初心者セッション：統計学（再）入門

BeginneR Session: Statistics 101 (Re)

22nd April 2023, Tokyo.R #105

Yuta Kanzawa @yutakanzawa

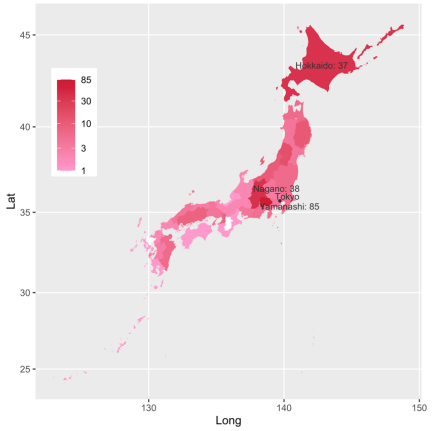
Data Scientist at Zurich Insurance Company Limited, Japan Branch

神沢雄大 Yuta Kanzawa

- データサイエンティスト@チューリッヒ保険会社
  - 日本支店
- Twitter: @yutakanzawa
- 好きなもの：オペラとワイン
  - ワーグナー
  - ブルゴーニュ (WSET Lv 3→?)
- 使用可能言語：7
  - 人間：日本語、英語、ドイツ語
  - コンピューター：R, Python, SAS, SQL

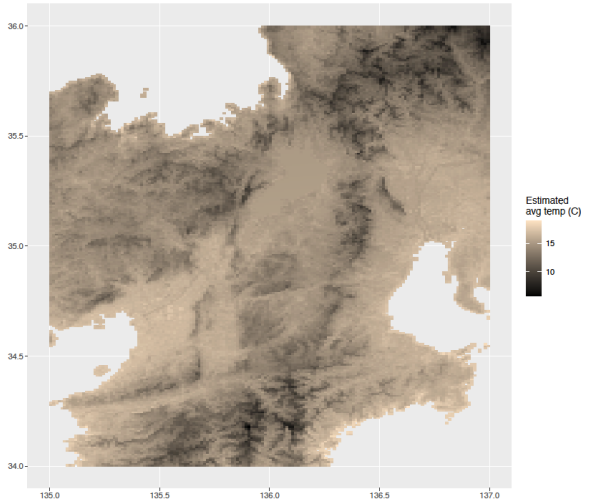
ポートフォリオ

Number of Wineries in Japan in 2019, by Prefecture

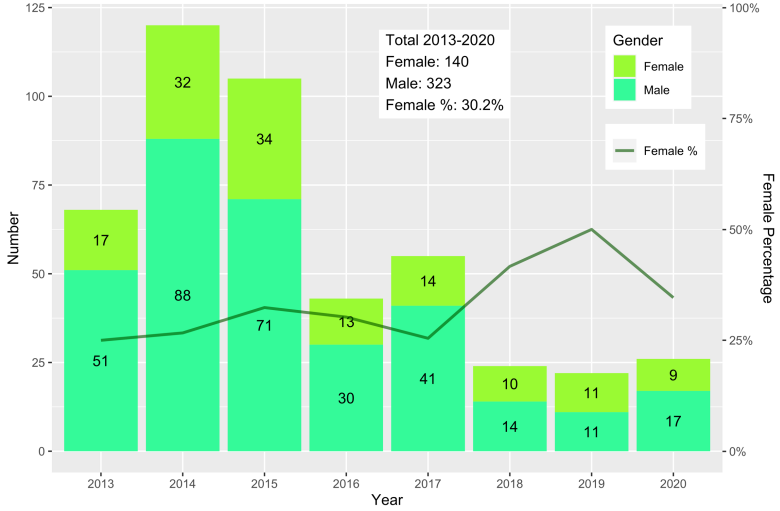


Source: https://www.nta.go.jp/taxes/sake/shiori-gaikyo/seizogaikyo/kajitsu/pdf/h30/30wine01.pdf

Estimated Avg Temperature around Kyoto

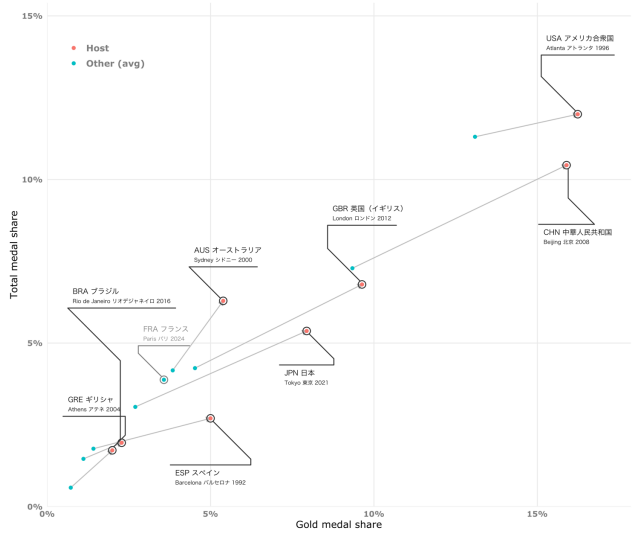


Number of Qualified JSA Sommelier Excellence and Equivalents* by Year and Gender, 2013-2020



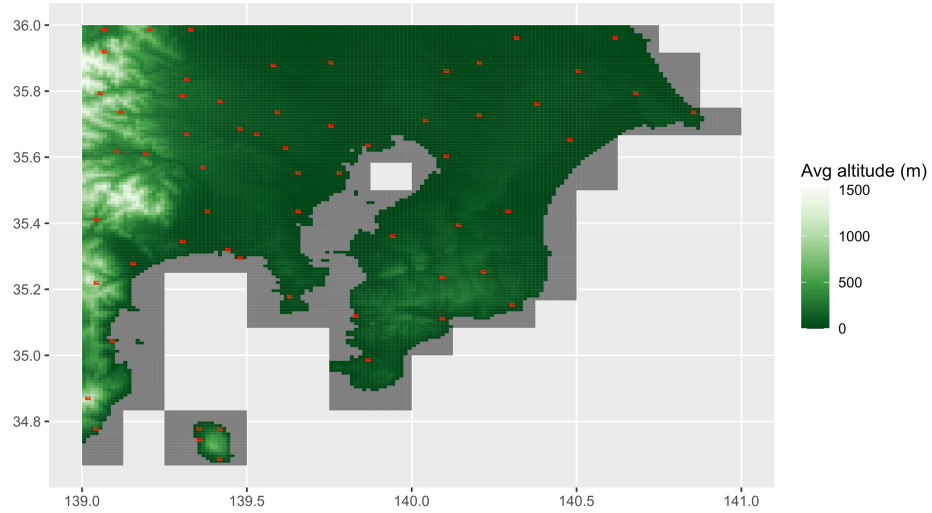
Source: Japan Sommelier Association https://www.sommelier.jp/exam/pdf/qualifiedholders.pdf
*Sommelier Excellence (2019-2020), Senior Sommelier (2013-2018), Senior Wine Adviser (2013-2015)

Medal shares of Olympic host countries in the past 30 years

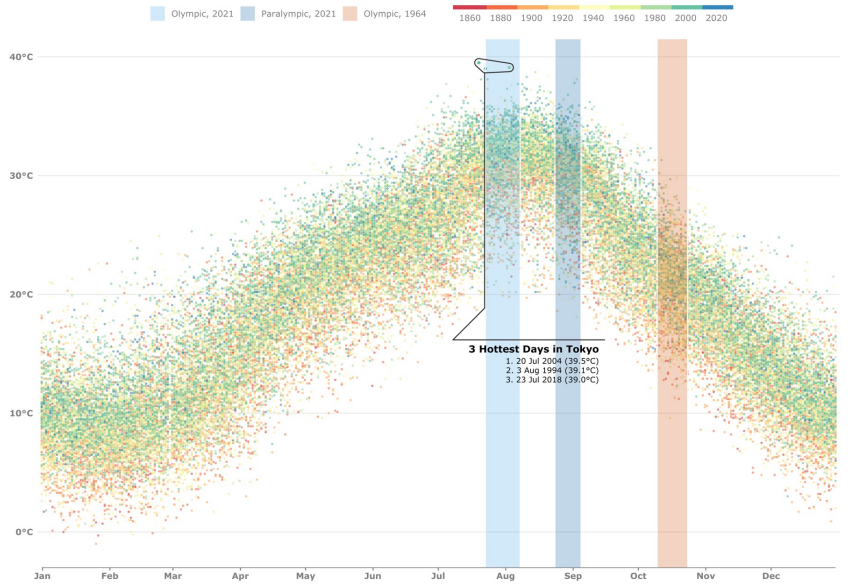


Data: International Olympic Committee via https://olympics.com & https://www.wikipedia.org · Graphic: Yuta Kanzawa

Avg Altitudes and Weather Observation Stations in Tokyo, Kanagawa, Chiba



Daily maximum temperature in Tokyo, 1875-2021



Data: Japan Meteorological Agency via https://www.jma.go.jp · Graphic: Yuta Kanzawa (inspired by Cédric Scherer)

アジェンダ

- 今日話すこと
  - 統計学の基礎知識
  - R
  - 気をつけるポイント
- 対象（以下のいずれか）
  - 統計学を使ってみたい人
  - 統計学を復習したい人

- 今日話さないこと
  - 統計解析の方法

おことわり

- 大まかな説明をします。厳密な定義は書籍を参照して下さい。

- 企業でビジネスやサービスに関するデータ分析を行うという文脈を前提において下さい。アカデミアやR&Dの分析には当てはまらない可能性があります。

- 自分で考えて試して下さい。

TL;DR

- 母集団と標本
  - 標本
  - サンプルサイズの決め方
  - バイアス
- 統計量
  - 相関≠因果
  - 散布図も描く！
- 仮説検定
  - 多重比較、p値ハックに気をつける！

Ch 1:
母集団と標本

Population & sample

「母集団」

- 分析の対象となる集団全体
  - (parent) population
  - Universe
  - 例：
    - 日本国民 → 特定可能
    - あるサービスの全契約者 → 特定可能
    - ある商品の購入者 → 特定可能/不特定多数
    - Rユーザー → 不特定多数

「標本」

- 母集団の一部（部分集合）
  - Sample
  - 母集団全体の把握、調査が困難な場合 → 標本を調査。
    - 母集団の特性を反映するよう無作為抽出（することが多い）。
  - 例：
    - 日本国民 → 内閣支持率の電話アンケート（Cf 国勢調査）
    - あるサービスの全契約者 → 顧客満足度調査
- 標本の要素数 →「標本サイズ」、「サンプルサイズ」
  - 俗に「n数」
  - 注：「標本数」、「サンプル数」→ 集合の数

サンプルサイズ

- どのくらいが適切なのか？
  - SurveyMonkeyの「標本サイズカルキュレータ」*
  - 山根の公式（信頼度95%）
    - $\frac{N}{1+Ne^2}$ （Nは母集団のサイズ、eは許容誤差）
  - 厳密には効果量と検出力から計算。
- 例
  - 日本の人口＝約1億2000万人 → サンプルサイズ＝約400人
    - 許容誤差5%

* https://jp.surveymonkey.com/mp/sample-size-calculator/

バイアス

- 標本の抽出に**偏り** → 母集団と異なる傾向
- 例（アンケート調査）
  - Rユーザーだけに聞く。
  - インターネットで調査する。

Ch 2:
統計量

Statistics

基本統計量

種類	英語	説明	Rの関数
平均値	mean, average	値の大小を均一にしたもの	mean()
中央値	median	値をソートした時の中央	median()
分散	variance	値のばらつき	var()
標準偏差	standard deviation	分散の平方根	sd()
最大値	maximum	最も大きい値	max()
最小値	minimum	最も小さい値	min()
最頻値	mode	同じ値の数が最も多いもの	table()

標本と不偏

- 不偏推定量：母集団の統計量を標本から推定したもの
  - 標本に**偏りがなければ**標本の統計量を母集団の統計量と見なせる。
    - 厳密には期待値が一致することが条件。
- 標準偏差と分散
  - 標本の分散の期待値は母集団とは異なる。
    →標本から計算した標準偏差（分散の平方根）≠母集団
    - **標本**標準偏差：$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ vs **不偏**標準偏差：$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
      - 注1：分母を要チェック！n-1で割る方（右）を「標本」と呼ぶことも。**sd()関数はn-1**で割る。
      - 注2：このn-1で割る「不偏標準偏差」は母集団の標準偏差の不偏推定量ではない。

相関係数

- 2つの変数の値の変動の一致度と方向を表す**順序**尺度。
  - **ピアソン**の積率相関係数
  - その他
    - スピアマンの順位相関係数
    - ケンドールの順位相関係数
    - 級内相関係数(ICC)、など。
  - 値の**比に意味はない**。
    - 誤り：「1.0は0.1の10倍相関が強い。」
  - 相関関係の**有無を決める閾値は存在しない**（自分で決める）。
- cor()関数
  - 引数methodに指定（デフォルトは"pearson"）。

相関≠因果

- 相関関係があっても**因果関係があるとは限らない**！
  - Correlation doesn’t imply causation!
  - 例：チョコレート消費量とノーベル賞受賞者数*
    - 擬似相関
    - 歴史的経緯？
    - 1人当たりGDP？

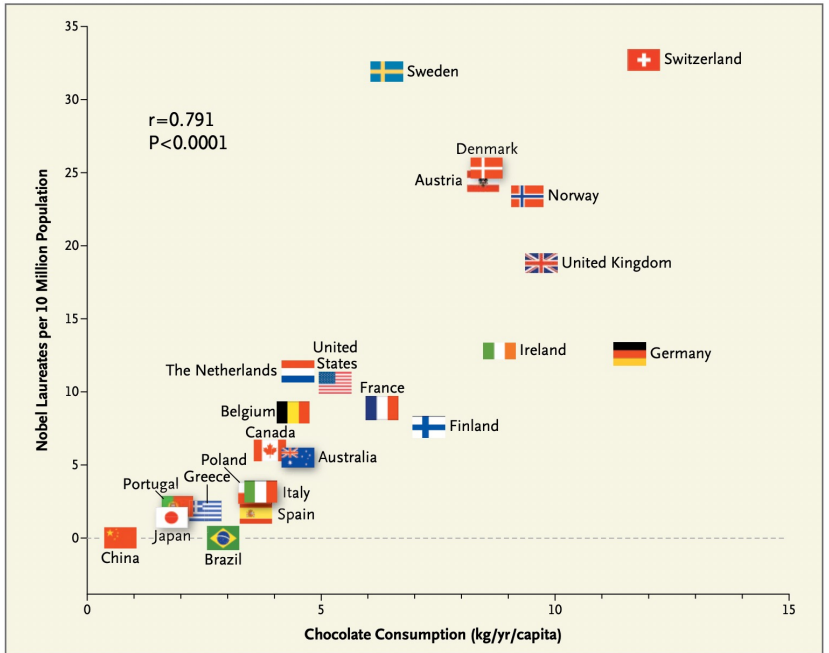
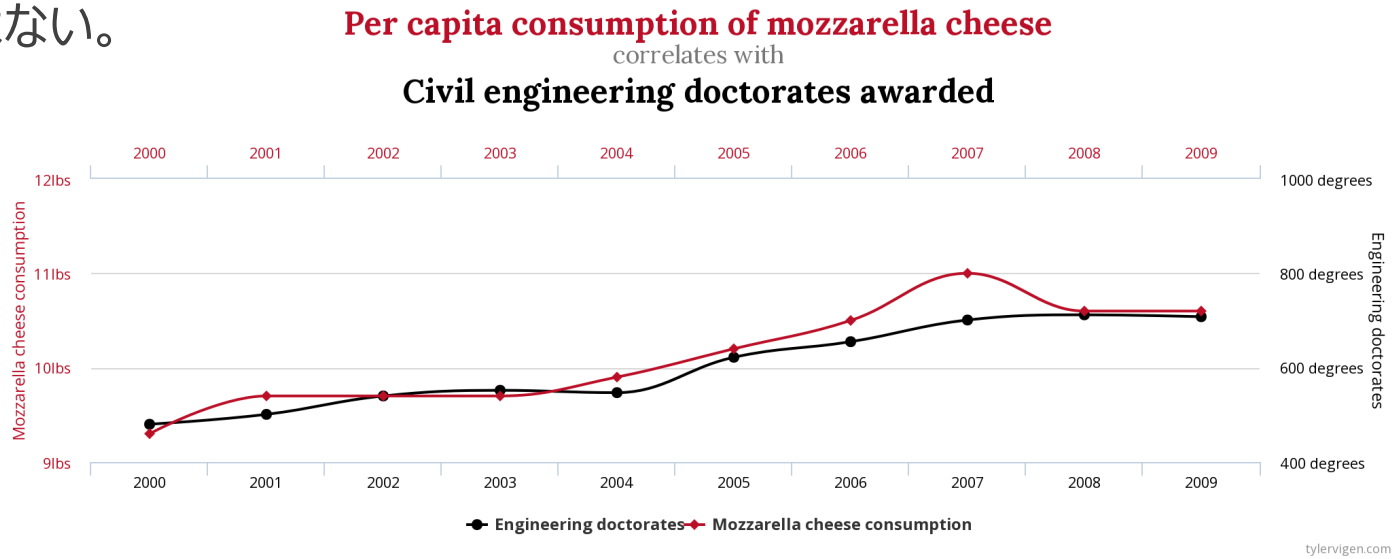


Figure 1. Correlation between Countries’ Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

* https://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laurates%20(NEJM).pdf

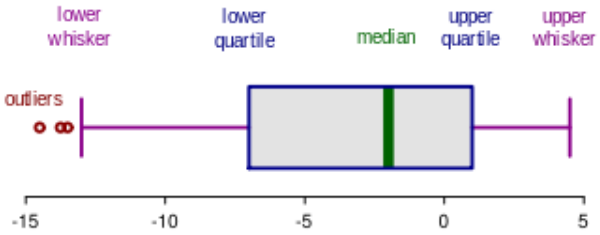
見せかけの回帰

- 相関関係も因果関係もないのに、そうであると誤認してしまう。
  - 時系列データの回帰
    - 単位根過程（ランダムウォーク）
    - 例：モッツァレラチーズの消費量と土木工学博士号の取得者数*
      - 厳密には見せかけの相関ではない。
      - 経済状況？



* https://www.tylervigen.com/spurious-correlations#stat012ac8597b4e674c25da6937ec9f649f

外れ値と分位点



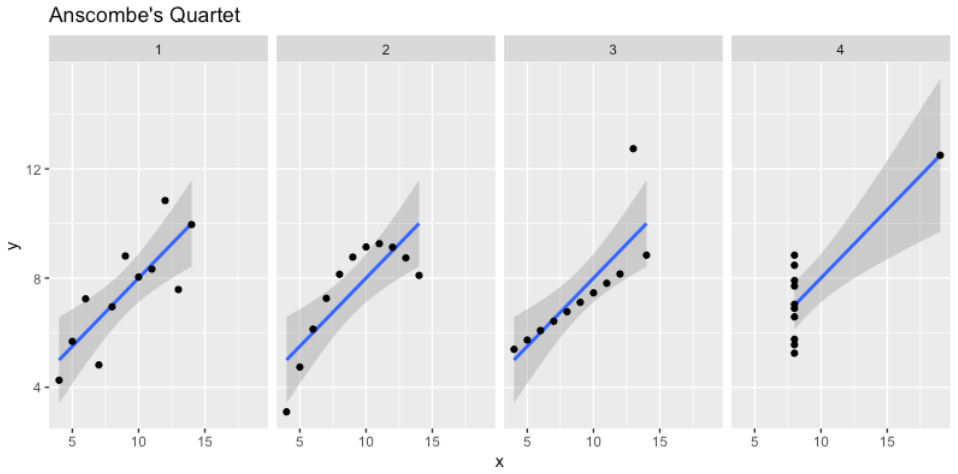
- 外れ値：データ全体の中で極端に大きいor小さい値
  - 定量的には**分位点**を計算して判断する。
    - 分位点：データを昇順ソートした時の位置（順位）を百分率で表したもの
      - percentile
      - 0%点 = 最小値、50%点 = 中央値（第2四分位点）、100%点 = 最大値
    - quantile()関数（引数probsに小数表記で渡す。）
  - 箱ひげ図：ggplot2のgeom_boxplot()関数
- 外れ値か否かを決める**閾値は存在しない**（自分で決める）。
  - 経験的には上下1%か5%とすることが多い（正規分布なら2σや3σ）。
  - 外れ値は上限値や下限値で**置き換える**（か除外する）。

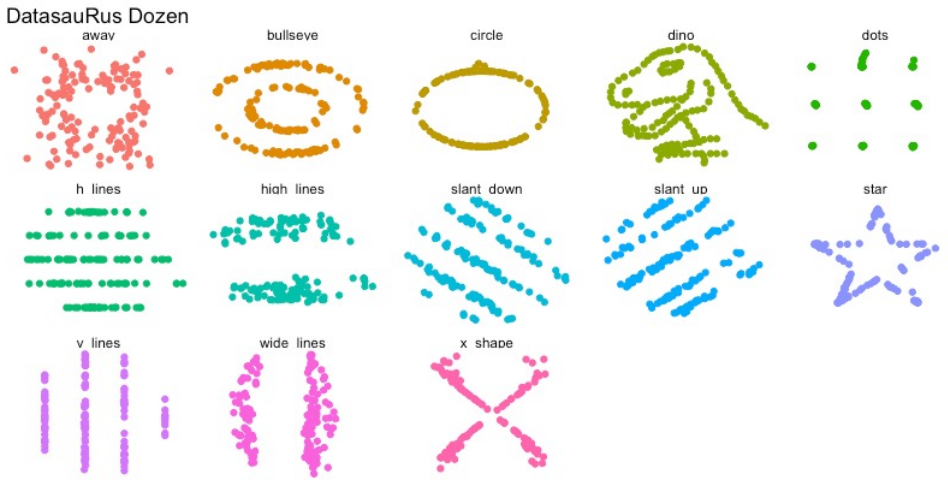
* https://commons.wikimedia.org/wiki/File:Elements_of_a_boxplot_en.svg

アンスコムの数値例

- 平均、分散、相関係数、回帰直線が同じになってしまう4つのデータ
  - 内蔵のanscombeデータセット（発展形：datasauRusパッケージ*）
  - 外れ値の影響が顕著。
- 定量的なことをするだけでなく**散布図も描きましょう！**

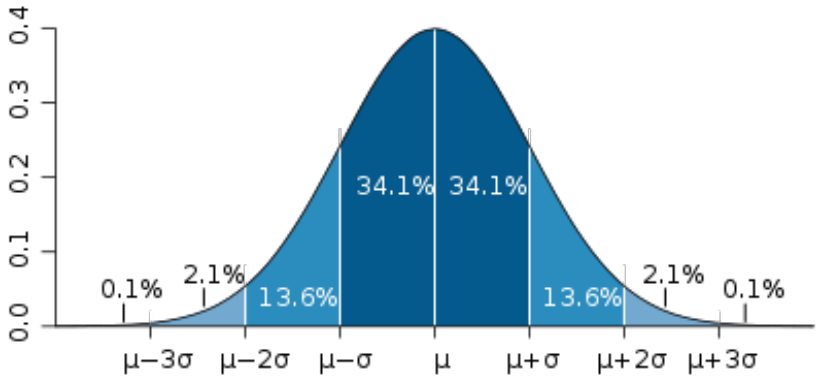
```
# アンスコムの数値例を描くコード
library(tidyverse)
library(ggplot2)
anscombe %>%
  rowid_to_column("id") %>%
  pivot_longer(!id) %>%
  separate(name, c("axis", "sample"), sep = 1) %>%
  pivot_wider(id_cols = c(id, sample), names_from = axis) %>%
ggplot(aes(x = x, y = y)) +
  geom_smooth(method = "lm") +
  geom_point() +
  facet_grid(cols = vars(sample)) +
  labs(title = "Anscombe's Quartet")
```





* https://cran.r-project.org/web/packages/datasauRus/

正規分布



- 正規分布
  - データがこの確率分布に従うことが多く、便利。
  - 正規分布を前提にしている手法がある。
    - 例：線形回帰（最小2乗法）→誤差（残差）が正規分布に従う。
- でも、自分のデータが正規分布かどうかは全く別の話。
  - 正規分布か否か調べる。→正規性の検定
    - シャピロ-ウィルク検定：shapiro.test()関数
      - 正規分布に従う母集団から抽出された標本か否か。
    - コルモゴロフ-スミルノフ検定：ks.test()関数
      - 2つの母集団の確率分布が異なるか否か。

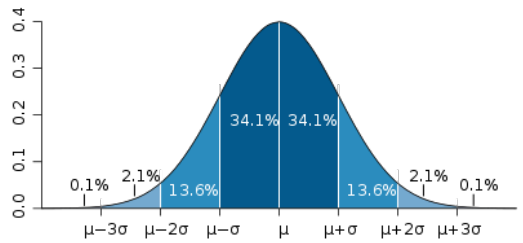
* https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram_micro.svg

Ch 3:
仮説検定

Hypothesis test

検定

- 帰無仮説：検定で棄却されることを前提とする仮説
  - 例：「2つの標本の平均値が等しい。」
  - p値：帰無仮説が真である確率
- 有意水準：帰無仮説を棄却する（採用しない）p値の閾値
  - 間違いである確率
  - 分野、内容に依る（経験的には0.1%～10%）。
- 両側検定と片側検定
  - 「等しい」という帰無仮説 → 両側
  - 「大きくない」、「小さくない」という帰無仮説 → 片側
    - 有意水準→両側検定の半分にする。



色々な検定

調べたいもの	検定	Rの関数
割合	母比率の差の検定	`prop.test()`
平均値	ウェルチのt検定	`t.test()`
度数分布	ピアソンのχ2乗検定	`chisq.test()`
正規性（再掲）	シャピロ-ウィルク検定	`shapiro.test()`
	コルモゴロフ-スミルノフ検定	`ks.test()`

アンチパターン

- 多重比較
  - 有意水準：本来有意でないはずが有意になる確率（間違いの確率）
  - **複数回比較（検定）**を行う。→「間違い」が起こりやすくなる。
  - 対処法：
    - 有意水準を下げる（キツくする）。
    - ダネットの検定、テューキーの範囲検定
- p値ハック
  - サンプルサイズが大きいほど有意になりやすい。
  - **禁止**：~~有意になるまでデータを増やす。~~

まとめ

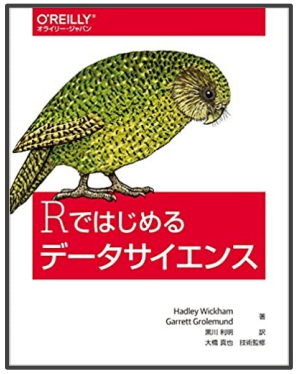
Long story short

Long story short

- 母集団と標本
  - 標本
  - サンプルサイズの決め方
  - バイアス
- 統計量
  - 相関≠因果
  - 散布図も描く！
- 仮説検定
  - 多重比較、p値ハックに気をつける！

参考書

- 『統計学入門』（東京大学教養学部統計学教室、1991年）
- 『戦略的データサイエンス入門』（Provost、Fawcett、2014年）
- 『Rではじめるデータサイエンス』（Wickham, Grolemund、2017年）
- 『ビジネスデータサイエンスの教科書』（Taddy、2020年）



Enjoy!