# 統計学第2/3講

明治大学情報ｺﾐｭﾆｹｰｼｮﾝ学部
後藤　晶
akiragoto@meiji.ac.jp

## Contents



### 今日のお話

# 前回の復習

## 統計学とは？

- データサイエンス：データから有用な情報・知識を引き出したり，新たな価値あるデータを創造するための基本な考え方
  - 平均・分散を算出したり，全体的な傾向を把握するためにヒストグラムを作るのは有用な情報・知識を引き出すため.
  - プログラムを組んで，新たに価値あるものを作る.
- 統計学を学ぶと何がどうなる？
  - 「データ」に基づいた思考方法が身につく
  - 「見せかけの類似性」に騙されなくなる.
  - 日常生活・ビジネスへの応用可能性が広がる

## 「血液型占い」：血液型によって性格が異なる.

- 「データ」に基づくと，血液型によって性格が異なるとはいえない（参考）.

**縄田 健悟** (2014).
血液型と性格の無関連性――日本と米国の大規模社会調査を用いた実証的論拠――
**心理学研究**, 85, 148-156. [ **PDF** ]

**【タイトル】**

**血液型と性格の無関連性――日本と米国の大規模社会調査を用いた実証的論拠――**

**【要約】**

日本の社会では，ABO式血液型と性格に関連があるという俗説が広く信じられている一方で，心理学の実証研究ではその関連性は認められていない。本研究は，血液型と性格が無関連であるより積極的な実証的根拠を提示することを目的としている。そのために，大規模調査の二次分析を行った。大規模な無作為標本抽出で日本とアメリカから合計10000以上 の標本を分析した。日本のデータセットは2004年 (N = 2878–2938)と2005年 (N = 3618–3692) であり，アメリカのデータセットは2004年 (N = 3037–3092)である。分析の結果，3つのデータセット全ての68項目中65項目で血液型間に有意な違いは確認されなかった。また，効果量$\eta^2$は.003以下であり，血液型の効果は全分散の0.3%以下しか説明しなかった。以上の結果は，血液型と性格の無関連であることを示している。

**【もっと簡単な説明】**

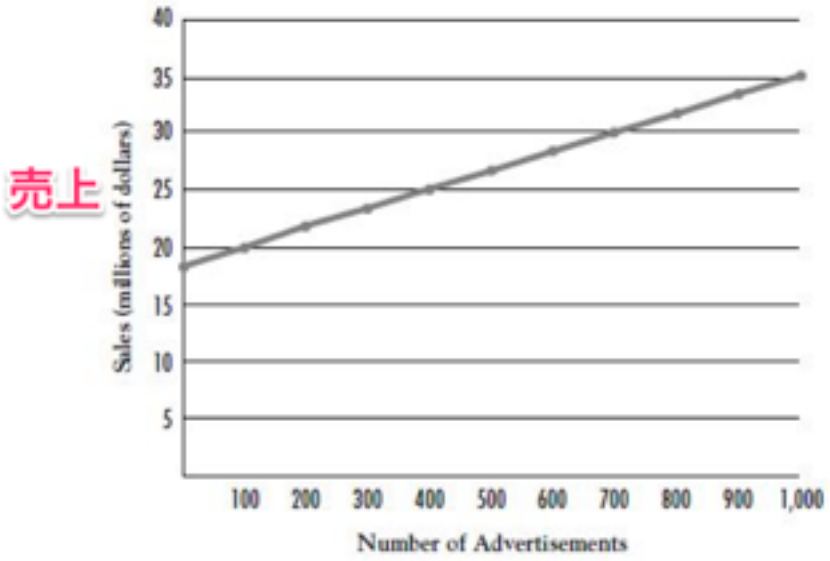
・これまでの心理学研究でも血液型と性格に関連性が見られないことは研究されてきた。
・本研究は血液型と性格の無関連性をより積極的に示すため，
　日本とアメリカの合計１００００人以上の規模のアンケートデータを元に，検討した。
・わずかな差でも検出できる大規模データにもかかわらず，一般的な生活に対する態度に関する**合計６８項目中６５項目**で血液型間に統計的に意味のある差が見られなかった。残り３項目も偶然の範囲。
・質問項目の個人差のうち血液型が説明した割合は，一番大きなもので**たった0.3%**。ほぼゼロだと見なせる。

## 見せかけの類似性：相関関係と因果関係.

- 「データ」はモノを考えるのに大事なことだが，「数字」や「見た目」に騙されてはいけない.
- その他の要因が影響している可能性がある（参考）.

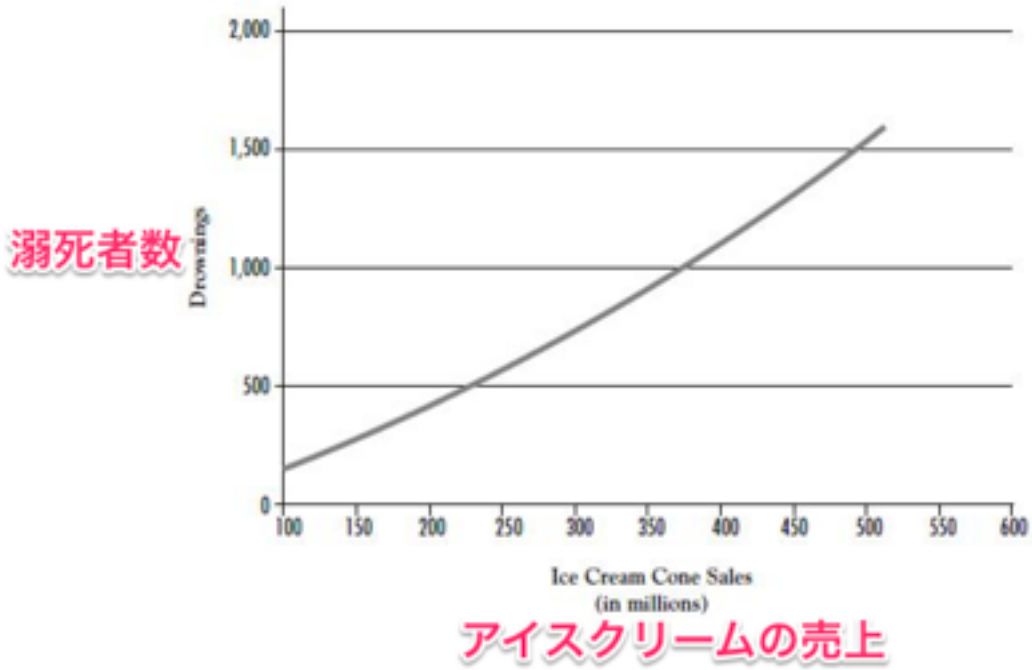
広告数と売上の関係

Relationship Between Advertisements and Sales




アイスクリームの売上と溺死者の関係

Relationship Between Ice Cream Sales and Drownings



## 日常生活・ビジネスへの応用可能性が広がる

- ビジネスにおいて，データを利用し，企業活動を改善・開拓するのに必要な3要素
- データを効率的に収集・処理する
  - 企業活動におけるデータは莫大であり，このようなデータを処理するためには情報技術の利用が必要
- データを適切に取り扱い，妥当かつ汎用的な成果を残す
  - データ全体から，その傾向の妥当性を検討する．
- データをビジネスの枠組みの中にうまく組み込む
  - 効率的に処理をし，有用な結論を導き出してビジネスへの応用をしていく．

## 加減乗除

```
123 + 123  #
```

```
## [1] 246
```

```
123 - 123   #
```

```
## [1] 0
```

```
23 * 123    #
```

```
## [1] 2829
```

```
123 / 123   #
```

```
## [1] 1
```

```
123 ^ 2  #
```

```
## [1] 15129
```

加減乗除の基本はこのような形です．少し演習問題を解いてみましょう．

## 注意です．

- 「#」から始まるとコメントアウトができます．
- 要は，「計算式」ではなく，「ただの文字」として認識されます．

## 演習問題1

- 113 + 987
- 2135 + 231
- 9832 - 3422
- 12348 - 8976
- 17 * 16
- 3298 * 5
- 285195 / 5
- 12387 * 33
- 324 ^ 2
- 89 ^ 4

## その他の計算

```
sqrt(144)#
```

```
## [1] 12
```

```
1234 %/% 123  #
```

```
## [1] 10
```

```
1234 %% 123  #
```

```
## [1] 4
```

### 演習問題2

- (34×2)+(43-12)
- 23×(92-9)
- (53+23)の5乗
- (334-56)÷90
- (34×2)+(43-12)
- (3221+239)÷(87+27)の整数商
- (751×90)÷(5412/32)の剰余

### オブジェクト指向

- 「オブジェクト」：データやモデル式などを入れる「何でも箱」
  - Rではモデル式，データなどをオブジェクトに入れて考える
  - 数式やデータをいちいち書くのは大変．．．
  - オブジェクトに入れることを「代入する」と言う

### オブジェクトに入れて計算する：

```
x <- 5
```

- xというオブジェクトに5を代入する

```
x
```

```
## [1] 5
```

- xの値を出力する

```
(y <- 3)
```

```
## [1] 3
```

- (　)に挟むと，一発で結果も出してくれる．

### オブジェクトに入れて計算する：

```
x + y
```

```
## [1] 8
```

```
x * y
```

```
## [1] 15
```

```
x - y
```

```
## [1] 2
```

### よくないオブジェクト：

```
x <- 5
x <- 16
x
```

```
## [1] 16
```

- これで出力すると，xに16が代入されてしまっている．
- 基本的には違う語を使うようにしたい．

```
# NA <- 3
```

「NA <- 3 でエラー: 代入の左辺が不正 (do_set) です」と怒られる．これは「NA」がデータがないことを示す記号として指定されているため． 他にも指定されている語がいくつかあるが，怒られたら違う文字を割り振れば良い．

### オブジェクトには文字を入れることが可能

```
A <- "Pen"
B <- "Pineapple"
C <- "Apple"
paste(A, B, C, A)
```

```
## [1] "Pen Pineapple Apple Pen"
```

- paste：複数の文字列を結合して，一つの文字列にする関数

## 記述統計量とは

### 平均値・分散・標準偏差とは？

- 平均値：全てのデータを足して割ったもの．一般的に代表値（データ全体を表している数値）として扱われる．
- 分散：平均値とそれぞれの値の差を求めて2乗して，合計したものをデータの個数で割ったもの．データの散らばり具合を示す数値であり，分散が大きければ大きいほど，データが散らばっていることを示す．
  - $\sigma^2$ という記号で表される．
  - $(\ ) = \Sigma\{(\ ) - (\ \ )\}^2/(\ \ )$
- 標準偏差：分散の平方根．通常の長さのばらつきを評価する際には同じ単位で理解したほうがわかりやすいために用いる．
  - $\sigma$ という記号で表される．

### その他，重要な指標

- 最小値：そのデータの中で最も小さい値
- 第一四分位数（25%パーセンタイル値）：最小値と中央値の間の中央値
- 中央値（第二四分位数）：データを大きい(小さい)順に並べたとき，真ん中の値のこと(median)．外れ値がある時に代表値として用いられる．
  - 奇数の場合：ちょうど真ん中が存在する．
  - 偶数の場合：真ん中の数字2つの平均値を中央値とする．
- 最頻値：データの中で最も多く出てくる値のこと(mode)．因子データの際に代表値として使われる．
- 第三四分位数（75%パーセンタイル値）：中央値と最大値の間の中央値
- 最大値：そのデータの中で最も大きい数

- 以下の2つは参考までに．
  - 平均偏差：「平均からの偏差」の絶対値の平均
  - 範囲：最大値から最小値の間．引き算で求められる．

## 平均値の計算

- 7人の学生の体重が50, 60, 85, 70, 80, 67, 66kgであったする．これらの学生の体重の平均値を求めよ．
- 平均値：全てのデータを足して割ったもの．一般的に代表値（データ全体を表している数値）として扱われる．

## 平均値の計算

- 7人の学生の体重が50, 60, 85, 70, 80, 67, 66kgであったする．これらの学生の体重の平均値を求めよ．

```
(50+60+85+70+80+67+66)/7
```

```
## [1] 68.28571
```

- 平均値=(それぞれのデータの値の合計)/(データの個数)

## データセットを作ろう

7人の学生の体重が50, 60, 85, 70, 80, 67, 66kgであったする．このデータを変数名“weight”に代入する．

```
weight<-c(50, 60, 85, 70, 80, 67, 66)
```

**演習問題3：**

同じ7人の学生の身長が155, 164, 182, 165, 177, 177, 172cmであったとする．このデータを変数名“height”に代入せよ．

## Rにおける「関数」とは？

- 関数：頻繁に用いられるデータ操作方法や，標準的な統計計算をまとめてオブジェクトにしたもの．正式には「関数オブジェクト」
  - 簡単に計算できるように，先人たちがまとめたものだと理解すれば良い．
  - これを用いることで，簡単に計算ができる．

## 基本的な関数：

- sum：合計
- mean：平均値
- max：最大値
- min：最小値
- range：範囲（最大値-最小値）
- median：中央値
- var：不偏分散
- sd：標準偏差
- quantile：四分位点
- IQR：四分位範囲
- summary：要約統計量

- sqrt：平方根

### 基本的な関数：

- abs：絶対値
- round：値の丸め
- floor：値の切り捨て
- ceiling：値の切り上げ
- log：自然対数
- log10：10を底とする対数
- log2：2を底とする対数
- log1p：1を加算した自然対数
- exp：指数関数
- sin, cos, tan：三角関数
- asin, acos, atan：三角関数の逆関数

### 記述統計量を色々出してみる.

```
sum(weight)/7
```

```
## [1] 68.28571
```

- sum()関数で合計を算出できる.

```
sum(weight)/length(weight)
```

```
## [1] 68.28571
```

- length()関数でデータの個数を数える.

```
mean(weight)
```

```
## [1] 68.28571
```

- 実はmean()という関数を使うと一発で出てしまう.

```
median(weight)
```

```
## [1] 67
```

- 中央値はmedian()という関数で出せる.

```
table(weight)
```

```
## weight
## 50 60 66 67 70 80 85
##  1  1  1  1  1  1  1
```

- 最頻値はtable()という関数を使って探し出す.
- ちなみに，“weight”の中に最頻値は存在していない.（全てが最頻値=1）

### 演習問題4：

変数名“height”の合計・個数・平均値・中央値・最頻値を求めよ.

### 体重の記述統計量をまとめて算出する.

```
summary(weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.00   63.00   67.00   68.29   75.00   85.00
```

- 左から順番に「最小値，第1四分位数，中央値，平均値，第3四分位数，最大値」を示しています.

### 演習問題5：

変数名"height"の最小値・第1四分位数・中央値・平均値・第3四分位数・最大値を求めよ.

## 分散と標準偏差を手計算で算出してみよう

### 分散を算出する

- 定義通りの算出方法

$$\sigma^2 = \Sigma((\ ) - (\quad))^2/(\ \ )$$

- 簡便に算出する際に用いられる数式

$$\sigma^2 = \Sigma((the\ value\ of data)^2/(a\ number\ of\ data))$$
$$-(\Sigma((the\ value\ of data)/(a number\ of\ data))^2)$$

- 「2乗の平均」-「平均の2乗」
- ここでは，定義通りの数式から分母と分子に分けて話を進めていきましょう.

## 分子を計算する

### 体重の平均値をオブジェクトに入れる

- "mean_weight"というオブジェクトを作って，体重の平均値を入れます.

```
mean_weight <- mean(weight)
mean_weight
```

```
## [1] 68.28571
```

### 平均からの偏差を求めて，オブジェクトに入れる

- (データの値)-(weightの平均値)をして「平均からの偏差」を求めます．結果は"hensa_weight"に代入します.

```
hensa_weight <- weight - mean_weight
hensa_weight
```

```
## [1] -18.285714  -8.285714  16.714286   1.714286  11.714286  -1.285714  -2.285714
```

### 「平均からの偏差」を2乗する

- 「平均からの偏差」を2乗します．“hensa_weight2”というオブジェクトを作って代入をしましょう．
  - 2乗しないとで全部足すと，数字は0になります．
  - ただし，小数点以下を四捨五入しているので，ここでは完璧に0にはなりませんが，限りなく0に近くなります．

```
hensa_weight2 <- hensa_weight^2
hensa_weight2
```

```
## [1] 334.367347  68.653061 279.367347   2.938776 137.224490   1.653061   5.224490
```

### 「平均からの偏差の2乗」を全部足してオブジェクトに入れる

- これらの5つの値を合計した「平均からの偏差の二乗和」を求めます．
  - “sum_hensa_weight2”という名前にしましょう．これで分子は完成です．

```
sum_hensa_weight2<-sum(hensa_weight2)
sum_hensa_weight2
```

```
## [1] 829.4286
```

## 分母を計算する

### データの個数を数えてオブジェクトに入れる

- 今度は分母を算出します．分母はデータ数です，“length_weight”というオブジェクトに代入しましょう．

```
length_weight<-length(weight)
length_weight
```

```
## [1] 7
```

## 最後の計算

### 分散の算出

- 分散は「平均からの偏差の二乗和」/「データ数」ですから，以下の通りに求められます．
  - 分散は“vari_weight”というオブジェクトに入れましょう

```
vari_weight<-sum_hensa_weight2/length_weight
vari_weight
```

```
## [1] 118.4898
```

## 標準偏差を算出する

### 標準偏差を算出する

$$\sigma^2 = \Sigma((the\ value\ of data)^2/(a\ number\ of\ data))$$
$$-(\Sigma((the\ value\ of data)/(a number\ of\ data))^2)$$

$$\sigma = \sqrt{\sigma^2}$$

- 標準偏差は分散の平方根です．
  - 平方根を求める関数は“sqrt()”であり，“hyohen_weight”というオブジェクトに入れてあげます．

## ルートをとります

```
hyohen_weight <- sqrt(vari_weight)
hyohen_weight
```

```
## [1] 10.8853
```

## 分散の別解

$$\sigma^2 = \Sigma((the\ value\ of data)^2/(a\ number\ of\ data))$$
$$-(\Sigma((the\ value\ of data)/(anumber\ of\ data))^2)$$

```
sum(weight^2/length(weight))-sum(mean(weight)^2)
```

```
## [1] 118.4898
```

## 演習問題6：

- “weight”と同様に，変数名“height”について分散・標準偏差を計算してください．
  - ヒント：ちょっと書き換えるだけですぐいけます．
  - “ctrl + f”(macでは“cmd + f”)で置換ができます．
  - 基本方針は「極力手抜きをしましょう」です．

## 関数を使って分散と標準偏差を算出する．

- 分散と標準偏差はよく算出します．当然Rでは関数が用意されています．

```
var(weight)
```

```
## [1] 138.2381
```

- 分散

```
sd(weight)
```

```
## [1] 11.75747
```

- 標準偏差
- 関数で求められるのは「不偏分散・不偏標準偏差」
- 手計算で求めたのは「標本分散・標本標準偏差」

# 2種類の分散と標準偏差

## 2種類の分散と標準偏差？？

「不偏分散・不偏標準偏差」と「標本分散・標本標準偏差」というものが出てきました．この話を理解するためには「母集団」と「標本」という話を理解する必要があります．ここでは簡単に，その2つの違いについてお話したいと思います．

私達が何かのデータを取る時は，全ての物事のデータを集めることが必ずしもできるとは限りません．例えば，「本学大学生1年生全員を対象としたアンケート」を実施すれば全てのデータを集めることができるかもしれませんが，「日本国民全てを対象としたアンケート」を集計するのは非常に困難です．

例えば，大学1年生の意見を調査することを目的として，1年生全員のデータをそのまま用いる分には問題ないのですが，「日本国民全てを対象としたアンケート」を実施するのはコストの面から考えても現実的ではありません．そのために，全体（母集団）の中から一部を取り出して（標本，サンプル），全体の意見・傾向を「推定」するという手法がとられるようになりました．

このような「推定」という手法を取る時に，"データ数"のままで分析するよりも"データ数-1"で計算してあげたほうがよりよい推定ができる，ということで"データ数-1"をするようになりました．

本当はもう少し細かな数学的な議論があるのですが，入り込むと帰って来れなくなるのでここまでにしておこうと思います．とりあえず，これからは「不偏分散・不偏標準偏差」が使われることが多い，とだけ覚えておいて下さい．

興味のある方はコチラをご参照ください．

## 関数と手計算が一致することを確認する．

### 分散

```
var(weight)
```

```
## [1] 138.2381
```

```
huhen_vari_weight<-sum_hensa_weight2/(length_weight-1)
huhen_vari_weight
```

```
## [1] 138.2381
```

## 関数と手計算が一致することを確認する．

### 標準偏差

```
sd(weight)
```

```
## [1] 11.75747
```

```
huhen_hyohen_weight<-sqrt(huhen_vari_weight)
huhen_hyohen_weight
```

```
## [1] 11.75747
```

## 演習問題7：

以下のデータ３つ数値の平均値、偏差平方和、分散，標準偏差を求めてください．なお，分散および標準偏差については不偏分散・不偏標準偏差でかまわない．

- 「2, 3, 3, 4」 なお，これらのデータは「data1」というオブジェクトに入れて算出すること

- 「1, 1, 1, 9」 なお，これらのデータは「data2」というオブジェクトに入れて算出すること
- 「43.6, 45.2, 45.4, 45.8, 47.2, 47.8, 48.2, 48.7, 48.8, 48.9, 49.0, 49.0, 49.4」 なお，これらのデータは「data3」というオブジェクトに入れて算出すること

## 次回の案内：

### 次回の案内：

- 次回は2種類の分散・標準偏差の説明をした後に，様々なデータの可視化を学びます．
- また，データの扱い方を含めた「実証分析の方法」についても紹介します．

## Rでデータを扱う時に注意すべきこと

### Rでデータを扱う時に注意すべきこと

- 必ず数字／文字は半角で入力する．
- 日本語は使わずにローマ字を使用する．
- コメントアウト（コードではなく，関係ないメモを入れること）をするときは半角の「#」から始める．
  - メモする内容は全角でもよい．
- ファイル名およびパスには決して全角の文字（ひらがな，カタカナ，漢字，全角スペースなど）を入れてはいけない．
  - 半角英数字だけにする．
- 慌てずに落ち着いて操作すれば，決して難しくない．
  - 1つずつ落ち着いて作業することを心がける．
- 「わからない」ことを恐れない
  - 周りの友人に聞いたり，教員に確認したりしよう．