

# 政治学方法論Ⅰ

## Rを使った統計分析の基礎

矢内 勇生

大学院法学研究科・法学部

2016年4月20日



神戸大学

# 今日の内容



## 1 数量分析の基礎

- データとは何か
- データの要約
- 統計量

## 2 確率と確率分布

- 確率
- 様々な確率分布

# R のコードと説明



授業のウェブページ：

URL <http://www.yukiyanai.com>

- 授業
- 政治学方法論 I
- 授業の内容
- 「R による統計分析の基礎」



データとは何か

# データ



データ : data (複) – datum (单)

調査、観察などによって集められた情報

- 数量データ : 年齢、年収、人口、GDP、etc.
- 質的データ : 性別、支持政党、投票参加、etc.

# どんなデータに興味がある？



- 観察の対象（観察単位：unit of observation）によって値が変わるもの：変数 (variable)
  - 年収や支持政党は人によって違う
- 値が変わらないもの（定数：constant）には興味がない
  - 女子高生の性別、学生の職業、etc.

データとは何か

# 変数に興味がある！



- 変数 (variable) とは
  - 数 (値) が一定でない = 変化する数 (変な数ではない)
  - 様々な (少なくとも 2 つ以上の) 値をとる : 分布する
- 値が一定のもの : 定数 (constant)

# 変数の分類



## ● 量的変数

- 比率 (ratio scale)
- 間隔 (interval scale)

## ● 質的変数

- 順序 (ordinal scale)
- 名義 (nominal scale)

データとは何か

## 変数の種類とその特性



変数の種類	異なる値の間の			
	異同	順序	差	比
質的変数	名義尺度	○	×	×
	順序尺度	○	○	×
量的変数	間隔尺度	○	○	○
	比率尺度	○	○	○

## データの要約

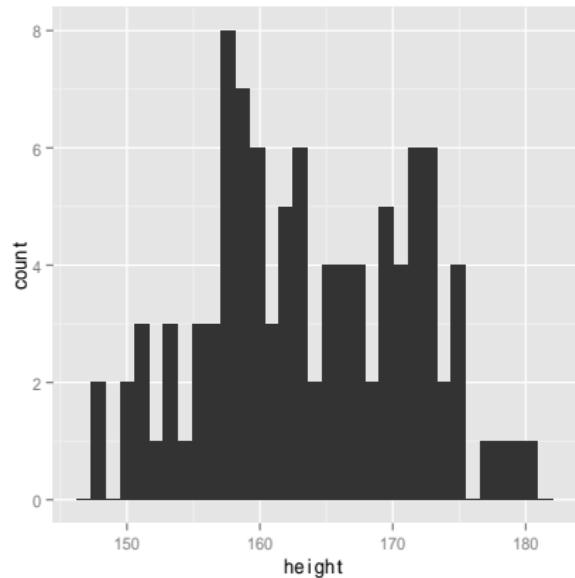
## 変数の視覚化：ヒストグラム



- ヒストグラム (histogram) を描いて変数の分布を確認する

- 分布の中心は？
- 左右対称か？
- 大きな山はいくつあるか？
- どれくらいの範囲に広がっているか？

- 右図：身長（授業用データ）の分布

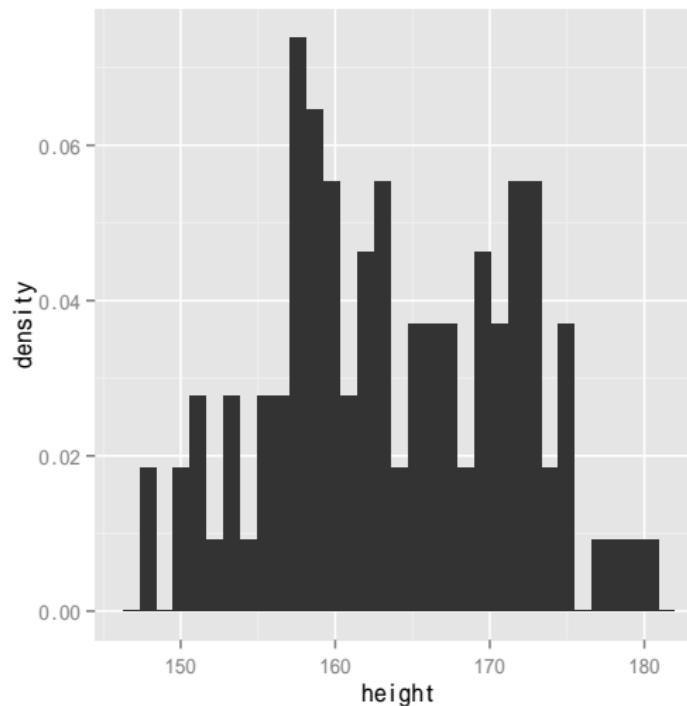


○○○○○  
○●○○○○○○○○  
○○○○○○

○○○○○○○  
○○○○○○○○○

## データの要約

## ヒストグラム：縦軸を確率密度に変える

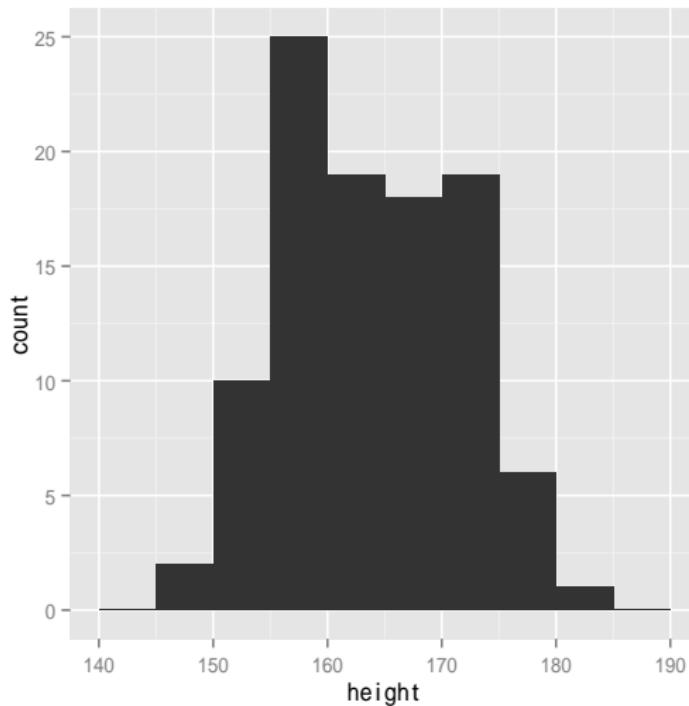


○○○○○  
○○●○○○○○○○○  
○○○○○○

○○○○○○○  
○○○○○○○○○○

## データの要約

## ヒストグラム：BINの幅を変える

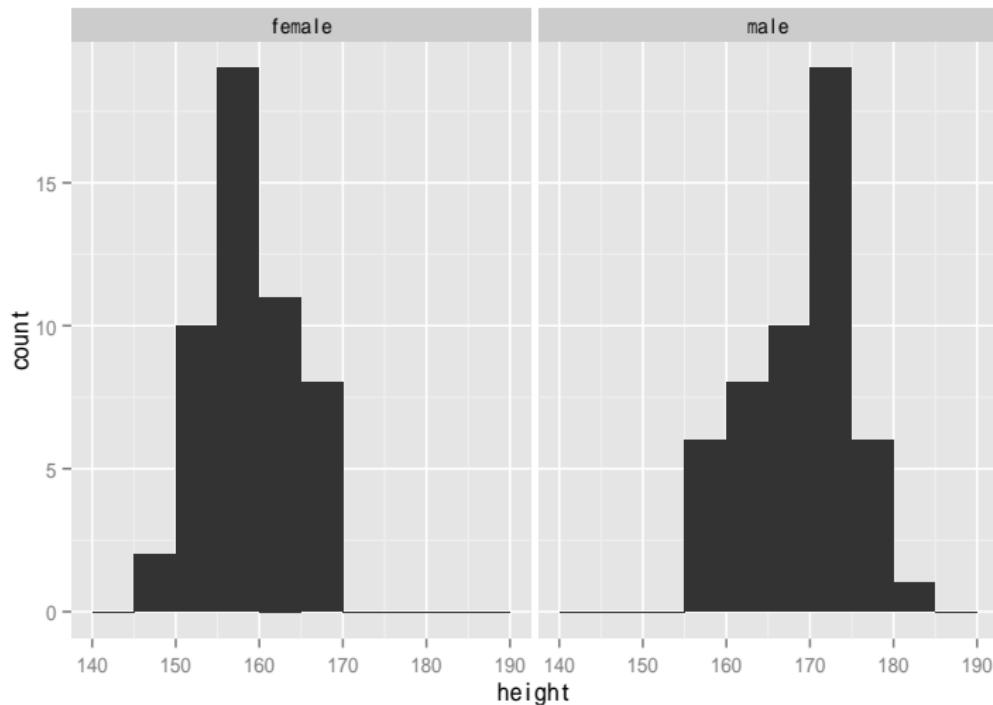


○○○○○  
○○○●○○○○○○  
○○○○○○

○○○○○○○  
○○○○○○○○○

## データの要約

## ヒストグラム：グループ分けする

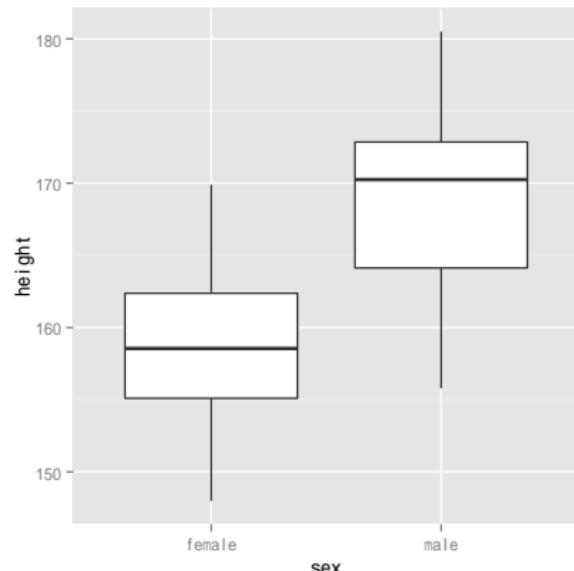


## データの要約

## 変数の視覚化：箱ひげ図



- 箱ひげ図 (box[and whiskers] plot) を描いて変数の分布を比較する
- 箱ひげ図を見ると、変数の最小値（外れ値を除く）、第1四分位点、中央値、第3四分位点、最大値（外れ値を除く）がわかる
- 右図：身長（授業用データ）の男女別分布

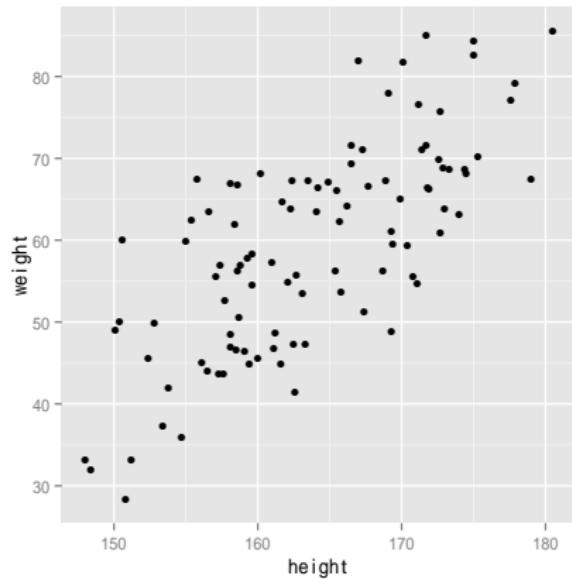


## データの要約

## 変数の視覚化：散布図



- 散布図 (scatter plot) を描いて 2 变数の関係を確認する
- 本当は無関係でも、パタンをあるように見えることがあることに注意する
- 右図：身長と体重（授業用データ）の散布図

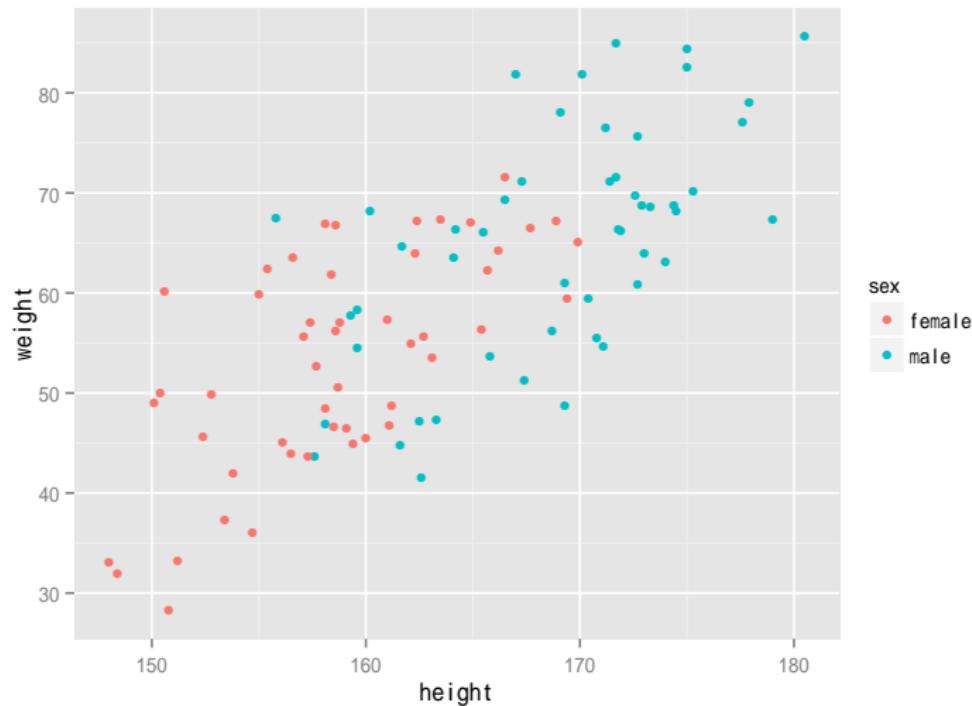


○○○○○  
○○○○○○●○○○  
○○○○○○

○○○○○○○  
○○○○○○○○

## データの要約

## 散布図：グループごとに色分けする

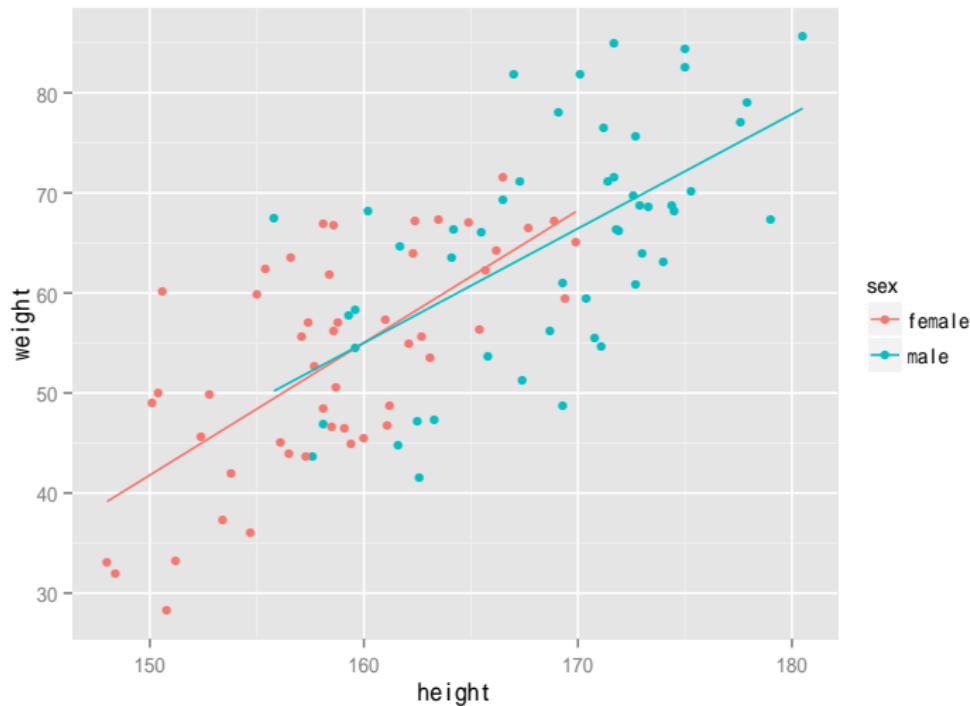


○○○○○  
○○○○○○○●○○  
○○○○○○

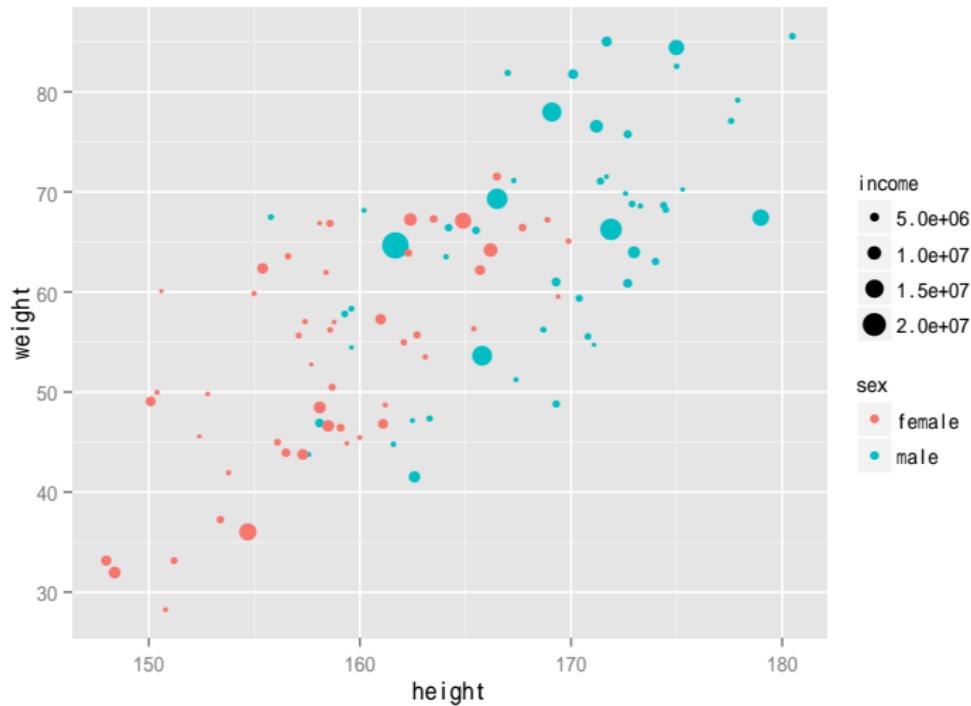
○○○○○○○  
○○○○○○○○

## データの要約

## 散布図：関係を直線で示す



散布図：次元を増やす



5 次元!

## 統計量

## 統計量



統計量 (statistic) とは

- 変数の「ある特徴」を表す数式
- 決められた方法（アルゴリズム）を使うことによって得られる
- 様々な統計量がある

## 統計量

## 中心的傾向を表す統計量：算術平均 (mean)



- 変数  $x$  の平均値（算術平均, 相加平均, **mean**）を  $\bar{x}$  と表し（「エックスバー」と読む）、

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

である

- 算術平均はヒストグラムの重心である
- 算術平均は外れ値 (outlier) の影響を受け易い

## 中心的傾向を表す統計量：中央値 (median)



- 変数  $x$  の中央値（中位値, **median**）とは

$$\int_{-\infty}^m dF(x) \geq \frac{1}{2} \quad \text{and} \quad \int_m^{\infty} dF(x) \geq \frac{1}{2}$$

を満たす  $m$  のことである ( $F(x)$  は  $x$  の累積分布関数)

- 簡単に言い換えると、変数を小さい順（大きい順）に並べ替えたとき、ちょうど真中にある値である
- 真中がないとき（変数の長さ  $n$  が偶数のとき）、真中を挟む2つの値の算術平均を用いる
- 中央値は外れ値の影響を受けにくい
- 分布が左右対称なとき、算術平均と中央値は一致する

## ばらつきを表す統計量：四分位範囲 (IQR)



- 変数  $x$  の四分位範囲 (inter-quartile range: **IQR**) は、第 3 四分位数点 ( $Q_{3/4}$ ) と第 1 四分位点 ( $Q_{1/4}$ ) の差、つまり

$$\text{IQR} = Q_{3/4} - Q_{1/4}$$

である

- 四分位点とは、変数を小さい順に 4 つの個数が等しいグループに分けたとき、グループの境界となる点のことである
  - 例： $x = \{0, 1, 1, 2, 4, 8, 9, 10\}$  のとき
  - 第 1 四分位点は 1、第 3 四分位点は 8.5 (2.5 と 7.5 ではない！)
- IQR は箱ひげ図の箱の高さ（横にした場合は幅）を表す
- IQR は外れ値の発見に利用される
  - $[Q_{1/4} - 1.5\text{IQR}, Q_{3/4} + 1.5\text{IQR}]$  の範囲外を外れ値とみなす

# ばらつきを表す統計量：分散



- 変数  $x$  の不偏分散（unbiased **variance**）を  $u_x^2$  と表し、

$$u_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

である

- 分散は変数のばらつき（分布の広がり）を表す：数値が大きいほどばらつきが大きい
- 分散は必ず正の値をとる（ただし、定数の分散は 0）

上の式で表されるものは不偏分散と呼び、分母の  $n - 1$  を  $n$  に変えたものを「分散」と呼ぶ場合がある

## ばらつきを表す統計量：標準偏差



- 変数  $x$  の分散の平方根を標準偏差（standard deviation: **sd**）と呼ぶ
- 標準偏差を  $u_x$  と表し、

$$u_x = \sqrt{u_x^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

である

- 分散の単位が「変数の単位の二乗」になってしまうのに対し、標準偏差の単位は変数そのものの単位と同じである

上の式で表されるものは「不偏分散の平方根」と呼び、分母の  $n-1$  を  $n$  に変えたものを「標準偏差」と呼ぶ場合がある

## 確率の定義



標本空間  $S$  における任意の事象  $A$  に対して  $\Pr(A)$  という数を与える。 $\Pr(A)$  が以下の 3 つの公理をみたすとき、それを確率と呼ぶ

- ① どの事象も、それが起きる確率は非負である

$$\Pr(A) \geq 0$$

- ② 全事象  $S$  の起きる確率は 1 である

$$\Pr(S) = 1$$

- ③ 排反事象に関して以下の和の法則が成り立つ

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$$

## 確率分布



- $\Pr(A)$  の  $A$  にあたる事象は色々ある（確率変数  $A$  は様々な値をとる）
- それぞれの事象 ( $A$  の各値) について確率を考える：**確率分布**
  - 例：「正しい」コインを 2 回投げ、表が出る回数を調べる
  - $S = \{0, 1, 2\}$
  - $\Pr(0) = 1/4, \Pr(1) = 1/2, \Pr(2) = 1/4,$
- 確率の分布の仕方は様々：確率質量関数または確率密度関数と累積分布関数で表す

## 確率質量関数 (PMF)



- 確率変数が離散型のとき、確率分布を表すために確率質量関数 (probability mass function: PMF) を用いる
- 確率変数  $X$  のとる値の集合が  $S = \{x_1, x_2, \dots\}$  のとき、確率質量関数は、

$$f_X(x_i) = \Pr(X = x_i) = \Pr(x_i)$$

と表すことができる

## 確率密度関数 (PDF)



- 確率変数  $X$  が連続型のとき、確率分布を表すために**確率密度関数 (probability density function: PDF)** を用いる
- 連続型の確率変数  $X$  が特定の値をとる確率  $\Pr(X = x_i) = 0$  である
- 代わりに区間を用い、 $X$  が区間  $[a, b]$  の値をとる確率を確率密度関数  $f_X$  を使って表すと

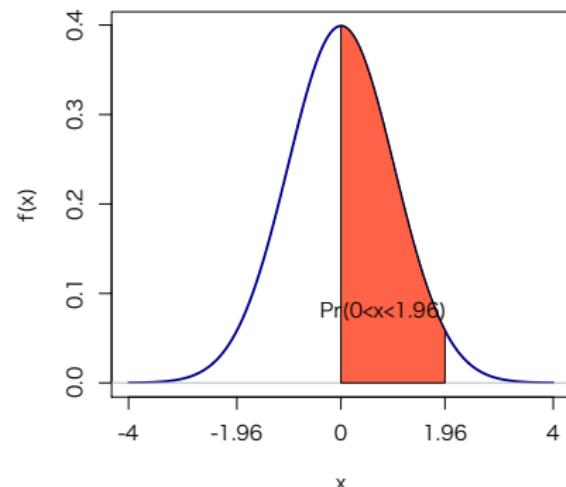
$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)dx$$

となる

## 確率密度関数の例：標準正規分布

- 標準正規分布の確率密度関数（右図）
- 横軸：変数の（とり得る）値
- 縦軸：確率密度
- 確率：確率密度関数と横軸の間の面積
  - $x$  が  $[0, 1.96]$  の値をとる確率 = 図中で赤く塗りつぶされた部分の面積

標準正規分布の確率密度関数



## 累積分布関数 (CDF)



- 確率分布を表すために使われるその他の関数として、**累積分布関数 (cumulative distribution function: CDF)** がある
- 累積分布関数  $F_X$  は

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f(x)dx$$

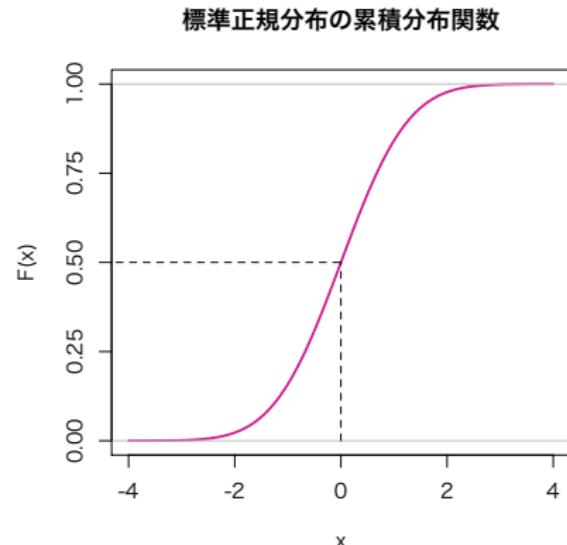
と表すことができる

- つまり、確率密度関数は累積分布関数の導関数である

## 累積分布関数の例：標準正規分布



- 標準正規分布の累積分布関数（右図）
- 横軸：変数の（とり得る）値
- 縦軸：確率  $\Pr(X \leq x)$ 
  - 図中の点線で示されている  $F(X = 0)$  は  $X$  が 0 以下の値をとる確率
  - 「 $X$  が 0 になる確率」ではない！

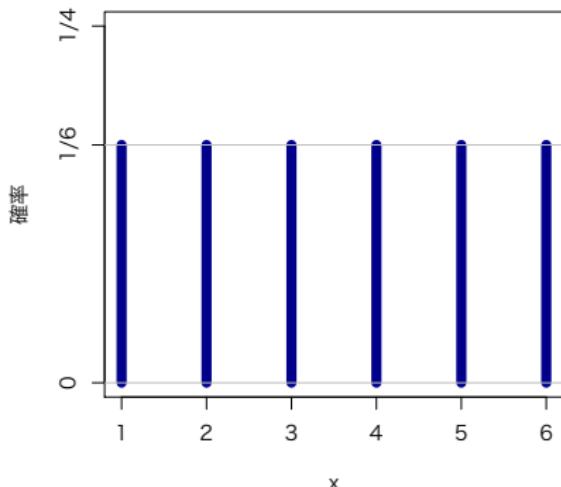


# 離散一様分布 (discrete uniform distribution)



- 確率変数が  $n$  個の値を同じ確率でとり得るときの分布
- 例：「正しい」サイコロを 1 回投げるときに出る目の確率分布（右図）

離散一様分布の確率質量関数：サイコロ



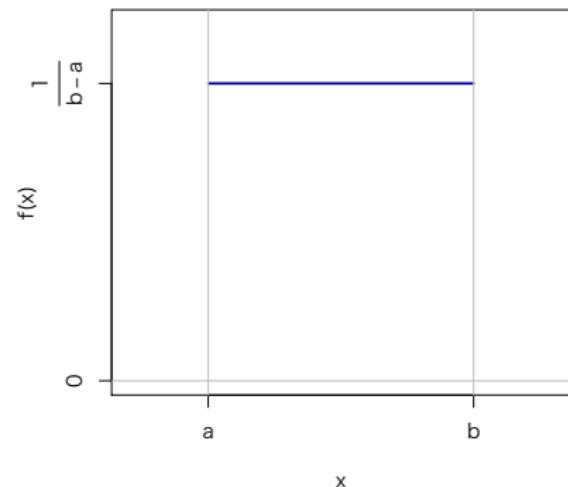
# 連続一様分布 (contiguous uniform distribution): $U(a, b)$

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a} \quad (a \leq x \leq b)$$

一様分布の確率密度関数

- 母数：最小値  $a$  と最大値  $b$
- 区間  $[a, b]$  内で確率密度一定（一様）



# ベルヌーイ分布 (Bernoulli distribution) : $\text{Ber}(p)$

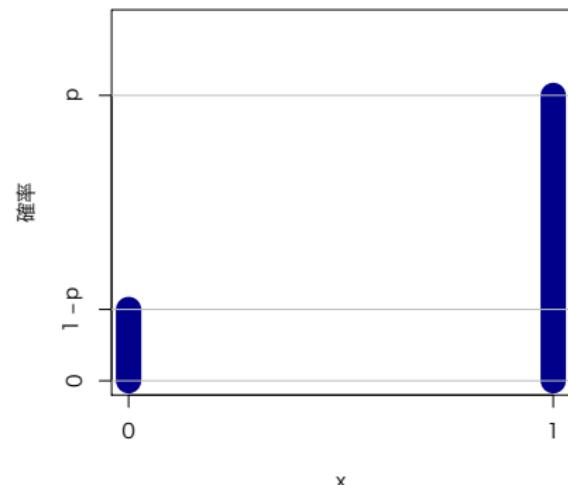


$$X \sim \text{Ber}(p)$$

$$f(x) = p^x(1-p)^{1-x}$$

ベルヌーイ分布の確率質量関数

- 確率  $p$  で 1、 $1-p$  で 0 になるような事象
- 例：「正しい」コイン投げ：  
 $p = 1/2$  のベルヌーイ試行
- 母数：成功確率  $p$
- $X$  は 1 (成功) または 0 (失敗) のいずれか



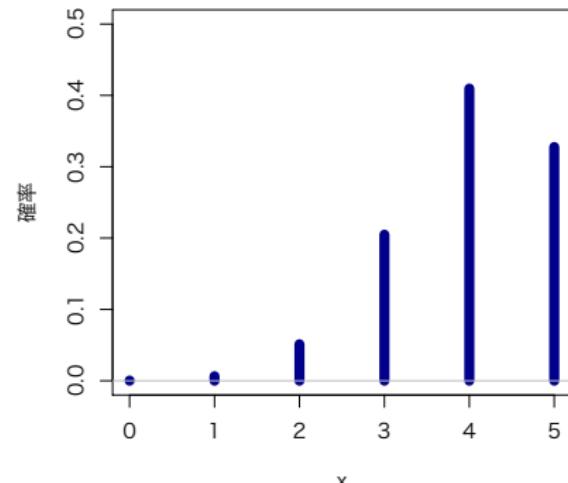
## 様々な確率分布

二項分布 (binomial distribution) :  $\text{Bin}(n, p)$ 

$$X \sim \text{Bin}(n, p)$$

$$\binom{n}{k} p^k (1-p)^{n-k}$$

- 確率  $p$  のベルヌーイ試行を  $n$  回繰り返したときの成功回数の分布
- 母数：試行回数  $n$  と 1 試行の成功確率  $p$
- 成功回数  $k = 0, 1, \dots, n$

二項分布 ( $n=5, p=0.8$ ) の確率質量関数

## 様々な確率分布

# 正規分布 (normal distribution) : $N(\mu, \sigma^2)$

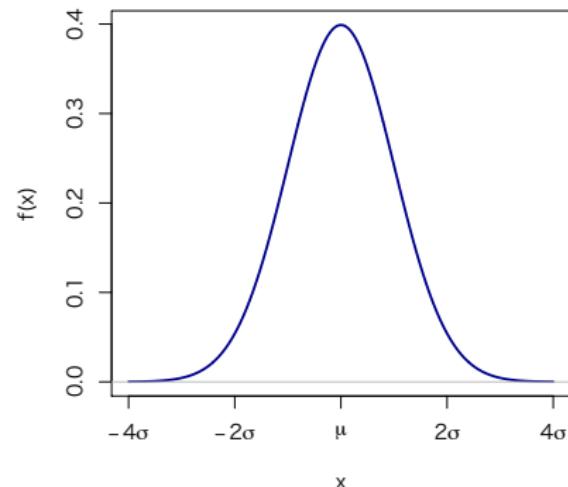


$$X \sim N(\mu, \sigma^2) \text{ (または } N(\mu, \sigma))$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

標準正規分布の確率密度関数

- 母数：平均  $\mu$  と分散  $\sigma^2$
- 中心  $\mu$  に関して左右対称
- $\mu \pm \sigma$  の間の値をとる確率が約 68%
- $\mu \pm 1.96\sigma$  の間の値をとる確率が約 95%



## 様々な確率分布

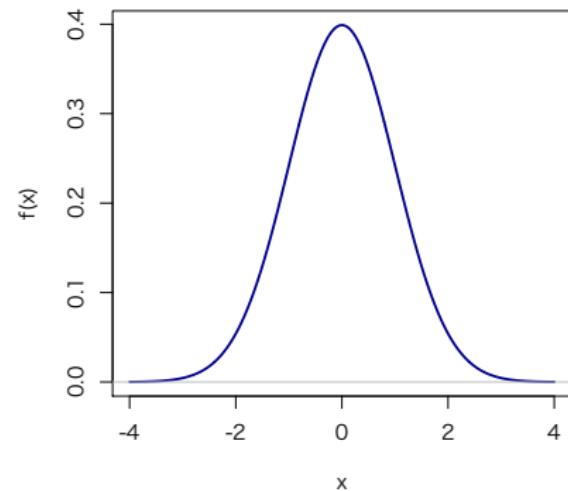
u 標準正規分布 (standard normal distribution) :  $N(0, 1)$

$$X \sim N(0, 1)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

標準正規分布の確率密度関数

- 正規分布のうち、  
 $\mu = 0, \sigma^2 = 1$  のもの
- 中心 0 に関して左右対称
- $[-1, 1]$  の間の値をとる確率  
が約 68%
- $[-1.96, 1.96]$  の間の値をと  
る確率が約 95%



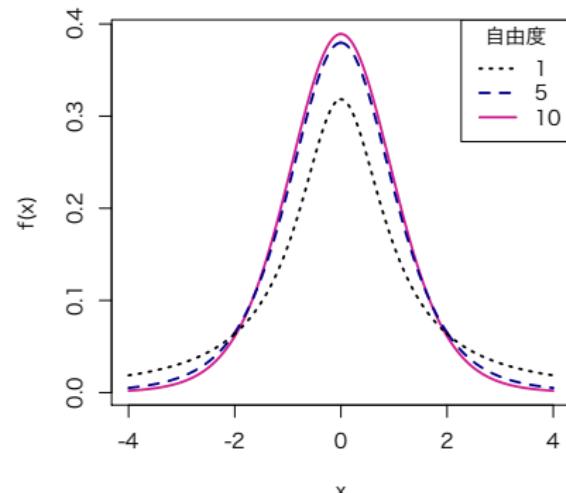
## 様々な確率分布

*t* 分布 (Student's *t* distribution)

$$X \sim t(v)$$

$$\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$

- 母数：自由度  $v$  (0 より大きい実数)
- 観測数  $n$  が十分大きくないとき、誤差の分布は *t* 分布に従う
- $v \rightarrow \infty$  で正規分布に近づく

*t* 分布の確率密度関数

## 様々な確率分布

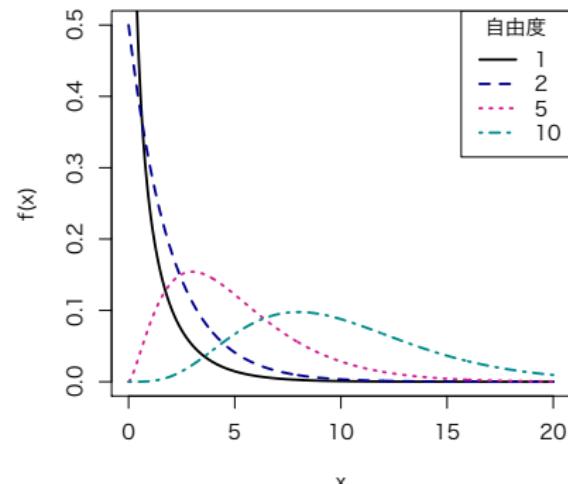
カイ二乗分布 ( $\chi^2$  distribution)

$$X \sim \chi^2(k)$$

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \quad (x \geq 0)$$

- 母数：自由度  $k$  (自然数)
- クロス表における変数間の独立性の検定などに使う
- $k \rightarrow \infty$  で正規分布に近づく  
(が、近づき方は遅い)

カイ二乗分布の確率密度関数



## その他によく出てくる分布



- $F$  分布
- ガンマ (gamma) 分布
- 負の二項 (negative binomial) 分布
- ポアソン (Poisson) 分布
- ディリクレ (Dirichlet) 分布
- コーシー (Cauchy) 分布
- etc.