# 基 礎 統 計 学

## 第2回　データの整理と正規性の検証

2017. 10. 13

# 統計分析のツールとして — R

## R Commander と EZR

- R とはフリーの統計ソフト.
- R Commander および EZR をインストールすればマウス操作可能になる.
- R 上で `library(Rcmdr)` で起動する. (Mac では X11 が必要)

# データファイルの形式

データファイルはExcelで作っておくのがよい.

一般には次のような形式でデータをまとめる.

| No | sex | height | weight | ... |
|---|---|---|---|---|
| 1 | M | 168 | 60 | ... |
| 2 | F | 173 | 65 | ... |
| 3 | F | 164 | 50 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ... |

ここでbodyMY.csvを読み込んでみる.

bodyMY.csvは青年の身長, 体重のデータ.

# 母集団と標本

情報を得たい対象全体 $\Omega$ を母集団（population）という．母集団から（無作為に）抽出された一部分を標本（sample）という．

**例**

日本人成人の平均身長を調査したいとすると，母集団は 20 歳以上の日本人全体である．

すべての統計的推測や検定には母集団が想定されている．

# 統計学とは

- 知りたいことは，母集団の情報
  例えば母平均，母分散，母標準偏差など
  これらを総称して母数 (パラメータ) という.
- 母集団から得られたサンプルの傾向・関連・分布の様子などを数値やグラフで表すのが記述統計である.
- 一方，サンプルの情報から母集団の母数を推測するのが推測統計である.（検定や推定）

# 統計分析の前提

## 無作為性

統計の分析では，標本は母集団から無作為（Random）に抽出されていることを前提とする.

現実には様々な要因により無作為性が崩れる場合がある．このような状況をバイアスがかかるなどという.

# 種々のバイアス

- 観察バイアス・・・観察者が特定を変数の過小・過大報告するとき
- 交絡バイアス・・・交絡因子が調整されてないとき
- 選択バイアス・・・母集団を代表していない標本を選んだとき（年齢等の背景因子に偏りがあるとき）

- 情報バイアス・・・測定値などに特定の偏りがあるとき（例：薬剤がプラセボであることを知ると治療効果が落ちる）
- 公表バイアス・・・肯定的な結果のみが公表されがちなことに由来するバイアス

バイアスを回避すために，治験などでは無作為化比較試験 (RCT) や二重盲検試験 (DBT) などが用いられる.

# データの種類

データ（変数，変量，尺度）は次のような分類がある．

## 質的変数

- 名義（カテゴリカル）変数 ・・・ 性別，血液型など
- 順序変数 ・・・ 大小，レベル，順位などの値を持つもの

## 量的変数

- 離散変数 ・・・ とびとびの値をとるもの．サイコロの目の数など
- 連続変数 ・・・ 連続的な数値をとるもの．身長，体重，血圧など

# 一変量の分布 – 記述統計

得られたデータの分布の様子を確認することが統計分析の第１歩である.

$X$: $x_1, x_2, \ldots, x_n$ を標本の $n$ 個の（連続または離散変数の）データとする.

## (標本) 平均 (mean)・・・分布の中心を表す

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

これは母集団の平均（母平均）の推定値となる.

# 平均に関する注意

1. 平均は，はずれ値（極端に大きいまたは小さい数値）に大きく影響される.
2. このような場合には平均値を分布の中心とすることは不適当である.

   そのかわりに，中央値（後述）を用いる方がよい.

# 分散と標準偏差

- 偏差　$x_i - \bar{x}$
- 偏差平方　$(x_i - \bar{x})^2$
- 偏差平方和　$S_X = \sum_{i=1}^{n} (x_i - \bar{x})^2$

## 分散 (variance)・・・分布の広がり具合を表す

$$s^2 = \frac{S_X}{n-1}$$

## 標準偏差 (standard deviation, SD)

$$s = \sqrt{\frac{S_X}{n-1}}$$　　分散のルート

# 分散についての注意

1. この分散は不偏分散といわれるのもので, $n-1$ を偏差平方和 $S_X$ の自由度という.
2. 不偏分散 $s^2$ は母分散 $\sigma^2$ の推定値である.
3. 統計ソフトにおいては, 標準偏差は上記の不偏分散を用いた式で計算している.
4. 平均と同様, はずれ値に影響を受けやすい.

# 中央値と四分位偏差

はずれ値のあるデータやいびつな分布をもつデータでは，平均や分散は不適切な場合がある.

**中央値 (median)=50%点, $Q_2$**

値の小さいものから順に並べちょうど真ん中にある値

**25%点, $Q_1$**

値の小さいものから順に並べちょうど 1/4 のところにある値

順序変数でも意味をもつ指標.

# 中央値と四分位偏差

順序変数でも意味をもつ.

## 75%点, $Q_3$

値の小さいものから順に並べちょうど 3/4 のところにある値

## 四分位偏差 $Q$

$$Q = (Q_3 - Q_1)/2$$

はずれ値があるときには標準偏差の代わりとして使える.

# 分布の視覚化 — ヒストグラムと箱ひげ図

- 分布の様子はヒストグラムや箱ひげ図を描くことにより視覚化される.
- 正規分布は左右対称の釣り鐘状（単峰性）の分布.
- 医学・生物関係では，正規分布でないデータも多い.

**例**

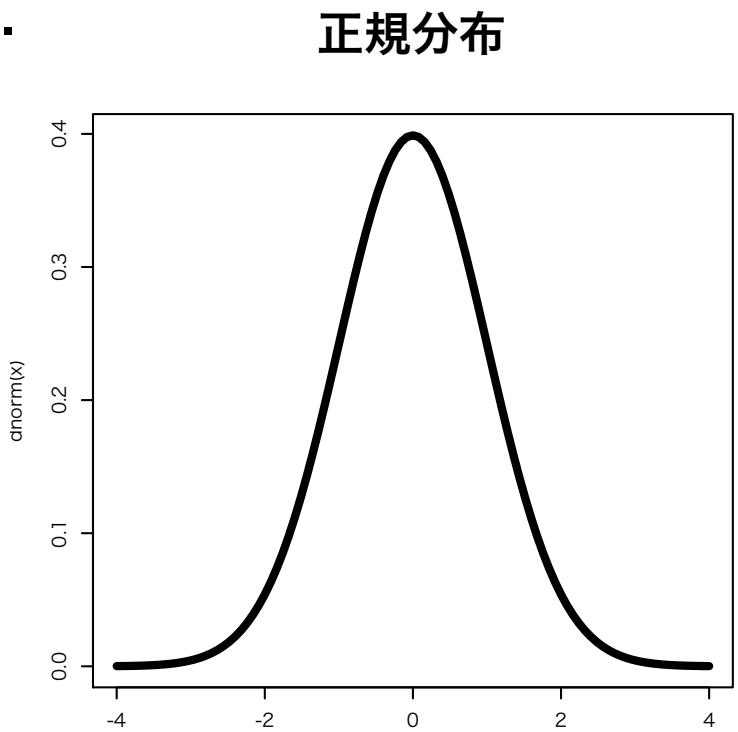
身長の分布はほぼ正規分布だが，体重はそうとは限らない.

# 正規分布

測定誤差の分析からガウス（世界３大数学者の一人）により発見された．ガウス分布ともいう．統計学の理論の基盤となる分布．

$N(\mu, \sigma^2)$ :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



# 標準正規分布

平均 $\mu$, 分散 $\sigma^2$ である正規分布を記号で $N(\mu, \sigma^2)$ で表す.

**標準正規分布**

$N(0, 1)$ を標準正規分布という.

正規分布はパラメトリック検定の理論的基盤となる分布である.

# 標準化

変数 $X$ が正規分布 $N(\mu, \sigma^2)$ に従うとき,

## $X$ の標準化

$$Z = \frac{X - \mu}{\sigma}$$

と変換すると $Z$ は標準正規分布 $N(0, 1)$ に従う.

この変換を標準化（規準化）という.

# 正規性の検証

データが正規分布に従っているか否かで検定の手法が変わるので，事前に正規性の検証することが重要である.

1. t 検定や分散分析などはデータの分布が正規分布であることを前提としている．これらはパラメトリックな検定といわれる.
2. 一方，ウィルコクスンの順位和検定やマン・ホイットニーの U 検定は正規分布であることを前提としないのでノンパラメトリックな検定といわれる.

# 正規性の利点

検定について考えると

- パラメトリック検定・・・データの数値そのものを利用する
- ノンパラメトリック検定・・・データの大小の順位のみを利用する

パラメトリック検定の方がより正確な検定ができる.

# 正規性の検証

## 検証の手段

1. ヒストグラム（視覚的に確かめる.）
2. PP プロット，QQ プロット（理論値と観測値を平面上にプロットしたもの．正規分布ならば直線上にプロットされる.）
3. 正規性の検定（Shapiro-Wilk 検定など）

# EZR でデータの分布を調べよう

ファイル bodyMY.cvs は青年の身長，体重のデータの表である.

青年男子の身長, 体重データの基本統計および正規性について調べたい.

# 問題

1. 正規性の検討1： それぞれのヒストグラムをかいてみる.
2. 正規性の検討2： QQプロットをみる. （「標準メニュー」→「グラフ」→「QQプロット」）

   qqnorm(x$x); qqline(x$X)
3. 正規性の検討3： Shapiro-Wilk検定をしてみる.

   (「標準メニュー」→「統計量」→「要約」→「シャピロ-ウィルクの正規性の検定」)

# 答

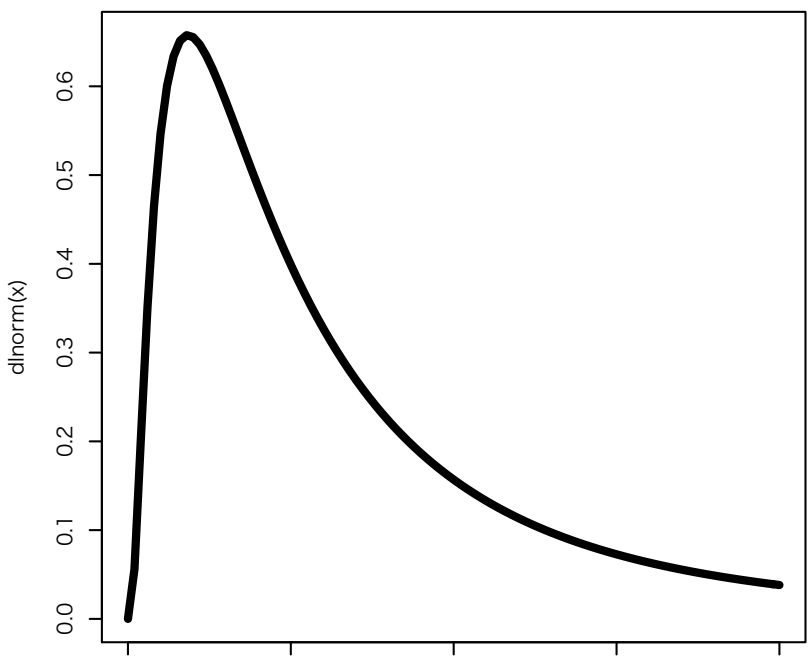
以上から，身長は正規分布に従っているが，体重はそうではないことが結論される.

注）　「統計解析」→「連続変数の解析」→「正規性の検定 (Kolmogorv-Smirnov 検定)」でも Shapiro-Wilk の検定が実行される.（サンプル数 5000 未満のとき）

# 正規分布でないときでも

## 対数変換 $\log X$

を行うと正規分布になることがある．このような分布 $X$ を対数正規分布といい，医学・生物関係のデータにしばしば現れる．



# 体重データは対数正規分布か？

bodyMY.csv の体重（Wt）の log は logWt の列データである.

Shapiro-Wilk 検定をしてみよ.

# 答

Shapiro-Wilk 検定で有意にならないので，logWt は正規分布といえる.