データの散布度

樋口さぶろお　https://hig3.net

龍谷大学理工学部数理情報学科

確率統計☆演習　L03(2020-10-12 Mon)

最終更新: Time-stamp: "2020-10-10 Sat 12:56 JST hig"

今日の目標

- 散布度:レンジ, 四分位範囲, 分散, 標準偏差を手で求められる 岩薩林 確率・統計 §1.2 高校 数学 I
- 平均値, 分散を 1 次式で変換できる
- データの標準得点, 偏差値を求められる

L02-Q1

Quiz 解答:平均値中央値最頻値

$n = 9$.

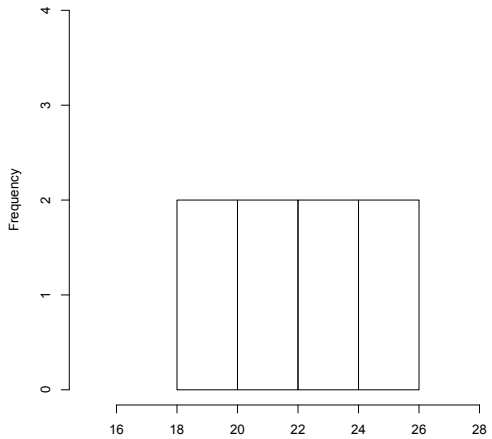
1. 中央値 $Q_2 = x_{\frac{1}{2}\cdot(9-1)} = x_4$. (0 番目から数え始めて 4 番目). 高校のりでいったら, 小さい順に並べた 9 個の真ん中の値. よって階級 $21 - -23$ に含まれる. 階級値 22 で近似できる.
2. もっとも度数の多い階級の階級値で, 10.
3. 階級値で近似して計算すると,
   $\frac{1}{9}(10 \times 3 + 22 \times 2 + 24 \times 2 + 26 \times 2) = 19.3$.

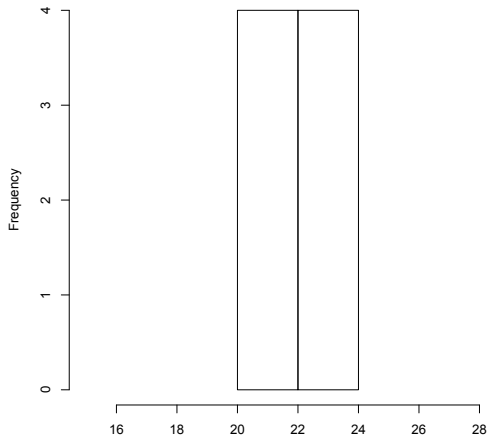
ここまで来たよ

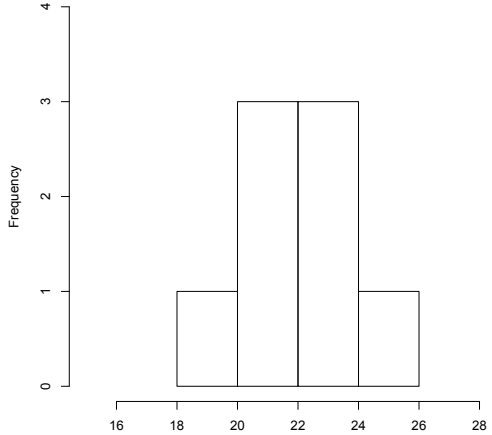
2. データの代表値

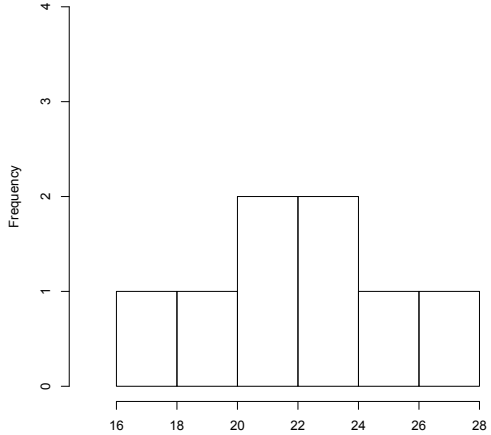
3. データの散布度
   - 分位数タイプ:レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 平均値タイプ:分散・標準偏差
   - 平均値・分散・標準偏差の 1 次式による変換
   - 標準得点・偏差値

平均値が同じでも分布はいろいろ









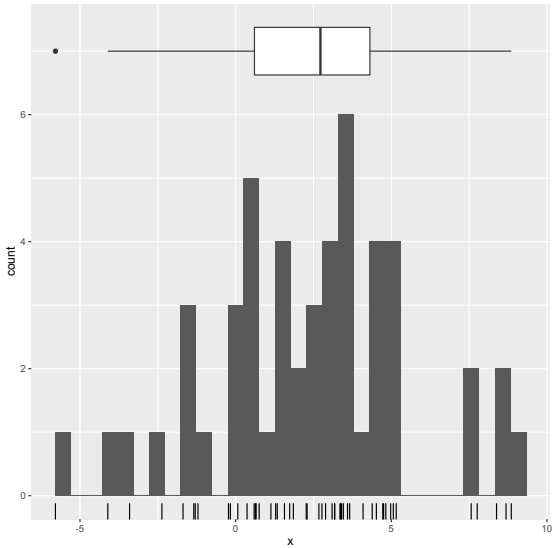
第 1,3 四分位数は?

散布度:散らばりの尺度が必要

レンジ・四分位範囲の定義 I

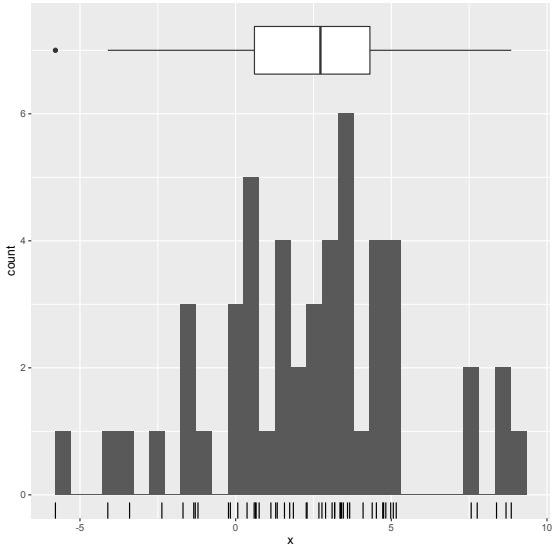
分位数タイプの量の定義 高校 数学 I 岩薩林 確率・統計なし

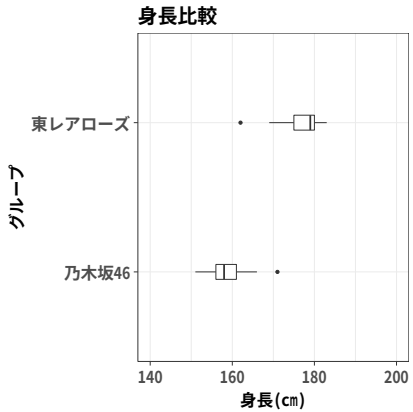
- 範囲 (レンジ) $=$ $Q_4 - Q_0$
- 四分位範囲 (interquartile range) IQR$=$ $Q_3 - Q_1$



- $Q_0 = x_0$:最小値
- $Q_1 = x_{\frac{1}{4}(n-1)}$:第 1 四分位数
- $Q_2 = x_{\frac{2}{4}(n-1)}$:中央値
- $Q_3 = x_{\frac{3}{4}(n-1)}$:第 3 四分位数
- $Q_4 = x_{n-1}$:最大値

箱ひげ図 (Box Plot, Box and Whisker diagram)





箱ひげ図を描く手順 高校 数学 I

- Q_0, Q_4 Q_1, Q_2, Q_3 と平均値 $\overline{x}$ を求める
- Q_2 に縦線をいれる
- Q_1, Q_3 を左右の端として箱を描く
- Q_0, Q_4 に短い縦線をいれ，点線のひげで箱とつなぐ
- (平均値に $+$ を 1 個描く)
- (「外れ値を○で描く」)

並べて，データ群の間の比較によく使う

外れ値=Q_1, Q_3 から外側に，IQR の $\alpha = 1.5$ 倍より離れているデータ

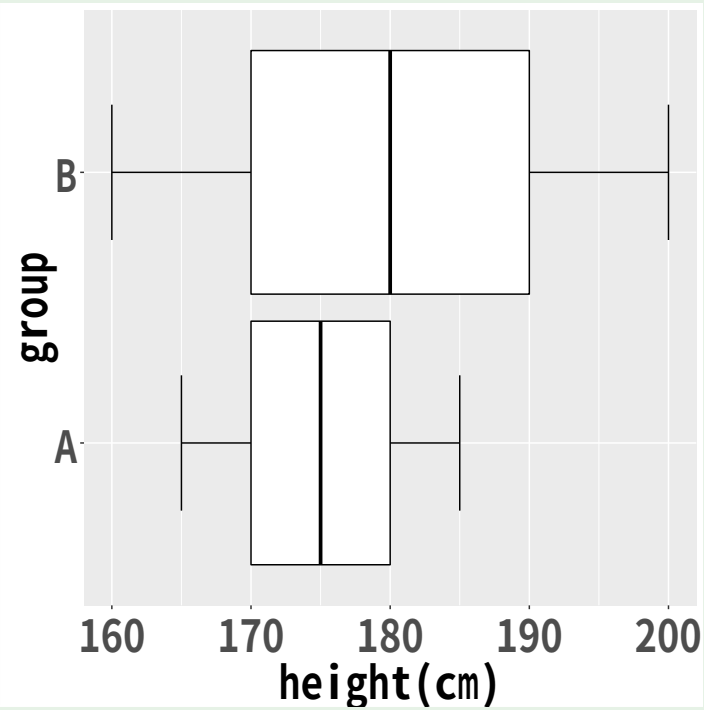
`https://www.geogebra.org/m/dbfbkews`

L03-Q1

Quiz(箱ひげ図の比較)

8000 人からなる group A と, 2000 人からなる group B の身長 (cm) を測定して箱ひげ図に表したものが次である.

1. group A を身長の大きい方から小さい方に並べたとき, 6000 位 (くらい) の人の身長を答えよう.
2. group B を身長の大きい方から小さい方に並べたとき, 500 位 (くらい) の人の身長を答えよう.
3. group A で身長が 170cm 以上 175cm 未満の人の人数は, group B で身長が 170cm 以上 190cm 未満の人の人数の何倍か. 小数で答えよう.



ここまで来たよ

2 データの代表値

3 データの散布度
- 分位数タイプ:レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
- 平均値タイプ:分散・標準偏差
- 平均値・分散・標準偏差の 1 次式による変換
- 標準得点・偏差値

分散・標準偏差の定義 高校 数学 I | 岩薩林 確率・統計 (1.4),(1.5)

データ: $x_1, x_2, \ldots, x_n$.

分散タイプの量の定義

- データの分散 (variance)

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

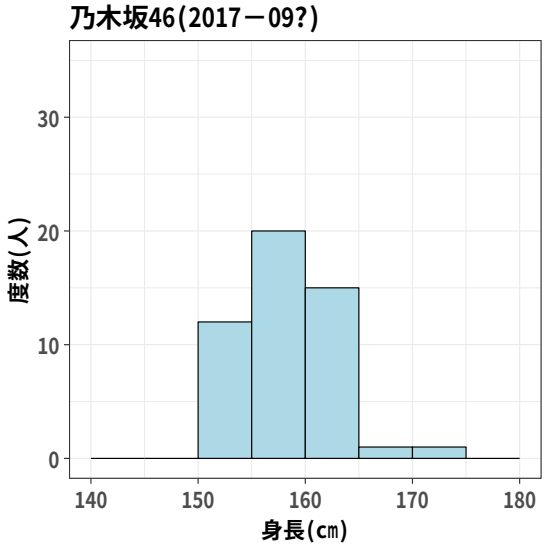
なぜそんな定義? 下の「平均値との差」の平均値見て

- データの標準偏差 (standard deviation)$= S = \sqrt{S^2} \geq 0$

岩薩林 確率・統計定理 1.1(p.6)

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

(例) グループ (49 人) の身長 I



$n-1=49-1$ で割りたくなった人もいるかも. ここは 49 で OK

そのうちちゃんと区別を説明します.

データの単位 $\neq$ 分散の単位

- 平均値 $\overline{x}=\frac{171+166+165+\cdots+151}{49}=158.7(\text{cm})$
- 分散 $S^2=\frac{(171-158.7)^2+(166-158.7)^2+\cdots+(151-158.7)^2}{49}=17.7\ (\text{cm}^2)$
- 標準偏差 $S=\sqrt{17.7}=4.21\ (\text{cm})$

158.7 cm を四捨五入で 159,160 にすると, 丸め誤差, 桁落ち の危険 数値計算法

度数分布表からの分散・標準偏差の求め方

高校 数学 I | 岩薩林 確率・統計 §1.2.3,(1.14)p.17

$$S^2 = \frac{1}{n}\sum_j (x_j - \overline{x})^2 f_j = \frac{\sum_j (x_j - \overline{x})^2 f_j}{\sum_j f_j}.$$

ヒストグラムからの標準偏差の読み取り方

長方形なら幅の 0.3 倍くらい

力学のりでの分散の求め方

質点系の慣性モーメント $I = \frac{\sum_{j=1}^{k}(x_j - x_G)^2 m_j}{\sum_j m_j}$ 力学

j 番目の質点の位置 x_j, 質量 m_j

分散公式 (便利な (こともある) 計算方法)

高校 数学 I 岩薩林 確率・統計定理 1.2(p.10)

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

岩薩林 確率・統計例題 1.3(p.7)

岩薩林 確率・統計例題 1.4(p.11)

L03-Q2

Quiz(平均値・分散・標準偏差)

データ 87kg, 93kg, 89kg, 91kg, 90kg の平均値・分散・標準偏差を求めよう.

ここまで来たよ

2. データの代表値

3. データの散布度
   - 分位数タイプ:レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 平均値タイプ:分散・標準偏差
   - 平均値・分散・標準偏差の 1 次式による変換
   - 標準得点・偏差値

平均値・分散・標準偏差の変換 岩薙林 確率・統計 §1.2.2

u から x への変換

データ $u_1, u_2, \ldots, u_n$, u の平均値 $\overline{u}$, 分散 S_u^2, 標準偏差 S_u がわかってるとする.
$x_i = bu_i + a$ で新しいデータを作る (a, b 定数).
データ $x_1, x_2, \ldots, x_n$, x の平均値 $\overline{x}$, 分散 S_x^2, 標準偏差 S_x はどうやって求める?

例: 身長の換算 $x = 1.8(\mathrm{m}) \leftarrow u = 80(\mathrm{cm})$

$x = bu + a$, $\boxed{b = 0.01, a = 1}$

平均値・分散・標準偏差の 1 次式による変換 岩薩林 確率・統計定理 1.3

$x = bu + a$ のとき

1. $\overline{x} = b\overline{u} + a$ 岩薩林 確率・統計 (1.9)
2. $S_x^2 = b^2 \times S_u^2$ 岩薩林 確率・統計 (1.10)
3. $S_x = |b| \times S_u$ 岩薩林 確率・統計 (1.11)

岩薩林 確率・統計例題 1.5(p.13), 問題 3(p.14)

L03-Q3

Quiz(平均値・分散・標準偏差の 1 次式による変換)

ある集団の身長 (みんな大人で 100cm 以上) を, cm で書いたものの下 2 桁 u cm の, 平均値は 60cm, 分散は 25cm^2 だった.
m で書いた身長 x m の平均値と分散と標準偏差を求めよう.

ここまで来たよ

2. データの代表値

3. データの散布度
   - 分位数タイプ:レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 平均値タイプ:分散・標準偏差
   - 平均値・分散・標準偏差の 1 次式による変換
   - 標準得点・偏差値

標準偏差の意味 I

L03-Q4

Quiz(分散の意味)

あるクラスで行われたテストで, 英語の平均点は 60 点, 標準偏差 10 点.
数学の平均点は 60 点, 標準偏差 20 点.
英語の 70 点と数学の 70 点, どちらのほうが価値ある (上位にいる可能性が高い)? 次のうちから正しいものを 1 つ選ぼう.

1. たぶん英語のほうが価値ある
2. たぶん数学のほうが価値ある
3. どちらも同じ
4. 追加の情報がないとわからない
5. 追加の情報があっても比べることはできない

標準得点

標準得点 (standard score, z-score, z 得点) 岩薩林 確率・統計 (1.13) 例 4(p.14)

$$(\text{値 } x_i \text{ の}) \text{標準得点 } z_i = \frac{x_i - \overline{x}}{S_x}$$

平均値から, 上下どちらに, 標準偏差の何倍離れているかを表す値.

$u = z$ と思うと, $b = S_x, a = \overline{x}$

例 $n = 5$

i	1	2	3	4	5	平均値	標準偏差
データ x_i	15	13	12	11	9	12	2
標準得点 z_i	1.50	0.5	0	-0.5	-1.50	0	1

L03-Q5

Quiz(標準得点と偏差値)

データ 87, 93, 89, 91, 90 で, 87 の標準得点と偏差値を求めよう.

標準得点の性質

標準得点 z の性質 岩薩林 確率・統計問題 4(p.15)

- $\overline{z} = \boxed{0}$
- $S_z^2 = \boxed{1}$, $S_z = \boxed{\sqrt{1} = 1}$
- z の単位は $\boxed{\frac{\mathrm{m}}{\mathrm{m}}}$, 無次元の数. 身長が 180cm, 80cm, 1.8m どれでも同じ結果.

なぜなら 岩薩林 確率・統計問題 1.4,

$$\overline{x} = b\overline{z} + a$$
$$S_x = |b| S_z.$$

偏差値

学力データ (テストの点数や成績?) によく使われる.
受験者 1 人 1 人の成績が, 平均値から上, または下に離れている程度を見られる.

偏差値

$$\begin{aligned}(\text{値 } x_i \text{ の}) \text{ 偏差値 } w_i =& 10z_i + 50 \\ =& \frac{x_i - \overline{x}}{S_x} \times 10 + 50.\end{aligned}$$

- 異なるテストでも比べられる.
- 偏差値の平均値は 50, 偏差値の標準偏差は 10
- 偏差値はまあ '無次元の数'(同じ受験者集団の, 1000 点満点と 100 点満点のテストを比較可能)