データのばらつきを表す値

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I　L03(2016-10-06 Thu)

最終更新: Time-stamp: "2016-10-07 Fri 17:09 JST hig"

今日の目標

- 高校 数学 I　塚田確率統計 1.5
- データから範囲, 四分位範囲, 四分位偏差, 分散, 標準偏差を求められる
- 標準得点, 偏差値が計算できる
- 1 次関数で平均値, 分散, 標準偏差を変換できる

http://hig3.net

L02-Q1

Quiz 解答:四分位値

$Q_2 = 17$, $Q_1 = 14.5$, $Q_3 = 18$.

L02-Q2

Quiz 解答:代表値

1. $Q_2 = 17\text{cm}, Q_1 = 14.5\text{cm}, Q_3 = 18\text{cm}$.
2. 最頻値は 18cm.
3. 平均値は $(14 + \cdots + 25)/8 = 17.25\text{cm}$.

L02-Q3

Quiz 解答:平均値中央値最頻値

1. 22
2. 10
3. 19.3

L02-Q4

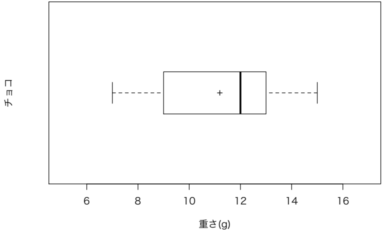
Quiz 解答:四分位数と箱ひげ図

$\overline{x} = 10.8\mathrm{g}.$

$\mathrm{Q}_2 = 12\mathrm{g}.$

$\mathrm{Q}_1 = \frac{1}{2}[8 + 10] = 9\mathrm{g}.$

$\mathrm{Q}_3 = \frac{1}{2}[12 + 14] = 13\mathrm{g}.$



ここまで来たよ

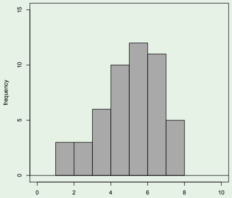
1. データの代表値
   - ヒストグラムと箱ひげ図の対応
2. データのばらつきを表す値
   - 範囲 (Range)・四分位範囲 (IQR)・四分位偏差
   - 分散
   - 分散の意味と平均値・分散・標準偏差の変換
   - 変動係数・標準得点・偏差値

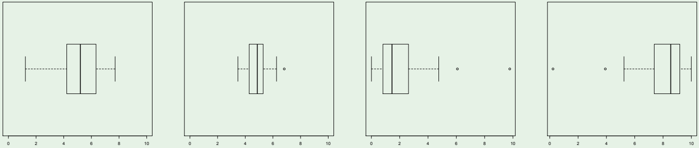
ヒストグラムと箱ひげ図の用法

箱ひげ図は, 複数の集団 (サンプル) の分布の比較に便利.
情報量: 度数分布表=ヒストグラム $>$ 箱ひげ図=第 i 四分位数
相互に変換できるようになろう.
cf. センター試験 (2015,2016)

Quiz(ヒストグラムと箱ひげ図)

このヒストグラムに対応する箱ひげ図はどれ?

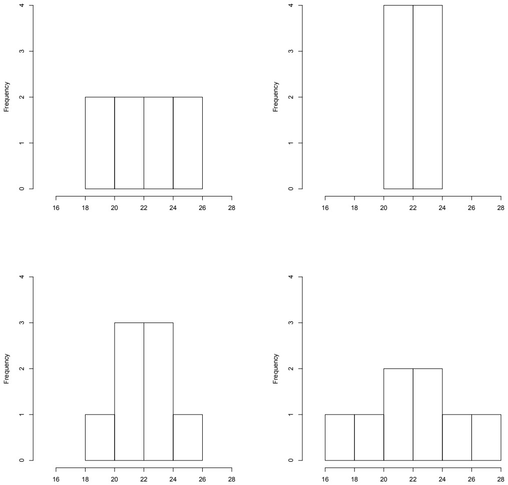




ここまで来たよ

1. データの代表値
   - ヒストグラムと箱ひげ図の対応
2. データのばらつきを表す値
   - 範囲 (Range)・四分位範囲 (IQR)・四分位偏差
   - 分散
   - 分散の意味と平均値・分散・標準偏差の変換
   - 変動係数・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?

データのばらつきを表す量

範囲タイプの量の定義 塚田確率統計 1.5.1 高校 数学 I

- 範囲 (range) $=\boxed{}$
- 四分位範囲 interquartile range

  IQR$=\boxed{}$
- 四分位偏差 quartile deviation $=\frac{1}{2}$ IQR

例 1: 30 50 55 55 60 70 70 70 75 100

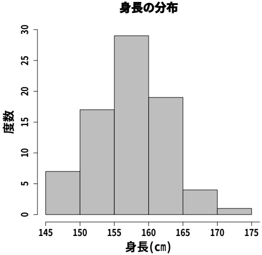
L03-Q1

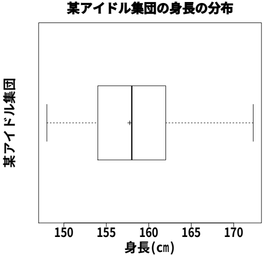
Quiz(範囲)

次のデータの, 範囲, 四分位範囲, 四分位偏差 を求めよう.

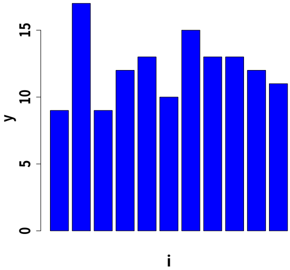
14 14 15 16 18 18 18 25

範囲・四分位範囲・四分位偏差の箱ひげ図・ヒストグラム的意味

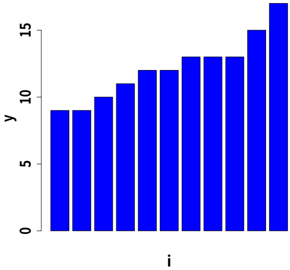




本来，箱ひげ図はAKBとBerrizと…を並べて比較に使うもの．



→ 並べかえ



ここまで来たよ

1. データの代表値
   - ヒストグラムと箱ひげ図の対応
2. データのばらつきを表す値
   - 範囲 (Range)・四分位範囲 (IQR)・四分位偏差
   - 分散
   - 分散の意味と平均値・分散・標準偏差の変換
   - 変動係数・標準得点・偏差値

分散 塚田確率統計 1.5.2 高校 数学 I

データ: $x_1, x_2, \ldots, x_n$.
データの平均値 (mean): $\overline{x}$ $(= m)$
x_i の偏差 (deviation)$= x_i - \overline{x}$

分散と標準偏差の定義

- データの分散 (variance): (偏差)^2 の平均

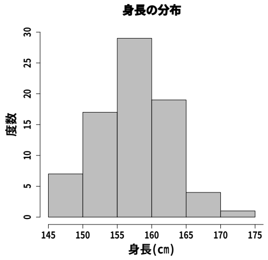
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- データの標準偏差 (standard deviation)$=$ $\boxed{}$

高校 数学 I の公式

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

(例) 某国民的アイドル集団 (77 人) の身長



- 平均値 $\overline{x} = \frac{148+148.5+\cdots+172.3}{77} = 158$(cm)
- 分散 $s^2 = \frac{(148-158)^2+(148.5-158)^2+\cdots+(172.3-158)^2}{77} = 26.0$ (cm^2)
- 標準偏差 $s = \sqrt{26.0} = 5.1$ (cm)

$n-1 = 77-1$ で割りたくなった人もいるかも．ここは 77 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

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.

度数分布表からの平均値と分散の (だいたいの) 求め方

$$\overline{x} \approx \frac{1}{n}\sum_{i=1}^{k} x_{(i)} f_i$$

塚田確率統計 p.25　確率統計☆演習 I(2016)L02

$$s^2 \approx \frac{1}{n}\sum_{i=1}^{k} (x_{(i)} - \overline{x})^2 f_i$$

塚田確率統計 p30 式 (1.5)

L03-Q3

Quiz(度数分布表から分散)

次の度数分布表で表されるデータの平均値と分散を (近似的に) 求めよう.

階級	度数
45cm 以上 55cm 未満	10
55cm 以上 65cm 未満	20
65cm 以上 75cm 未満	20

ここまで来たよ

1. データの代表値
   - ヒストグラムと箱ひげ図の対応
2. データのばらつきを表す値
   - 範囲 (Range)・四分位範囲 (IQR)・四分位偏差
   - 分散
   - 分散の意味と平均値・分散・標準偏差の変換
   - 変動係数・標準得点・偏差値

分散の意味 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? 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. これだけの情報ではまったくわからない
5. 平均点が 60 点だと再テストがあるだろう

平均値・分散・標準偏差の変換 塚田確率統計式 (1.2), (1.5) の前

x から y への変換

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 がわかってるとする.
$y_i = ax_i + b$ で新しいデータを作る (a, b 定数).
データ $y_1, y_2, \ldots, y_n$, y の平均値 $\overline{y}$, 分散 s_y^2, 標準偏差 s_y はどうやって求める?

例: 身長の換算 $y = 1.8$(m) $\leftarrow x = 80$(cm)

$y = ax + b$, $\boxed{}$

平均値, 分散, 標準偏差の変換

$y = ax + b$ のとき

1. $\overline{y} = a\overline{x} + b$ 塚田確率統計式 (1.2)
2. $s_y^2 = |a|^2 \times s_x^2$ 塚田確率統計 p.30 のいちばん上の式
3. $s_y = |a| \times s_x$

L03-Q5

Quiz(平均値・分散・標準偏差の換算)

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 x cm の, 平均値は 60cm, 分散は 25cm^2 だった.
m で書いた身長 y 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

1. データの代表値
   - ヒストグラムと箱ひげ図の対応
2. データのばらつきを表す値
   - 範囲 (Range)・四分位範囲 (IQR)・四分位偏差
   - 分散
   - 分散の意味と平均値・分散・標準偏差の変換
   - 変動係数・標準得点・偏差値

身長と靴のサイズじゃ標準偏差の意味が違う! 塚田確率統計 1.6

Berriz 工房内で, 身長の標準偏差は 20cm くらいだけど, 靴のサイズの標準偏差は 3cm くらい.

標準偏差が大きい = いろんな体格の人がいる

みたいに思いたいけど, 身長と靴のサイズじゃ標準偏差の意味が違う.

変動係数 (coefficient of variation) 塚田確率統計 (1.6)

$$(\text{データ } x \text{ 全体の}) \ \text{変動係数} = \frac{s_x}{\bar{x}} \times 100$$

- これは無次元の数. すなわち単位がない量.
  $\boxed{}$.
- $\dfrac{\text{分散}}{\text{平均値}}$ だと無次元の数にはならない.

標準得点

標準得点 (standard score)

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

- 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.
- z-得点 (z-score) などともいう.

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

標準得点の性質

標準得点 z の性質

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

なぜなら… いま $\boxed{}$.

$$\overline{z} = a\overline{x} + b = \frac{1}{s_x} \cdot \overline{x} - \frac{\overline{x}}{s_x} = 0.$$

$$s_z = |a| s_z = \left| \frac{1}{s_x} \right| s_x = 1.$$

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

$$\text{(値 } x_i \text{ の) 偏差値 } w = 10z_i + 50 = \frac{x_i - \overline{x}}{s_x} \times 10 + 50.$$

$a = \boxed{}$, $b = \boxed{}$

- 異なるテスト, クラスでも比べられる.
- 偏差値の平均値は $\boxed{}$, 偏差値の標準偏差は $\boxed{}$
- 偏差値はまあ '無次元の数'(1000 点満点と 100 点満点を比較可能)

L03-Q6

Quiz(偏差値)

(学力) 偏差値について, 次のうち正しいのはどれ (とどれ)?

1. 偏差値の最低値は 0 である
2. 偏差値の最高値は 75 である
3. 平均点 (をとった人) の偏差値は 50 である
4. 100 点のテストで満点を取った場合の偏差値は, 他の人の成績しだいである
5. 偏差値 50 の人の順位は上から 1/2 程度である
6. 偏差値 60 の人の順位は上から 15% 程度である.

連絡

欠席届 毎回出席を前提に進めます．やむを得ず欠席して，ピーナッツ的に考慮されたい場合は，専用用紙に事情を説明する書類を貼って，授業前後各 5 分に提出 (事前事後とも可．ファイナルトライアルが締切)．欠席に事前連絡は原則不要．何回欠席してもファイナルトライアル参加資格を失うことはありません．

- 配布資料は 1-503 向かいの引出, `http://hig3.net` で再配布．
- 加減乗除と平方根 (ルート) の使える電卓持ってきてね．関数電卓でなくてもいいです．携帯電話の機能・アプリでもかまいません．
- 樋口オフィスアワー木 6 金昼 (1-502), Math ラウンジ月-木昼 (1-614)
- 次回は 塚田確率統計 1.6 , 塚田確率統計 1.7 , 臨時教室変更．

`https://manaba.ryukoku.ac.jp`