

データの分析

中西 渉

watayan@meigaku.ac.jp
名古屋高等学校

2022.08.16

2022 年度高等学校情報科教員研修

(このスライドのデータは <https://watayan.net/doc/> にあります)

- ① はじめに
- ② グラフ
- ③ 量的データ
- ④ 質的データ

お願い

説明で用いるソフトウェアには偏りがあります
他の環境に「翻訳」しつつ読み取ってください

- ① はじめに
- ② グラフ
- ③ 量的データ
- ④ 質的データ

はじめに

尺度

- 名義尺度…等しいか等しくないか
例：都道府県，血液型
- 順序尺度…大小に意味がある
例：評定，「満足」「やや満足」「やや不満足」「不満足」
- 間隔尺度…差に意味がある
例：気温，偏差値
- 比例尺度…0に（絶対的な）意味がある
例：絶対温度，個数

意味がある計算

	大小	差・平均	比
名義尺度	×	×	×
順序尺度	○	×	×
間隔尺度	○	○	×
比例尺度	○	○	○

尺度で気になること

- 「満足」～「不満足」の平均って？
- 評定平均値って？

データの入手

- オープンデータ
- Web API
- スクレイピング

オープンデータで気になること

- データを読むのは誰?

[illegible]

- ① はじめに
- ② グラフ
- ③ 量的データ
- ④ 質的データ

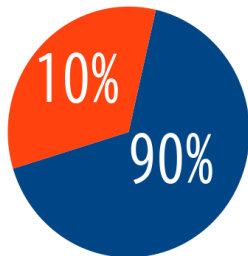
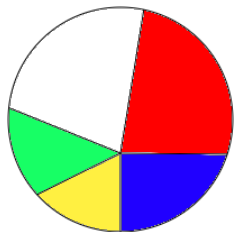
グラフ

- 折れ線グラフ
- 棒グラフ
- ヒストグラム
- 箱ひげ図
- 円グラフ
- 帯グラフ
- 散布図
- ...

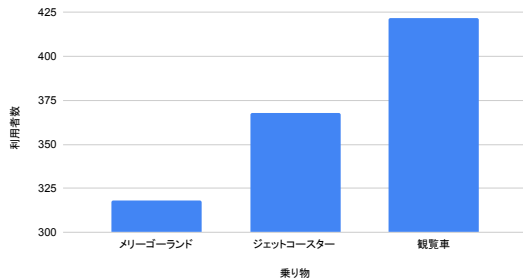
適切なグラフの選択 正しい表示方法の選択

ひどい グラフ

検索



利用者数 と 乗り物



円グラフに関して (R の help から引用)



Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

- ① はじめに
- ② グラフ
- ③ 量的データ
- ④ 質的データ

量的データ

さまざまな統計量

- 平均値
- 中央値・四分位数
- 分散・標準偏差
- 共分散・相関係数
- ...

四分位数のいろいろ

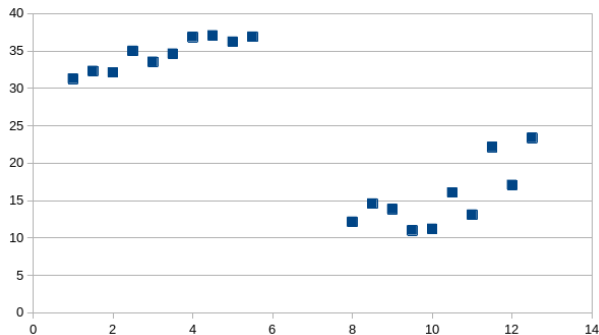
- Tukey の Hinge
- 中高数学の教科書にでてくるもの
- 表計算ソフトの QUARTILE.EXC, QUARTILE.INC

例：第一四分位数の違い

	1～9	1～10
Hinge	3	3
教科書	2.5	3
QUARTILE.EXC	2.5	2.75
QUARTILE.INC	3	3.25

値を盲信しない

例：下の散布図の相関係数は -0.79



検定の基本的な考え方

- ① 中立的な仮定（帰無仮説）を立てる
- ② 帰無仮説のもとで「実際に起こった事象＋それより珍しい事象」の確率を計算する
- ③ それが一定の水準（有意水準・危険率）より小さいかどうか
→帰無仮説を棄却するか棄却しないかを判断する

検定で気になること：

- 「主張したいことの逆を帰無仮説とする」？
- 「帰無仮説を採択」？
- 「 $p = 0.05$ とは，帰無仮説が正しい確率が 95%」？

よく使われる検定

- 二項検定
- t 検定
- χ^2 検定

信頼区間

95%信頼区間… p 値 ≥ 0.05 に相当する区間

※「真の値が95%の確率で含まれる区間」ではない

※※両側 p 値なのか，片側 p 値の2倍なのか…

- ① はじめに
- ② グラフ
- ③ 量的データ
- ④ 質的データ

質的データ

テキストの場合…形態素解析してから

- ワードクラウド
- 出現頻度グラフ
- 共起ネットワーク
- Word2Vec
- ...

たとえばワードクラウド



こころ



山月記

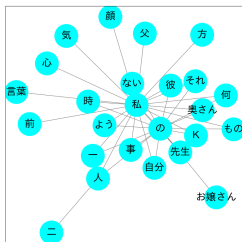


檸檬

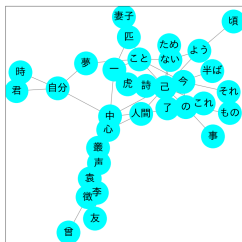
wordcloud harmful

検索

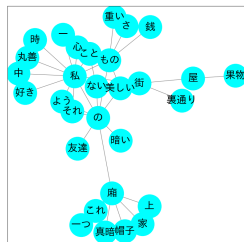
たとえば共起ネットワーク



こころ



山月記



檸檬

たとえば Word2Vec...