

情報科学（講義資料）

第1回

今回の講義で使用するファイル

- 第1回（講義資料）.pptx
- 第1回（演習・課題）.xlsx

講義内容

-
- 動画1
(この動画)
- 1. はじめに：統計とは
 - 2. 質的データの度数分布
 - 1変数の度数分布
 - 2変数の度数分布
 - 演習1
- 動画2
- 3. 量的データの度数分布
 - 度数分布表とヒストグラム
 - 演習2
- 動画3
- 4. ヒストグラムから分かること
 - 分布の見方
 - 分布の形状
 - 5. 課題
 - 6. まとめ
 - 7. Appendix

1. はじめに：統計とは



統計 = ^す統べる & ^{はか}計る



① たくさんのものを一つにまとめる

② ある基準をもとに度合いを調べる

たくさんの個別データ

一つにまとめられたデータ

度合い

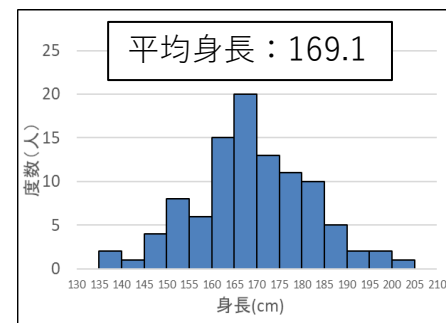
| 回答者ID | 性別 | 出身県 | 生まれ月 | 身長 (cm) | 体重 (kg) |
|-------|----|-----|------|---------|---------|
| 1 | 男 | 栃木 | 4 | 173 | 60 |

| 回答者ID | 性別 | 出身県 | 生まれ月 | 身長 (cm) | 体重 (kg) |
|-------|----|-----|------|---------|---------|
| 2 | 女 | 東京 | 12 | 165 | 53 |

| 回答者ID | 性別 | 出身県 | 生まれ月 | 身長 (cm) | 体重 (kg) |
|-------|----|-----|------|---------|---------|
| 3 | 男 | 茨城 | 9 | 182 | 73 |

⋮

| 回答者ID | 性別 | 出身県 | 生まれ月 | 身長 (cm) | 体重 (kg) |
|-------|----|-----|------|---------|---------|
| 1 | 男 | 栃木 | 4 | 173 | 60 |
| 2 | 女 | 東京 | 12 | 165 | 53 |
| 3 | 男 | 茨城 | 9 | 182 | 73 |
| 4 | 男 | 大阪 | 1 | 169 | 59 |
| 5 | 女 | 北海道 | 6 | 153 | 48 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |



統計分析でやること：

たくさんの情報をデータとして一つにまとめ、そこからデータ全体の傾向や特徴を示す量 (=度合い) を調べ、それを可視化する。

今回の講義の概要

例1 質的データ 個人データの例

| 回答者ID | 性別 | 出身県 | 生まれ月 |
|-------|----|-----|------|
| 1 | 男 | 栃木 | 4 |
| 2 | 女 | 東京 | 12 |
| 3 | 男 | 茨城 | 9 |
| 4 | 男 | 大阪 | 1 |
| 5 | 女 | 北海道 | 6 |
| ⋮ | ⋮ | ⋮ | ⋮ |

例2 量的データ

| 身長 (cm) | 体重 (kg) |
|---------|---------|
| 173 | 60 |
| 165 | 53 |
| 182 | 73 |
| 169 | 59 |
| 153 | 48 |
| ⋮ | ⋮ |

データの可視化：
各変数の値を、カテゴリーや階級（数値の範囲）に分けて、そこにいくつデータが含まれるか（=度数）を集計し、グラフ化する。

例1:性別

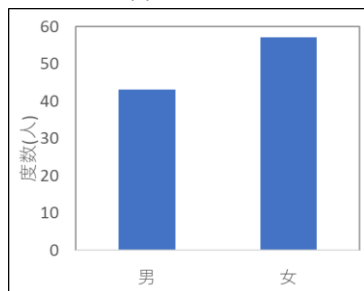
カテゴリー

度数分布表

| 性別 | 度数(人) |
|----|-------|
| 男 | 43 |
| 女 | 57 |



棒グラフ



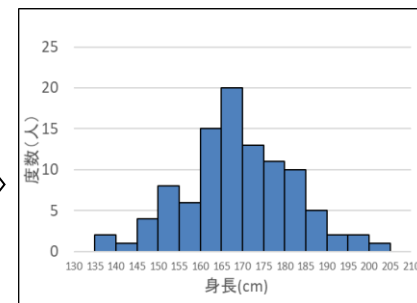
例2:身長(cm)

度数分布表

| 身長の階級(cm) | 度数 (人) |
|-----------|--------|
| 130-135 | 0 |
| 135-140 | 2 |
| 140-145 | 1 |
| ⋮ | ⋮ |



ヒストグラム



今回の講義では、質的データ・量的データの度数分布表の作成方法と、それらをグラフ化したものから、データの全体像を把握する方法について学ぶ。

2. 質的データの度数分布

達成目標：1変数の場合と2変数の場合の度数分布表とそのグラフを作成できるようになる。

1 変数の度数分布

■ 分析の手順：

1. データから1つの変数の度数をカウントし、度数分布表を作成する。
2. 度数分布表を棒グラフにし、データ全体の様子を可視化する。

成績データの例

| 学籍番号 | 成績 |
|--------|----|
| 190281 | 良 |
| 190509 | 可 |
| 190832 | 優 |
| 191214 | 良 |
| ⋮ | ⋮ |

ここにカウント

ここにカウント

ここにカウント

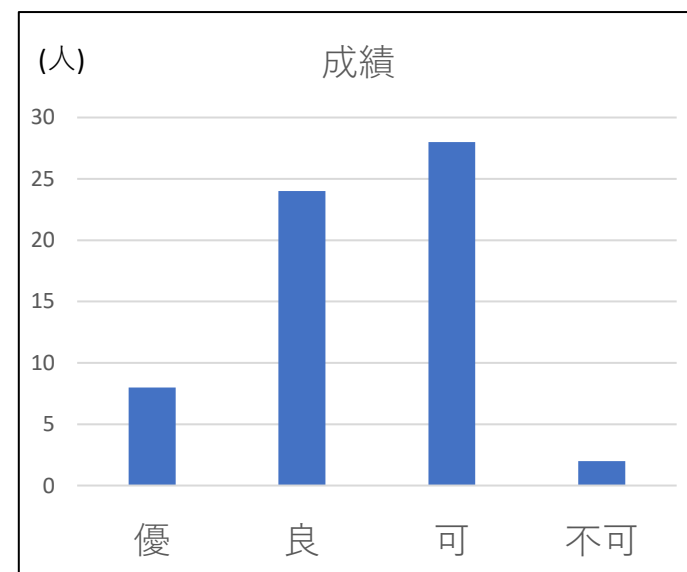
ここにカウント

度数分布表

| 成績 | 度数 (人) |
|----|--------|
| 優 | 8 |
| 良 | 24 |
| 可 | 28 |
| 不可 | 2 |



棒グラフ



元の数値データからは分からない成績データの全体像が一目で把握できるようになった。

■ Excel関数：=COUNTIF関数(データ範囲,条件) (詳細はAppendix参照)

2変数の度数分布

■ 分析の手順：

1. 2変数のすべての組み合わせに対する度数をカウントし、クロス集計表にまとめる。
2. クロス集計表を棒グラフにし、2変数間の関係を可視化する。

成績データの例

| 学籍番号 | 学部 | 成績 |
|--------|-----|----|
| 190281 | C学部 | 良 |
| 190509 | B学部 | 可 |
| 190832 | A学部 | 優 |
| 191214 | B学部 | 良 |
| ⋮ | ⋮ | ⋮ |

ここにカウント

ここにカウント

ここにカウント

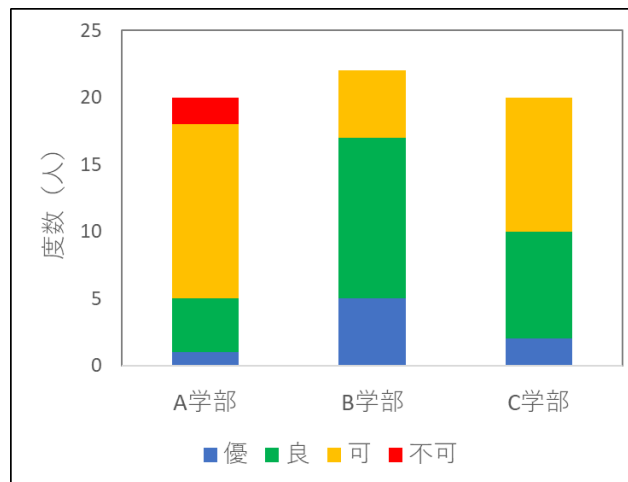
ここにカウント



「学部」と「成績」の2変数の関係として、「学部ごとの成績の傾向に違いがあること」がわかった。（B学部は他の学部と比べて優と良の受講者の割合が大きい。）

クロス集計表

| | A学部 | B学部 | C学部 | 合計 |
|----|-----|-----|-----|----|
| 優 | 1 | 5 | 2 | 8 |
| 良 | 4 | 12 | 8 | 24 |
| 可 | 13 | 5 | 10 | 28 |
| 不可 | 2 | 0 | 0 | 2 |
| 合計 | 20 | 22 | 20 | 62 |



演習1 質的データの度数分布表と棒グラフ

■ 度数分布（演習・課題）.xlsxの「質的データ」シートのデータの度数分布表とその棒グラフを作成する。

1. 「学部」の度数分布表と棒グラフを作成する。
 - 各学部の度数を出すため、F2のセルに「=COUNTIF(\$B\$2:\$B\$63,E2)」と入力し(※)、オートフィル（書式なしコピー）によりF4のセルまでコピーする。
 - E2:F4を選択状態にし、「挿入」タブの「縦棒/横棒のグラフの挿入」の中の「2-D 縦棒」の「集合縦棒」を選ぶ。縦軸を左クリックした後、右クリックし、「軸の書式設定(E)」を左クリックで開く。「軸のオプション」内の「境界値」の「最小値」を「0」にする。
2. 「成績」の度数分布表と棒グラフについても同様に作成する。
3. クロス集計表とその積み上げ縦棒グラフを作成する。
 - まず、L2のセルに「=COUNTIFS(\$C\$2:\$C\$63,\$K2,\$B\$2:\$B\$63,L\$1)」と入力し、オートフィル（書式なしコピー）により、L5のセルまでコピーする（COUNTIFS関数についてはAppendixを参照すること）。さらにL2:L5を選択状態にし、オートフィル（書式なしコピー）により、N2:N5までコピーする。
 - K1:N5を選択状態にし、「挿入」タブの「縦棒/横棒のグラフの挿入」の中の「2-D 縦棒」の「積み上げ縦棒」を選ぶ。横軸が「成績」になっている場合は、グラフをクリックし、「デザイン」タブの中の「行/列の切り替え」をクリックすると、横軸が「学部」となる。

\$の付け方に注意

(※ \$ マークは絶対参照（絶対番地）を意味し、F4キーで付けたり外したりできます。詳細を知りたい方は、教科書を参照してください。)

3. 量的データの度数分布

達成目標：量的データの度数分布表とヒストグラムを作成できるようになる。

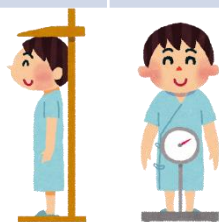
度数分布表とヒストグラム

■ 分析の手順：

1. データのある1つの変数（例えば身長）に対し、それを階級（クラス分けした値の範囲）ごとに度数を数えて度数分布表を作る。
2. 度数分布表をヒストグラムにし、データ全体の様子を可視化する。

身長・体重データ

| 個人ID | 身長(cm) | 体重(kg) |
|------|--------|--------|
| 1 | 182.1 | 104.4 |
| 2 | 162.6 | 67.3 |
| 3 | 146.5 | 49.2 |
| ⋮ | ⋮ | ⋮ |
| 100 | 160.8 | 66.1 |

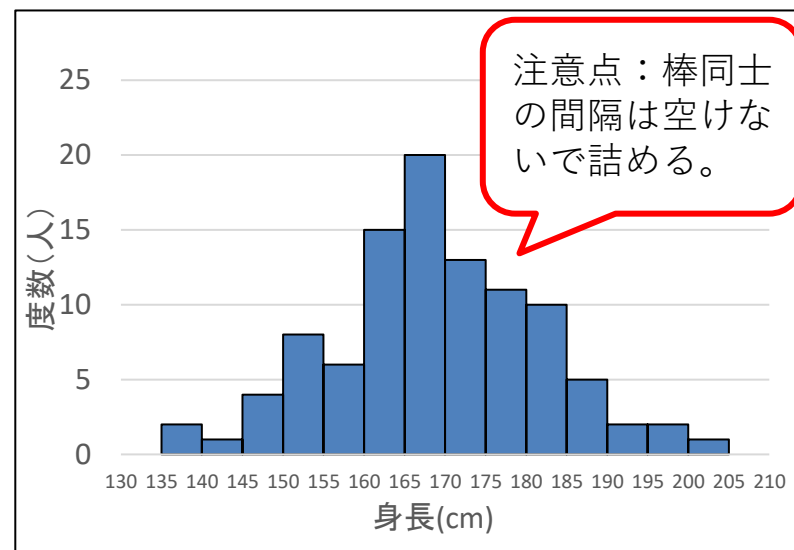


度数分布表

| 身長の階級(cm) | 度数(人) |
|-----------|-------|
| 130-135 | 0 |
| 135-140 | 2 |
| 140-145 | 1 |
| 145-150 | 4 |
| 150-155 | 8 |
| 155-160 | 6 |
| 160-165 | 15 |
| 165-170 | 20 |
| 170-175 | 13 |
| 175-180 | 11 |
| 180-185 | 10 |
| 185-190 | 5 |
| 190-195 | 2 |
| 195-200 | 2 |
| 200-205 | 1 |
| 205-210 | 0 |

(各階級は、下限値は含まず上限値は含む。)

ヒストグラム




身長の分布の様子が一目で分かるようになった。(165cm~170cmの階級の度数が一番大きく、160cm~185cmの間の身長の人が約7割を占める。)

演習2 量的データの度数分布表とヒストグラム

■ 「量的データ」シートのデータの度数分布表とヒストグラムを作成する。

1. B列（身長）の度数分布表とヒストグラムの作成

- まず、階級を決めるために、身長の最小値と最大値を、Excelの関数のMIN(データ範囲)とMAX(データ範囲)を用いてF2とF3セルに入力する。
- 階級幅を5 cmとし、階級を(130,135], (135,140], ..., (205,210]とする。(それぞれの階級で下限値は含まず、上限値は含む。)
- 階級の上限値135, 140, ..., 210をI2~I17のセルにオートフィルを使って入力する。
- J列にはグラフに表示させる階級を入力する。まずJ2には「130-135」と入力する。J3には「=I2&"-"&I3」と入力すると「135-140」と表示されるので、それをJ3をJ17までオートフィルを使ってコピーする。
- 度数を計算する。まず、K2:K17のセルを選択状態にし、数式バーに「=FREQUENCY(B2:B101,I2:I17)」と入力し、**Ctrl+Shift+Enter**を同時に押すと、度数が全て自動で計算されて入力される。(K2:K17の各セルには、配列数式として「{=FREQUENCY(B2:B101,I2:I17)}」と入力されている状態になっている。)
- J1:K17を選択し、「挿入」タブから、棒グラフのマークをクリックし、「2-D縦棒」の中の「集合縦棒」を選ぶ。グラフ内の棒をクリックし、棒の角に○が付いた状態になったら、右クリックして現れたボックスから「データ系列の書式設定(F)」を選ぶ。表示された「系列オプション」の中の「要素の間隔(W)」を「0%」にすると、棒同士の隙間がなくなる。さらに、グラフの横軸の数値を左クリックで選択した状態で右クリックし、「軸の書式設定(F)」を選択する。現れたボックスの「サイズとプロパティ ()」の中の「文字列の方向(F)」で「右へ90度回転」を選択する。

2. C列（体重）の度数分布表とヒストグラムも同様に作成する。ただし、階級は(20,30],(30,40],..., (120,130]とすること。

4. ヒストグラムから分かること

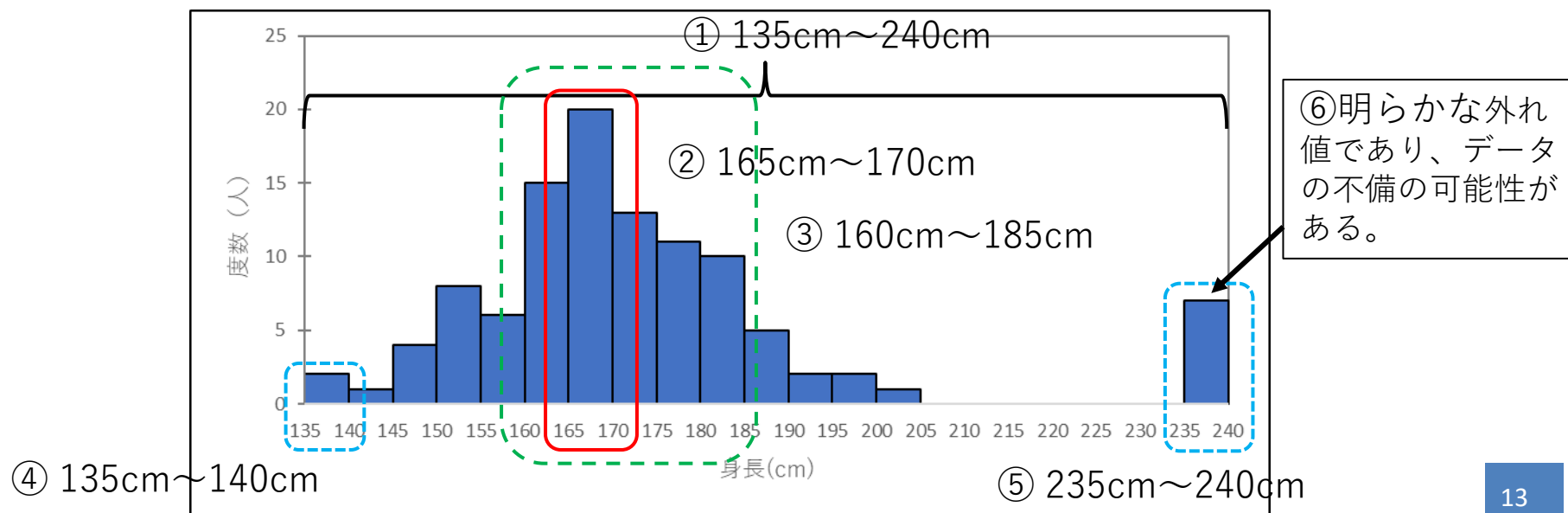
達成目標：ヒストグラムで表現された分布（＝データのばらつきの様子）の見方と、分布の形状として代表的な「単峰性・多峰性」、「左右対称・左右非対称」のそれぞれについて、形状の特徴から読み取れるデータの性質について理解する。

分布の見方

■ ヒストグラムから読み取る情報：

- ① 階級の範囲は？
- ② どの階級の度数が一番大きい？
- ③ データはどこに集中している？
- ④ 最小値はどのくらい？
- ⑤ 最大値はどのくらい？
- ⑥ データに異常値や外れ値はないか？

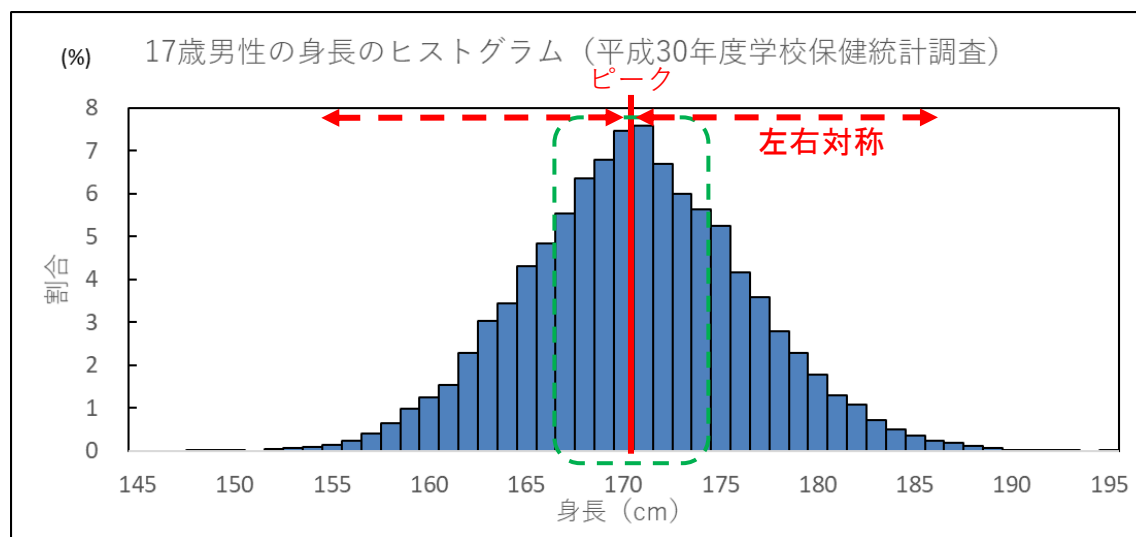
ヒストグラムの例



分布の形状

- 単峰性：山になった部分（ピーク、峰）が1箇所の形

単峰性・左右対称のヒストグラムの例



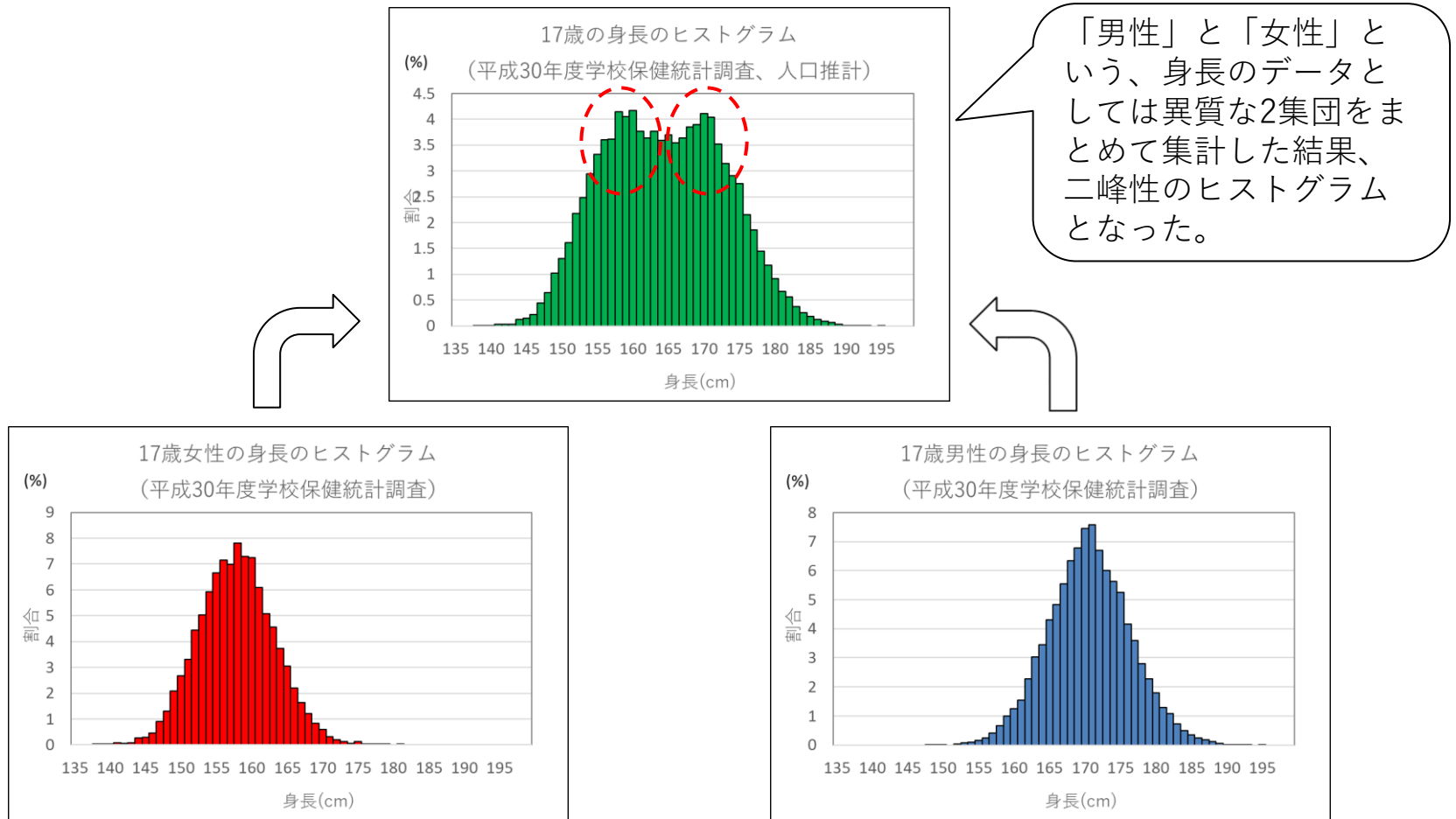
(割合=度数÷全データ数×100)

特徴

- ピークのある階級付近にデータが集中しており、その付近の値をとるデータが多いことを示している。
- データのもとになっている集団が同じ性質を持っていれば（同質であれば）、単峰性でピークを中心とした左右対称の形になることが多い。（上記の例では「17歳男性」という性質が同じ集団になっており、ヒストグラムはピークを中心にほぼ左右対称である。）

■ 多峰性：ピークが複数箇所ある形

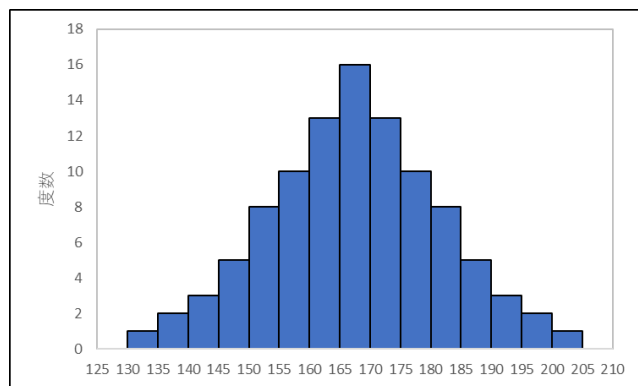
多峰性（二峰性）のヒストグラムの例



ヒストグラムが多峰性となった場合、異質な集団のデータが混在している可能性がある。もし混在していれば、データを同質な集団ごとに分析する等の処理が必要となる。

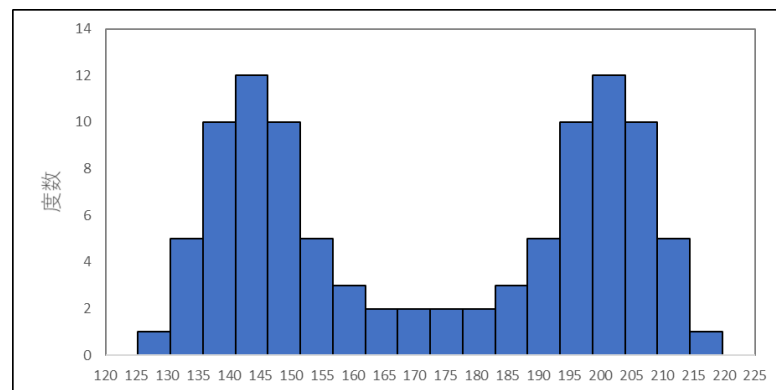
■ 左右対称のヒストグラムの例

左右対称（単峰性）



実データの例：年齢別・男性（女性）の身長や体重

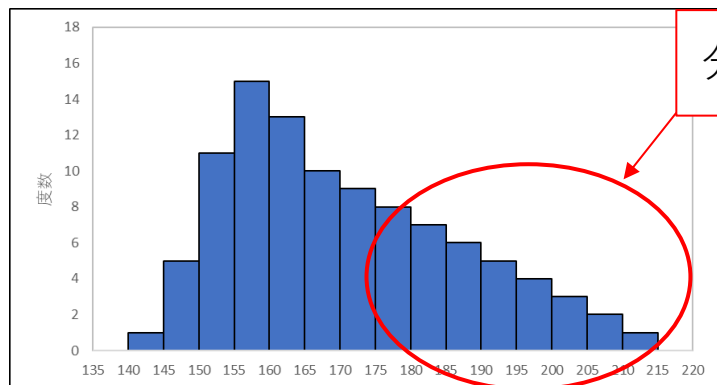
左右対称（二峰性）



実データの例：年齢別・男女混合の身長や体重

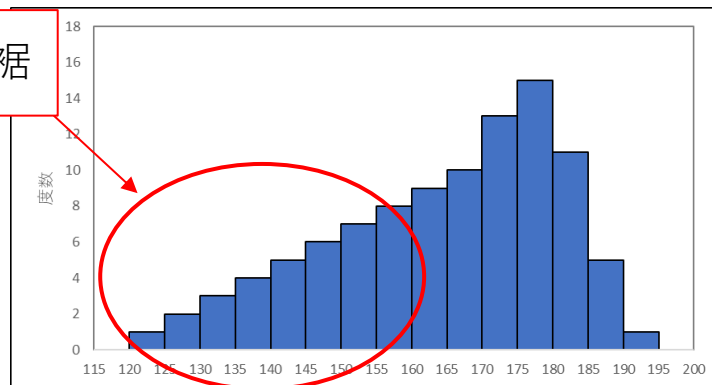
■ 左右非対称のヒストグラムの例

右に裾の長い分布



実データの例：市区町村の人口、個人所得

左に裾の長い分布



実データの例：赤ちゃんの出生時の身長

5. 課題

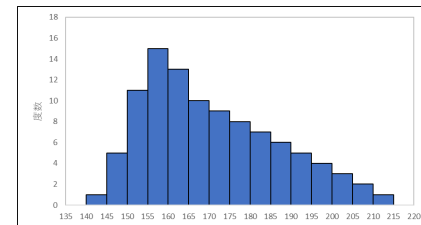
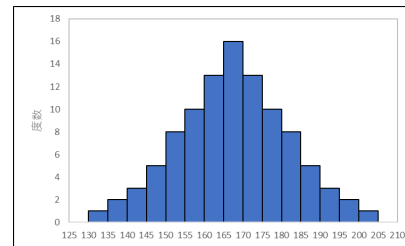
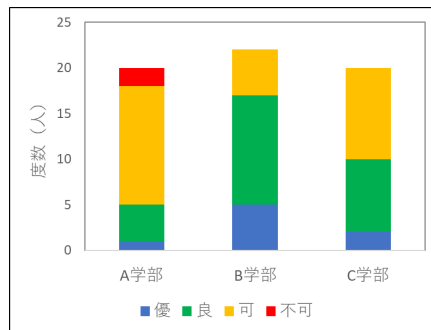
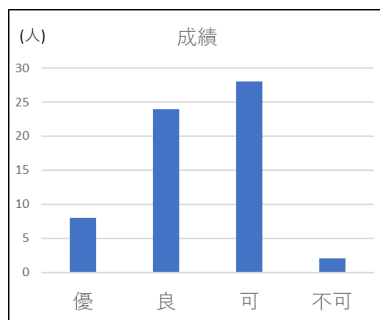
「課題」シート内の日本の市区町村の平均所得のデータ（2021年）を用いてヒストグラムを作成する。

- 市区町村の平均所得のヒストグラムを作成する。
 - E2とE3のセルに2021年の平均所得の最小値と最大値をExcelの関数を用いて入力する。
 - G列に階級の上限值が既に入力されているので、H列にオートフィル機能を使って階級を入力する（例：150-200）。
 - I列にその度数をExcelの関数を用いて入力する。
 - H列とI列を用いてヒストグラムを作成する。（ヒストグラムの作成方法は、演習2を参照すること。）
 - 作成したヒストグラムの形状を、E6のプルダウンから選択する。

提出方法は、「データサイエンス入門」の担当の先生の指示に従って下さい。

6. まとめ

- 質的データ（1変数、2変数）の度数分布表とその棒グラフの作成方法について学んだ。
- 量的データの度数分布表とヒストグラムの作成方法について学んだ。
- ヒストグラムで表現された分布の見方と、分布の形状として代表的な「単峰性・多峰性」、「左右対称・左右非対称」のそれぞれについて、形状の特徴から読み取れるデータの性質について学んだ。



Appendix

- 比較演算子
- 記号入力方法
- COUNTIF関数
- COUNTIFS関数

比較演算子

- `"= 10"` : 「10に等しい」
- `"<10"` : 「10より小さい」
- `">10"` : 「10より大きい」
- `"<=10"` : 「10以下」
- `">=10"` : 「10以上」
- `"<>10"` : 「10に等しくない」

記号の入力方法

半角英数で以下を入力する。

| | | | | |
|---|-----|---------|---|------------|
| " | ... | ⇧ Shift | + | " 2 ふ |
| < | ... | ⇧ Shift | + | < 、 , ね |
| > | ... | ⇧ Shift | + | > 。 . る |

COUNTIF関数

- 意味：指定した範囲で「条件」に合うセルの個数を求める。
- 使い方： = COUNTIF(範囲, 条件)

- 例1： = COUNTIF(A2:A101, "=1")

「A2:A101の範囲のセルの中で、数値が1に等しいセルの個数」

- 例2: =COUNTIF(E2:E145,"宇都宮市")

「 E2:E145の範囲のセルの中で、宇都宮市と入力されているセルの個数」

COUNTIFS関数

- 意味：COUNTIF関数の範囲と条件を複数にしたもの。
- 使い方： = COUNTIFS(範囲1, 条件1, 範囲2, 条件2, . . .)
- 例1： = COUNTIFS(A2:A101, "=1", B2:B101, "=99")
「A2:A101の範囲のセルの中で、数値が1に等しいセル**かつ**、
B2:B101の範囲のセルの中で、数値が99に等しいセルの個数」
- 例2： = COUNTIFS(A2:A101, ">=20", A2:A101, "<=29")
「A2:A101の範囲のセルの中で、数値が20以上のセル**かつ**、
B2:B101の範囲のセルの中で、数値が29以下のセルの個数」