, \quad \text{最大値}$$

です。同値に、

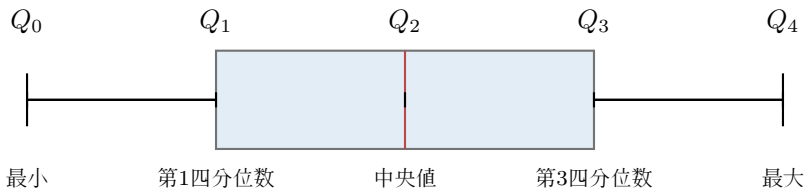
$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4.$$

箱ひげ図は、中心傾向とばらつきを同時に示すので便利です。中央値はデータの中心を示し、箱はデータの中央半分を示し、ひげはデータの外側の部分を示します。

3.2 箱ひげ図の構造

この講義で使う単純な箱ひげ図では、

- 左のひげは最小値 Q_0 から始まります。
- 箱の左端は Q_1 です。
- 箱の中の線は中央値 Q_2 です。
- 箱の右端は Q_3 です。
- 右のひげは最大値 Q_4 で終わります。



箱は Q_1 から Q_3 まで伸びる。ひげは Q_0 から Q_1、および Q_3 から Q_4 まで伸びる。

定義：箱ひげ図

この講義で使う単純な形では、**箱ひげ図**とは、最小値、Q_1、中央値、Q_3、最大値からなる五数要約を用いてデータセットを表示するグラフです。

3.3 箱ひげ図の作り方

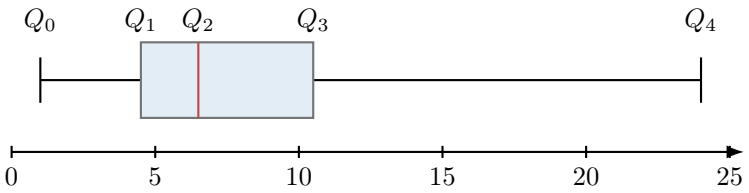
次のデータセットを考えます。

$$1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24.$$

すでに五数要約を次のように求めました。

$$Q_0 = 1, \qquad Q_1 = 4.5, \qquad Q_2 = 6.5, \qquad Q_3 = 10.5, \qquad Q_4 = 24.$$

その結果として得られる箱ひげ図は次の通りです。



右のひげが左のひげよりもかなり長いことに注意してください。これは、データの上側の端が下側の端よりも大きく散らばっていることを示しています。