データリテラシー

第5回　代表値と散らばりの尺度

メディア情報コース
平居　悠（ひらい　ゆたか）

到達目標

- データサイエンス・AIが社会でどのように活用されているのかを理解する。
- データを適切に読み解き理解する力をつける。
- データを適切な可視化手法で他者に説明できるようになる。
- 表計算ソフトを使ってデータを集計・加工できるようになる。

前回の目標

度数分布表を用いてヒストグラムを描けるようになる。

前回学んだこと

1. データとは
2. 度数分布表
3. ヒストグラム

データとは

何らかの実験・観測・調査を行って得られる観測値の集合

生データ

実験を行って記録したデータやアンケート調査で得られたデータで、集計などの加工を行う前のデータ

完全データと不完全データ

- **完全データ**：全ての観測値がそろっているデータ
- **不完全データ**：所々の観測値が欠けているデータ
- **欠損値（欠測値）**：観測値が欠けているところ

表による表現

例１：ある集団の性別、年齢、身長の生データ

番号	性別	年齢[歳]	身長[cm]
1	女	22	160.6
2	男	19	171.6
3	男	21	173.2
⋮	⋮	⋮	⋮

変数（変量）

分布

データがどういう値でどのように散らばっているかを**分布**という。

分布を読み取る方法

1. **グラフを描いて視覚的に読み取る**
2. **代表値と散らばりの尺度で分布の特徴を数量的に読み取る**

データの分布をグラフとして表現するために、度数分布表を作成してグラフ（**ヒストグラム**）に描くことがよく行われる。

度数分布表とは

生データをいくつかの**階級**というグループに分け、それぞれの階級に入る観測値がいくつか数えて**度数（頻度）**で表した表

度数分布表の作成方法

①全体の範囲を決める

データの最小値と最大値から全体の範囲を決める。階級幅と合わせてきりがよくなるように決めた方が分かりやすい。

度数分布表の作成方法

②階級数と階級幅を決める

それぞれの階級の**階級幅**は上限値から下限値を引いたものに等しい。普通は階級幅を等幅にとるため、全体の範囲を階級数で割ったものが階級幅である。

度数分布表の作成方法

③各階級の下限値と上限値を設定する

階級数と階級幅をもとに各階級の下限値と上限値を設定する。*i*番目の階級を代表する値を階級値v_iという。階級値v_iは普通下限値l_iと上限値u_iの中央の値にとる。

$$v_i = \frac{l_i + u_i}{2}$$

度数分布表の作成方法

④観測値がどの階級に属するか調べて度数として数える

i番目の階級にいくつの観測値が入るかを数えて度数f_iとする。通常**下限値以上、上限値未満**の観測値を数える。

（下限値）≦（観測値）＜（上限値）

度数分布表の作成方法

⑤度数の合計が観測値の総数と一致することを確認する

観測値はいずれかの階級に属するため、度数f_iの合計は観測値の総数nに一致する。

$$n = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i$$

度数分布表の作成方法

⑥累積度数や相対度数を求める

必要なら度数から累積度数や相対度数を求める。

ヒストグラムとは

度数分布表を柱状グラフで表したもの

ヒストグラム

ヒストグラムを描くことで、分布の中心位置や散らばり具合などの特徴を知ることができる。

前回の問い

- 度数分布表とは何か？
  - データをいくつかの階級に分け、それぞれの階級に入る観測値がいくつか数えて度数で表した表。
- ヒストグラムを描くと何がわかるか？
  - 分布の中心位置や散らばり具合などの特徴を知ることができる。

タイピング練習スケジュール

第１回　　ホームポジション
第２回　　ホームポジション
第３回　　ローマ字
第４回　　英語初級
第５回　　日本国憲法（trr試験）
第６回　　ホームポジション
第７回　　ローマ字
第８回　　英語初級
第９回　　日本国憲法（trr試験）
第１０回　ホームポジション
第１１回　ローマ字
第１２回　英語初級
第１３回　日本国憲法（trr試験）

trr試験 (日本国憲法) 試験時間：15分

1. ブラウザを起動し、https://www.koeki-prj.org/trr/ に繋ぐ。
2. 学籍番号（Cは大文字、省略なし８桁）を入力する。
3. Koeki MAILに届いたパスコードをPasscode: 欄に入力する。

今回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析

今回の目標

代表値と散らばりの尺度から数量的にデータの分布の特徴を読み取れるようになる。

今回学ぶこと

1. 変数
2. 代表値
3. 散らばりの尺度

今回の問い

- 量的変数と質的変数の違いは何か？
- 代表値とは何か？
- 散らばりの尺度には何があるか？

今回学ぶこと

1. 変数
2. 代表値
3. 散らばりの尺度

変数

様々な値に変化するデータ

変数の種類と尺度

変数の種類	変数の尺度
量的変数	比率尺度
	間隔尺度
質的変数	順序尺度
	名義尺度

量的変数

長さや温度など定量的な値（数値）で観測されるデータを扱う変数

量的変数

単位を持つ場合が多い
例：100 g, 10 ℃

割合や指数のように単位を
持たない場合もある。

量的変数の種類

- 離散型変数
- 連続型変数

離散型変数

変数の値としてとびとびの
値しかとらない変数

離散型データ

離散型変数で表されるデータ

例

- サイコロの目：1, 2, 3, 4, 5, 6
- 人数：1人, 2人, 3人
- 個数：1個, 2個, 3個

連続型変数

変数の値として連続的な値をとりうる変数

連続型データ

連続型変数で表されるデータ

例

- 身長：168.3 cm
- 体重：60.1 kg
- 温度：36.2 ℃
- 時間：36分49.2秒

質的変数

定量的な値では観測できないデータを扱う変数

例

- 性別：男, 女
- 血液型：A型, B型, O型, AB型
- GP：S, A, B, C, D

質的変数の注意点

便宜上、数値を割り当てて扱うこと（**数値化**）ができるが、量的変数にはならない。

例：血液型でA型=0, B型=1, AB型=2, O型=3と数字を割り当てられるが、AB型=1, A型=2, B型=3, O型=4のように割り当てることもできる。

変数の尺度

変数を測定する基準

変数の尺度の種類

- 比率尺度
- 間隔尺度
- 順序尺度
- 名義尺度

比率尺度（比例尺度、比尺度）

原点が決まっており、間隔と比率に意味がある。比率尺度の変数同士では四則演算（足し算、引き算、掛け算、割り算）ができる。

例

- 長さ：175 cm, 2236 m
- 重さ：10 g, 1.5 kg
- 価格：750 円, $150

間隔尺度

数値の間隔に意味がある。比率尺度の変数同士では足し算と引き算ができる。

例

- 暦年：西暦2021年, 令和5年（和暦）
- 温度：36.5 ℃

比較尺度と間隔尺度の違い

比率に意味があるかどうか

例：間隔尺度である暦年では「西暦2000年は西暦1000年の2倍である」という言い方はできない。

順序尺度

順序に意味がある。順序尺度の変数同士の演算には意味がない。

例

- 順位：1位, 2位, 3位
- 科目の評価：優, 良, 可, 不可

順序尺度の例：順位

試験の結果

名前	点数	順位
Aさん	100	1
Bさん	92	2
Cさん	8	3

順位自体は数の上では引き算ができるが、1位と2位の点差と2位と3位の点差が同じとは言えないので、順位の引き算には意味がない。

名義尺度

単に分類するためだけに使われる。便宜上、数値化はできるが、順序や演算には意味がない。

例

- 血液型：A型, B型, AB型, O型
- コース：経営, 政策, 地域経営, 観光まちづくり, メディア情報

今回の問い

- **量的変数と質的変数の違いは何か？**
  - 定量的な値で観測されるデータを扱う変数を**量的変数**といい、定量的な値では観測できないデータを扱う変数を**質的変数**という。
- **代表値とは何か？**
- **散らばりの尺度には何があるか？**

今回学ぶこと

1. 変数
2. 代表値
3. 散らばりの尺度

代表値

分布の中心を表す値

例：平均値、中央値、最頻値

平均値（算術平均）

それぞれのデータの総和をデータの個数で割ったもの

算術平均の表記

変数xの場合、通常算術平均は$\bar{x}$のように変数の上に横棒（バー）を付けて表す。

生データの平均値

生データ$x_i(\mathrm{i} = 1, 2, \ldots, \mathrm{n})$の平均$\bar{x}$は以下で定義される

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

度数分布の平均値

度数分布の階級値v_i、度数　$f_i(i = 1, 2, \ldots, n)$とすると、平均は以下で定義される。

$$\bar{x} = \frac{v_1 f_1 + v_2 f_2 + \cdots + v_k f_k}{n} = \frac{1}{n} \sum_{i=1}^{k} v_i f_i$$

ここで

$$n = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i$$

度数分布では、v_1の値がf_1個、 v_2の値がf_2個という考え方をしている。

度数分布の平均

相対度数 f_i/n を使うと次のような表し方もできる。

$$\bar{x} = \sum_{i=1}^{k} v_i \left(\frac{f_i}{n} \right)$$

生データの平均と、その生データの度数分布の平均は一致するとは限らない。

中央値（メディアン）

データを**小さい順（もしくは大きい順）に並べたとき**のちょうど中央の値

生データの中央値

生データ$x_i(\mathrm{i} = 1, 2, \ldots, \mathrm{n})$が小さい順（または大きい順）に並んでいるとすると、中央値x_Mの定義はnが奇数か偶数かで異なる。

$$x_M = \begin{cases} x_{(n+1)/2} & (n\text{が奇数}) \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & (n\text{が偶数}) \end{cases}$$

生データの中央値

例えば、$x_1, x_2, x_3, x_4, x_5 (n = 5)$の場合の中央値は以下の通りである。

$$x_M = x_{(5+1)/2} = x_3$$

x_M ↓				
x_1	x_2	x_3	x_4	x_5

生データの中央値

例えば、$x_1, x_2, x_3, x_4, x_5, x_6 (n = 6)$の場合の中央値は以下の通りである。

$$x_M = \frac{x_{6/2} + x_{(6/2)+1}}{2} = \frac{x_3 + x_4}{2}$$

x_M ↓					
x_1	x_2	x_3	x_4	x_5	x_6

度数分布の中央値

次の手順で求める。

①i番目の階級において下限値l_i、上限値u_i、累積度数F_iとすると、$F_{j-1} < \frac{n}{2} \leq F_j$を満たす$j$番目の階級に中央値が含まれるので、まずはその階級を見つける。

度数分布の中央値

②中央値を含む階級が見つかったら、一つの階級では観測値が一様に分布すると考え、線形補間で中央値を求める。

$$x_M = l_j + q(u_j - l_j)$$

ここで

$$q = \frac{\frac{n}{2} - F_{j-1}}{F_j - F_{j-1}}$$

最頻値

観測値の中で最も多く現れる観測値

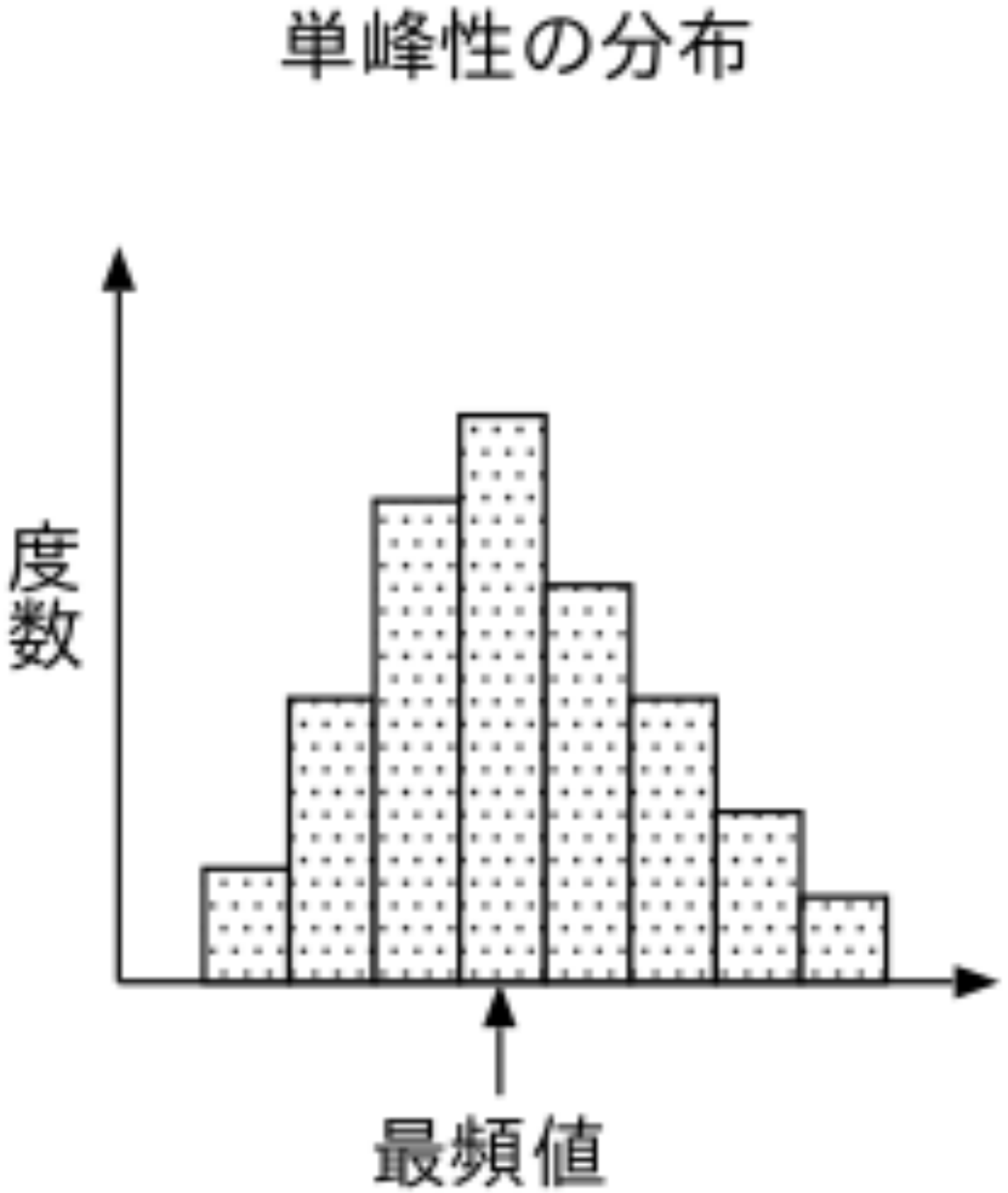
生データの最頻値

離散型データの場合は同じ観測値がいくつ現れるのかを数えて最頻値を求める。

連続型データの場合は度数分布表を作ってその最頻値を求める。

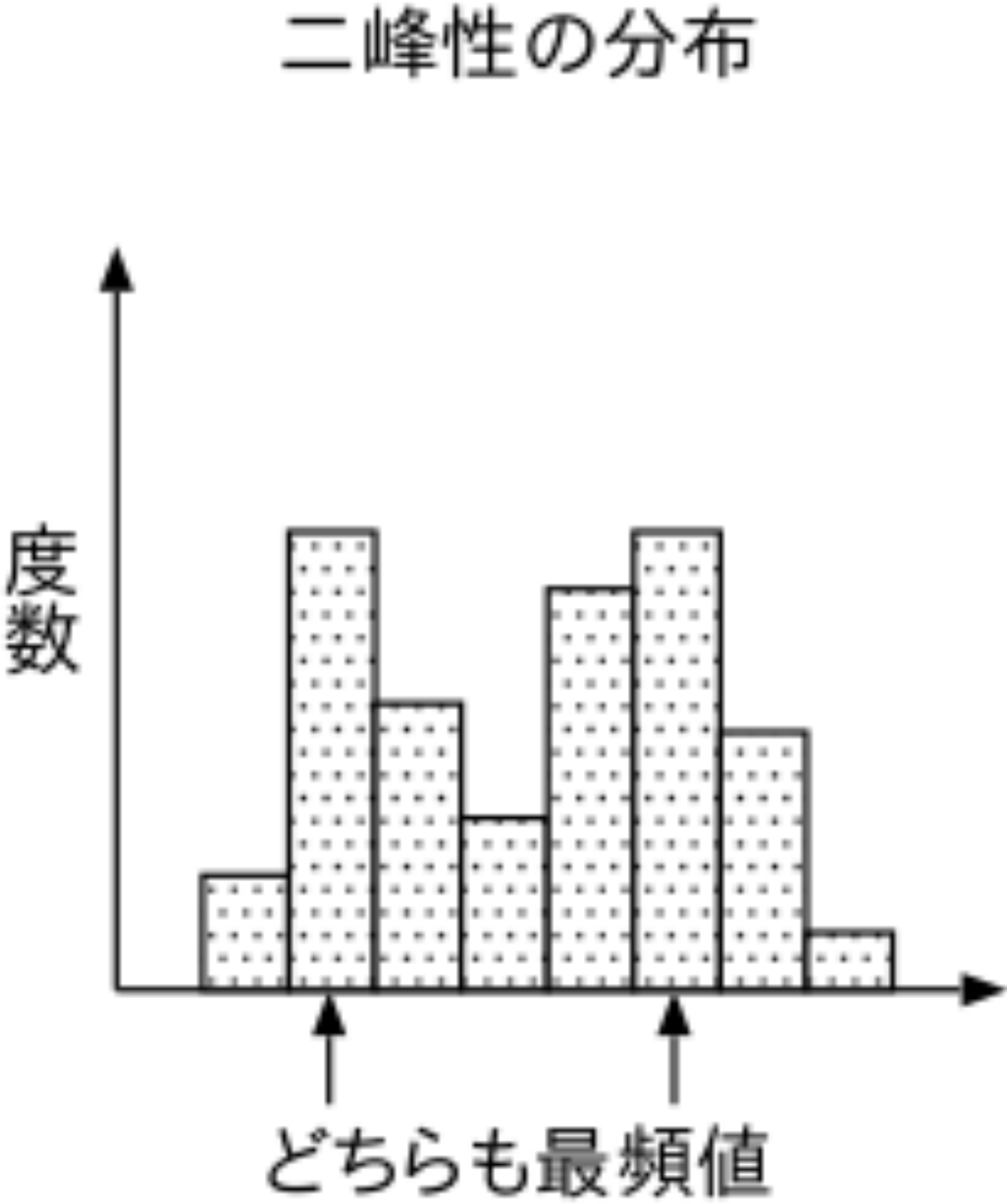
度数分布の最頻値

度数が最大となる階級値が最頻値で、ヒストグラムにすると**最も高い峰の値**。元が同じデータでも階級の取り方で最も高い峰は変わりうるので、最頻値が異なることがある。



度数分布の最頻値

最も高い峰が二つ以上ある場合は、最頻値を一義的に定義することができない。



演習：代表値の計算（生データ）

dataset1.csv列Dの「身長」のデータから代表値を求めよう。

平均値

・空のセルに「**=AVERAGE(D2:D501)**」と入力する。

中央値

・空のセルに「**=MEDIAN(D2:D501)**」と入力する。

最頻値

・空のセルに「**=MODE(D2:D501)**」と入力する。

・この生データで最頻値を見てもあまり意味はない。

演習：代表値の計算（度数分布）

前回の「演習：度数分布表の作成」で作成した身長の度数分布から代表値を求めよう。

平均値

・空のセルに「**=SUMPRODUCT(J3:J12,K3:K12)/500**」と入力する。

中央値

・$\frac{n}{2} = 250$なので、5番目の階級に中央値が含まれる。

・空のセルに「**=H7+(250-L6)/(L7-L6)*(I7-H7)**」と入力する。

最頻値

・度数が最大となる階級値が最頻値である。

・空のセルに「**=INDEX(J3:J12, MATCH(MAX(K3:K12), K3:K12, 0))**」と入力する。

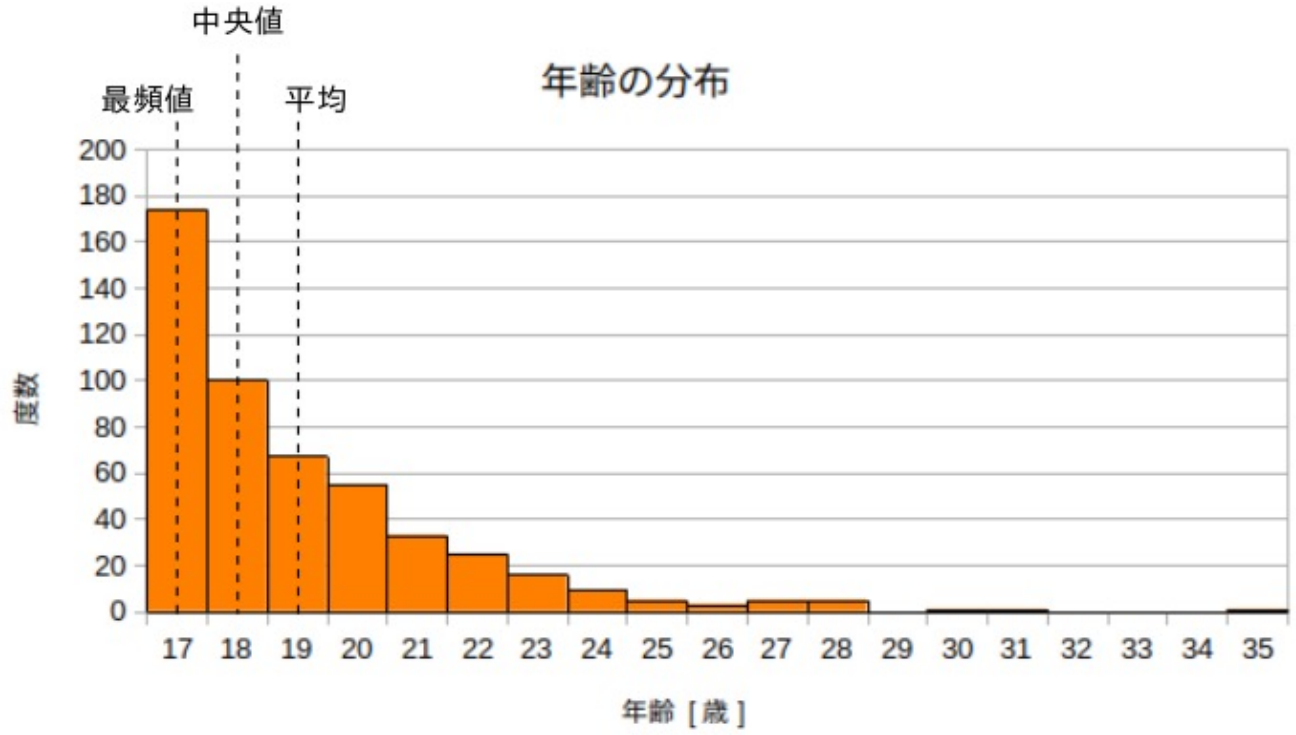
代表値の違い

平均、中央値、最頻値は分布の中心を表すが、分布によってそれぞれ同じ値になるとは限らない。

- **左右対称の分布「平均＝中央値＝最頻値」**
- **右に歪んだ分布「最頻値＜中央値＜平均」**
- **左に歪んだ分布「平均＜中央値＜最頻値」**

代表値の違いの例

dataset1.csvの年齢の分布のグラフ。代表値を計算すると、平均が19.1歳、中央値が18歳、最頻値が17歳。



このように右に歪んだ分布（値が小さい方に度数が偏った分布）では、「最頻値＜中央値＜平均」となる。このような分布では、平均値より最頻値を分布の中心とみなした方が分かりやすい。

外れ値

極端に大きい値や小さい値。平均値は外れ値に影響されやすいが、中央値は外れ値に影響されにくい。

外れ値の例

以下の国語の点数のデータでは、$x_5 = 98$は他の値より極端に値が大きい外れ値である。

	x_1	x_2	x_3	x_4	x_5
点数	3	8	11	15	98

- 平均 $\bar{x} = \frac{3+8+11+15+98}{5} = 27$
- 中央値 $x_M = x_3 = 11$

パーセンタイルと四分位数

パーセンタイル（分位数）

データを**小さい順に並べたとき**の小さい方から$100p\%$のところの値を$100p$ パーセンタイルという $(0 \leq p \leq 1)$。データを$p:1-p$に分ける点ともいえる。

中央値＝50パーセンタイル $(p = 0.5)$

四分位数

パーセンタイルのうち、特に25%, 50%, 75%パーセンタイルを**四分位数**という。これらはデータをちょうど四分割する点にあたるため、**四分位点**ともいう。

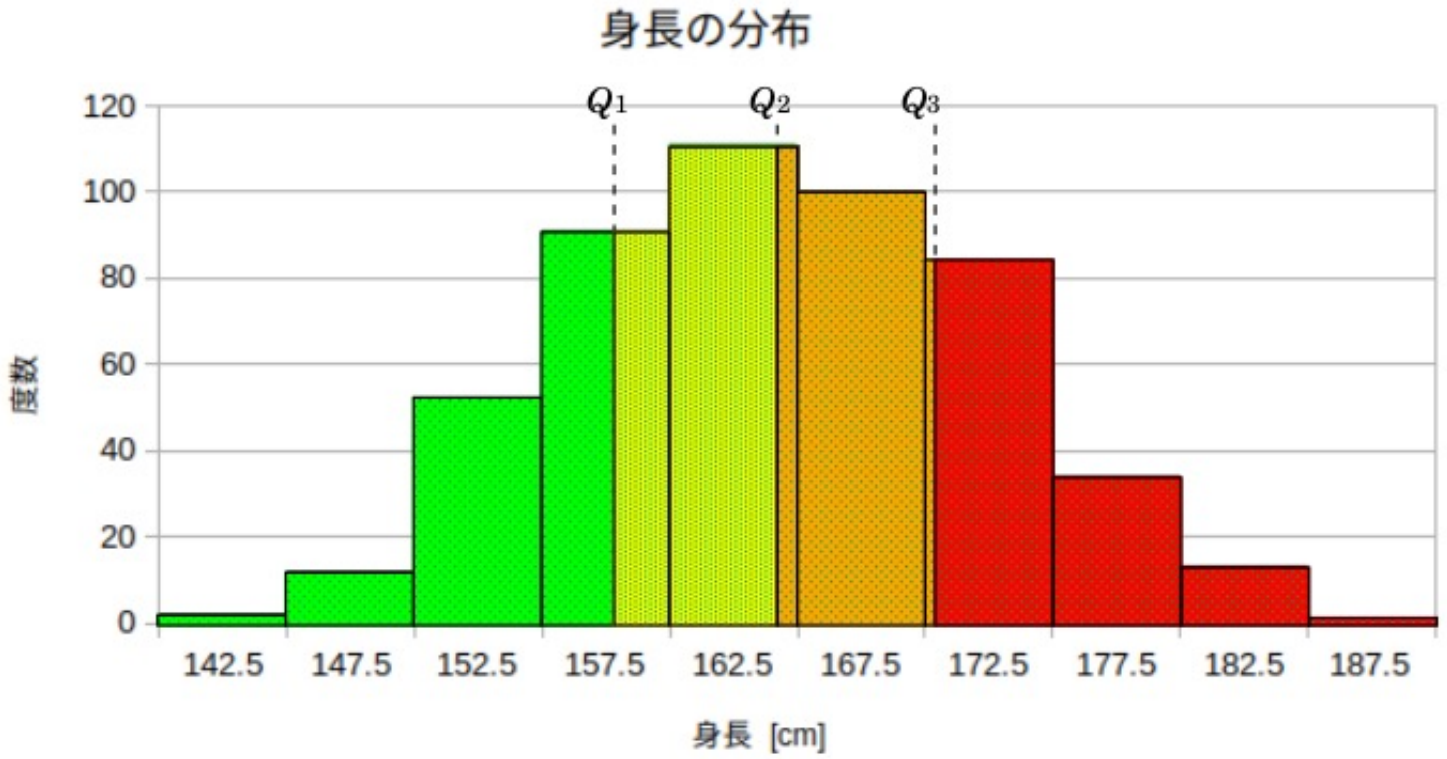
- 第1四分位数 Q_1 = 25パーセンタイル
- 第2四分位数 Q_2 = 50パーセンタイル（= 中央値）
- 第3四分位数 Q_3 = 75パーセンタイル

これに最小値Q_1と最大値Q_4を加えてQ_0, Q_1, Q_2, Q_3, Q_4の5つの指標を**五数要約**という。

四分位数の例

dataset1.csvの身長のデータでは四分位数は以下のようになる。

- $Q_1 = 157.775[\mathrm{cm}]$
- $Q_2 = 164.1[\mathrm{cm}]$
- $Q_3 = 170.3[\mathrm{cm}]$



演習：四分位数の計算

dataset1.csv列Dの「身長」のデータから四分位数を求めよう。

第1四分位数

・空のセルに「**=QUARTILE.INC(D2:D501, 1)**」と入力する。

第2四分位数

・空のセルに「**=QUARTILE.INC(D2:D501, 2)**」と入力する。

第3四分位数

・空のセルに「**=QUARTILE.INC(D2:D501, 3)**」と入力する。

今回の問い

- 量的変数と質的変数の違いは何か？
  - 定量的な値で観測されるデータを扱う変数を**量的変数**といい、定量的な値では観測できないデータを扱う変数を**質的変数**という。
- **代表値とは何か？**
  - **平均値、中央値、最頻値など、分布の中心を表す値。**
- 散らばりの尺度には何があるか？

今回学ぶこと

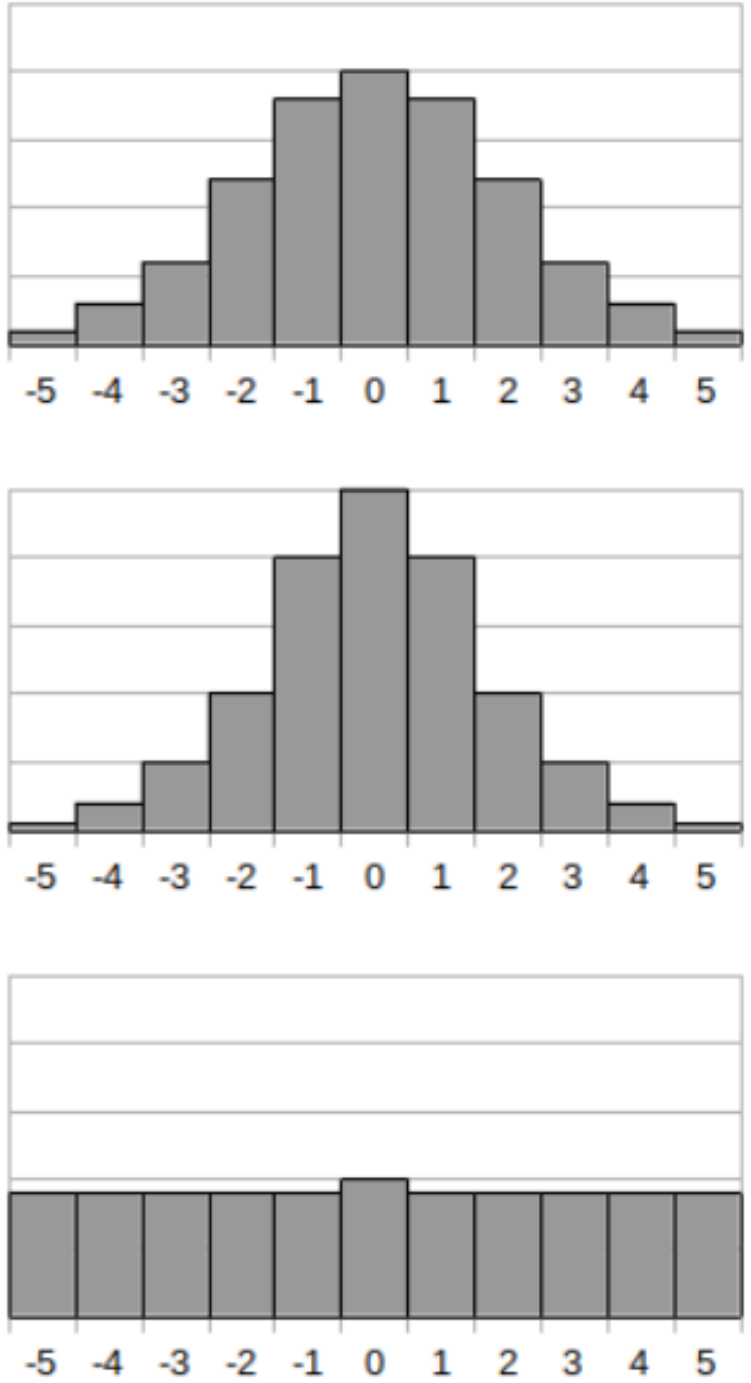
1. 変数
2. 代表値
3. 散らばりの尺度

散らばりの尺度

分布の広がりを表す指標

散らばりの尺度の必要性

分布の代表値が同じであっても分布の形が同じとは限らない。右の三つの分布は全て左右対称で「平均値＝中央値＝最頻値＝0」であるが、分布の形は様々。



偏差

分布の散らばりは平均からのずれで表せる。
個々のデータの平均からのずれ

$$x_i - \bar{x}$$

を**偏差**という。

偏差

個々の偏差を全て取り入れて一つの値で分布の散らばりを表すことを考える。単純に偏差の総和を計算すると

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

のように必ず零になってしまうため、うまくいかない。これは、個々の偏差は正の値や負の値があり、全て足すと打ち消し合うためである。

分散

偏差の総和が 0 になるのは偏差に負の値があるせいなので、偏差の負の値を無くすために、**偏差を二乗**して平均をとったものを**分散**s^2という。分散が分布の散らばりを表す尺度になる。

分散の単位

分散はデータを二乗するため、単位もデータの単位の二乗になってしまう。

例えば、データが長さの単位[m]（メートル）を持つ場合、分散の単位は面積の単位[m^2]（平方メートル）となるため、分散をデータやその平均とは直接比べられない。

生データの分散

生データ$x_i(i = 1,2, \dots, n)$の分散S^2：

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

度数分布の分散

度数分布の階級値v_i、度数$f_i(i = 1,2, \dots, k)$とすると、分散は以下の通り。

$$S^2 = \frac{(v_1 - \bar{x})^2 f_1 + (v_2 - \bar{x})^2 f_2 + \cdots + (v_k - \bar{x})^2 f_k}{n} = \frac{1}{n} \sum_{i=1}^{k} (v_i - \bar{x})^2 f_i$$

不偏分散

分数の分母のnを$n-1$に変えたもの。標本に対して使う分散で、主に推定や仮説検定で使用する。

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

標準偏差

分散S^2の平方根（ルート）をとったもの

$$S = \sqrt{S^2}$$

データの単位と同じになるため、データへ平均と直接比べられる。

不偏標準偏差

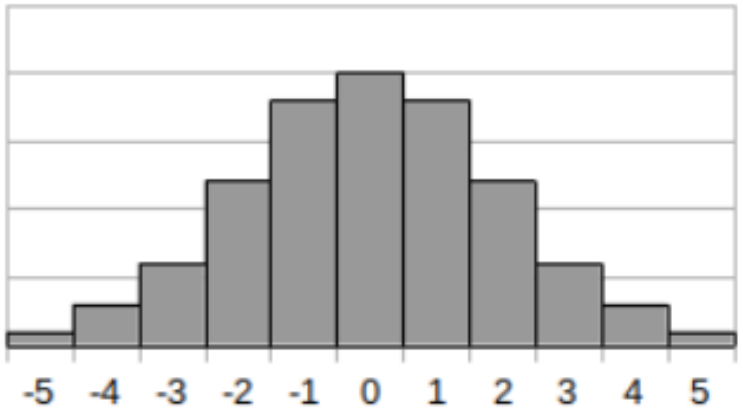
不偏分散の平方根をとったもの

$$s = \sqrt{s^2}$$

標本に対して使う標準偏差

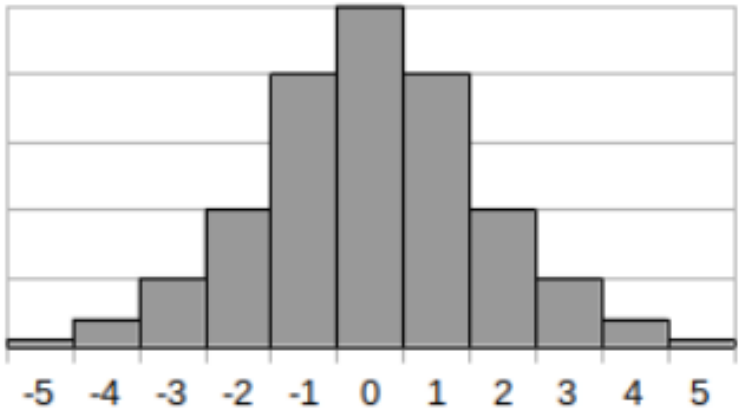
散らばりの尺度の違い

先ほどの「平均値＝中央値＝最頻値＝0」の分布の分散S^2と標準偏差Sは右のようになる。



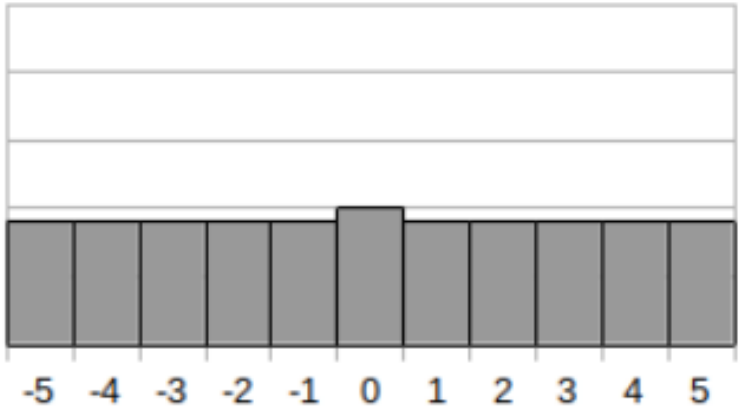
$S^2 = 3.86$

$S = 1.96$



$S^2 = 2.99$

$S = 1.73$



$S^2 = 9.90$

$S = 3.15$

偏差値（偏差値得点）

データx_iを次の式で変換したもの

$$T_i = \frac{10(x_i - \bar{x})}{S} + 50$$

ここで$\bar{x}$はx_iの平均、Sはx_iの標準偏差

この変換によってT_iは平均が50、標準偏差が10のデータとなる。

演習：散らばりの尺度の計算（生データ）

dataset1.csv列Dの「身長」のデータから散らばりの尺度を求めよう。

分散

・分散は、空のセルに「**=VAR.P(D2:D501)**」と入力する。
・不偏分散は、空のセルに「**=VAR.S(D2:D501)**」と入力する。

標準偏差

・標準偏差は、空のセルに「**=STDEV.P(D2:D501)**」と入力する。
・不偏標準偏差は、空のセルに「**=STDEV.S(D2:D501)**」と入力する。

標準偏差は、分散を求めた後に「**=SQRT(分散のセル)**」と入力して求めても良い。
SQRT関数は引数の平方根を計算する関数である。

演習：散らばりの尺度の計算（度数分布）

前回の「演習：度数分布表の作成」で作成した身長の度数分布から散らばりの尺度を求めよう。

分散

- 空のセルに「**=SUMPRODUCT((J3:J12-平均)^2, K3:K12)/500**」と入力する。「平均」には前に求めた「平均」のセルもしくは値を入力する。

標準偏差

- 「**=SQRT(度数分布表の分散)**」と入力する。「度数分布表の分散」には、上で求めた「度数分布表の分散」セルもしくは値を入力する。

今回の問い

- 量的変数と質的変数の違いは何か？
  - 定量的な値で観測されるデータを扱う変数を**量的変数**といい、定量的な値では観測できないデータを扱う変数を**質的変数**という。
- 代表値とは何か？
  - 平均値、中央値、最頻値など、分布の中心を表す値。
- 散らばりの尺度には何があるか？
  - 分散、標準偏差などがある。

代表値と散らばりの尺度の関数

関数	説明
MIN(引数)	引数の最小値を求める。
MAX(引数)	引数の最大値を求める。
SUM(引数)	引数の合計を求める。
AVERAGE(引数)	引数の平均を求める。
MEDIAN(引数)	引数の中央値を求める。
MODE(引数) MODE.SNGL(引数)	引数の最頻値を（一つ）求める。
VAR.P(引数)	引数の分散を求める。
VAR.S(引数)	引数の不偏分散を求める。
STDEV.P(引数)	引数の標準偏差を求める。
STDEV.S(引数)	引数の不偏標準偏差を求める。

パーセンタイルと四分位数の関数

関数	説明
PERCENTILE(セル範囲, p) PERCENTILE.INC(セル範囲, p)	セル範囲のデータの $100p$ パーセンタイルを求める（定義1）. p は $0 \leq p \leq 1$ の実数である。
PERCENTILE.EXC(セル範囲, p)	セル範囲のデータの 100p パーセンタイルを求める（定義2）. p は $\frac{1}{n+1} \leq p \leq \frac{n}{n+1}$ の実数である。
QUARTILE(セル範囲, タイプ) QUARTILE.INC(セル範囲, タイプ)	セル範囲のデータの四分位数を求める（定義1）。 タイプは 0,1,2,3,4 のいずれかで、Q_0, Q_1, Q_2, Q_3, Q_4 に対応する。
QUARTILE.EXC(セル範囲, タイプ)	セル範囲のデータの四分位数を求める（定義2）。 タイプは 1,2,3 のいずれかで、Q_1, Q_2, Q_3 に対応する。

今回学んだこと

1. 変数
2. 代表値
3. 散らばりの尺度

今回の目標

代表値と散らばりの尺度から数量的にデータの分布の特徴を読み取れるようになる。

課題

dataset1.csvの「体重」のデータから代表値、散らばりの尺度、四分位数を求めなさい。作成したファイルはODF表計算ドキュメント（拡張子.ods）で提出すること。

提出先：yutaka.hirai@koeki-u.ac.jp

件名：データリテラシー第5回課題 [学籍番号]

締め切り：11月6日 (金)

次回

第１回　社会におけるデータ・AI利活用
第２回　表計算ソフトの基本操作
第３回　セルの参照と集計
第４回　度数分布表とヒストグラム
第５回　代表値と散らばりの尺度
第６回　データのグラフによる表現とデータ比較
第７回　データの分布と標本抽出
第８回　推定
第９回　分割表とクロス集計
第１０回　２変量データと相関
第１１回　時系列データ
第１２回　仮説検定
第１３回　回帰分析