

データの分布

樋口さぶろお

龍谷大学理工学部数理情報学科

使える統計! L01(2013-09-25 Wed)

今日の目標

- 1 変数の量的データから箱ひげ図が描ける
- 2 1 変数の量的データから度数分布表が作れる
- 3 1 変数の量的データからヒストグラムが描ける



<http://hig3.net>

この授業ののり

本名

生活の中の統計技術

知るというより、できることが目標の科目です。

成績計算

面倒な科目かも…科目の成績 100 ピーナッツは

- 20 ピーナッツ/14 回授業での活動. 演習, 携帯で投稿, 自宅で予習復習, など毎回内容は違うので, 先週と違う, ってのは言いっこなしで. ただし配点の大きいことをやる時は十分事前に言います.
- 30 ピーナッツ: プチテスト (予定日 2013-11-13)
- 50 ピーナッツ: ファイナルトライアル (予定日 2014-01-29)

現在の点数は e ラーニングサイトで見られるようになる予定.

欠席届

授業のようなミーティングに欠席するときは事前に連絡, が世の中の常識ですが, この授業に関しては, 自分で情報を取得して復習して補うことを前提に, 欠席の事前連絡は不要です. 何回欠席しても期末試験受験資格を失うことはありません.

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).

授業のページ

<http://hig3.net> > (左コラム) 樋口の授業 > 生活の中の統計技術. eラーニングシステムにもページがあります.

オフィスアワー

予約なしで科目について質問相談会話できる時間です. 火昼 (1号館5階 1-502), 木6 (1号館5階 1-539). 月金昼も在室時は訪問歓迎. お弁当可.

ここまで来たよ

① はじめに

② データの分布

- データとは？
- 箱ひげ図
- 度数分布表
- ヒストグラム
- ヒストグラムと箱ひげ図の対応

1 変数の量的データ

某アイドル集団 (77 名)+某バレーボール選手 (1 名) の身長データ.

148cm
148.5cm
149cm
⋮
185cm

ps3id_raicho_1182 さん (最終更新日時:2012/3/20) 投稿日:
2012/2/15 AKB48 身長 まとめ (研究生は 12.5 期まで)

<http://note.chiebukuro.yahoo.co.jp/detail/n32745>

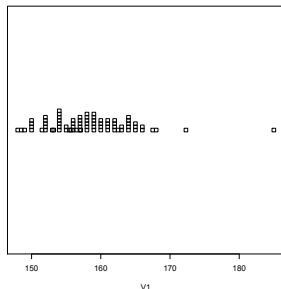
このコースの最後までいくと問えること (正確な表現ではありません)

- オーディションにおいて、身長は考慮されているか?
- チーム編成において、身長は考慮されているか?
- ⋮

自分の専攻分野に置きかえてみると?

ストリップチャート

データを図解して直観的につかもう！



横軸:身長 (cm)

ドット 1 個=データ 1 個. ぴったり重なったら上に積み重ねていく.

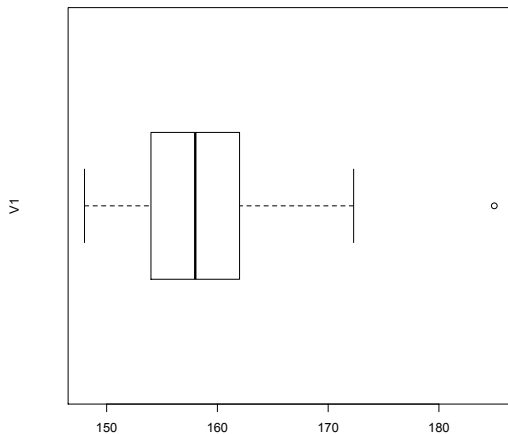
ここまで来たよ

① はじめに

② データの分布

- データとは?
- **箱ひげ図**
- 度数分布表
- ヒストグラム
- ヒストグラムと箱ひげ図の対応

箱ひげ図 (Box Plot)



横軸:身長 (cm)

箱ひげ図の描き方

まず四分位数 (quartile) Q_2, Q_1, Q_3 を求める

Q_1 第1四分位数 下から数えて全体の $1/4$ のデータ

Q_2 第2四分位数 下から数えて全体の $2/4$ のデータ = 中央値

Q_3 第3四分位数 下から数えて全体の $3/4$ のデータ

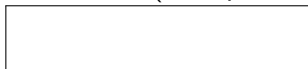
- ① データを小さい順に並べる.
- ② Q_2 を求める.
 - ▶ データが奇数個なら真ん中の値. データが偶数個な

ら

- ③ Q_1 を求める. Q_2 より小さいデータの中央値
- ④ Q_3 を求める. Q_2 より大きいデータの中央値

- 四分位範囲 (interquartile range) 差 $Q3 - Q1$

-



- ▶ $Q1$ から下に, 四分位範囲の 1.5 倍以上離れたデータ
- ▶ $Q3$ から上に, 四分位範囲の 1.5 倍以上離れたデータ

手順

- $Q1, Q2, Q3$ を求める
- $Q2$ に縦線をいれる
- $Q1$ から $Q3$ まで箱を描く
- 外れ値を除いた最大値, 最小値までひげを描く
- 外れ値を \circ で描く

Q1

Quiz(ヒストグラムと箱ひげ図を描こう)

次のデータから作ろう.

- ① 箱ひげ図
- ② 度数分布表
- ③ ヒストグラム

14 14 15 16 18 18 18 25

ここまで来たよ

① はじめに

② データの分布

- データとは?
- 箱ひげ図
- 度数分布表
- ヒストグラム
- ヒストグラムと箱ひげ図の対応

度数分布表の作り方

n = データの個数

階級 = 一定間隔で区切った区間

階級値 = その階級のまん中の値

- 階級の個数 $= 1 + \log_2 n = 1 + 3.3 \times$ (データの個数の桁数) くらいにとる.
- 最大値と最小値の差を, この個数くらいにわけるとよい. きりのよい階級幅 (1 とか 5 とか 10 とか) に調節してよい
- 度数 = その範囲に入ってるデータの個数
- 相対度数 = 度数 / データ全体の個数 (%で書くことも)

階級	度数	相対度数
145 より大きく 150 以下	7	0.09
150 より大きく 155 以下	17	0.22
155 より大きく 160 以下	29	0.37
160 より大きく 165 以下	19	0.24
165 より大きく 170 以下	4	0.05
170 より大きく 175 以下	1	0.01

- 見にくかったら外れ値は除いてもいい
- 階級の幅は一定で

-

- ▶ 以下, 以上, 未満=より小さい, より大きい

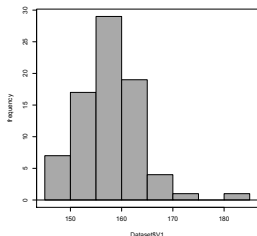
ここまで来たよ

① はじめに

② データの分布

- データとは?
- 箱ひげ図
- 度数分布表
- **ヒストグラム**
- ヒストグラムと箱ひげ図の対応

ヒストグラム



- ‘度数分布表を棒グラフにしたもの’
- 階級の個数:見やすければそれが正義
 - ▶ 階級の幅=超大きい ⇨ 長方形 1 個
 - ▶ 階級の幅=超小さい ⇨
- 階級の取り方で印象はずいぶん変わっちゃう…
- ヒストグラムのほうが箱ひげ図より多くの情報を持っていることが多い

ここまで来たよ

① はじめに

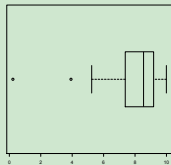
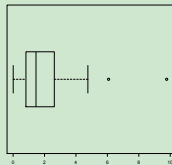
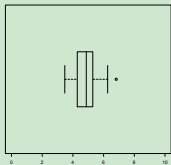
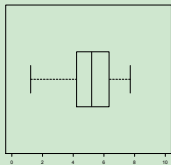
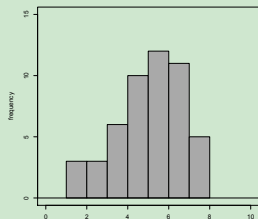
② データの分布

- データとは?
- 箱ひげ図
- 度数分布表
- ヒストグラム
- ヒストグラムと箱ひげ図の対応

Q2

Quiz(ヒストグラムと箱ひげ図)

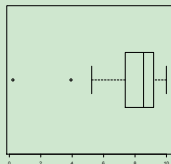
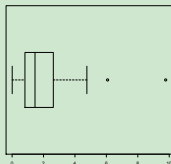
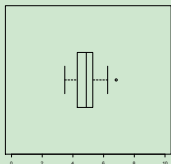
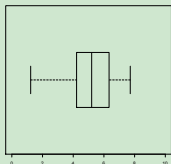
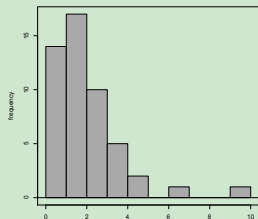
このヒストグラムに対応する箱ひげ図はどれ?



Q3

Quiz(ヒストグラムと箱ひげ図)

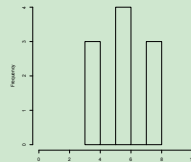
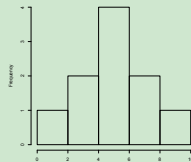
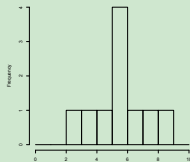
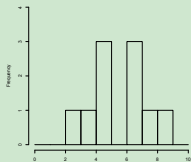
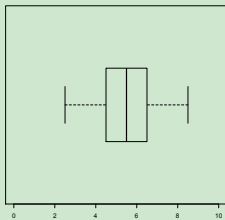
このヒストグラムに対応する箱ひげ図はどれ?



Q4

Quiz(ヒストグラムと箱ひげ図の対応)

この箱ひげ図に対応するヒストグラムはどれ?



Q5

Quiz(ヒストグラムと箱ひげ図)

次のデータから作ろう

- ① 箱ひげ図
- ② 度数分布表 (30 以上 40 未満, みたいな感じで)
- ③ ヒストグラム

30 50 55 55 60 70 70 70 75 100

連絡

- 来週 2013-10-02 水 3 は臨時教室変更. 3-B105 計算機実習室. Windows のパスワード思い出しておいてね.
- 今週は授業内で紙を 1 枚提出
- 今回は必須の '宿題' はありません.
- 来週は平均値 (と分散?) やります. 余裕ある人は e ラーニングシステム <https://moodle.media.ryukoku.ac.jp/> > リメディアル統計学 > 平均 で予習しておこう.
- 来週からは, 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.
- 学期初めなので急な教室変更があるかも. 掲示やメールに注意.