# 統計学ノートNo. 2
# 記述統計（1変量）

**担当教員：蛭川雅之**
**研究室：紫英館3階345研究室**
**メール：** hirukawa@econ.ryukoku.ac.jp
**オフィスアワー：毎週金曜日3講時**



10/1/2021

# 1　和の記号Σ

# 和の記号

- **統計学では、しばしばデータ**

$$\{X_i\}_{i=1}^{n} = X_1, X_2, \ldots, X_n$$

**の和を**

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

**と書く。**

- **記号Σは**sum（**和**）**の頭文字**S**に相当するギリシャ文字で、シグマと呼ぶ。**

# Σの性質

1. **定数$c$に対し、**

$$\sum_{i=1}^{n} c = \underbrace{c + c + \cdots + c}_{n\text{個}} = nc$$

$$\sum_{i=1}^{n} cX_i = cX_1 + cX_2 + \cdots + cX_n$$

$$= c(X_1 + X_2 + \cdots + X_n)$$

$$= c\sum_{i=1}^{n} X_i$$

# Σの性質（つづき）

2. 2つのデータ$\{X_i\}_{i=1}^n$および$\{Y_i\}_{i=1}^n$に対し、

$$\sum_{i=1}^{n}(X_i + Y_i)$$

$$= (X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_n + Y_n)$$

$$= (X_1 + X_2 + \cdots + X_n) + (Y_1 + Y_2 + \cdots + Y_n)$$

$$= \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

# Σの性質（つづき）

3. $2$**つの添え字**$(i,j)$**を持つデータ**$\left\{\left\{X_{ij}\right\}_{i=1}^{n}\right\}_{j=1}^{m}$**に対し、**

$$\sum_{i=1}^{n}\sum_{j=1}^{m}X_{ij}$$

$$= (X_{11} + X_{12} + \cdots + X_{1m}) + (X_{21} + X_{22} + \cdots + X_{2m}) + \cdots + (X_{n1} + X_{n2} + \cdots + X_{nm})$$

$$= (X_{11} + X_{21} + \cdots + X_{n1}) + (X_{12} + X_{22} + \cdots + X_{n2}) + \cdots + (X_{1m} + X_{2m} + \cdots + X_{nm})$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n}X_{ij}$$

# 2　度数分布表

# 生データから何がわかるか？

- **次の表は、女子大生80人の身長に関する*生データ*である。**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 151 | 154 | 154 | 164 | 158 | 146 | 162 | 151 |
| 154 | 162 | 152 | 158 | 151 | 166 | 167 | 156 |
| 160 | 156 | 161 | 150 | 155 | 161 | 159 | 166 |
| 160 | 162 | 160 | 155 | 155 | 143 | 153 | 159 |
| 163 | 157 | 160 | 157 | 165 | 156 | 146 | 157 |
| 156 | 162 | 153 | 161 | 165 | 156 | 156 | 156 |
| 158 | 162 | 155 | 168 | 154 | 149 | 160 | 159 |
| 156 | 169 | 163 | 162 | 148 | 162 | 151 | 156 |
| 154 | 150 | 160 | 153 | 169 | 159 | 151 | 156 |
| 160 | 162 | 159 | 154 | 158 | 164 | 157 | 161 |

# 生データから何がわかるか？（つづき）

- 女子大生の身長が**まちまち**の値をとることはわかる。
  - 身長の**分布**が存在する。
  - **不確実性**がまちまちの値を生じさせる。
- データを一見しただけで、**分布の特性（癖）**を知ることができるか？
  - データに相当慣れていなければ、ほぼ無理！

# 統計 ＝ 縮約

- **統計は分布の特性を引き出す手法である。**
- **統計では縮約を用いる。**
  - **データをある基準で整理整頓し、意味のある情報を取り出す。**
    1. **データをグラフ化する。**
    2. **データを一つの数字（統計量）で代表させる。**

# 度数分布表

- データの特徴を分かり易くするため、**度数分布表**を作成する。

| 階　級 | 階級値 | 度　数 | 相対度数 | 累積度数 | 累積比率 |
|---|---|---|---|---|---|
| 140 ～ 145 | 142.5 | 1 | 0.0125 | 1 | 0.0125 |
| 145 ～ 150 | 147.5 | 4 | 0.0500 | 5 | 0.0625 |
| 150 ～ 155 | 152.5 | 17 | 0.2125 | 22 | 0.2750 |
| 155 ～ 160 | 157.5 | 27 | 0.3375 | 49 | 0.6125 |
| 160 ～ 165 | 162.5 | 23 | 0.2875 | 72 | 0.9000 |
| 165 ～ 170 | 167.5 | 8 | 0.1000 | 80 | 1.0000 |
| 合　計 | — | 80 | 1.0000 | — | — |

# 度数分布表：作成手順１

- データの最大値・最小値を求める。
  - Excelの関数maxおよびminを使う。
  - このデータの最大値は169、最小値は143である。

# 度数分布表：作成手順２

- **おおよそ最大値から最小値になるような区切りのよい範囲を決める。**
  - **最大値169と最小値143に近い区切りの良い数字として、例えば、170と140を選ぶ。**
- **決めた範囲を5～8の等間隔の小区間（階級）に区切る。**
  - **データを“140～145”，“145～150”,…，“165～170”の6階級に区切る。**
  - **階級“140～145”は「140cm以上145cm未満」を意味する（他の階級も同様）。**

# 度数分布表：作成手順３

- **各階級を代表する値（階級値）を決める。**
  - **通常、階級の中央値を用いる。**
  - “140～145”, “145～150”,…, “165～170”**の**6**階級の階級値をそれぞれ**142.5,147.5,…,167.5**とする。**

# 度数分布表：作成手順４

- 各階級に含まれるデータ数を数える。
  - このデータ数のことを**度数**という。
- 度数は、データ数がごく少数であれば、直接手で数える、あるいは「正」か "tally" を使って数えることができる。

# 度数分布表：作成手順４（つづき）

- データ数が多い場合はどうするか？
  - 累積度数を先に求める。
  - 累積度数とは、ある階級までの度数の累計である。
  - 最後の階級の累積度数はデータ数に一致する。

# 度数分布表：作成手順４（つづき）

- **累積度数はExcelの関数countifを利用して求めることができる。**
  - **各階級の度数は**

    **(その階級までの累積度数)**
    **－(直前の階級までの累積度数)**

    **として求められる。**
  - **度数の合計はデータ数に一致する。**

# 度数分布表：作成手順５

- **各階級の度数の全データ数に対する割合を計算する。**
  - **この割合を相対度数という。**
  - **相対度数の合計は1である。**
- **同様に、各階級の累積度数の全データ数に対する割合を計算する。**
  - **この割合を累積比率（累積相対度数）という。**
  - **最後の階級の累積比率は1である。**

# 度数分布表：完成品

| 階　級 | 階級値 | 度　数 | 相対度数 | 累積度数 | 累積比率 |
|---|---|---|---|---|---|
| 140 ~ 145 | 142.5 | 1 | 0.0125 | 1 | 0.0125 |
| 145 ~ 150 | 147.5 | 4 | 0.0500 | 5 | 0.0625 |
| 150 ~ 155 | 152.5 | 17 | 0.2125 | 22 | 0.2750 |
| 155 ~ 160 | 157.5 | 27 | 0.3375 | 49 | 0.6125 |
| 160 ~ 165 | 162.5 | 23 | 0.2875 | 72 | 0.9000 |
| 165 ~ 170 | 167.5 | 8 | 0.1000 | 80 | 1.0000 |
| 合　計 | – | 80 | 1.0000 | – | – |

# 度数分布表：何を失い何を得たか？

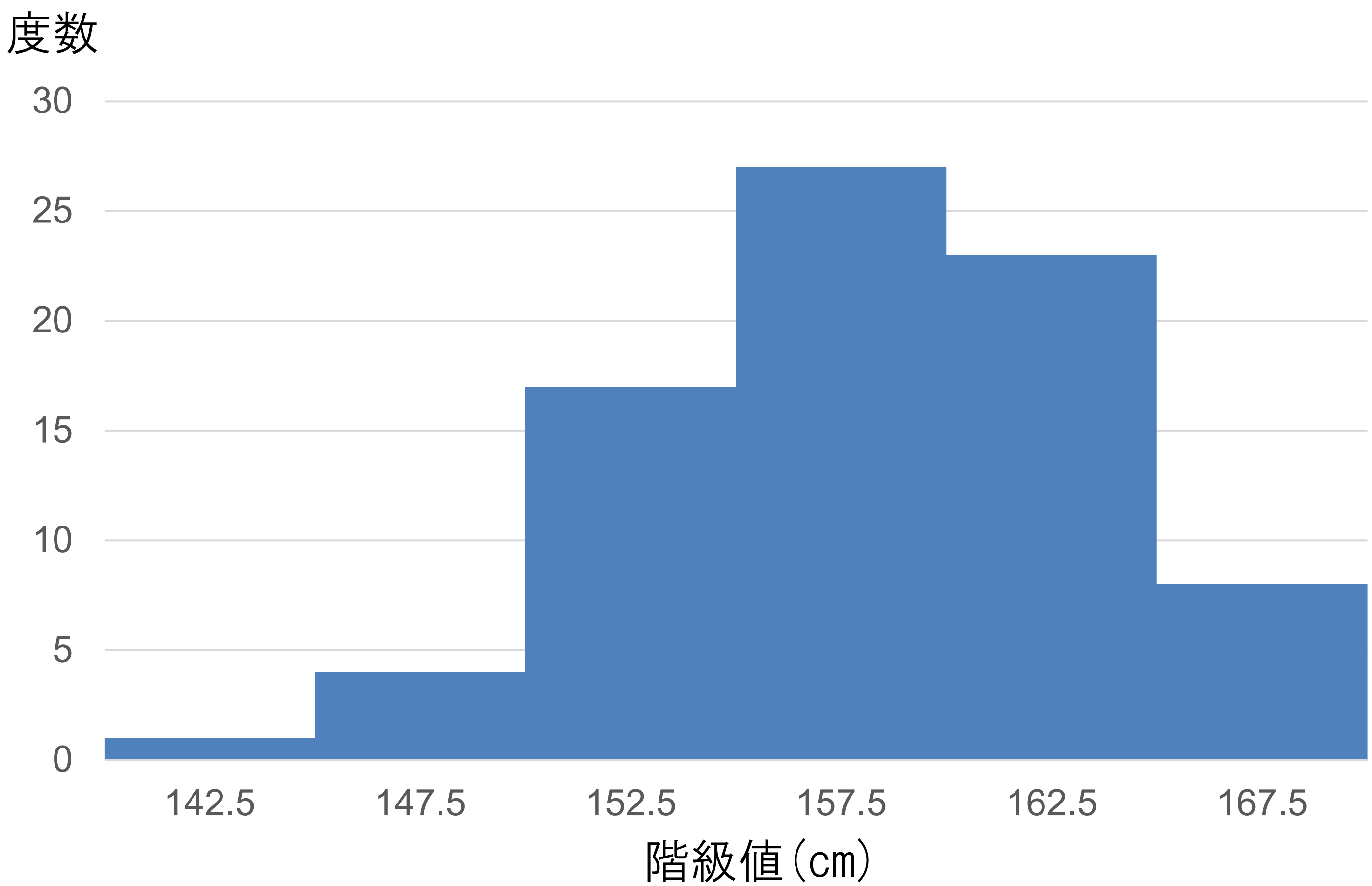
- **縮約により個別のデータの数値は失われる。**
- **生データを一見しただけではわからない、次のような情報を得る。**
  1. **身長は均等に分布しているのでなく、ある階級（“155〜160”）に集中している。**
  2. **分布の「山」は一つである（単峰分布）。**
  3. **分布はおおよそ左右対称に見える。（157.5cm近辺が中心か？）**

# 3　ヒストグラム

# 度数分布表のグラフ化

- ヒストグラムとは、階級（値）を横軸、度数を縦軸に取ったグラフである。
- Excelでは“2-D縦棒” グラフに手を加えることにより作図可能である。
  - “ヒストグラム”を利用すると、階級を自動的に決められてしまうようである。

# ヒストグラム：完成品



# ヒストグラムから読みとれること

- **度数分布表の情報がより視覚的に明瞭になる！**

1. **身長は階級値**157.5cm**で代表される階級に集中している。**
2. **単峰分布である。**
3. **分布はおおよそ左右対称に見える。**

# Computer Exercise 1

- データファイル"data1.xlsx"を講義ウェブページからダウンロードし、Excelを利用して以下のものを作成せよ。

1. 度数分布表
2. ヒストグラム

# 4　平均値とメディアン

# 度数分布表、ヒストグラムは有益だが…

1. グラフから受ける印象は人によって違う。
   - グラフの特徴を口頭で説明するのは容易でない。
2. 表・グラフはスペースをとる。
   - 特に論文・レポートの場合、大量の図表を掲載してスペースを浪費するのは好ましい見せ方ではない。

# 統計量：もう一つの縮約

- １つの数字による縮約であれば、スペースをとらずに済む。
- **統計量**の例：
  - 位置（中心）：
    - 平均値
    - メディアン
  - ちらばりの程度：
    - 分散
    - 標準偏差

# 平均値

- **平均値**（mean）は

$$(\text{平均値}) = \frac{(\text{データ合計})}{(\text{全データ数})}$$

で計算できる。

- $n$個のデータ$X_1, X_2, \ldots, X_n$に対し、

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

と表現できる。

# 平均値（つづき）

- **女子大生の身長の平均値：**

$$(\text{全データ数}) = n = 80$$

$$(\text{データ合計}) = \sum_{i=1}^{n} X_i = 12606$$

**であるから、**

$$(\text{平均値}) = \overline{\mathrm{X}} = \frac{12606}{80} = 157.575$$

  - Excel**の関数**average**を利用しても同じ結果が得られる。**

# 度数分布表から平均値を求める

- 度数分布表しか与えられていない場合、どのようにして平均値を求めるか？
  - 政府機関、民間調査機関が公表するデータの多くは予め度数分布表に集計されている。

| 階　級 | 階級値 | 度　数 | 相対度数 | 累積度数 | 累積比率 |
|---|---|---|---|---|---|
| 140 ~ 145 | 142.5 | 1 | 0.0125 | 1 | 0.0125 |
| 145 ~ 150 | 147.5 | 4 | 0.0500 | 5 | 0.0625 |
| 150 ~ 155 | 152.5 | 17 | 0.2125 | 22 | 0.2750 |
| 155 ~ 160 | 157.5 | 27 | 0.3375 | 49 | 0.6125 |
| 160 ~ 165 | 162.5 | 23 | 0.2875 | 72 | 0.9000 |
| 165 ~ 170 | 167.5 | 8 | 0.1000 | 80 | 1.0000 |
| 合　計 | － | 80 | 1.0000 | － | － |

# 度数分布表から平均値を求める（つづき）

- **階級値**157.5cm**で代表される階級を例にとり、度数分布表を以下のように解釈する。**
  - **「身長**157.5cm**の学生が**80**人中**27**人いる。」**
  - **「身長**157.5cm**の学生が全体に占める割合は**0.3375**である。」**

# 度数分布表から平均値を求める（つづき）

- この解釈をすべての階級にあてはめ、次の平均値を得る。
  - 平均値を求める際、個別のデータの数値を階級値に置き換えていることに注意せよ。

| 階級値(a) | 相対度数(b) | (a×b) |
|---|---|---|
| 142.5 | 0.0125 | 1.7813 |
| 147.5 | 0.0500 | 7.3750 |
| 152.5 | 0.2125 | 32.4063 |
| 157.5 | 0.3375 | 53.1563 |
| 162.5 | 0.2875 | 46.7188 |
| 167.5 | 0.1000 | 16.7500 |
| | 平均値 | 158.1875 |

# 度数分布表から平均値を求める（つづき）

- **この計算には、**Excel**の関数**sumproduct**を利用するのが便利である。**
  - **具体的には、**

$$sumproduct\left((\text{階級値}),(\text{相対度数})\right)$$

**とセル指定する。**

  - **以下のように計算してもよい。**

$$\frac{sumproduct\left((\text{階級値}),(\text{度数})\right)}{(\text{全データ数})}$$

# 生データから計算した平均値との比較

- **生データから計算した平均値は157.575cmであった。**
- **度数分布表から計算した平均値158.19cmは大きく外れていない。**
  - **度数分布表は個別のデータの情報を失っているにもかかわらず、数値にあまり差がない！**
  - **実用上この数値で問題はなさそう。**

    **⇒なぜ？**

# なぜ度数分布表から計算した平均値は大きくズレないのか？

- **階級値**157.5cm**の階級について、**

  「**階級値**157.5**というデータが**80**個中**27**個ある**」

  **と解釈し直す。**

$$157.5 \times 27 \div 80$$

$$= 157.5 \times \frac{27}{80}$$

$$= (\text{階級値}) \times (\text{相対度数})$$

# なぜ度数分布表から計算した平均値は大きくズレないのか？（つづき）

- **この作業を全階級にわたる和にあてはめる。**

$$(\text{平均値}) \approx \frac{\text{“}(\text{階級値}) \times (\text{度数})\text{”の和}}{(\text{全データ数})}$$

$$= \text{“}(\text{階級値}) \times \frac{(\text{度数})}{(\text{全データ数})}\text{”の和}$$

$$= \text{“}(\text{階級値}) \times (\text{相対度数})\text{”の和}$$

# どんな場合に生データの平均値と一致するか？

- **階級値を階級平均（＝その階級内の平均値）に置き換えると、度数分布表からの平均値は生データの平均値に一致する。**

| 階級値(a) | 相対度数(b) | (a×b) |
|---|---|---|
| 143.0000 | 0.0125 | 1.7875 |
| 147.2500 | 0.0500 | 7.3625 |
| 152.3529 | 0.2125 | 32.3750 |
| 156.8519 | 0.3375 | 52.9375 |
| 161.4783 | 0.2875 | 46.4250 |
| 166.8750 | 0.1000 | 16.6875 |
| | 平均値 | 157.5750 |

# どんな場合に生のデータの平均値と一致するか？（つづき）

- **なぜ？**

$$(\text{平均値}) = \frac{(\text{データ合計})}{(\text{全データ数})} = \frac{\text{“(階級内のデータ合計)” の和}}{(\text{全データ数})}$$

$$= \frac{\text{“}\dfrac{(\text{階級内のデータ合計})}{(\text{度数})} \times (\text{度数})\text{” の和}}{(\text{全データ数})}$$

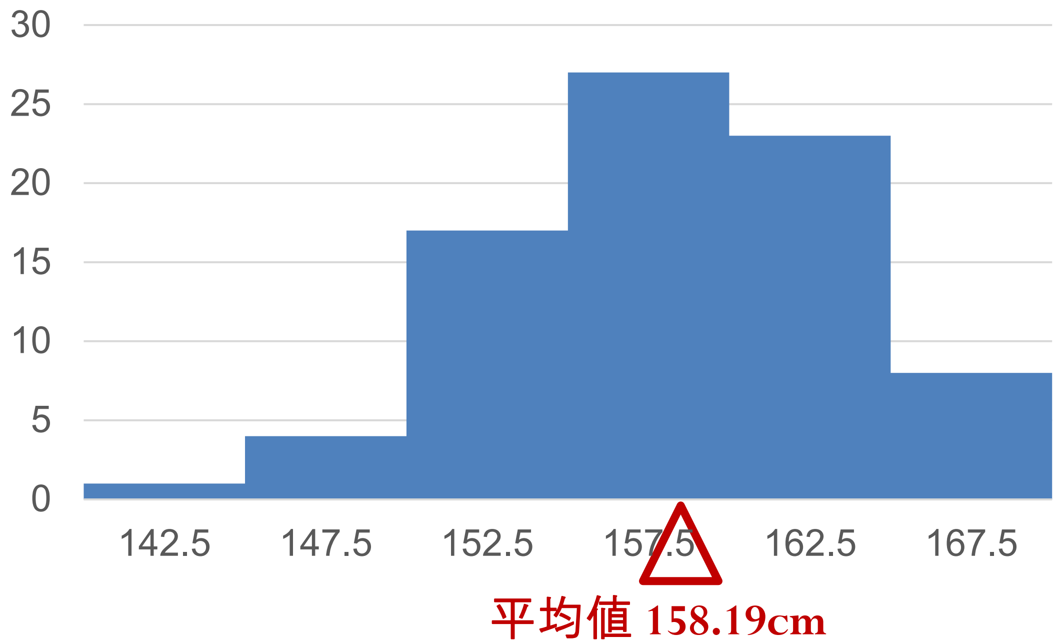
$$= \frac{\text{“(階級平均)} \times (\text{度数})\text{” の和}}{(\text{全データ数})}$$

$$= \text{“(階級平均)} \times \frac{(\text{度数})}{(\text{全データ数})}\text{”の和}$$

$$= \text{“(階級平均)} \times (\text{相対度数})\text{”の和}$$

# 平均値の意味

- **度数分布表から計算した平均値は、ヒストグラム全体を「やじろべえ」と見なした場合の支点に相当する。**



# 平均値の性質

1. データは平均値の周辺に分布している。
2. 平均値は多く現れるデータに影響を受ける。
   - 平均値は極端なデータにも影響を受ける。
3. ヒストグラムが左右対称である場合、その対称軸の通る点が平均値である。

# メディアン

- 「中心」を表す統計量は平均値だけではない。
- **メディアン**（median; **中央値、中心値**ともいう）も中心を表す統計量としてよく利用される。
- 平均値と違い、メディアンは極端なデータの影響を受けにくい。

# 生データからメディアンを求める

- メディアンは、データを大きさの順に並べ替えたとき、ちょうど真ん中に位置するデータを指す。
  - 全データ数が奇数なら何の問題もない。

# 生データからメディアンを求める（つづき）

- 全データ数が偶数の場合どうするか？
  - 女子大生の身長の例の場合、全データ数が80であるため、40番目と41番目のデータの平均をとる。
  - 40番目と41番目のデータが157と158であるため、

$$(\text{メディアン}) = \frac{157+158}{2} = 157.5$$

  - Excelの関数medianを利用しても同じ結果が得られる。

# 生データからメディアンを求める（つづき）

- **生データから計算した平均値**157.575cm**とメディアン**157.5cm**とは非常に近い。**
  - **身長の分布を左右対称と考えてよい一つの根拠となる。**

# 度数分布表からメディアンを求める

- 度数分布表しか与えられていない場合、どのようにしてメディアンを求めるか？

| 階　級 | 階級値 | 度　数 | 相対度数 | 累積度数 | 累積比率 |
|---|---|---|---|---|---|
| 140 ～ 145 | 142.5 | 1 | 0.0125 | 1 | 0.0125 |
| 145 ～ 150 | 147.5 | 4 | 0.0500 | 5 | 0.0625 |
| 150 ～ 155 | 152.5 | 17 | 0.2125 | 22 | 0.2750 |
| 155 ～ 160 | 157.5 | 27 | 0.3375 | 49 | 0.6125 |
| 160 ～ 165 | 162.5 | 23 | 0.2875 | 72 | 0.9000 |
| 165 ～ 170 | 167.5 | 8 | 0.1000 | 80 | 1.0000 |
| 合　計 | – | 80 | 1.0000 | – | – |

# 度数分布表からメディアンを求める（つづき）

- **40番目と41番目のデータは階級値157.5の階級の18番目と19番目に位置する。**

■**40番目：** $155+\frac{160-155}{27}\times 18\approx 158.333$

■**41番目：** $155+\frac{160-155}{27}\times 19\approx 158.519$

$$\therefore (\text{メディアン})=\frac{158.333+158.519}{2}\approx 158.43$$

# 5　分散と標準偏差

# なぜ散らばり具合を知るのか？

- 平均値は「中心」を示す統計量であった。
  - 平均値からデータの散らばり具合（分布の広がり具合）を知ることはできない。
- 現実には、データの散らばり具合に関心のある場合が多い。
  - 所得分布：ほぼ平等か、貧富の差が顕著か？
  - バスの運行状況：誤差は±2分か±10分か？

# 偏差を縮約する

- **個別のデータについて偏差（deviation; 平均値からの乖離）を計算する。**

  **(偏差) = (生データの値) − (平均値)**

- **偏差を縮約したい。**
  - **偏差の平均値は常にゼロになる。**

    **⇒単純に平均値をとってはダメ！**

# 偏差を縮約する（つづき）

- 偏差の平均値がゼロとなる理由は、プラスとマイナスの偏差が打ち消しあうためである。
  - 偏差の符号に興味はない。
  - 偏差の大きさに関心がある。
- 偏差の２乗の平均値を散らばり具合の尺度としてはどうか？
  - 偏差の２乗をとれば「符号の打ち消しあい」は起こらない。

# 分散と標準偏差

- 散らばり具合の尺度として、**分散**（variance）を

$$(\text{分散}) = \frac{(\text{偏差２乗和})}{(\text{全データ数})}$$

と定義する。

# 分散と標準偏差（つづき）

- **$n$個のデータ$X_1, X_2, \ldots, X_n$に対し、分散$\hat{\sigma}^2$は**

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

**と表現できる。**

# 分散と標準偏差（つづき）

- **以下の関係を用いると、分散の計算が簡単になる場合が多い。**

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \bar{X}^2$$

$$= \frac{(\text{生データの２乗和})}{(\text{全データ数})} - (\text{平均値の２乗})$$

# 分散と標準偏差（つづき）

- **分散の単位は、生データの単位の２乗となる。**
  - **女子大生の身長の例では、単位は**cm$^2$??
- **分散の正の平方根をとれば、単位の問題は解決する。**
  - **この統計量を標準偏差**（standard deviation; **略称**S.D.）**と呼ぶ。**
  - **分散**$\hat{\sigma}^2$**に対し、標準偏差**$\hat{\sigma}$**は**$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$**と定義される。**

# 分散と標準偏差：数値例

- **女子大生の身長の例：**

1. **分散：**

$$\hat{\sigma}^2 = \frac{(\text{偏差2乗和})}{(\text{全データ数})} = \frac{2307.55}{80} \approx 28.844$$

2. **標準偏差**

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \approx 5.371$$

# 分散と標準偏差：数値例（つづき）

- Excelの関数varpおよびstdevp（またはsqrt(varp)）を利用しても同じ結果が得られる。
  - <u>注意</u>：関数varおよびstdev（またはsqrt(var)）を利用すると、これらより大きい数値（正確に言うと、不偏分散およびその平方根）が得られる。

# 分散と標準偏差：数値例（つづき）

- 女子大生の身長のデータから、次の統計量が得られた。
  - 平均値：157.575cm
  - 標準偏差：5.371cm
- この結果を次のように解釈することができる。
  - 平均的な女子大生の身長はほぼ157.6cmだが、実際の身長はその前後に5.4cm程度散らばっている。

# 度数分布表から標準偏差を求める

- **度数分布表しか与えられていない場合、どのようにして分散・標準偏差を求めるか？**

| 階　級 | 階級値 | 度　数 | 相対度数 | 累積度数 | 累積比率 |
|---|---|---|---|---|---|
| 140 ~ 145 | 142.5 | 1 | 0.0125 | 1 | 0.0125 |
| 145 ~ 150 | 147.5 | 4 | 0.0500 | 5 | 0.0625 |
| 150 ~ 155 | 152.5 | 17 | 0.2125 | 22 | 0.2750 |
| 155 ~ 160 | 157.5 | 27 | 0.3375 | 49 | 0.6125 |
| 160 ~ 165 | 162.5 | 23 | 0.2875 | 72 | 0.9000 |
| 165 ~ 170 | 167.5 | 8 | 0.1000 | 80 | 1.0000 |
| 合　計 | – | 80 | 1.0000 | – | – |

# 度数分布表から標準偏差を求める（つづき）

- **基本的な考え方は、度数分布表から平均値を求める場合と同様である。**
  - **具体的には、偏差を**

    **(偏差)**
    **＝ (階級値)**
    **– (度数分布表から計算した平均値)**

    **と定義すればよい。**

# 度数分布表から標準偏差を求める（つづき）

| 階級値(a) | 相対度数(b) | (a×b) | 偏差(c) | 偏差2乗 (d = c^2) | (d×b) |
|---|---|---|---|---|---|
| 142.5 | 0.0125 | 1.7813 | -15.6875 | 246.0977 | 3.0762 |
| 147.5 | 0.0500 | 7.3750 | -10.6875 | 114.2227 | 5.7111 |
| 152.5 | 0.2125 | 32.4063 | -5.6875 | 32.3477 | 6.8739 |
| 157.5 | 0.3375 | 53.1563 | -0.6875 | 0.4727 | 0.1595 |
| 162.5 | 0.2875 | 46.7188 | 4.3125 | 18.5977 | 5.3468 |
| 167.5 | 0.1000 | 16.7500 | 9.3125 | 86.7227 | 8.6723 |
| | 平均値 | 158.1875 | | 分散 | 29.8398 |
| | | | | 標準偏差 | 5.4626 |

- Excel**でこの分散を求めるには、**

$$sumproduct((\text{偏差2乗}),(\text{相対度数}))$$

**とセル指定する。**

# ６　標準偏差の持つ意味

# 標準偏差を知ることで…

1. 一組のデータの中の特定の1つのデータの持つ意味がわかる。
   - そのデータは特殊かどうか？
2. 複数のデータセットを比較して出てくる違いがわかる。
   - 違うデータセットを違う性質のものと評価できるか？

# データの特殊性の評価：試験の例

- ある試験の平均点が60点のところ、自分の点数は75点であった。
  - どの程度胸を張ってよいか？
- 標準偏差の大きさ次第で胸の張り具合も変わる！
  1. 標準偏差が12点なら、偏差（＝平均点からの乖離）は標準偏差1個分程度（普通）。
  2. 標準偏差が8点なら、偏差（＝平均点からの乖離）は標準偏差の2倍近い（良い）。

# 標準化

- **試験の点数の例では、**

$$\frac{(\text{偏差})}{(\text{標準偏差})} = \frac{(\text{データ}) - (\text{平均値})}{(\text{標準偏差})}$$

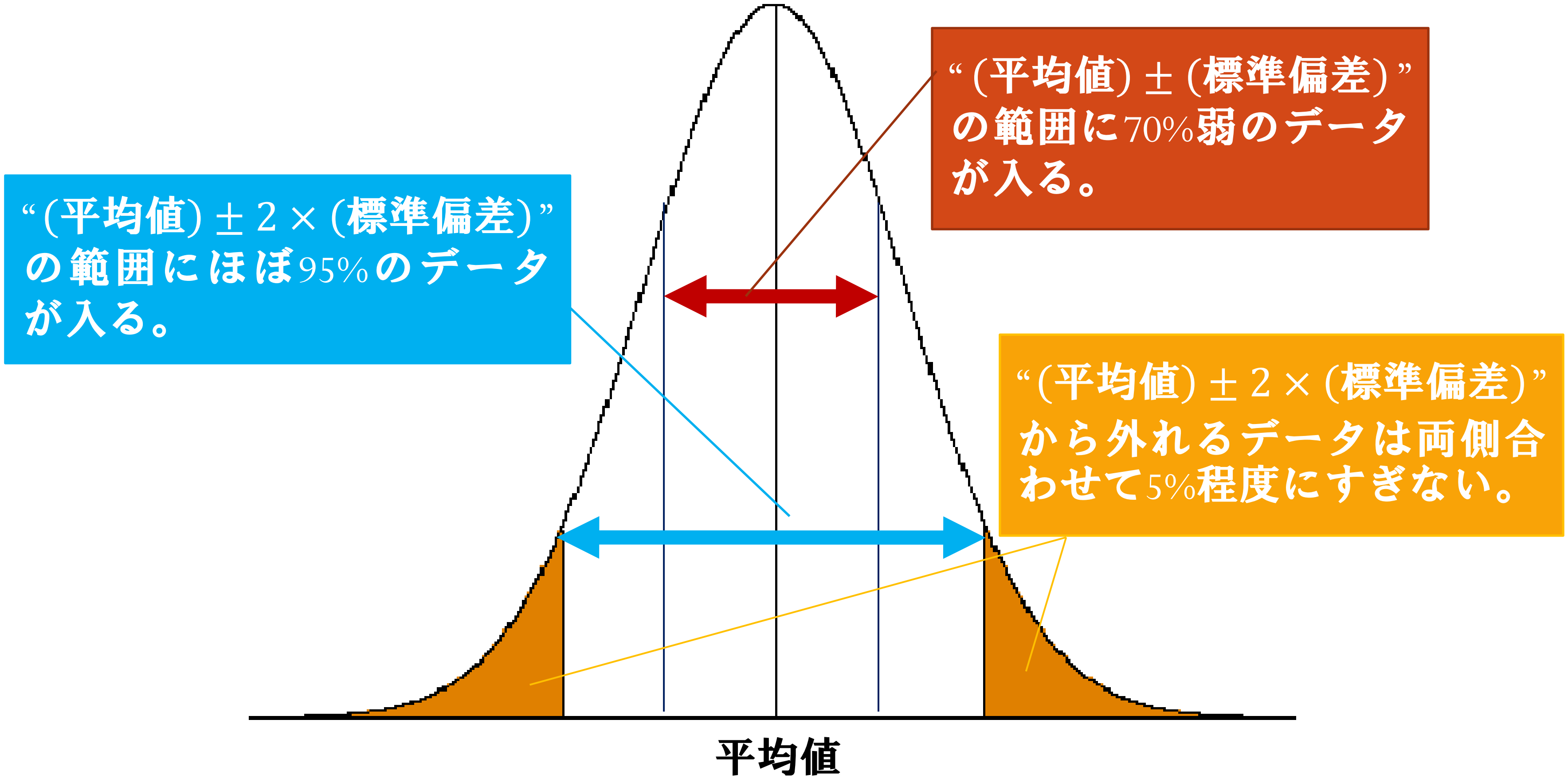
**を「普通の出来」か「良い出来」かの判断基準としていた。**

- **このようなデータの変形を標準化（基準化）という。**

# 特殊性の根拠

- **データセット（例：試験の点数）の分布は正規分布に近いと仮定する。**
- **正規性の下では、**
  - **(平均値)±(標準偏差)の範囲に70%弱のデータが入る。**
  - **(平均値)±2×(標準偏差)の範囲にほぼ95%のデータが入る。**
    - **この範囲を外れるデータは5%にすぎない。**

# 特殊性の根拠（つづき）



“(平均値) ± (標準偏差)” の範囲に70%弱のデータが入る。
“(平均値) ± 2 × (標準偏差)” の範囲にほぼ95%のデータが入る。
“(平均値) ± 2 × (標準偏差)” から外れるデータは両側合わせて5%程度にすぎない。
平均値

# データの性質の違い：試験の例

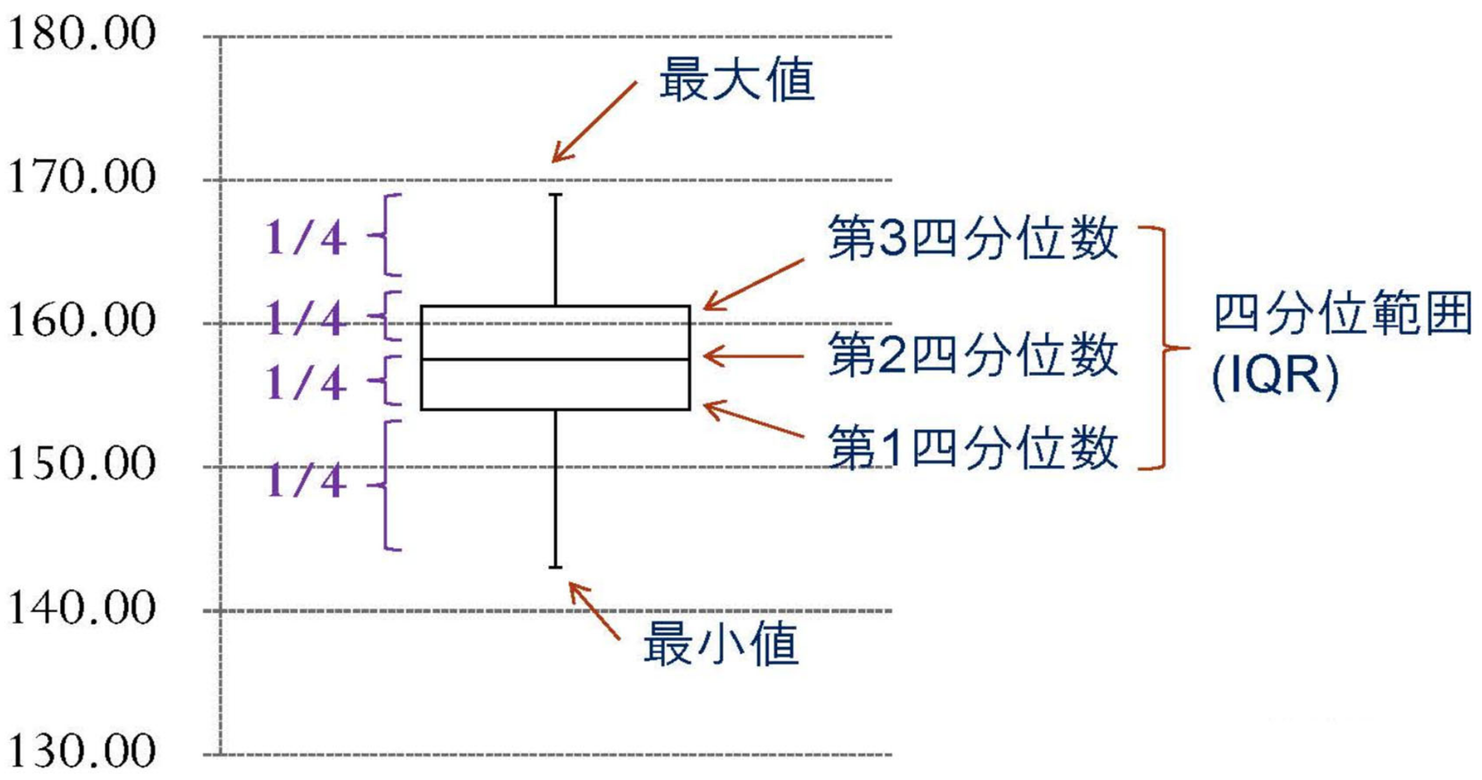
- **模擬試験を**10**回受験した結果、**
  1. X**君は平均**60**点、標準偏差**10**点であった。**
  2. Y**君は平均**50**点、標準偏差**30**点であった。**
- X**君の方が優秀か？**
  - **標準偏差**1**個分の幅をとる：**
    1. X**君は**50~70**点をとる学生である（安定）。**
    2. Y**君は**20~80**点をとる学生である（ムラ）。**
  - **合格ラインが**80**点だとしたらどうなるか？**

# 7　箱ひげ図

# 箱ひげ図とは？

- ①最小値、②第1四分位数、③第2四分位数（＝中央値）、④第3四分位数、⑤最大値を使い、データのばらつきをわかりやすく表現するためのグラフである。
- これらはExcelの関数quartile.incを使って計算可能である。

# 箱ひげ図とは？（つづき）


180.00
170.00
160.00
150.00
140.00
130.00
最大値
1/4
1/4
1/4
1/4
第3四分位数
第2四分位数
第1四分位数
四分位範囲
(IQR)
最小値

# Excelを利用した作図

- Excel 2016から“統計グラフ”のメニューに箱ひげ図が加わった。
  - 箱ひげ図オプションでは自動的に「外れ値」が計算・表示されてしまう。
  - 旧バージョンExcelの手順に従い、積み上げ棒グラフを使うと外れ値なしの箱ひげ図を作成できる。

# Excelを利用した作図（つづき）

1. **最小値（Min）、第1四分位数（Q1）、第2四分位数（Q2）、第3四分位数（Q3）、最大値（Max）を求める。**
2. **1の結果をもとに次のような表を作成する。**

| | |
|---|---|
| Min | 143.00 |
| Q1 | 154.00 |
| Q2 | 157.50 |
| Q3 | 161.25 |
| Max | 169.00 |

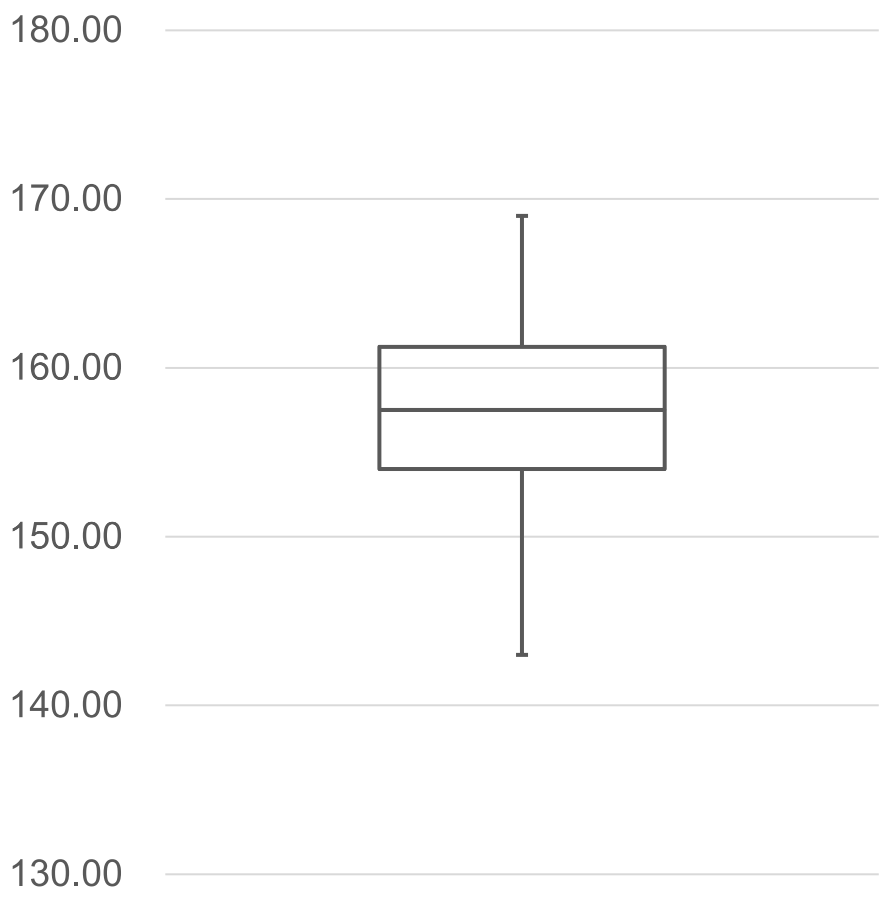
--->

| | |
|---|---|
| Q1 - Min | 11.00 |
| Q1 | 154.00 |
| Q2 - Q1 | 3.50 |
| Q3 - Q2 | 3.75 |
| Max - Q3 | 7.75 |

# Excelを利用した作図（つづき）

3. **2のデータから箱ひげ図を作成する大まかな手順は次のとおりである【詳細は実演】。**
   a. Q1、Q2-Q1、Q3-Q2**の3つの数値から積み上げ縦棒グラフを描き、一番下の部分を「塗りつぶしなし」とし「箱」の部分を作成する。**
   b. Q1-Min、Max-Q3**の数値を誤差範囲として追加し「ひげ」の部分を描く。**
   c. **箱の部分を白抜きにし、横幅を調整して箱ひげ図を完成させる。**

# 箱ひげ図：完成品



# Computer Exercise 2

- ファイル"data1.xlsx"にある女子大生の身長データについて、Excelを利用して
  ①平均
  ②メディアン
  ③分散
  ④標準偏差
  を計算せよ。ただし、いずれの統計量も生データと度数分布表双方から計算し、結果を比較すること。