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

確率統計☆演習 I　L01(2014-09-19 Fri)

今日の目標

- データから 手で/Excel で 度数分布表が作れる
- データから 手で/Excel で ヒストグラムが描ける

http://hig3.net

ここまで来たよ

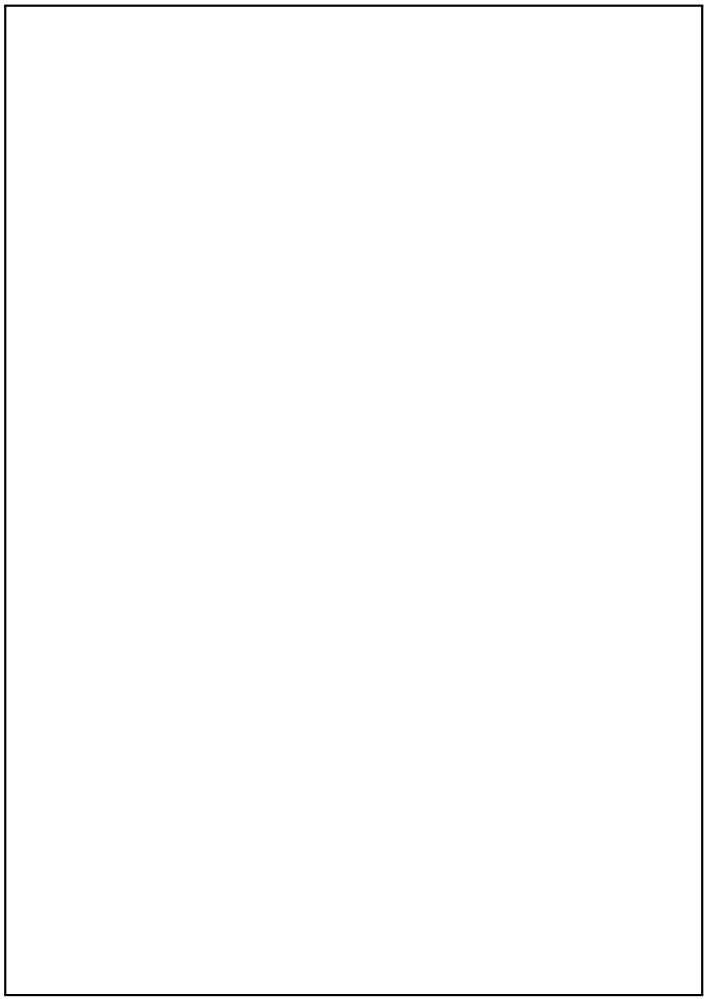
学習目標

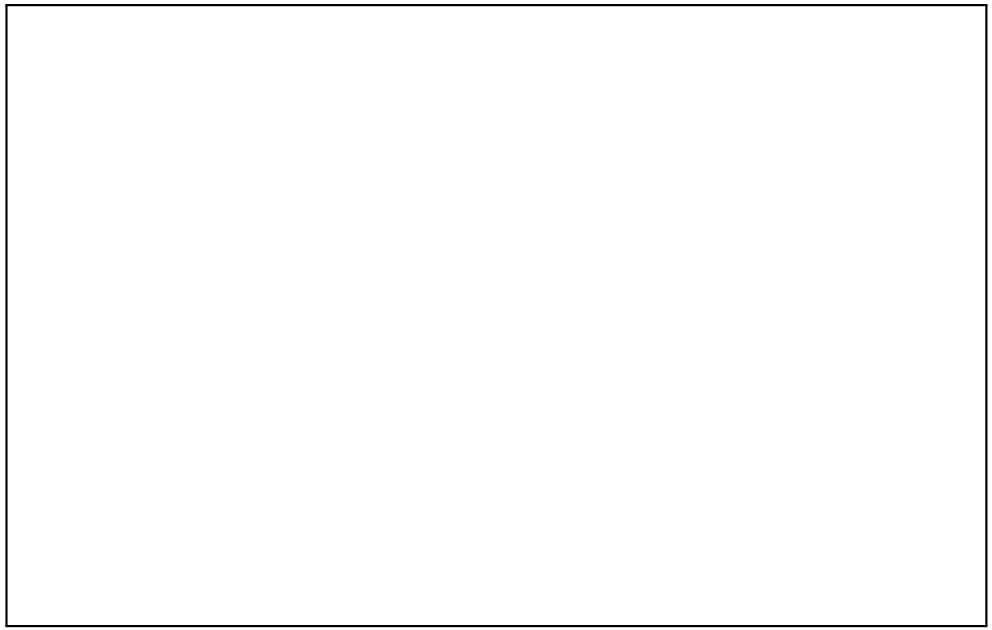
講義概要 → シラバス
現実世界の現象を理解し, 数理モデルとの関係を明らかにするためには, 観察・実験により取得した現象のデータを整理・解析することが必要です. データを表現する記述統計, 限られたデータから現象の性質を推測する推測統計を学びます. ただし, 量的 1 変量の場合を主に扱います. これに必要な範 囲で確率論を学びます. 数式を用いた解析, ソフトウェアによる解析の両方に習熟します.
到達目標 → シラバス
実験・観察により取得した (質的, 量的, 1 変量, 2 変量) データを統計的に整理して, 他者に対して表現できる. データから仮説を立てて検証し, 他者を説得 できる.

確率統計☆演習 I を履修してはいけない理由

次のどれも響かない人は履修しないことを奨めます.

- 数学の教員免許に必要
- コア M
- (3 年前期) 確率統計☆演習 II, 計算科学☆演習 II の前提科目
- 中高の数学で統計はすでに強化されてる
- 教育の評価に統計は必要
- いま, 統計学が熱い! ← CPU パワー, インターネット上でのデータ集積
- いま, ビッグデータ, 人工知能 (AI), 機械学習 (machine learning) が熱い!! ← CPU パワー, インターネット上でのデータ集積
- 統計は科学技術の言葉 ⇝ 数理卒は当然期待されてる
- 確率統計を使ってる数理の教員: 松木平 (確率セルオートマトン), 馬, 佐野, 高橋 (性能評価), 飯田 (物理シミュレーション), 樋口 (確率過程, 教育評価), 他にもいるかも
- 統計検定 2 級





こんなことに答えます

1. 高校の数学で, こういう教え方導入したら, ちょっとだけ平均点が上がった. これ効果あったって言っていいの? (Evidence-based teaching)
2. YouTube から猫の動画を見つけるアルゴリズム, こう改良して, 100 個の入力画像で試したら, 判定精度がちょっとあがった. これで結論していいの? 10000 個でやり直すべき?

確率統計☆演習 I ののり

成績計算難しくないけどとにかく注文の多い科目です…
科目の成績 100 ピーナッツは

- 30 ピーナッツ:毎回授業での quiz, 授業時間外の予習復習, 授業時間内の活動など
- 30 ピーナッツ:プチテスト (11 月)
- 40 ピーナッツ:ファイナルトライアル (定期試験期間)
- その他追加ピーナッツ. その時に説明.

その時点のピーナッツにかかわらず, ファイナルトライアルに参加しないと合格にはなりません. ファイナルトライアル時点で 20 ピーナッツ未満の人も, (平均点を上げるために) 参加をすすめますが, 追試験はなし.
欠席届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.

担当者ののり

- なまえ: 樋口さぶろお hig-compsci2@math.ryukoku.ac.jp
- へや: 1-502
- オフィスアワー: 木 6(1-502/1-539), 金昼 (この教室/1-502). 訪問歓迎な時間: 月火昼. お弁当持参歓迎. お湯あげます.
- Web ページ: http://hig3.net 演習の指示や, スケジュールもここから.

1 週間のタイムライン

模索中です. 金 2 が講義メインじゃないとき (ex. 初回) はちょっと違うかも

1. 水 09:20 まで RaMMoodle で予習問題 (=Quiz 予想問題) に解答
2. 水 09:20 ごろ 予習問題解答公開
3. 金 2 Quiz(=テスト) 参照不可 相談不可
4. いつ採点返却?
5. 金 2 来週の Quiz の予告
6. 金 15:30 ごろ 予習問題公開

(自前の)e ラーニングサイトを使ってみよう
`http://hig3.net` → RaMMoodle (全学認証) → 確率統計☆演習 I
(ブックマークすると楽)

ここまで来たよ

- データが与えられたとき
- 指示された階級で または 自分で階級を決めて
- 度数分布表とヒストグラムが作れる

ここまで来たよ

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) の身長のデータ.

148cm
148.5cm
149cm
⋮
185cm

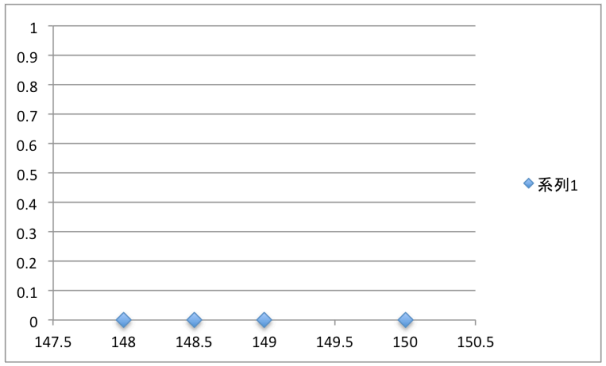
ps3id_raicho_1182 さん（最終更新日時:2012/3/20）投稿日：
2012/2/15 AKB48　身長　まとめ（研究生は 12.5 期まで）
http://note.chiebukuro.yahoo.co.jp/detail/n32745

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?
- ⋮

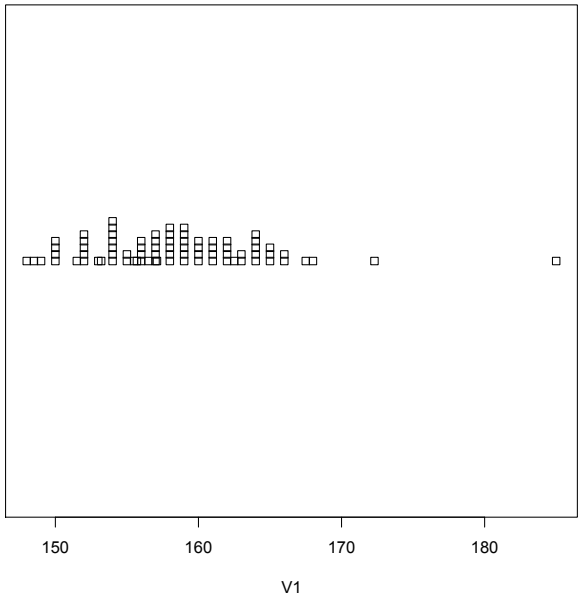
1 次元散布図

(説明のためのものです. あまり実用されません)
実軸上に, データに対応する点をマークする.



もっとデータが多くなったらどうするの?

→ ストリップチャート



→ ヒストグラム

ここまで来たよ

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, 下品な?言葉 'bin' ビン
階級幅=区間の幅
階級値=その階級のまん中の値

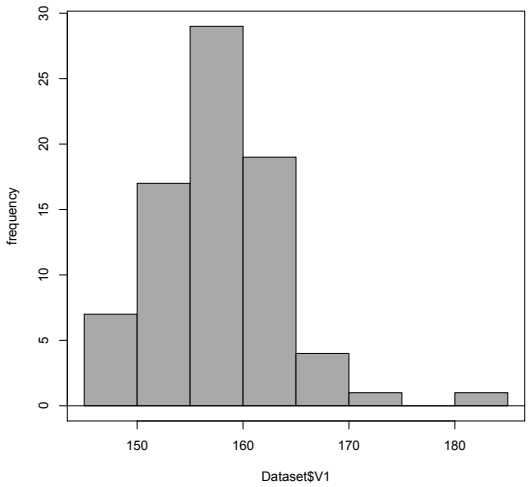
- スタージスの公式:
  $$\boxed{\text{階級の個数}=1+\log_2 n = 1+3.3\log_{10} n}$$
  ぐらいが適切.
- 最大値と最小値の差 を, この個数くらいにわける. きりのよい階級幅 (1 とか 5 とか 10 とか) に調節してよい
- 度数=その範囲に入ってるデータの個数
- 相対度数=度数/データ全体の個数 (%で書くことも)

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
175 より大きく 180 以下	0	0.00
180 より大きく 185 以下	1	0.01
185 より大きく 190 以下	0	0.00
合計	78	1.00

- 見にくかったら外れ値[　　]は除いてもいい
- 階級幅は一定で
- [　　]
  - 以下, 以上, 未満=より小さい, より大きい

ここまで来たよ

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…
- 必ず階級幅は一定
-

手と Excel でやってみよう.

さいしょの練習用データ

兒玉遥	18	島崎遥香	21
山本彩	21	渡辺麻友	21
中野郁海	14	指原莉乃	22
大島涼花	16	横山由依	22
川本紗矢	16	松井玲奈	23
加藤玲奈	17	柏木由紀	23
宮脇咲良	17	宮澤佐江	24
小嶋真子	17	小嶋陽菜	26
白間美瑠	17	大和田南那	15
高橋朱里	17	向井地美音	17
渋谷凪咲	18	森保まどか	17
田野優花	18	松井珠理奈	18
矢倉楓子	18	木崎ゆりあ	19
入山杏奈	19	渡辺美優紀	21
生駒里奈	19	峯岸みなみ	22
川栄李奈	20	須田亜香里	23
武藤十夢	20	高橋みなみ	23

連絡

- 次回は 7-002 講義室
- 配布資料は 1-503 向かいの引出, `http://hig3.net` で再配布しています.
- 次回からは, 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.
- 最初のころはいろいろ変更あるかも. メールに注意.
- 週のタイムラインで見たように, 予習問題を RaMMoodle に金 15:30 までに公開. 翌週水 09:20 までにやってね. それまで何回でも「受験」できます. 最後の受験が点数になります.