

数理科学統論 I/J (2019 年度版)

統計データ解析 確率と統計

吉田朋広 (東京大学)
小池祐太 (東京大学)
村田 昇 (早稲田大学・東京大学)

September 4, 2019

東京大学大学院数理科学研究科
統計データ解析教育研究グループ[°]

1 極限定理

データ解析の際に、分析の対象としたい集団全体のデータが入手できることは多くの場合稀である。例えばテレビ番組の視聴率を厳密に調べようとした場合、全世帯の視聴状況を確認しなくてはならないが、こうした調査を行うことは現実的には困難であろう。また、データ解析の目的が「将来の予測」や「集団の背後にある共通の法則の発見」であった場合、「(現時点では手に入らない)将来のデータ」や「(必ずしも現時点の集団に含まれているとは限らない)実現する可能性のあるデータ」が対象としたい集団に含まれてしまうため、目的にかなうような集団全体のデータの入手はそもそも不可能である。このような理由から、多くの場合分析対象の集団の一部のデータのみを対象としてデータ解析を行い、その結果を敷衍して集団全体の性質についての知見を得る必要が生じる。そのための方法論を体系的に研究する分野を**推測統計**と呼ぶ。

分析対象の集団の一部から解析のためのデータ収集を行う際、データの集め方に偏りがあると、データ解析の結果から集団全体の性質を推測することが難しくなることは直感的には理解できるであろう。具体的な例で説明すると、日本全体の平均気温を計測したい場合に、沖縄県の各地点の気温のみ計測してデータとしてしまうと、得られたデータから計算された平均気温は真に知りたい値より明らかに高くなってしまうため、何らかの方法で補正する必要が生じる。このような問題を回避するためには、データを「ランダムに」収集すれば良いことは、直感的にも経験的にも理解できるであろう。しかし、「ランダムに」データを収集することでなぜ問題が解決できるのかということの数学的・論理的な根拠は、実際にはそれほど自明でない。この問い合わせに厳密な意味での解答を与えるためには、数学の一分野である「(測度論的)確率論」を学習する必要があるが、そのためには位相空間論や測度論・関数解析など他の数学の分野にも習熟する必要がある。本講義ではそのような問題を避けて、ランダムネスによって上述のサンプリングの問題などのデータ解析上の困難が解決できることを直感的に理解するために、乱数を使ったシミュレーションによってランダムネスから結論される種々の数学的結果を観察する。

1.1 確率論の基礎事項の復習

本節は次節以降の極限定理の説明で必要となる、高校数学で学習する範囲の確率論の基礎事項を復習する。

1.1.1 確率変数

数学的には、乱数は**確率変数**という概念でモデル化される。確率変数とは、値がランダムに決定される変数で、すべての実数 $a \leq b$ に対して、その値が区間 $[a, b]$ に含まれる確率があらかじめ定められているような変数のことをいう¹。

X が確率変数ならば、定義より X が区間 $[a, b]$ ($a \leq b$) に含まれる確率が定まるから、その確率を $P(a \leq X \leq b)$ で表すことにする。特に $a = b$ のとき $P(a \leq X \leq b)$ は $X = a$ となる確率を表すから、それを $P(X = a)$ で表することにする。

以下本章では、記述の簡単のために主として有限個の値のみをとる確率変数のみを考える。無限個の値、特に連続的な値をとる確率変数については次章で説明する。

¹この定義は数学的には厳密性を欠くが、本講義ではこの定義を採用する。

1 極限定理

1.1.2 平均と分散

X を (とりうる値が有限個である) 確率変数として, X の取りうる値を x_1, x_2, \dots, x_N とする. このとき, 以下で定義される量

$$E[X] := \sum_{i=1}^N x_i P(X = x_i) \quad (1.1)$$

を X の**平均**もしくは**期待値**と呼ぶ. $E[X]$ は X の「理論上の平均値」に対応する量とみなせる. より一般に, X の関数 $\varphi(X)$ に対して $\varphi(X)$ の期待値を

$$E[\varphi(X)] := \sum_{i=1}^N \varphi(x_i) P(X = x_i) \quad (1.2)$$

で定義する. 例えば整数 p に対して

$$E[X^p] = \sum_{i=1}^N x_i^p P(X = x_i) \quad (1.3)$$

といったものを考えることができるが, これは X の p 次の**モーメント**と呼ばれる量となる.

測定誤差の大きさによって統計的推定の精度は左右されるため, 確率変数の値のばらつき具合を定量化する指標が必要である. そのような指標の1つとして, 平均からのはらつき具合を定量化した指標

$$\text{Var}[X] := E[(X - E[X])^2] = \sum_{i=1}^N (x_i - E[X])^2 P(X = x_i) \quad (1.4)$$

があり, これを X の**分散**と呼ぶ. 分散はもとの確率変数を二乗したスケールをもつため, 平均と単位をあわせる意味で分散の平方根

$$\sqrt{\text{Var}[X]} \quad (1.5)$$

を考えることがある. この量を X の**標準偏差**と呼ぶ.

分散の計算には次の恒等式が便利である:

$$\text{Var}[X] = E[X^2] - (E[X])^2 \quad (1.6)$$

演習 1.1. (1.6) 式の成立を確認せよ ($\sum_{i=1}^N P(X = x_i) = 1$ に注意).

1.1.3 独立性と同分布性

統計の文脈では, 確率変数は観測データの一つ一つに対応する. 統計学では多数の観測データを扱うため, 確率変数の列 X_1, X_2, \dots, X_n に対する考察が重要となる. 以下「 X_1 が x_1 という値をとり, X_2 が x_2 という値をとり, ..., X_n が x_n という値をとる」という事象が起きる確率を, 記号

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad (1.7)$$

で表す. 観測データが「ランダムに」サンプリングされた状況を表現するために, 次の概念を導入する.

1 極限定理

定義 1.1 (確率変数列の独立性). 確率変数列 X_1, X_2, \dots, X_n が**独立**であるとは、任意の n 個の実数 x_1, x_2, \dots, x_n に対して

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n) \quad (1.8)$$

が成り立つことをいう。

X_1, X_2, \dots, X_n が独立であるというのは、直感的にはすべての $i = 1, \dots, n$ について、 X_i がとる値は $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ がとる値と無関係に定まるということである。

独立性と並んで重要な概念に、確率変数列の同分布性がある。これは、観測データが同一の法則に従って生成された集団からサンプルされたということを数学的に表現した概念である。

定義 1.2 (確率変数列の同分布性). 確率変数列 X_1, X_2, \dots, X_n が**同分布**であるとは、任意の実数 x に対して

$$P(X_1 = x) = P(X_2 = x) = \cdots = P(X_n = x) \quad (1.9)$$

が成り立つことをいう。

独立かつ同分布な確率変数列を**独立同分布**もしくは*i.i.d.* であるという (*i.i.d.* は independent and identically distributed の略)。

1.2 大数の法則

分析対象の集団の平均値を求めたい場合、分析対象の一部を「ランダムに」収集したデータのみを用いて(標本) 平均を計算しても、データのサンプル数が十分大きければ、その標本の平均値は集団全体の平均値に近い値であることは、経験則としてよく知られている。例えば視聴率の調査などがそうである。また、表裏の出方に偏りがないコインを繰り返し投げ続けると、投げた回数に対して表が出た回数の割合は理論上の平均値 $1/2$ に近づいて行くであろう。このように、同一の法則に従って生成された(と仮定された) 集団に対して「ランダムな」観測を多数繰り返すと、観測値の平均は「真の平均値」(集団全体の平均値や理論上の平均値のこと) に近づくことは経験上よく知られている。この法則を数学的に定式化した定理は**大数の法則**として知られている。

数学的には「サンプル数が十分大きい」という状況を「サンプル数を無限大にしたときの極限」として定式化する。従って確率変数の無限列 X_1, X_2, \dots を考察する必要が生じる。この場合の独立性および同分布性を再定義しておく。

定義 1.3 (無限列の場合の独立性と同分布性).

- X_1, X_2, \dots が**独立**であるとは、任意の正整数 n に対して X_1, X_2, \dots, X_n が独立であることをいう。
- X_1, X_2, \dots が**同分布**であるとは、任意の正整数 n に対して X_1, X_2, \dots, X_n が同分布であることをいう。
- X_1, X_2, \dots が**独立同分布**もしくは*i.i.d.* であるとは、 X_1, X_2, \dots が独立かつ同分布であることをいう。

以上の定義のもと、大数の法則は以下のように述べられる。

定理 1.4 (大数の強法則). X_1, X_2, \dots を独立同分布な確率変数列とし、その平均を μ とする。このとき X_1, \dots, X_n の標本平均

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (1.10)$$

1 極限定理

が $n \rightarrow \infty$ のとき μ に収束する確率は 1 である。このことを「 \bar{X}_n は $n \rightarrow \infty$ のとき μ に概収束する」という。

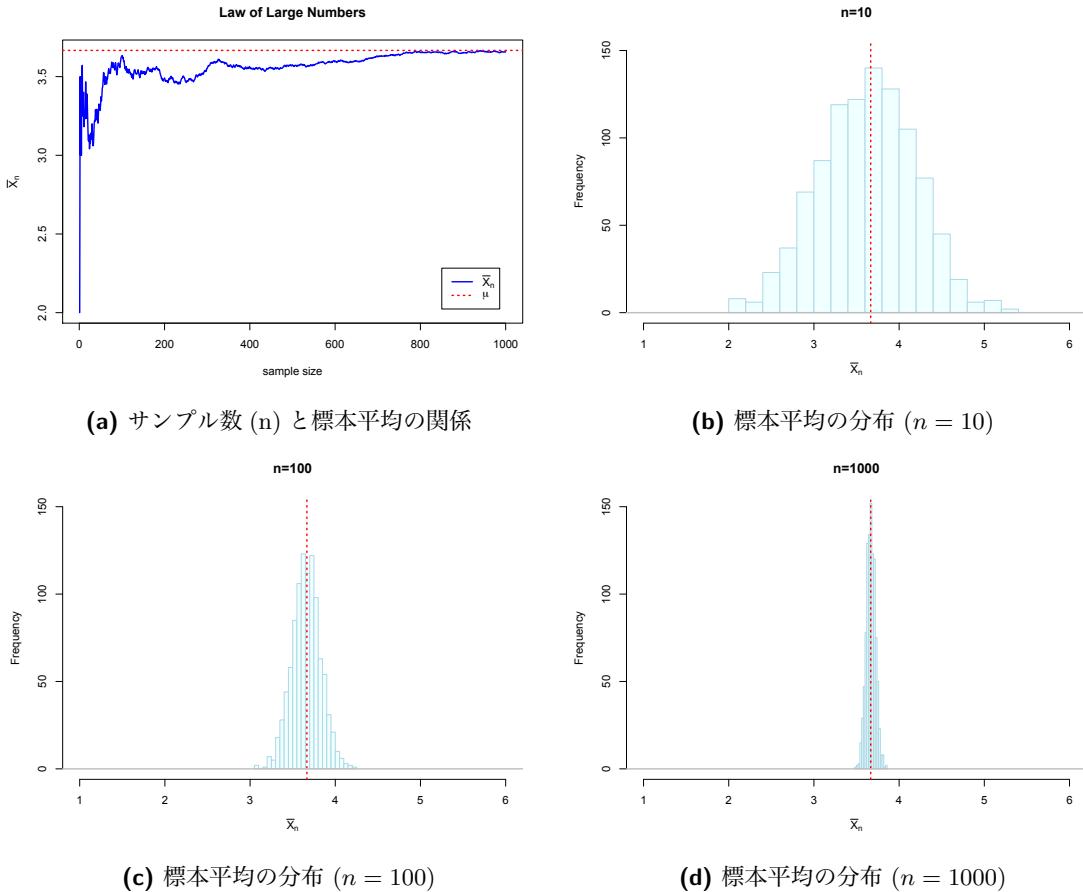


Figure 1.1: 大数の法則のシミュレーション例

[Figure 1.1 を参照]

```
> ### 大数の法則
> ### 偶数が出る確率が奇数の出る確率の 2倍となるサイコロを振ったときに
> ### 出る目の平均値を実験で確かめる
> set.seed(123)      # 乱数のシード値の指定
> omega <- 1:6       # サイコロの出る目の集合
> p <- rep(1:2, 3)  # 出現確率の比 (奇数 1:偶数 2, 正規化していなくてもよい)
> (mu <- weighted.mean(omega, p)) # 理論上の平均
[1] 3.666667

> ### 1回の実験
> ### サンプル数を大きくすると標本平均が理論上の平均に近づくことを確認
> n <- 1000 # 実験回数の最大値
> x <- sample(omega, size=n, prob=p, replace=TRUE) # 実験
> xbar <- cumsum(x)/(1:n) # サンプル数ごとの標本平均の計算 # help(cumsum) 参照
> plot(xbar, type="l", col="blue", lwd=2,
+       xlab="sample size", ylab=expression(bar(X)[n]),
+       main="Law of Large Numbers") # 標本平均の推移のプロット
> abline(h=mu, col="red", lty="dotted", lwd=2) # 理論平均を表す点線
```

```

> legend("bottomright", inset=1/20, # 位置をキーワードで指定
+       legend=c(expression(bar(X)[n]), expression(mu)),
+       col=c("blue", "red"), lty=c("solid", "dotted"), lwd=2) # 凡例の追加
> #### 複数回の実験
> #### サンプル数が大きければ、標本平均は理論平均に近い値を取ることの確認
> mymean <- function(n) # n回サイコロを振って出た目の標本平均を計算する関数
+   mean(sample(omega, size=n, prob=p, replace=TRUE))
> mc <- 1000 # Monte-Carlo 実験の繰り返し回数
> for(n in c(10, 100, 1000)){ # サンプル数を変えて実験
+   xbars <- replicate(mc, mymean(n)) # 同じ実験を mc 回繰り返し標本平均を記録
+   hist(xbars, breaks=20, col="azure", border="lightblue",
+         xlim=c(1,6), ylim=c(0,150), # 描画範囲を指定
+         xlab=expression(bar(X)[n]), main=paste0("n=",n))
+   abline(v=mu, col="red", lwd=2, lty="dotted") # 理論平均
+   abline(h=0, col="grey", lwd=2, lty="solid")
+ }

```

(mc-lln.r)

演習 1.2. 大数の法則について調べてみよう.

1. 様々な乱数を用いて、大数の法則を確認してみなさい.
2. 概収束、確率収束、平均収束、法則収束といった言葉を調べてみなさい.

1.3 中心極限定理

前節で説明した大数の法則は、サンプル数 n を大きくするに従い標本平均 \bar{X}_n が真の平均 μ に限りなく近づくことを保証している。言い換えると「推定誤差」 $\bar{X}_n - \mu$ は n を大きくすると限りなく 0 に近づく。しかし、実際に推定誤差 $\bar{X}_n - \mu$ がどの程度の大きさになるのか定量的に評価する手段は与えていない。統計学では、推定誤差がある区間 $[\alpha, \beta]$ に入る確率

$$P(\alpha \leq \bar{X}_n - \mu \leq \beta) \quad (1.11)$$

を計算することによって推定誤差を定量的に評価する（詳しい手順については「推定」の章で説明する）。確率 (1.11) の正確な計算は一般には困難であるが、サンプル数が十分大きい場合には、ある関数の定積分で近似できることが知られている。このことを具体的に述べたのが次の**中心極限定理**である。

定理 1.5 (中心極限定理). X_1, X_2, \dots を独立同分布な確率変数列とし、その平均を μ 、標準偏差を σ とする。このとき、すべての実数 $a < b$ に対して

$$P\left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \quad (n \rightarrow \infty) \quad (1.12)$$

が成り立つ。

中心極限定理より、サンプル数 n が十分大きければ、確率

$$P\left(a \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq b \frac{\sigma}{\sqrt{n}}\right) \quad (1.13)$$

は定積分

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \quad (1.14)$$

1 極限定理

によって近似できる。積分(1.14)の被積分関数 $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ は**標準正規密度(関数)**と呼ばれており、Rでは関数 `dnorm()` で計算できる。また、定積分(1.14)の値は関数 `pnorm()` を用いてコマンド `pnorm(b)-pnorm(a)` で計算できる。

注意 1.6. 詳細は次章で説明するが、各実数 $a \leq b$ について区間 $[a, b]$ に値が入る確率が定積分(1.14)で与えられるような確率変数を**標準正規確率変数**と呼び、標準正規確率変数の分布の仕方を**標準正規分布**と呼ぶ。中心極限定理の意味するところは、 X_i の分布が何であってもサンプル数 n が十分大きければ、標本平均を正規化した量 $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ の分布は標準正規分布で近似できるということである。

中心極限定理のシミュレーションによる確認は、ヒストグラムによる可視化を用いる方法がよく利用される。これは、 $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ をシミュレーションした際のヒストグラムのビン $[a, b]$ における高さが

$$\frac{1}{b-a} P \left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b \right) \quad (1.15)$$

で与えられることを利用する(関数 `hist()` でオプション `freq` を `FALSE` に指定した場合)。ビンの幅 $b - a$ が十分小さければ、中心極限定理が正しい限りビン $[a, b]$ におけるヒストグラムの高さは $\phi(a)$ で近似できるはずである。従って $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ のヒストグラムに標準正規密度 $\phi(x)$ を重ね書きすることで、近似の度合いを評価することができる。

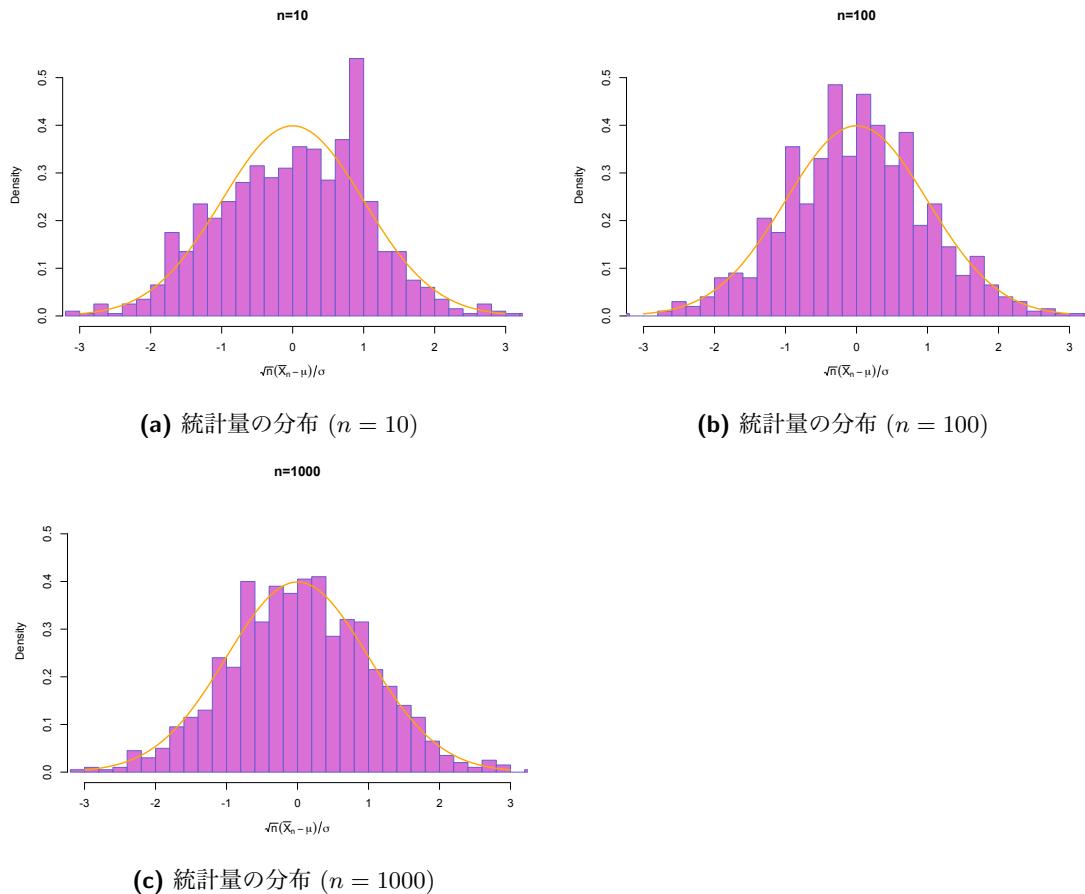


Figure 1.2: 中心極限定理のシミュレーション例

[Figure 1.2 を参照]

```
> ### 中心極限定理
> ### 偶数が出る確率が奇数の出る確率の 2倍となるサイコロを振ったときに
> ### 出る目の標本平均の分布を実験で確かめる
> set.seed(123)    # 亂数のシード値の指定
> omega <- 1:6     # サイコロの出る目の集合
> p <- rep(1:2, 3) # 出現確率の比 (奇数 1:偶数 2, 正規化していなくてもよい)
> (mu <- weighted.mean(omega, p)) # 理論上の平均
[1] 3.666667

> (sigma <- sqrt(weighted.mean(omega^2, p)-mu^2)) # 理論上の標準偏差
[1] 1.699673

> mymean <- function(n) # n回サイコロを振って出た目の標本平均を計算する関数
+   mean(sample(omega, size=n, prob=p, replace=TRUE))
> mc <- 1000 # Monte-Carlo 実験の繰り返し回数
> for(n in c(10, 100, 1000)){ # サンプル数を変えて実験
+   xbars <- replicate(mc, mymean(n)) # 実験を mc 回繰り返し標本平均を記録
+   hist(sqrt(n)*(xbars - mu)/sigma, breaks=25, freq=FALSE,
+         xlim=c(-3, 3), ylim=c(0, 0.55),
+         col="orchid", border="slateblue",
+         xlab=expression(sqrt(n)*(bar(X)[n]-mu)/sigma), main=paste0("n=", n))
+   curve(dnorm, add=TRUE, col="orange", lwd=2) # 理論曲線を重ねる
+ }

```

(mc-clt.r)

演習 1.3. 様々な乱数を用いて、中心極限定理を確認してみなさい。

1.4 重複対数の法則

これ以外の興味深いものとして、標本平均の振幅の挙動に関して**重複対数の法則**が知られている。

定理 1.7 (重複対数の法則). X_1, X_2, \dots を独立同分布な確率変数列とし、その平均を μ 、標準偏差を σ とする。このとき、

$$\limsup_{n \rightarrow \infty} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log \log n}} = 1 \quad a.s., \quad (\text{上極限}) \quad (1.16)$$

$$\liminf_{n \rightarrow \infty} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log \log n}} = -1 \quad a.s., \quad (\text{下極限}) \quad (1.17)$$

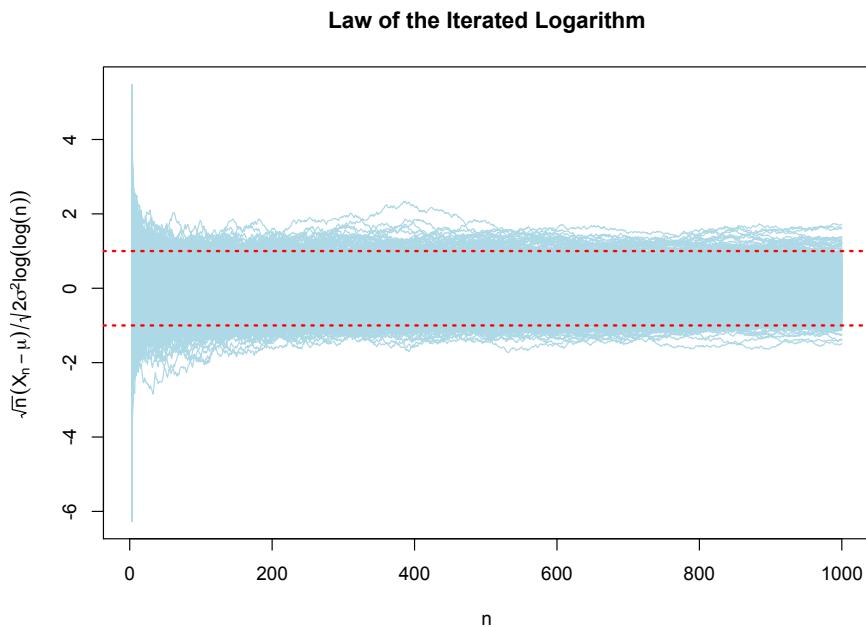
が成り立つ。より一般に、列

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log \log n}} \quad (n = 3, 4, \dots) \quad (1.18)$$

のある部分列の収束先となるような実数全体の集合を C とすると、 C が閉区間 $[-1, 1]$ に一致する確率は 1 である。

定理 1.7 の後半の主張²は **Hartman-Wintner の定理**として知られている。

²重複対数は $n \geq 3 (> e)$ でしか計算できないことに注意。

**Figure 1.3:** 重複対数の法則のシミュレーション例

[Figure 1.3 を参照]

```
> #### 重複対数の法則
> #### 偶数が出る確率が奇数の出る確率の2倍となるサイコロを振ったときに
> #### 出る目の標本平均に対する重複対数の法則の確認
> set.seed(111)      # 亂数のシード値の指定
> omega <- 1:6      # サイコロの出る目の集合
> p <- rep(1:2, 3) # 出現確率の比
> (mu <- weighted.mean(omega, p)) # 理論上の平均
[1] 3.666667
> (v <- weighted.mean(omega^2, p) - mu^2) # 理論上の分散
[1] 2.888889
> n <- 1000 # サンプル数
> mc <- 1000 # Monte-Carlo 実験の繰り返し回数
> x <- sample(omega, size=n*mc, prob=p, replace=TRUE) # 亂数のシミュレーション
> x <- matrix(x, n, mc) # 各列が1つの実験に対応
> n0 <- 3 # n>=n0からプロット
> y <- sqrt(n0:n)*(apply(x, 2, "cumsum")[n0:n, ]/(n0:n) - mu) /
+   sqrt(2*v*log(log(n:n))) # 列ごとにプロットする量を計算
> matplot(n0:n, y, type="l", lty=1, col="lightblue",
+           xlab=expression(n),
+           ylab=expression(sqrt(n)*(bar(X)[n]-mu)/sqrt(2*sigma^2*log(log(n)))), 
+           main="Law of the Iterated Logarithm") # 複数のパスを同時にプロット
> abline(h=c(-1,1), lty="dotted", col="red", lwd=2) # y=+-1のプロット
```

(mc-lil.r)

演習 1.4. 様々な乱数を用いて、重複対数の法則を確認してみなさい。

1.5 少数の法則

少数の法則とは、滅多に起こらない事象が起こる回数の分布に関する法則である。例えば、ある製品の不良品率 p はとても小さいとする。1日に n 個(非常に多数とする)生産するとき、不良品の個数を S_n と書くことにする。不良品は平均的には1日 $\lambda = np$ 個発生するが、日によって不良品の個数 S_n には多少のばらつきが生じる。従って S_n は確率変数であるが、 S_n がとる値の確率法則は、強度 λ の **Poisson 分布**で近似できることが知られている。これを正確に述べたのが次の**少数の法則**である。

定理 1.8 (少数の法則). X_1, X_2, \dots, X_n を独立な確率変数列とし、各 $i = 1, 2, \dots, n$ について X_i は確率 $p_{n,i}$ で 1 を、確率 $1 - p_{n,i}$ で 0 をとるとする:

$$P(X_i = 1) = p_{n,i}, \quad P(X_i = 0) = 1 - p_{n,i} \quad (i = 1, 2, \dots, n). \quad (1.19)$$

このとき、ある正の実数 λ が存在して、 $n \rightarrow \infty$ のとき

$$\max_{i=1,2,\dots,n} p_{n,i} \rightarrow 0, \quad \sum_{i=1}^n p_{n,i} \rightarrow \lambda \quad (1.20)$$

が成り立つならば、任意の 0 以上の整数 k に対して

$$P\left(\sum_{i=1}^n X_i = k\right) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!} \quad (n \rightarrow \infty) \quad (1.21)$$

が成り立つ。

上の定理において、 $\sum_{i=1}^n X_i$ が本小節冒頭の例の S_n に対応する。

注意 1.9. 詳細は次章で説明するが、取りうる値が 0 以上の整数全体で、値が整数 $k \geq 0$ となる確率が

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (1.22)$$

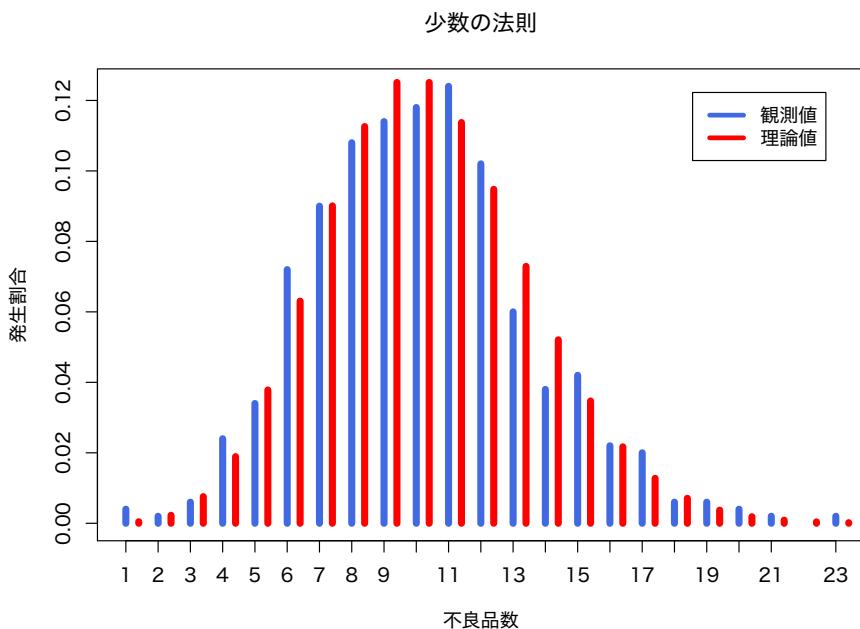
で与えられる確率変数 X を強度 λ の **Poisson型確率変数**と呼び、その値の分布の仕方を強度 λ の Poisson 分布と呼ぶ。式 (1.22) は関数 `dpois()` で計算できる。

[Figure 1.4 を参照]

```
> ### 少数の法則
> ### 大きさ 1 の 2 項分布を使う
> set.seed(123) # 亂数のシード値の指定
> n <- 5000    # 1日の総生産量
> p <- 0.002   # 不良品の発生確率
> mc <- 5*50*2 # 実験回数(週 5 日 x 50 週間 x 2 年操業に対応)
> x <- replicate(mc, sum(rbinom(n, 1, p))) # 実験を mc 回実行して不良品数を記録
> (A <- table(x)) # 不良品数の度数分布表を作成

x
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 23
 2  1  3 12 17 36 45 54 57 59 62 51 30 19 21 11 10  3  3  2  1  1

> ### それぞれの不良品数が生じた日数の割合のグラフを作成
> par(family = "HiraginoSans-W4") # 日本語フォントの指定
> plot(A/mc, type="h", col="royalblue", lwd=6,
+       xlab="不良品数", ylab="発生割合", main="少数の法則")
```

**Figure 1.4:** 少数の法則のシミュレーション例

```
> lines(min(x):max(x)+0.4, dpois(min(x):max(x), n*p), type="h",
+       col="red", lwd=6) # 理論上の割合を上書き
> legend("topright", inset=1/20, # 位置をキーワードで指定
+        legend=c("観測値", "理論値"),
+        col=c("royalblue", "red"), lwd=4) # 凡例を作成
                                                 (mc-lsn.r)
```

演習 1.5. 少数の法則について調べてみよう.

1. どのような事例がこの法則にあてはまるか調べてみなさい.
2. 個々の確率が比較的大きな場合に個数の分布がどのようなになるか調べてみなさい.

1.6 補遺

1.6.1 参考文献

確率論に関する参考文献をいくつか追記しておく.

- [1] Patrick Billingsley. *Convergence of probability measures*. 2nd. New York: Wiley, 1999.
- [2] 福島正俊. **確率論 (第 5 版)**. 東京: 裳華房, 2006.
- [3] U. リゲス (石田基広訳). **R の基礎とプログラミング技法**. 東京: 丸善出版, 2012.
- [4] 竹村彰通. **統計 (第 2 版)**. 東京: 共立出版, 2007.

1 極限定理

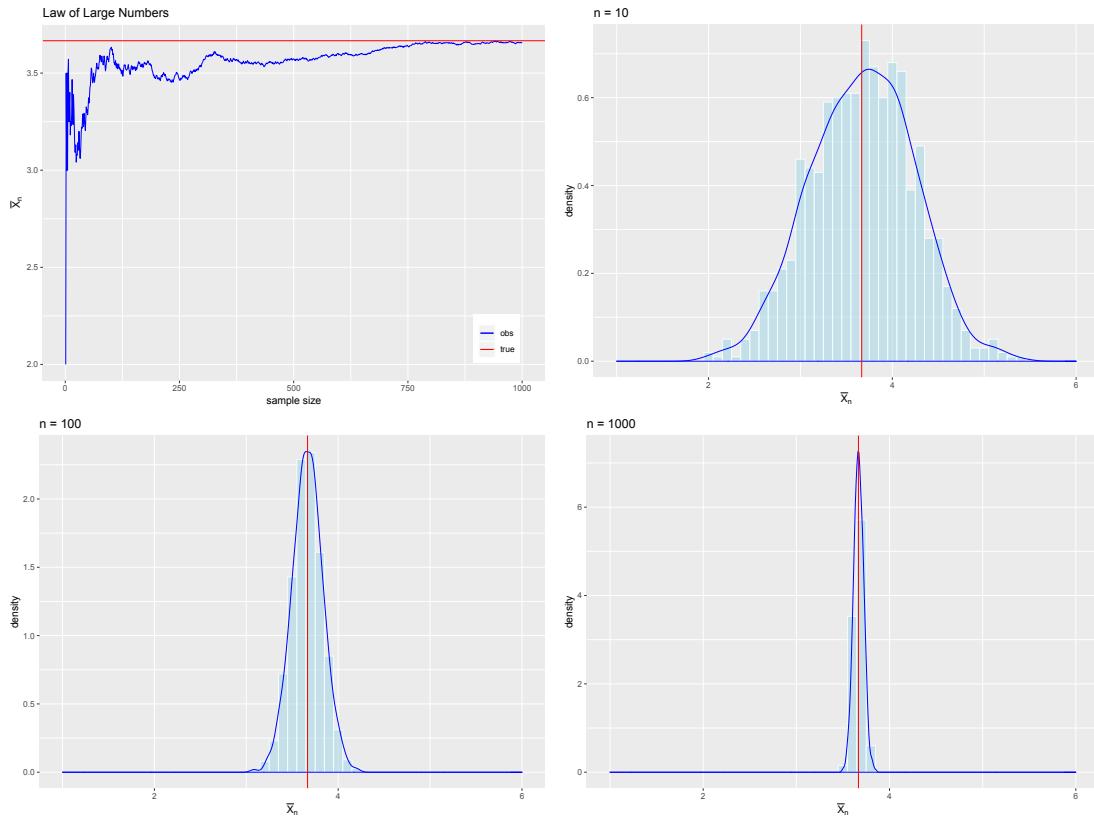


Figure 1.5: 大数の法則のシミュレーション例

1.6.2 ggplot2によるシミュレーション例

以下にはパッケージ `ggplot2` を用いて視覚化したシミュレーションの例を掲載しておく。

[Figure 1.5 を参照]

```
> #### 大数の法則
> #### ggplot2での描画
> require(tidyverse) # パッケージの読み込み
> set.seed(123) # 亂数のシード値の指定
> omega <- 1:6 # サイコロの出る目の集合
> p <- rep(1:2, 3) # 出現確率の比 (奇数 1:偶数 2, 正規化していなくてもよい)
> (mu <- weighted.mean(omega, p)) # 理論上の平均
[1] 3.666667

> #### 1回の実験
> #### サンプル数を大きくすると標本平均が理論上の平均に近づくことを確認
> n <- 1000 # 実験回数の最大値
> x <- sample(omega, size=n, prob=p, replace=TRUE) # 実験
> mydf <- data.frame(n=1:n,
+                      xbar=cumsum(x)/(1:n)) # サンプル数ごとの標本平均の計算
> ggplot(mydf,aes(x=n, y=xbar)) +
+   geom_line(aes(colour="obs")) +
+   geom_hline(aes(yintercept=mu, colour="true")) +
+   labs(x="sample size", y=expression(bar(X)[n]),
+        title="Law of Large Numbers") +
+   scale_color_manual(name=NULL, values=c("obs"="blue", "true"="red")) +
```

1 極限定理

```

+     theme(legend.position=c(.95,.05), legend.justification=c(1, 0))
> ### 複数回の実験
> ### サンプル数が大きければ、標本平均は理論平均に近い値を取ることの確認
> mymean <- function(n) # n回サイコロを振って出た目の標本平均を計算する関数
+   mean(sample(omega, size=n, prob=p, replace=TRUE))
> mc <- 1000 # Monte-Carlo 実験の繰り返し回数
> for(n in c(10,100,1000)){ # サンプル数を変えて実験
+   mydf <- data.frame(
+     xbar=replicate(mc, mymean(n))) # 同じ実験を mc 回繰り返し標本平均を記録
+   gg <- ggplot(mydf, aes(x=xbar)) +
+     geom_histogram(aes(y=..density..), binwidth=0.1,
+                   colour="azure", fill="lightblue", alpha=.6) +
+     geom_density(colour="blue", alpha=.3) + # 実験値の分布の概形
+     geom_vline(xintercept=mu, colour="red") + # 理論平均
+     xlim(1,6) + # 描画範囲を指定
+     labs(x=expression(bar(X)[n]), title=paste("n =", n))
+   print(gg)
+ }

```

(mcg-lln.r)

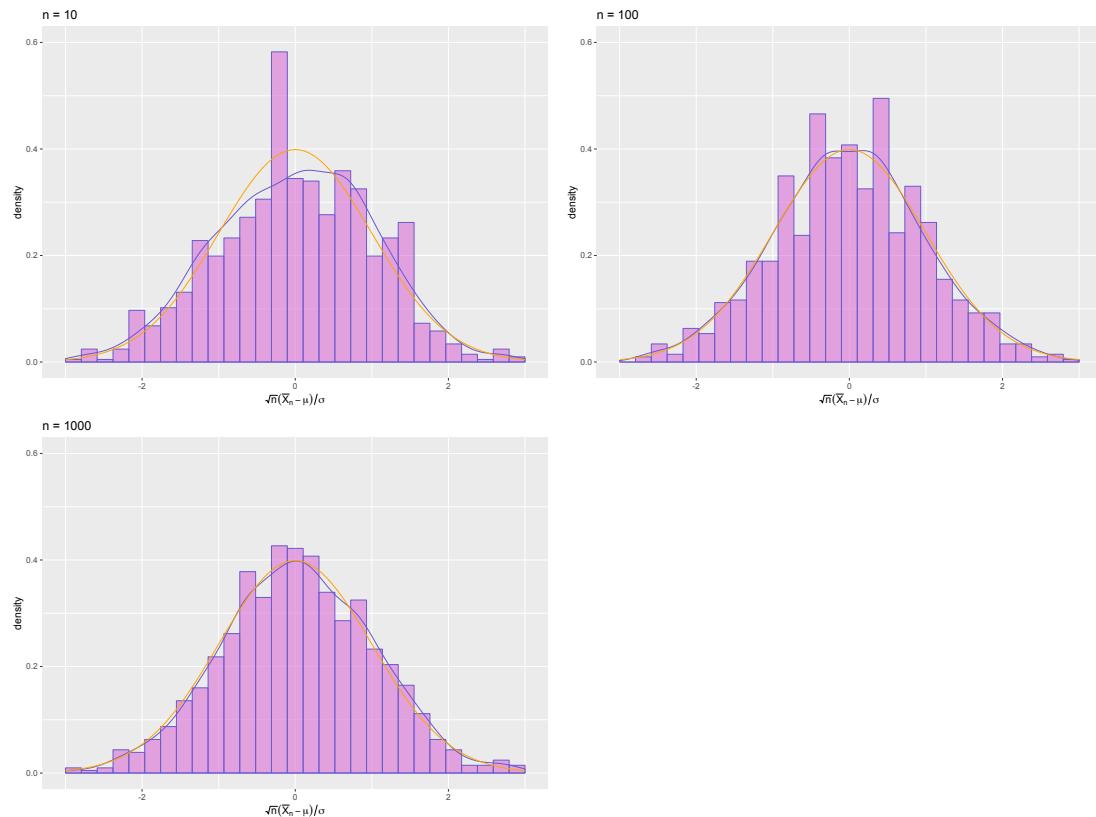


Figure 1.6: 中心極限定理のシミュレーション例

[Figure 1.6 を参照]

```

> ### 中心極限定理
> ### ggplot2での描画

```

1 極限定理

```

> require(tidyverse) # パッケージの読み込み
> set.seed(123)      # 亂数のシード値の指定
> omega <- 1:6       # サイコロの出る目の集合
> p <- rep(1:2, 3)   # 出現確率の比 (奇数 1:偶数 2, 正規化していなくてもよい)
> (mu <- weighted.mean(omega, p)) # 理論上の平均
[1] 3.666667

> (sigma <- sqrt(weighted.mean(omega^2, p)-mu^2)) # 理論上の標準偏差
[1] 1.699673

> mymean <- function(n) # n回サイコロを振って出た目の標本平均を計算する関数
+     mean(sample(omega, size=n, prob=p, replace=TRUE))
> mc <- 1000 # Monte-Carlo 実験の繰り返し回数
> for(n in c(10,100,1000)){ # サンプル数を変えて実験
+     mydf <- data.frame(
+         xbar=replicate(mc, mymean(n))) # 同じ実験を mc 回繰り返し標本平均を記録
+     gg <- ggplot(mydf, aes(x=sqrt(n)*(xbar - mu)/sigma)) +
+         geom_histogram(aes(y=..density..),# binwidth=0.2,
+                         colour="slateblue", fill="orchid", alpha=.6) +
+         geom_density(colour="slateblue", alpha=.3) + # 実験値の分布の概形
+         stat_function(fun=dnorm, colour="orange") +
+         xlim(-3, 3) + ylim(0,0.6) + # 描画範囲を指定
+         labs(x=expression(sqrt(n)*(bar(X)[n]-mu)/sigma),
+              title=paste("n =", n))
+     print(gg)
+ }

```

(mcg-clt.r)

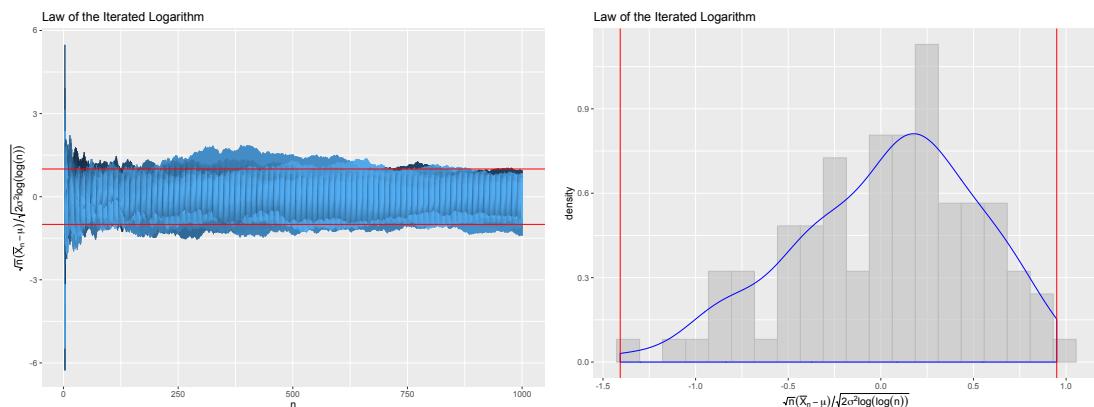


Figure 1.7: 重複対数の法則のシミュレーション例

[Figure 1.7 を参照]

```

> #### 重複対数の法則
> ####  ggplot2 での描画
> require(tidyverse) # パッケージの読み込み
> set.seed(111)      # 亂数のシード値の指定
> omega <- 1:6       # サイコロの出る目の集合
> p <- rep(1:2, 3)   # 出現確率の比
> (mu <- weighted.mean(omega, p)) # 理論上の平均

```

1 極限定理

```
[1] 3.666667
> (v <- weighted.mean(omega^2, p) - mu^2) # 理論上の分散
[1] 2.888889

> n <- 1000 # サンプル数
> mc <- 100 # 実験回数
> x <- sample(omega, size=n*mc, prob=p, replace=TRUE) # 亂数のシミュレーション
> x <- matrix(x, n, mc) # 各列が1つの実験に対応
> n0 <- 3 # n>=n0からプロット
> y <- sqrt(n0:n) * (apply(x, 2, "cumsum")[n0:n, ]/(n0:n) - mu) /
+   sqrt(2*v*log(log(n0:n))) # 列ごとにプロットする量を計算
> mydf <- gather(data.frame(x=n0:n, y), path, value, -x)
> ggplot(mydf) +
+   geom_line(aes(x=x, y=value, colour=as.numeric(factor(path)))) +
+   geom_hline(yintercept=c(-1,1), colour="red") + # y=+/-1 のプロット
+   labs(x=expression(n),
+         y=expression(sqrt(n)*(bar(X)[n]-mu)/sqrt(2*sigma^2*log(log(n)))),
+         title="Law of the Iterated Logarithm") +
+   theme(legend.position="none")
> #### 別の描き方の例
> myobs <- data.frame(obs=y[nrow(y),])
> ggplot(myobs, aes(x=obs)) +
+   geom_histogram(aes(y=..density..), bins=20,
+                 colour="gray", fill="gray", alpha=.6) +
+   geom_density(colour="blue") +
+   geom_vline(xintercept=with(myobs, max(obs)), colour="red") +
+   geom_vline(xintercept=with(myobs, min(obs)), colour="red") +
+   labs(x=expression(sqrt(n)*(bar(X)[n]-mu)/sqrt(2*sigma^2*log(log(n)))),
+        title="Law of the Iterated Logarithm")
```

(mcg-lil.r)

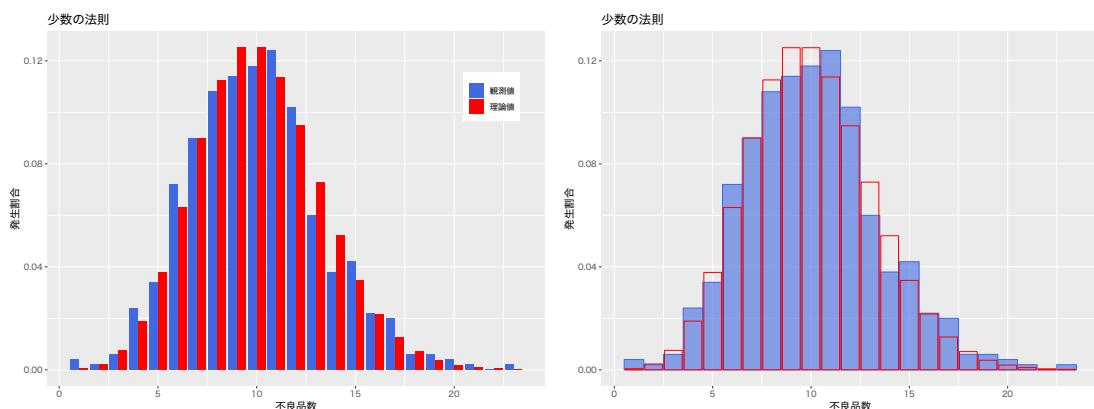


Figure 1.8: 少数の法則のシミュレーション例

[Figure 1.8 を参照]

```
> #### 少数の法則
> #### ggplot2での描画
> require(tidyverse) # パッケージの読み込み
> set.seed(123)
```

1 極限定理

```
> n <- 5000 # 1日の総生産量
> p <- 0.002 # 不良品の発生確率
> mc <- 5*50*2 # 実験回数(週5日x50週間x2年操業に対応)
> x <- replicate(mc, sum(rbinom(n, 1, p))) # 実験を mc 回実行して不良品数を記録
> num <- min(x):max(x)
> ### それぞれの不良品数が生じた日数の割合のグラフを作成
> mydf <- gather(data.frame(num=num,
+                               観測値=sapply(num, function(n)sum(x==n))/mc,
+                               理論値=dpois(num, n*p)),
+                               type, value, -num)
> ggplot(mydf, aes(x=num, y=value, fill=type)) +
+   geom_bar(stat="identity", position="dodge") +
+   scale_fill_manual(name=NULL, values=c("観測値"="royalblue", "理論値"="red")) +
+   labs(x="不良品数", y="発生割合", title="少数の法則") +
+   theme(legend.position=c(.95,.75), legend.justification=c(1, 0),
+         text=element_text(family="HiraginoSans-W4"))
> ### 別の描き方の例
> myobs <- data.frame(obs=x) # 観測値
> mythm <- data.frame(x=num, y=dpois(num, n*p)) # 理論値
> ggplot() +
+   geom_histogram(data=myobs, aes(x=obs,y=..density..), binwidth=1,
+                 colour="royalblue", fill="royalblue", alpha=.6) +
+   geom_bar(data=mythm, aes(x=num,y=y),stat="identity",
+            colour="red", alpha=0) +
+   labs(x="不良品数", y="発生割合", title="少数の法則") +
+   theme(text=element_text(family="HiraginoSans-W4"))
```

(mcg-lsn.r)

2 確率分布

2.1 確率変数と確率分布

前章で述べたように、乱数（ランダムに生成された数列）の数学的なモデル化には確率変数が用いられる。確率変数とは、値がランダムに決定される変数で、すべての実数 $a \leq b$ に対して、その値が区間 $[a, b]$ に含まれる確率があらかじめ定められている変数のことである。 X を確率変数とすると、定義より X が区間 $[a, b]$ ($a \leq b$) に含まれる確率が定まるから、その確率を $P(a \leq X \leq b)$ で表す。特に $a = b$ のとき、 $P(a \leq X \leq b)$ は $X = a$ となる確率を表すからそれを $P(X = a)$ で表すことにする。

確率変数 X に対して、各区間 $[a, b]$ ($a \leq b$) と、 X が区間 $[a, b]$ に含まれる確率 $P(a \leq X \leq b)$ との対応を示したものを、 X の**確率分布**または単に**分布**といい、 X はこの分布に**従う**という。観測の結果として定まる確率変数の実現値はランダムに決定されるため、その値自体には格別の意味はなく、現象の理解のためには値の出現しやすさの方にこそ興味がある。そのため、統計学では、確率分布の数学的モデリングを通じて現象の理解を試みることとなる。本章では、いくつかの基本的な確率分布の数理モデルを、Rにおけるシミュレーション方法とあわせて説明する。

2.2 離散分布

取りうる値が有限個、もしくは可算無限個（例えば整数値のみとする場合）であるような確率変数は**離散型**であるといい、対応する確率分布を**離散分布**または**離散確率分布**と呼ぶ。離散分布は、その分布に従う確率変数 X が取りうる値 x のそれぞれに対して、 $X = x$ となる確率 $P(X = x)$ を対応させる関数 $f(x) = P(X = x)$ を考えることで完全に決定される。この関数 f を**確率質量関数**あるいは単に**確率関数**と呼ぶ。

前章と同様に、離散型の確率変数に対してその平均は統計学において重要な概念である。確率変数の取りうる値が無限個あるかもしれない場合は、その定義には少し注意を要する。まず、有限もしくは可算無限個の要素をもつ集合 \mathcal{X} とその上で定義された実数値関数 φ に対して、級数

$$\sum_{x \in \mathcal{X}} \varphi(x) \tag{2.1}$$

を定義することから始める¹。まず \mathcal{X} が n 個の要素をもつ有限集合の場合、 \mathcal{X} の要素の適当な番号づけを x_1, \dots, x_n とし、

$$\sum_{x \in \mathcal{X}} \varphi(x) := \sum_{i=1}^n \varphi(x_i) \tag{2.2}$$

と定義する。加法の可換性より、この定義は \mathcal{X} の要素の番号付けの仕方によらない。次に、 \mathcal{X} が可算無限集合の場合、 \mathcal{X} の要素のある番号付け x_1, x_2, \dots に対して級数 $\sum_{i=1}^{\infty} \varphi(x_i)$ が絶対収束するとき、

$$\sum_{x \in \mathcal{X}} \varphi(x) := \sum_{i=1}^{\infty} \varphi(x_i) \tag{2.3}$$

¹ \mathcal{X} は空集合でないとする。

2 確率分布

と定義し、絶対収束しない場合は定義できないとする。級数 $\sum_{i=1}^{\infty} \varphi(x_i)$ が絶対収束する場合、正項級数の性質よりこの級数は \mathcal{X} の要素の番号付けの仕方によらずに絶対収束し、上式右辺の級数の値は番号付けの仕方によらずに定まる。

以上の準備の下、離散型の確率変数 X の平均を以下のように定義する。確率変数 X の取りうる値全体からなる集合を \mathcal{X} とする。級数 $\sum_{x \in \mathcal{X}} xP(X = x)$ が定義できるとき、 X の平均を

$$E[X] := \sum_{x \in \mathcal{X}} xP(X = x) \quad (2.4)$$

で定義する。平均は**期待値**とも呼ばれる。級数 $\sum_{x \in \mathcal{X}} xP(X = x)$ が定義できない場合、 X は平均をもたない。より一般に、 X の関数 $\varphi(X)$ に対して、級数 $\sum_{x \in \mathcal{X}} \varphi(x)P(X = x)$ が定義できるとき、 $\varphi(X)$ の期待値を

$$E[\varphi(X)] := \sum_{x \in \mathcal{X}} \varphi(x)P(X = x) \quad (2.5)$$

で定義する。特に、正の整数 p に対して

$$E[X^p] = \sum_{x \in \mathcal{X}} x^p P(X = x) \quad (2.6)$$

であり、これを p 次の**モーメント**あるいは**積率**と呼ぶ。級数 $\sum_{x \in \mathcal{X}} x^p P(X = x)$ が定義できないとき、 X は p 次のモーメントをもたない。一般に、ある正整数 p に対して X が p 次のモーメントをもてば、 $q \leq p$ なるすべての正整数 q に対して X は q 次のモーメントをもつことが知られている。

前章と同様に、離散型の確率変数 X に対して、平均からのはらつき具合を定量化した指標として、 X の**分散**を

$$\text{Var}[X] := E[(X - E[X])^2] = \sum_{x \in \mathcal{X}} (x - E[X])^2 P(X = x) \quad (2.7)$$

で定義する(分散は X が 2 次のモーメントをもつときのみ定義できる)。分散の平方根 $\sqrt{\text{Var}[X]}$ を**標準偏差**と呼ぶ。取りうる値が有限個の場合と同様に、次の恒等式が成り立つ:

$$\text{Var}[X] = E[X^2] - (E[X])^2. \quad (2.8)$$

離散分布の平均、モーメント、分散、標準偏差は、その分布に従う確率変数の平均、モーメント、分散、標準偏差で定義する。定義より明らかのように、離散型の確率変数の平均、モーメント、分散、標準偏差はその分布のみに依存して定まるため、この定義は確率変数の選び方によらない。むしろ、確率変数の平均、モーメント、分散、標準偏差はその確率変数が従う分布のものとみなす方が本質的である。

前章では有限個の値をとる確率変数列のみについて大数の法則、中心極限定理および重複対数の法則を説明したが、実際にはそれらの主張は、確率変数たちが 2 次のモーメントをもつ限り、離散型の確率変数の列についても成り立つ²。ただし、一般の場合の確率変数列の独立性と同分布性は以下のように定義する。まず、 n 個の確率変数 X_1, X_2, \dots, X_n が**独立**であるとは、 $a_i \leq b_i$ ($i = 1, \dots, n$) なる任意の実数 $a_1, b_1, \dots, a_n, b_n$ に対して、

$$\begin{aligned} P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \dots, a_n \leq X_n \leq b_n) \\ = P(a_1 \leq X_1 \leq b_1)P(a_2 \leq X_2 \leq b_2) \cdots P(a_n \leq X_n \leq b_n) \end{aligned} \quad (2.9)$$

² 2 次のモーメントをもたない場合、中心極限定理と重複対数の法則は成立しない(そもそも分散が定義できない)。大数の強法則は平均が存在すれば成立する。

2 確率分布

が成り立つことをいう。ここに、(2.9) の左辺は「 X_1 が区間 $[a_1, b_1]$ に値をとり、 X_2 が区間 $[a_2, b_2]$ に値をとり、…、 X_n が区間 $[a_n, b_n]$ に値をとる」という事象が起きる確率を表す。次に、 X_1, X_2, \dots, X_n が**同分布**であるとは、 $a \leq b$ なる任意の実数 a, b に対して

$$P(a \leq X_1 \leq b) = P(a \leq X_2 \leq b) = \dots = P(a \leq X_n \leq b) \quad (2.10)$$

が成り立つことをいう。確率変数の無限列に対する独立性および同分布性の定義は前章と同様のため省略する。離散型の確率変数列の場合、これらの定義は $a = b$ の場合のみ確認すれば十分であることがわかるため、前章での定義と同じ形式となる。

以下に代表的な離散分布を列挙する。

2.2.1 離散一様分布

x_1, \dots, x_n を相異なる実数とする。取りうる値が x_1, \dots, x_n であり、確率関数が

$$f(x) = \frac{1}{n}, \quad x \in \{x_1, \dots, x_n\} \quad (2.11)$$

で与えられる離散分布を、集合 $\{x_1, \dots, x_n\}$ 上の**離散一様分布**と呼ぶ。平均は $\bar{x} := 1/n \sum_{i=1}^n x_i$ 、分散は $1/n \sum_{i=1}^n (x_i - \bar{x})^2$ で与えられる。

例えば、歪みのないサイコロを 1 回投げたときに出る目の分布は、集合 $\{1, \dots, 6\}$ 上の離散一様分布に従う。

離散一様分布に従う乱数の発生には、前章で説明した関数 `sample()` が利用できる（オプション `replace` を `TRUE` に指定して使う）。

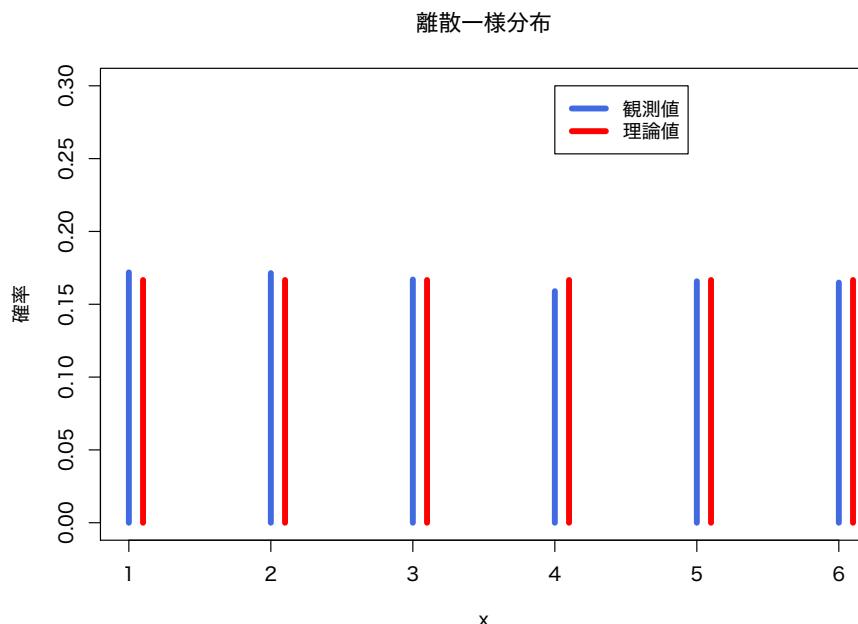


Figure 2.1: 離散一様分布の例

[Figure 2.1 を参照]

```
> #### 離散一様分布
> a <- 1:6 # サンプリング対象の集合をベクトルとして定義
```

```

> set.seed(123) # 亂数の初期値を指定
> sample(a, size=20, replace=TRUE) # 20個の離散一様分布のシミュレーション

[1] 3 6 3 2 2 6 3 5 4 6 6 1 2 3 5 3 3 1 4 1

> ## 統計的性質の確認
> x <- sample(a, size=10000, replace=TRUE)
> mean(x) # mean(a) = 3.5に近い(大数の法則)

[1] 3.4701

> (A <- table(x)/10000) # 出現確率ごとの表(度数分布表)を作成

x
 1      2      3      4      5      6 
0.1719 0.1714 0.1670 0.1590 0.1658 0.1649

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> plot(A, type="h", lwd=5, col="royalblue", ylab="確率",
+       main="離散一様分布", ylim=c(0, 0.3))
> lines(a+0.1, rep(1/length(a), length(a)), type="h", col="red", lwd=5)
> legend(4, 0.3, legend=c("観測値", "理論値"),
+         col=c("royalblue", "red"), lwd=5) # 凡例を作成

```

(dist-dunif.r)

演習 2.1. 離散一様分布の平均と分散の計算式が正しいことを定義に従って確認せよ。

2.2.2 二項分布

n を正の整数, p を 0 以上 1 以下の実数とする。取りうる値が $0, 1, \dots, n$ であり, 確率関数が

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n \quad (2.12)$$

で与えられる離散分布を, 試行回数 n , 成功確率 p の**二項分布**と呼ぶ。平均は np , 分散は $np(1-p)$ で与えられる。特に, 試行回数 1 の二項分布を **Bernoulli 分布**と呼ぶ。

例えば, 表が出る確率が p のコインを n 回投げたときに表が出る回数は試行回数 n , 成功確率 p の二項分布に従う。

前章でも述べたように, 二項分布に従う乱数の発生には関数 `rbinom()` を用いる。なお, 原則として, ある確率分布に従う乱数を生成するための R の関数の命名規則は, 「r + その乱数が従う分布の名前の省略形」となっている(離散一様分布など一部例外がある)。また, 離散分布の場合, その確率関数を計算するための関数が, 同じ省略形の文頭に d をつけることで得られる。例えば, 二項分布の確率関数は `dbinom()` で計算できる。

[Figure 2.2 を参照]

```

> ### 二項分布
> set.seed(123) # 亂数の初期値を指定
> rbinom(10, size=1, prob=0.5) # Bernoulli 分布のシミュレーション

[1] 0 1 0 1 1 0 1 1 1 0

> rbinom(10, size=1, prob=0.2) # 成功確率を小さくしてみる

[1] 1 0 0 0 0 1 0 0 0 1

```

2 確率分布

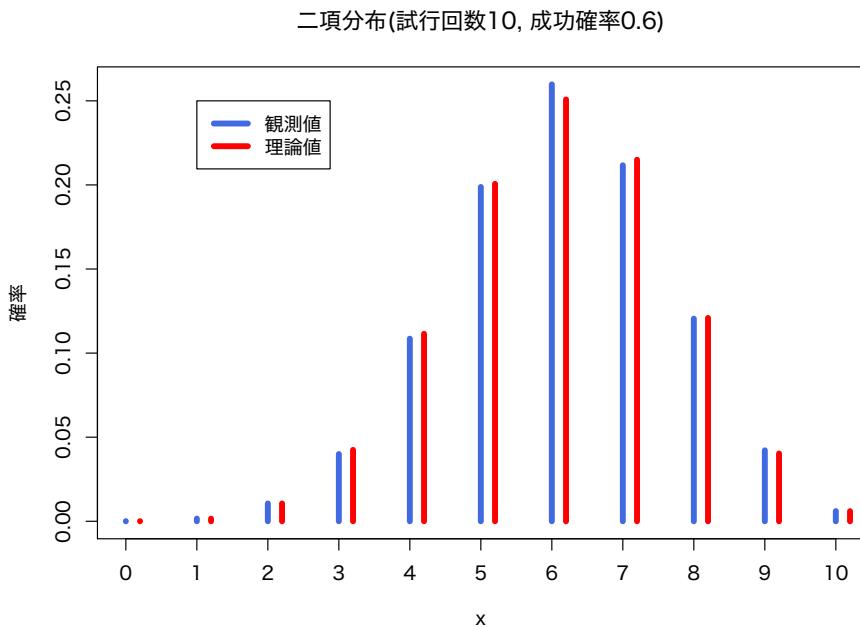


Figure 2.2: 二項分布の例

```
> rbinom(20, size=5, prob=0.6) # 20個の二項分布のシミュレーション
[1] 2 2 3 0 3 2 3 3 4 4 1 2 2 2 5 3 2 4 4 4
> ## 統計的性質の確認
> m <- 10
> p <- 0.6
> x <- rbinom(10000, size=m, prob=p)
> mean(x) # 10 * 0.6 = 6 に近い(大数の法則)
[1] 6.0167
> (A <- table(x)/10000) # 出現確率ごとの表(度数分布表)を作成
x
0      1      2      3      4      5      6      7      8      9      10 
0.0001 0.0016 0.0106 0.0400 0.1086 0.1988 0.2598 0.2117 0.1205 0.0422 0.0061 
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> plot(A, type="h", lwd=5, col="royalblue", ylab="確率",
+       main=paste0("二項分布(試行回数", m, ", 成功確率", p, ")"))
> lines(0:10 + 0.2, dbinom(0:10, size=m, prob=p),
+        type="h", col="red", lwd=5) # 理論上の出現確率
> legend(1, 0.25, legend=c("観測値", "理論値"),
+         col=c("royalblue", "red"), lwd=5) # 凡例を作成

```

(dist-binom.r)

演習 2.2. 二項分布の平均と分散の計算式が正しいことを定義に従って確認せよ。また、確率関数が $\sum_{x=0}^n f(x) = 1$ を満たすことを確認せよ。

2.2.3 Poisson 分布

λ を正の実数とする。取りうる値が 0 以上の整数であり、確率関数が

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots \quad (2.13)$$

で与えられる離散分布をパラメータ λ の **Poisson 分布** と呼び、記号 $P_o(\lambda)$ で表す。 λ は**強度**と呼ばれることがある。平均、分散はともに λ で与えられる。

例えば、放射性物質から一定時間に放射される粒子の数や、一定期間に起こる交通事故の数などは Poisson 分布に従うことが知られている。また、前章で観察したように、発生確率が低い事象が十分長い期間のあいだに起こる回数の分布は Poisson 分布で近似できる。

Poisson 分布に従う乱数の発生には関数 `rpois()` を用いる。

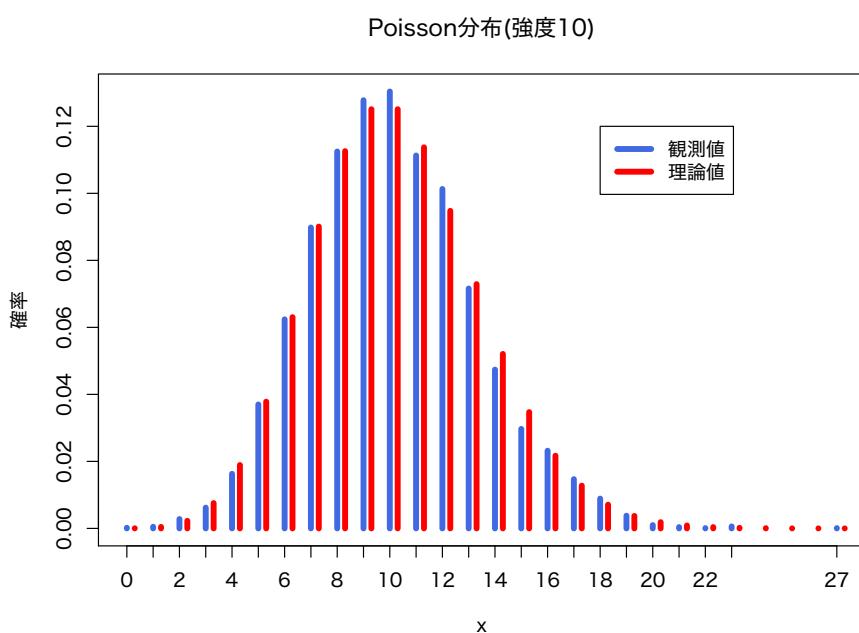


Figure 2.3: Poisson 分布の例

[Figure 2.3 を参照]

```
> ### Poisson 分布
> set.seed(12345) # 亂数の初期値を指定
> rpois(10, lambda=1) # 強度 1 の Poisson 分布に従う乱数を 10 個発生
[1] 1 2 2 2 1 0 0 1 1 4
> rpois(20, lambda=10) # 強度 10 の Poisson 分布に従う乱数を 20 個発生
[1] 4 11 7 8 11 9 10 9 11 13 10 12 14 7 4 15 8 11 11 9
> ## 統計的性質の確認
> lambda <- 10
> x <- rpois(10000, lambda=lambda)
> mean(x) # lambda=10 に近い (大数の法則)
[1] 10.0125
```

2 確率分布

```
> (A <- table(x)/10000) # 出現確率ごとの表(度数分布表)を作成
x
  0      1      2      3      4      5      6      7      8      9      10
0.0002 0.0005 0.0028 0.0062 0.0163 0.0370 0.0624 0.0898 0.1125 0.1278 0.1304
  11     12     13     14     15     16     17     18     19     20     21
0.1113 0.1013 0.0716 0.0474 0.0297 0.0232 0.0147 0.0089 0.0038 0.0010 0.0004
  22     23     27
0.0001 0.0006 0.0001

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> plot(A, type="h", lwd=5, col="royalblue", ylab="確率",
+       main=paste0("Poisson 分布(強度", lambda, ")"))
> lines(min(x):max(x) + 0.3, dpois(min(x):max(x), lambda=lambda),
+        type="h", col="red", lwd=5) # 理論上の出現確率
> legend(18, 0.12, legend=c("観測値", "理論値"),
+         col=c("royalblue", "red"), lwd=5) # 凡例を作成

```

(dist-pois.r)

演習 2.3. Poisson 分布の平均と分散の計算式が正しいことを定義に従って確認せよ。また、確率関数が $\sum_{x=0}^{\infty} f(x) = 1$ を満たすことを確認せよ。

2.2.4 幾何分布

$0 < p \leq 1$ とする。取りうる値が 0 以上の整数であり、確率関数が

$$f(x) = p(1-p)^x, \quad x = 0, 1, \dots \quad (2.14)$$

で与えられる離散分布を成功確率 p の**幾何分布**と呼ぶ。平均は $(1-p)/p$ 、分散は $(1-p)/p^2$ で与えられる。

例えば、表が出る確率が p のコインを投げ続けて初めて表が出るまでに出た裏の回数は、成功確率 p の幾何分布に従う。

幾何分布に従う乱数の発生には関数 `rgeom()` を用いる。

[Figure 2.4 を参照]

```
> #### 幾何分布
> set.seed(777) # 亂数の初期値を指定
> rgeom(20, prob=0.1) # 成功確率 0.1 の幾何分布に従う乱数を 20 個発生
[1] 3 19 0 7 2 2 2 9 12 8 4 5 6 7 10 8 0 8 7 3

> ## 統計的性質の確認
> p <- 0.4
> x <- rgeom(10000, prob=p)
> mean(x) # (1-p)/p=1.5 に近い(大数の法則)
[1] 1.4956

> (A <- table(x)/10000) # 出現確率ごとの表(度数分布表)を作成
x
  0      1      2      3      4      5      6      7      8      9      10
0.4025 0.2404 0.1416 0.0845 0.0541 0.0302 0.0178 0.0115 0.0067 0.0045 0.0029
  11     12     13     14     16     17
0.0014 0.0006 0.0006 0.0005 0.0001 0.0001
```

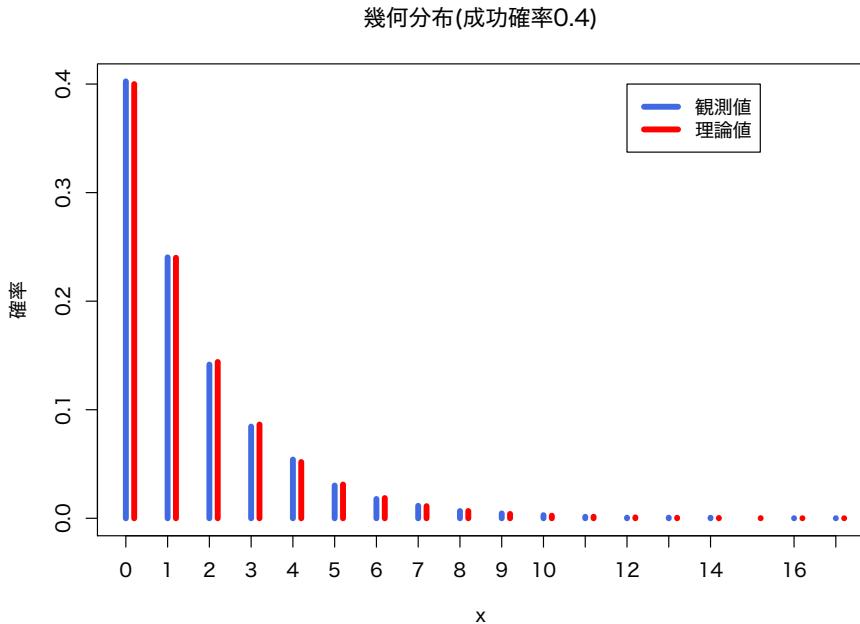


Figure 2.4: 幾何分布の例

```
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> plot(A, type="h", lwd=5, col="royalblue", ylab="確率",
+       main=paste0("幾何分布 (成功確率", p, ")"))
> lines(min(x):max(x) + 0.2, dgeom(min(x):max(x), prob=p),
+        type="h", col="red", lwd=5) # 理論上の出現確率
> legend(12, 0.4, legend=c("観測値", "理論値"),
+         col=c("royalblue", "red"), lwd=5) # 凡例を作成
```

(dist-geom.r)

演習 2.4. 幾何分布について調べてみよう.

1. 幾何分布の平均と分散の計算式が正しいことを定義に従って確認せよ. また, 確率関数が $\sum_{x=0}^{\infty} f(x) = 1$ を満たすことを確認せよ.
2. 幾何分布の一般化として負の二項分布と呼ばれる離散分布がある. この離散分布の確率関数, 平均, 分散および乱数の生成法について調べてみよ.

2.3 連続分布

実際のデータでは, 取りうる値が任意の実数またはある範囲の実数である場合, もしくは取りうる値のパターンが数多いため近似的にすべての実数値またはある範囲の実数値をとりうると考えられる場合が頻繁にある. 具体例としては, 株価, 気温, 風速などがある. このようなデータのモデル化には, しばしば連続分布に従う確率変数が用いられる. 一般に, 確率変数 X が**連続型**であるとは, 非負の値をとる実数直線上の関数 f があって, $a \leq b$ なるすべての実数 a, b に対して

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad (2.15)$$

2 確率分布

が成り立つことをいい、対応する確率分布を**連続分布**または**連続確率分布**と呼ぶ。また、関数 f をこの確率分布の**確率密度関数**あるいは単に**密度**と呼ぶ。

離散分布の場合と同様に、連続分布に対しても平均、分散、標準偏差の概念が定義される。まず、 X を連続型の確率変数、 f を X の分布の確率密度関数とする。積分 $\int_{-\infty}^{\infty} xf(x)dx$ が絶対収束するとき、 X の**平均**を

$$E[X] := \int_{-\infty}^{\infty} xf(x)dx \quad (2.16)$$

で定義する。平均は**期待値**とも呼ばれる。積分 $\int_{-\infty}^{\infty} xf(x)dx$ が絶対収束しないとき、 X は平均をもたない。より一般に、 X の関数 $\varphi(X)$ に対して、積分 $\int_{-\infty}^{\infty} \varphi(x)f(x)dx$ が絶対収束するとき、 $\varphi(X)$ の期待値を

$$E[\varphi(X)] := \int_{-\infty}^{\infty} \varphi(x)f(x)dx \quad (2.17)$$

で定義する。特に、正の整数 p に対して

$$E[X^p] = \int_{-\infty}^{\infty} x^p f(x)dx \quad (2.18)$$

を p 次の**モーメント**あるいは**積率**と呼ぶ。積分 $\int_{-\infty}^{\infty} x^p f(x)dx$ が絶対収束しないとき、 X は p 次のモーメントをもたない。離散型の確率変数の場合と同様に、一般にある正整数 p に対して X が p 次のモーメントをもてば、 $q \leq p$ なるすべての正整数 q に対して X は q 次のモーメントをもつことが知られている。

X が 2 次のモーメントをもつとき、 X の**分散**を

$$\text{Var}[X] := E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)dx \quad (2.19)$$

で定義する。分散の平方根 $\sqrt{\text{Var}[X]}$ を**標準偏差**と呼ぶ。離散型の確率変数の場合と同様に、次の恒等式が成り立つ：

$$\text{Var}[X] = E[X^2] - (E[X])^2. \quad (2.20)$$

連続分布の平均、モーメント、分散、標準偏差は、その分布に従う確率変数の平均、モーメント、分散、標準偏差で定義する。定義より明らかのように、連続型の確率変数の平均、モーメント、分散、標準偏差もその分布のみに依存して定まるため、この定義は確率変数の選び方によらない。離散分布の場合と同様に、むしろ確率変数の平均、モーメント、分散、標準偏差はその確率変数が従う分布のものとみなす方が本質的である。

離散型の確率変数の場合と同様に、大数の法則、中心極限定理および重複対数の法則は、確率変数が 2 次のモーメントをもつ限り、連続型の確率変数の列についても成り立つ³。

以下に代表的な連続分布を列挙する。

2.3.1 一様分布

$a < b$ とする。確率密度関数が

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \text{ のとき}, \\ 0 & \text{上記以外のとき} \end{cases} \quad (2.21)$$

³離散型の確率変数列の場合と同様に、2 次のモーメントをもたない場合、中心極限定理と重複対数の法則は成立しない（そもそも分散が定義できない）。大数の強法則は平均が存在すれば成立する。

2 確率分布

で与えられる連続分布を、区間 (a, b) 上の一様分布と呼び、記号 $U(a, b)$ で表す。平均は $(a + b)/2$ 、分散は $(b - a)^2/12$ で与えられる。

前章でも述べたように、一様分布に従う乱数の発生には関数 `runif()` を用いる。なお、連續分布の場合、分布の省略形の文頭に `d` をつけることで、確率密度関数を計算するための関数が得られる。例えば、一様分布の確率密度関数は関数 `dunif()` で計算できる。

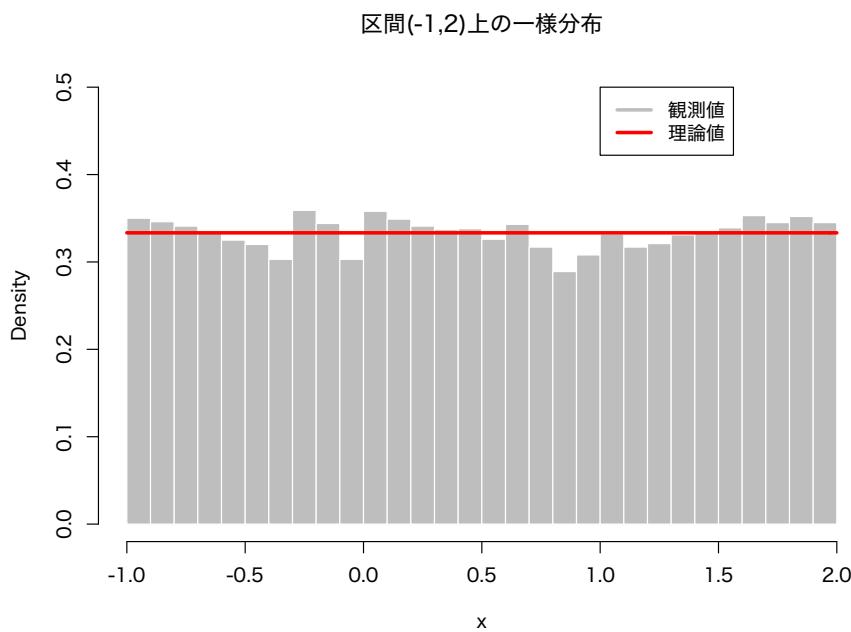


Figure 2.5: 一様分布の例

[Figure 2.5 を参照]

```
> ### 一様分布
> set.seed(1) # 亂数の初期値を指定
> runif(10) # 区間 (0, 1) 上の一様乱数を 10 個発生
[1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193 0.89838968
[7] 0.94467527 0.66079779 0.62911404 0.06178627

> ## 統計的性質の確認
> a <- -1
> b <- 2
> x <- runif(10000, min=a, max=b)
> mean(x) # (a+b)/2=0.5 に近い (大数の法則)
[1] 0.5001657

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=25, col="gray",
+       border="white", ylim=c(0, 0.5),
+       main=paste0("区間 (", a, ", ", b, ") 上の一様分布")) # ヒストグラム (密度表示)
> curve(dunif(x, min=a, max=b), add=TRUE,
+         col="red", lwd=3) # 理論上の確率密度関数
> legend(1, 0.5, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
```

(dist-unif.r)

2 確率分布

演習 2.5. 一様分布の平均と分散の計算式が正しいことを定義に従って確認せよ。また、確率密度関数が $\int_{-\infty}^{\infty} f(x)dx = 1$ を満たすことを確認せよ。

2.3.2 正規分布

μ を実数、 σ を正の実数とする。確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2.22)$$

で与えられる連続分布を、平均 μ 、分散 σ^2 の**正規分布**または**Gauss 分布**と呼び、記号 $N(\mu, \sigma^2)$ で表す。言葉通り、平均は μ 、分散は σ^2 で与えられる。特に、平均 0、分散 1 の正規分布を**標準正規分布**と呼ぶ。

例えば、物理実験等の観測誤差はしばしば正規分布でモデル化される。また、前章で観察したように、真の平均を標本平均で推定した際の推定誤差の確率分布は、サンプル数が大きくなるに従って正規分布に近づいていく(中心極限定理)。

正規分布に従う乱数の発生には関数 `rnorm()` を用いる。

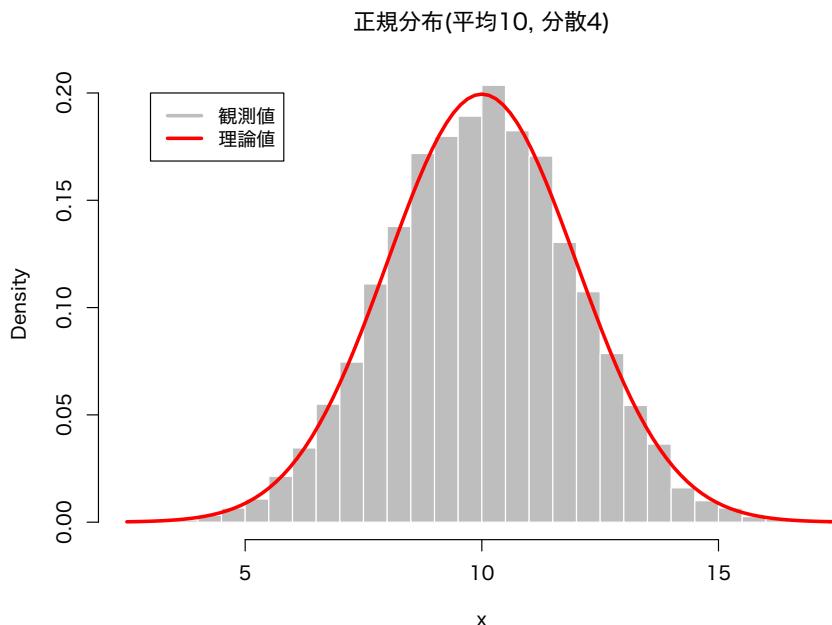


Figure 2.6: 正規分布の例

[Figure 2.6 を参照]

```
> ### 正規分布
> set.seed(20) # 亂数の初期値を指定
> rnorm(10) # 標準正規乱数を 10 個発生

[1] 1.1626853 -0.5859245 1.7854650 -1.3325937 -0.4465668 0.5696061
[7] -2.8897176 -0.8690183 -0.4617027 -0.5555409

> ## 統計的性質の確認
> mu <- 10
```

2 確率分布

```

> sigma <- 2
> x <- rnorm(10000, mean=mu, sd=sigma)
> mean(x) # mu=10に近い(大数の法則)
[1] 9.986371

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=25, col="gray", border="white",
+       main=paste0("正規分布(平均", mu, ", 分散", sigma^2, ")"))
# ヒストグラム(密度表示)
> curve(dnorm(x, mean=mu, sd=sigma), add=TRUE,
+         col="red", lwd=3) # 理論上の確率密度関数
> legend(3, 0.2, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成

```

(dist-norm.r)

Y を試行回数 n , 成功確率 p の二項分布に従う確率変数とすると, n が十分大きいとき $(Y - np)/\sqrt{np(1-p)}$ の分布は標準正規分布で近似できる。これは **de Moivre-Laplace の定理**として知られているが, 中心極限定理の特別な場合である。実際 X_1, \dots, X_n を成功確率 p の Bernoulli 分布に従う独立同分布な確率変数列とすると, $\sum_{i=1}^n X_i$ は試行回数 n , 成功確率 p の二項分布に従う⁴。Bernoulli 分布は平均 p , 分散 $p(1-p)$ であったから,

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\frac{1}{n} \sum_{i=1}^n X_i - p)}{\sqrt{p(1-p)}} \quad (2.23)$$

の分布は中心極限定理によって標準正規分布で近似できる。

[Figure 2.7 を参照]

```

> ### 二項分布の極限: 離散分布から連続分布へ
> set.seed(123)
> p <- 1/(7*pi)
> for(i in 1:4){
+   n <- 3*10^i
+   x <- (rbinom(1000000,n,prob=p)-n*p)/sqrt(n*p*(1-p))
+   hist(x,breaks=c(-Inf,seq(-3,3,0.25),Inf),freq=FALSE,
+         xlim=c(-3,3),xlab=n,col="lightblue",border="white")
+   curve(dnorm(x, mean=0, sd=1), add=TRUE,
+         col="red", lwd=3) # 理論上の確率密度関数
+ }

```

(dist-binom-norm.r)

ただし, p が非常に小さい場合, 特に np がそれほど大きくならない程度に p が小さい場合は, $(Y - np)/\sqrt{np(1-p)}$ の分布の正規近似よりも, Y の分布のパラメータ np の Poisson 分布による近似の方が精度がよい(後者は前章で述べた少数の法則の特殊な場合である)。

[Figure 2.8 を参照]

⁴ 実際, 確率変数 X_i は, 確率 p で表が出るコインを投げて表が出たら 1, 裏が出たら 0 を記録する試行に対応すると考えられるので, 確率変数列 X_1, \dots, X_n はこの試行を n 回独立に繰り返して記録される数字の列とみなせる。従って $\sum_{i=1}^n X_i$ はこのコインを n 回投げたときに表が出る回数に対応するから, 試行回数 n , 成功確率 p の二項分布に従う。

2 確率分布

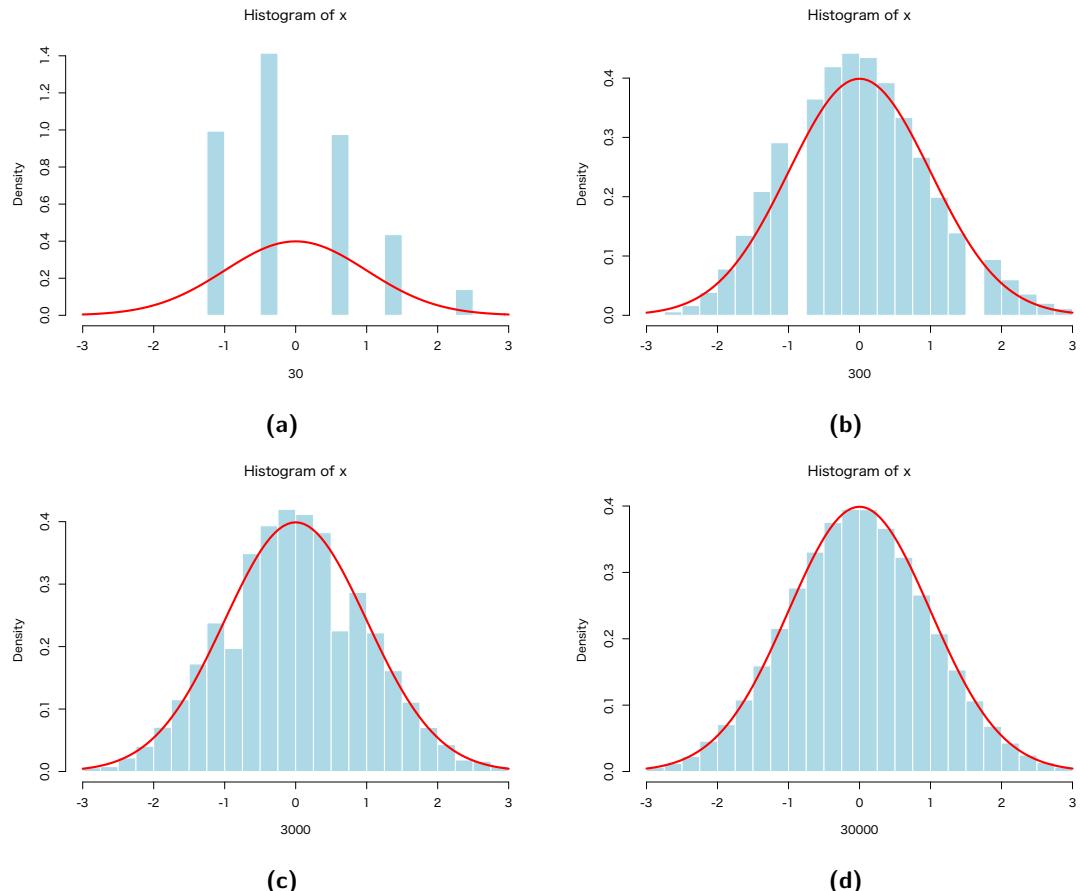


Figure 2.7: 二項分布の極限の例

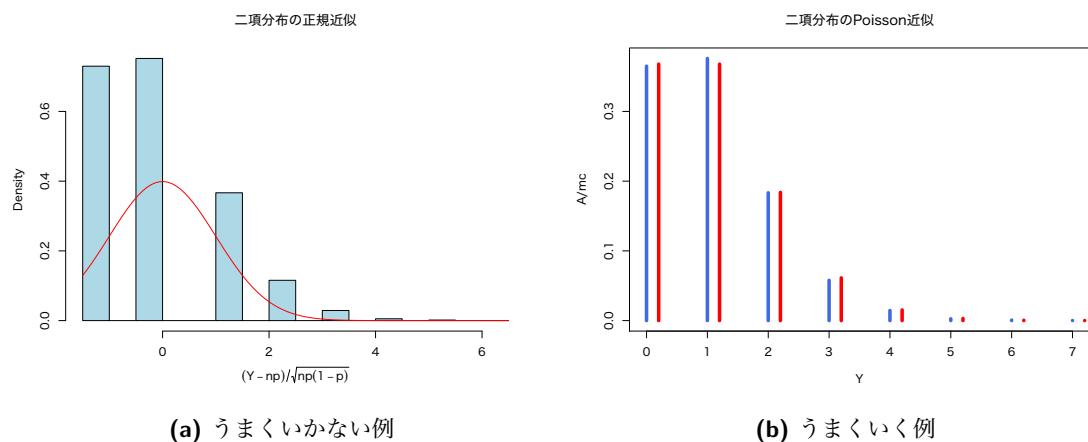


Figure 2.8: 正規近似と Poisson 近似の比較の例

```
> ### 二項分布の極限: Poisson 近似が有効な場合
> set.seed(123)
> n <- 1000
> p <- 0.001
> mc <- 10000
```

2 確率分布

```

> y <- rbinom(mc, n, p)
> ## まず正規近似を試す
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist((y - n * p) / sqrt(n * p * (1 - p)), freq=FALSE,
+       col="lightblue", main="二項分布の正規近似",
+       xlab=expression((Y-n*p)/sqrt(n*p*(1-p))))
> curve(dnorm, add=TRUE, col="red") # うまくいかない
> ## 次に Poisson 近似を試す
> (A <- table(y))

y
 0    1    2    3    4    5    6    7
3649 3760 1833  578  145   26    8    1

> plot(A/mc, type="h", lwd=5, col="royalblue",
+       main="二項分布の Poisson 近似", xlab=expression(Y))
> lines(min(y):max(y)+0.2, dpois(min(y):max(y), n * p), type="h",
+         col="red", lwd=5) # うまくいく

```

(dist-binom-pois.r)

上述のように、連続分布に従う独立同分布な確率変数列の標本平均も、2次モーメントが存在する限りは正規分布で近似できる。

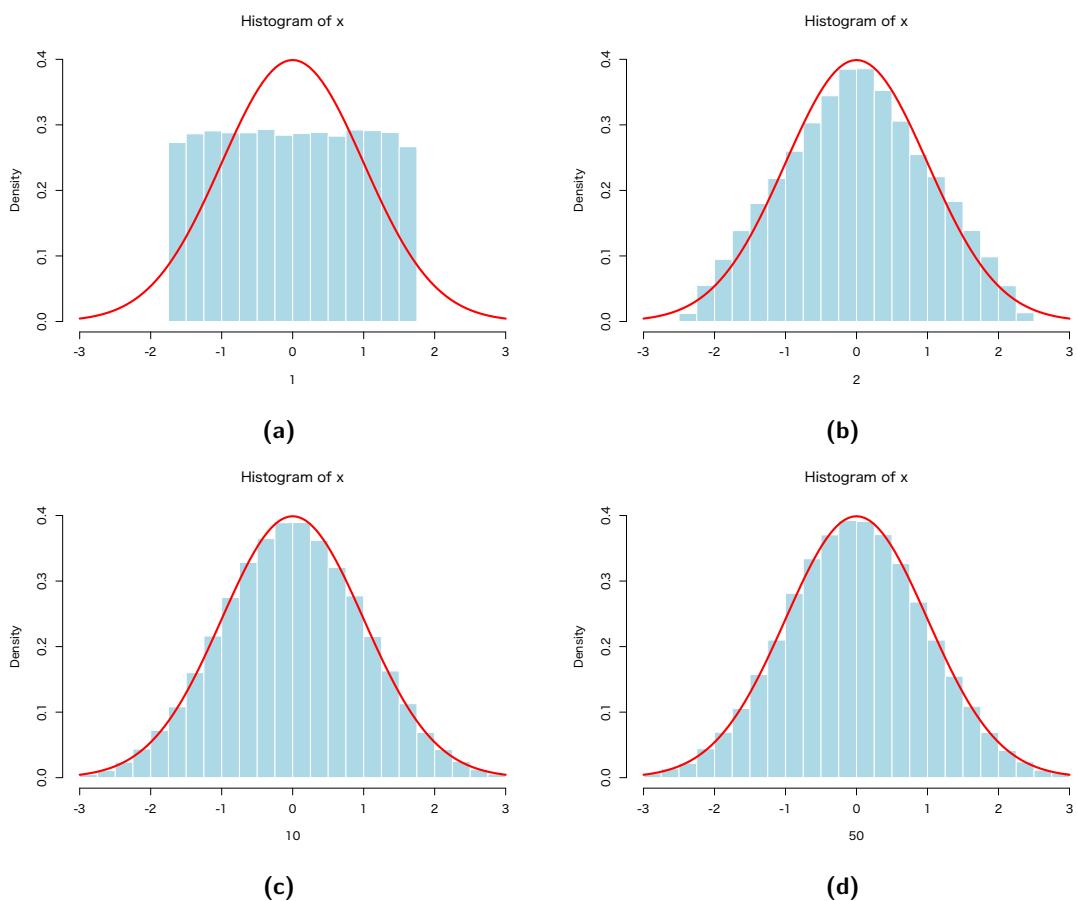


Figure 2.9: 一様分布の極限の例

[Figure 2.9 を参照]

```
> ### 一様乱数の標本平均に対する中心極限定理
> set.seed(111)
> mymean <- function(n) # n 個の一様乱数の標本平均を計算する関数
+   mean(runif(n))
> mc <- 100000 # シミュレーション回数
> for(n in c(1, 2, 10, 50)){
+   xbar <- replicate(mc, mymean(n))
+   x <- sqrt(n) * (xbar - 1/2)/sqrt(1/12)
+   hist(x, breaks=c(-Inf, seq(-3, 3, 0.25), Inf), freq=FALSE,
+         xlim=c(-3, 3), ylim=c(0, 0.4), xlab=n, col="lightblue", border="white")
+   curve(dnorm(x, mean=0, sd=1), add=TRUE,
+         col="red", lwd=3) # 理論上の確率密度関数
+ }
```

(dist-unif-norm.r)

演習 2.6. 正規分布について調べてみよう.

1. 正規分布の平均と分散の計算式が正しいことを定義に従って確認せよ. また, 確率密度関数 $f(x)$ が

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (2.24)$$

を満たすことを確認せよ.

2. U_1, U_2 を 2 つの独立な確率変数とし, ともに $(0, 1)$ 上の一様分布に従うとする. このとき

$$\begin{cases} X_1 = \sqrt{-2 \log(U_1)} \cos(2\pi U_2), \\ X_2 = \sqrt{-2 \log(U_1)} \sin(2\pi U_2) \end{cases} \quad (2.25)$$

とおくと, X_1, X_2 は独立かつともに標準正規分布に従うことが知られている (この変換を Box-Müller 変換と呼ぶ). このことをシミュレーションによって確かめてみよ.

2.3.3 ガンマ分布

ν, α を正の実数とする. 確率密度関数が

$$f(x) = \frac{1}{\Gamma(\nu)} \alpha^\nu x^{\nu-1} e^{-\alpha x} \quad (x > 0), \quad f(x) = 0 \quad (x \leq 0) \quad (2.26)$$

で与えられる連続分布を, パラメータ ν, α の**ガンマ分布**と呼び, 記号 $\Gamma(\nu, \alpha)$ や $G(\alpha, \nu)$ で表す. ただし, $\Gamma(\nu)$ はガンマ関数

$$\Gamma(\nu) = \int_0^{\infty} x^{\nu-1} e^{-x} dx \quad (2.27)$$

を表す. ν, α はそれぞれ**形状パラメータ**, **レート**と呼ばれることがある. 平均は ν/α , 分散は ν/α^2 で与えられる. 例えば, 体重の分布はガンマ分布に従うといわれている.

ガンマ分布に従う乱数の発生には関数 `rgamma()` を用いる.

2 確率分布

ガンマ分布 $\Gamma(4, 2)$

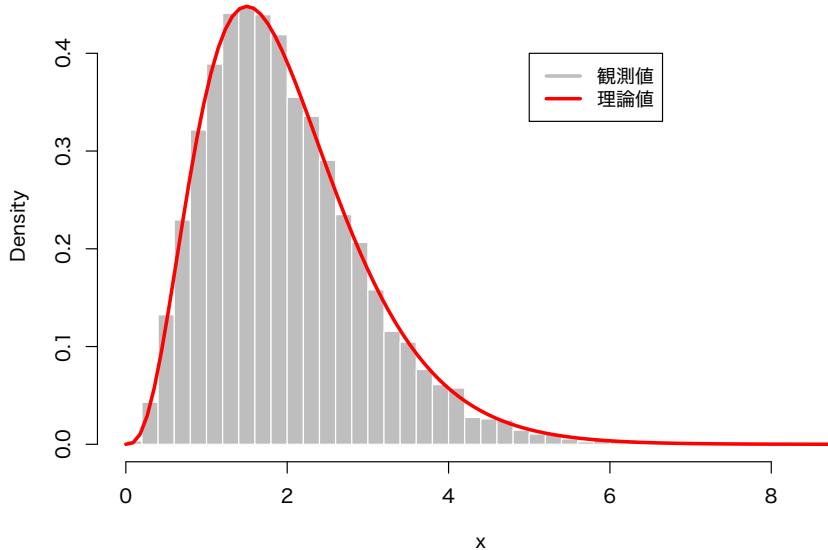


Figure 2.10: ガンマ分布の例

[Figure 2.10 を参照]

```
> #### ガンマ分布
> set.seed(123) # 亂数の初期値を指定
> rgamma(10, shape=3, rate=1) # ガンマ分布に従う乱数を 10 個発生
[1] 1.6923434 4.7360299 0.5422275 2.7086007 5.9471178 3.2818834 0.8998575
[8] 0.5148113 4.8100373 3.1012821

> ## 統計的性質の確認
> nu <- 4
> alpha <- 2
> x <- rgamma(10000, shape=nu, rate=alpha) # ガンマ乱数を 10000 個発生
> mean(x) # nu/alpha=2 に近い (大数の法則)
[1] 1.980431

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=50, col="gray", border="white",
+       main=bquote(paste("ガンマ分布 ", Gamma(.(nu), .(alpha))))) # ヒストグラム (密度表示)
> curve(dgamma(x, shape=nu, rate=alpha), add=TRUE,
+        col="red", lwd=3) # 理論上の確率密度関数
> legend(5, 0.4, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
```

(dist-gamma.r)

上の実行例におけるタイトルの作成では、文字列・数式・R オブジェクトを組み合わせた文字列を作成するために関数 `bquote()` を利用している。表現`.()`は数式と R オブジェクトを区別するために使われている。

ガンマ分布はいくつかの応用上重要な確率分布を特殊な場合として含む。正の実数 λ に対して、 $\Gamma(1, \lambda)$ をパラメータ λ の**指数分布**と呼び、記号 $\text{Exp}(\lambda)$ で表す。 λ は**レート**と呼ばれる

2 確率分布

ことがある。指数分布の平均、分散はそれぞれ λ^{-1} , λ^{-2} で与えられる。また、正の実数 k に対して、 $\Gamma(k/2, 1/2)$ を自由度 k の χ^2 分布と呼び、記号 $\chi^2(k)$ で表す⁵。自由度 k の χ^2 分布の平均、分散はそれぞれ k , $2k$ で与えられる。

χ^2 分布および指数分布はガンマ分布の特殊な場合であるから関数 `rgamma()` によって乱数を発生させられるが、便宜のためそれぞれ専用の乱数発生関数 `rchisq()` および `rexp()` が用意されている。

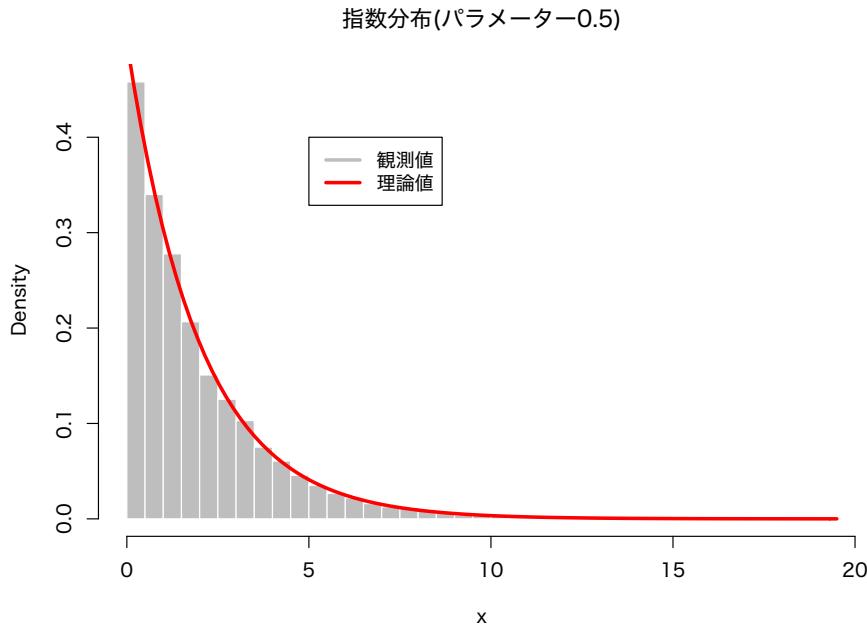


Figure 2.11: 指数分布の例

[Figure 2.11 を参照]

```
> ### 指数分布
> set.seed(20) # 亂数の初期値を指定
> rexp(10) # レート 1 の指数分布に従う乱数を 10 個発生
[1] 0.19336251 0.05832739 0.06330693 2.21143320 1.00352299 1.17344535
[7] 0.43105511 0.51559271 6.37169900 0.98173630

> ## 統計的性質の確認
> lambda <- 0.5
> x <- rexp(10000, rate=lambda) # レート 0.5 の指数乱数を 10000 個発生
> mean(x) # 1/lambda=2 に近い (大数の法則)
[1] 1.962623

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=50, col="gray", border="white",
+       main=paste0("指数分布 (パラメーター=", lambda, ")"))
# ヒストグラム (密度表示)
> curve(dexp(x, lambda), add=TRUE, col="red", lwd=3) # 理論上の確率密度関数
> legend(5, 0.4, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
```

(dist-exp.r)

⁵ χ^2 は「カイ二乗」と読む。

2 確率分布

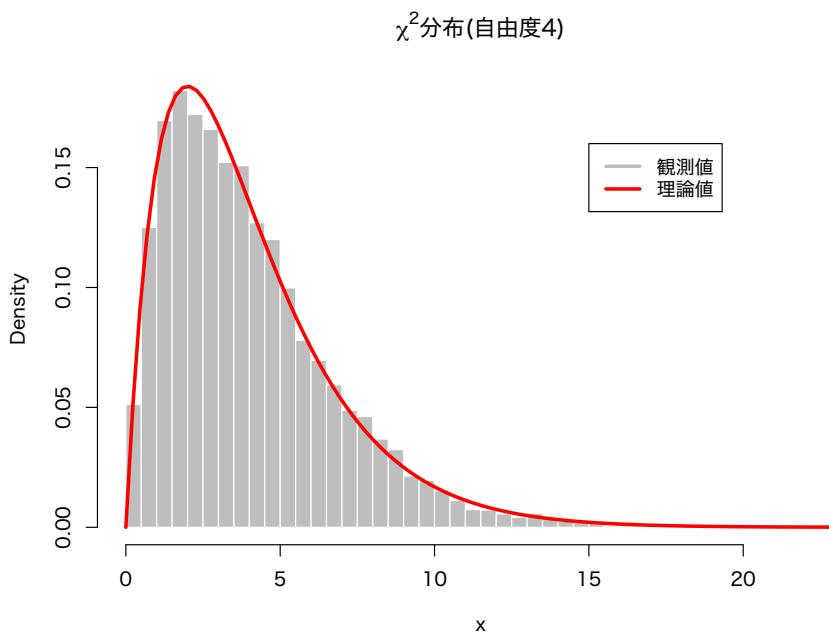


Figure 2.12: χ^2 分布の例

[Figure 2.12 を参照]

```
> ###  $\chi^2$  分布
> set.seed(20) # 亂数の初期値を指定
> rchisq(10, df=1) # 自由度 1 のカイ二乗分布に従う乱数を 10 個発生
[1] 2.47564812 0.38394375 1.60988258 0.29093644 0.67851954 0.01357661
[7] 1.27772421 0.56221273 0.63248955 0.18637919

> ## 統計的性質の確認
> k <- 4 # 自由度
> x <- rchisq(10000, df=k) # 自由度 4 のカイ二乗乱数を 10000 個発生
> mean(x) # k=4 に近い (大数の法則)
[1] 4.01317

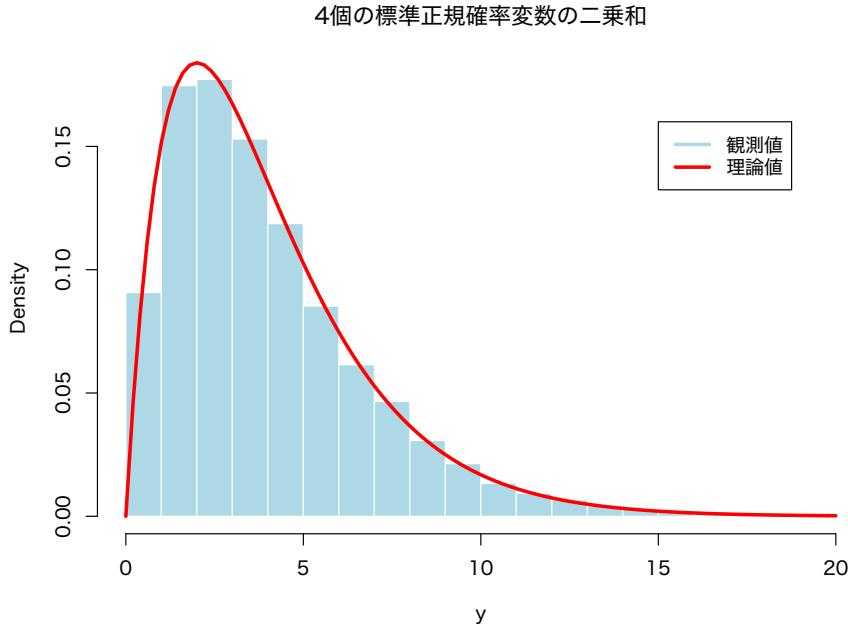
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=50, col="gray", border="white",
+       main=bquote(paste(chi^2, "分布 (自由度", .(k), ")")))) # ヒストグラム (密度表示)
> curve(dchisq(x, k), add=TRUE, col="red", lwd=3) # 理論上の確率密度関数
> legend(15, 0.16, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
```

(dist-chisq.r)

標準正規分布に従う k 個の独立な確率変数の 2 乗和は、自由度 k の χ^2 分布に従うことが知られている。

[Figure 2.13 を参照]

```
> ###  $\chi^2$  分布の特徴付け：標準正規確率変数の二乗和
> set.seed(123) # 亂数の初期値を指定
```

**Figure 2.13:** 正規分布と χ^2 分布の関係

```

> n <- 30000
> k <- 4
> y <- colSums(matrix(rnorm(n*k, 0, 1)^2, k, n))
> # 標準正規分布に従う乱数を nk 個発生し, k 個の 2 乗和を n 個作る。
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(y, freq=FALSE, breaks=40, col="lightblue", xlim=c(0,20),
+       border="white",
+       main=paste0(k, "個の標準正規確率変数の二乗和")) # ヒストグラム(密度表示)
> curve(dchisq(x, df=k), add=TRUE, xlim=c(0,20),
+         col="red", lwd=3) # 理論上の確率密度関数
> legend(15, 0.16, legend=c("観測値", "理論値"),
+         col=c("lightblue", "red"), lwd=3) # 凡例を作成

```

(dist-chisq-norm.r)

演習 2.7. ガンマ分布と χ^2 分布について調べてみよう。

1. ガンマ分布の平均と分散の計算式が正しいことを定義に従って確認せよ。また、確率密度関数が

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (2.28)$$

を満たすことを確認せよ。

2. 自由度が非常に大きい χ^2 分布はどのような分布に近づくか確認せよ。

2.3.4 ベータ分布

α, β を正の実数とする。確率密度関数が

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (0 < x < 1), \quad f(x) = 0 \quad (x \notin (0, 1)) \quad (2.29)$$

2 確率分布

で与えられる連続分布を、パラメータ α, β の**ベータ分布**と呼び、記号 $B_E(\alpha, \beta)$ で表す。ただし、 $B(\alpha, \beta)$ はベータ関数

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \quad (2.30)$$

を表す。平均は $\alpha/(\alpha + \beta)$ 、分散は $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ で与えられる。
ベータ分布に従う乱数の発生には関数 `rbeta()` を用いる。

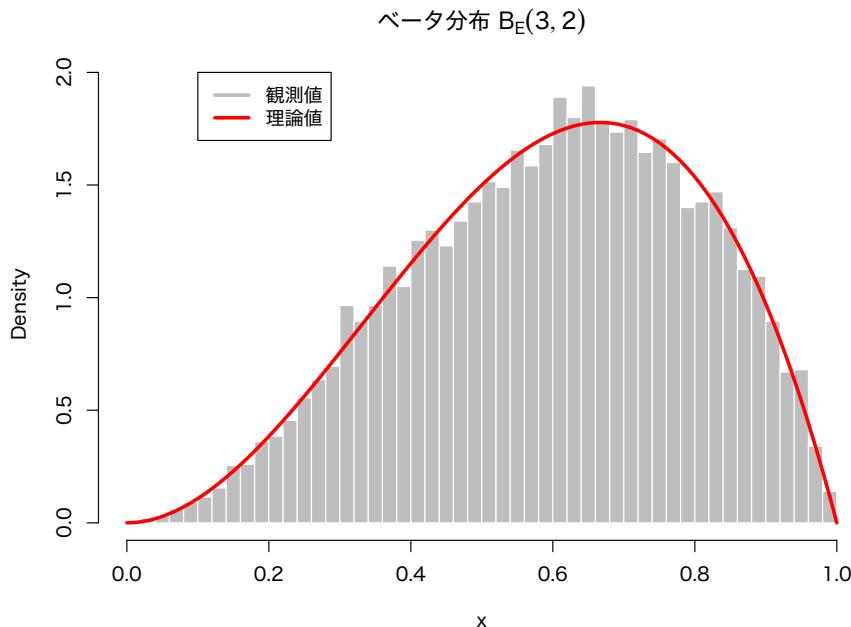


Figure 2.14: ベータ分布の例

[Figure 2.14 を参照]

```
> ### ベータ分布
> set.seed(123) # 亂数の初期値を指定
> rbeta(10, 0.5, 0.5) # パラメーター 0.5, 0.5 のベータ分布に従う乱数を 10 個発生
[1] 0.859887668 0.676206530 0.003991051 0.443966073 0.398207168 0.002031146
[7] 0.184634326 0.903712761 0.216118436 0.412547219

> ## 統計的性質の確認
> a <- 3
> b <- 2
> x <- rbeta(10000, a, b) # パラメーター a, b のベータ乱数を 10000 個発生
> mean(x) # a/(a+b)=0.6 に近い (大数の法則)
[1] 0.6000056

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=50, col="gray", border="white",
+       main=bquote(paste("ベータ分布 ", B[E](.(a), .(b))))) # ヒストグラム (密度表示)
> curve(dbeta(x, a, b), add=TRUE, col="red", lwd=3) # 理論上の確率密度関数
> legend(0.1, 2, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
```

(dist-beta.r)

2 確率分布

演習 2.8. ベータ分布の平均と分散の計算式が正しいことを定義に従って確認せよ。また、確率密度関数が

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (2.31)$$

を満たすことを確認せよ。

2.3.5 t 分布

ν を正の実数とする。確率密度関数が

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \quad (2.32)$$

で与えられる連続分布を、自由度 ν の (Student の) t 分布と呼び、記号 $t(\nu)$ で表す⁶。平均は $\nu > 1$ のときに限り存在し、自由度によらず 0 となる。分散は $\nu > 2$ のときに限り存在し、 $\nu/(\nu - 2)$ で与えられる。

t 分布に従う乱数の発生には関数 `rt()` を用いる。

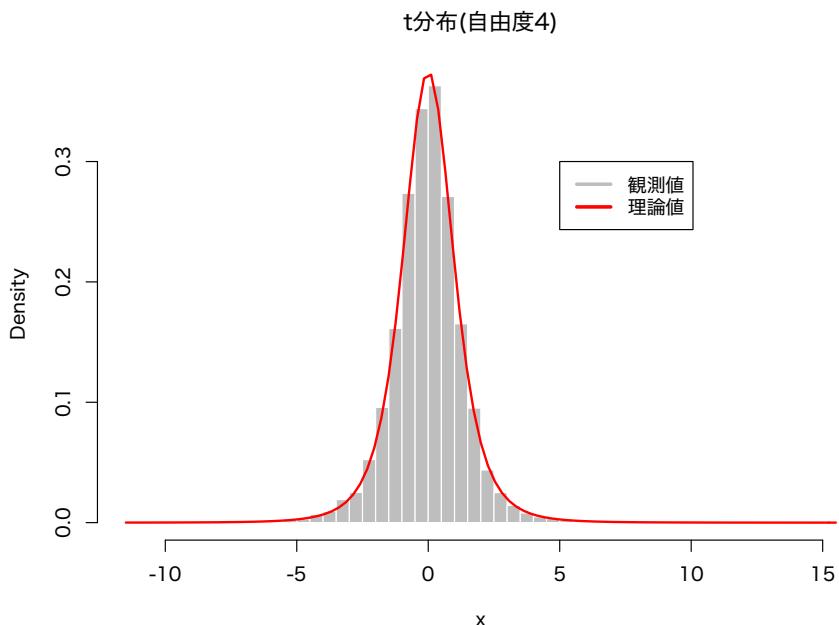


Figure 2.15: t 分布の例

[Figure 2.15 を参照]

```
> ### t 分布
> set.seed(3)
> rt(10, df=1) # 自由度 1 の t 分布に従う乱数を 10 個発生
[1] -1.4924670  1.2401246  0.1284078 -0.7243351 -3.4487104 -1.1283652
[7] -2.2026233  1.2624960 -1.9968796 -2.5450296
```

⁶Student は t 分布を導入した統計学者 Gosset のペンネームである。

2 確率分布

```

> ### 0から大きく離れた値が現れている(裾が重い)
> mean(rt(10000, df=1)) # 自由度1のt分布は平均をもたないため、大数の法則が成立しない
[1] 2.129077

> ## 統計的性質の確認
> nu <- 4
> x <- rt(10000, df=nu) # 自由度4のt乱数を10000個発生
> mean(x) # 0に近い(大数の法則)

[1] -0.006082694

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=50, col="gray", border="white",
+       main=paste0("t分布(自由度", nu, ")")) # ヒストグラム(密度表示)
> curve(dt(x, df=nu), add=TRUE,
+        col="red", lwd=2) # 理論上の確率密度関数
> legend(5, 0.3, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
> ### 0から大きく離れた値が現れている(裾が重い)

```

(dist-t.r)

Z を標準正規分布に従う確率変数, Y を自由度 k の χ^2 分布に従う確率変数とし, Z, Y は独立であるとする。このとき, 確率変数

$$\frac{Z}{\sqrt{Y/k}} \quad (2.33)$$

は自由度 k の t 分布に従うことが知られている。

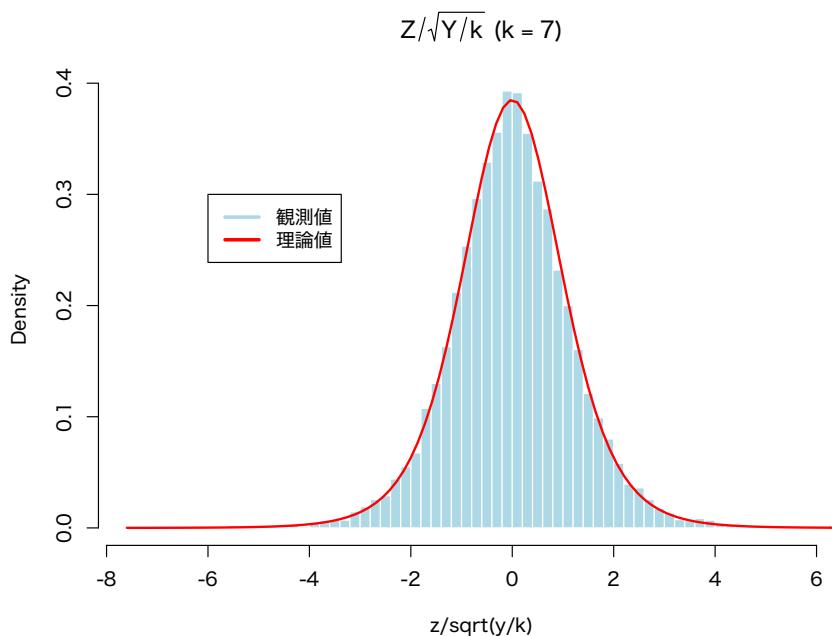


Figure 2.16: t 分布と正規分布の関係

[Figure 2.16 を参照]

```
> ### t 分布の特徴付け： 正規分布とカイ二乗分布から生成
> set.seed(11111)
> k <- 7
> y <- rchisq(10000, df=k) # 自由度 7 のカイ 2 乗分布に従う乱数
> z <- rnorm(10000) # 標準正規乱数
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(z/sqrt(y/k), freq=FALSE, breaks=50, col="lightblue",
+       border="white",
+       main=bquote(paste(Z/sqrt(Y/k), " (", k==.(k), ")")))
# ヒストグラム (密度表示)
> curve(dt(x, df=k), add=TRUE, col="red", lwd=2) # 理論上の確率密度関数
> legend(-6, 0.3, legend=c("観測値", "理論値"),
+         col=c("lightblue", "red"), lwd=3) # 凡例を作成

```

(dist-t-norm.r)

演習 2.9. 自由度が非常に大きい t 分布はどのような分布に近づくか確認せよ.

2.3.6 F 分布

ν_1, ν_2 を正の実数とする. 確率密度関数が

$$f(x) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} \frac{x^{\nu_1/2-1}}{(1+\nu_1 x/\nu_2)^{(\nu_1+\nu_2)/2}} \quad (x > 0), \quad f(x) = 0 \quad (x \leq 0) \quad (2.34)$$

で与えられる連続分布を, 自由度 ν_1, ν_2 の F 分布と呼び, 記号 $F(\nu_1, \nu_2)$ で表す. 平均は $\nu_2 > 2$ のときに限り存在し, $\nu_2/(\nu_2 - 2)$ で与えられる. 分散は $\nu_2 > 4$ のときに限り存在し, $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ で与えられる.

F 分布に従う乱数の発生には関数 `rf()` を用いる.

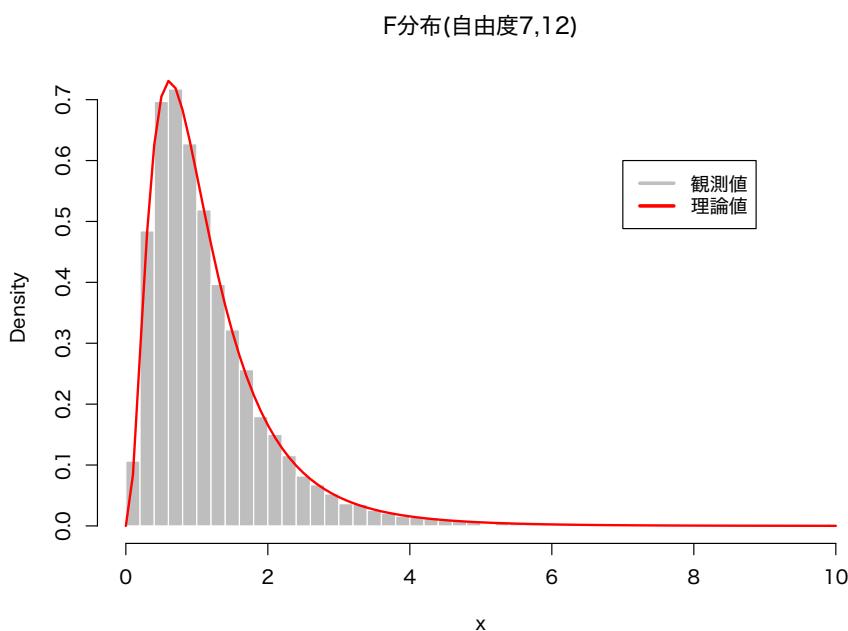


Figure 2.17: F 分布の例

2 確率分布

[Figure 2.17 を参照]

```
> ### F 分布
> set.seed(1) # 亂数の初期値を指定
> rf(10, df1=4, df2=7) # 自由度 4,7 の F 分布に従う乱数を 10 個発生
[1] 0.2530757 1.6113500 1.5400491 1.6052632 0.5898581 0.6921258 0.2311880
[8] 0.8437135 1.4594894 1.3044407

> ## 統計的性質の確認
> nu1 <- 7
> nu2 <- 12
> x <- rf(10000, df1=nu1, df2=nu2) # 自由度 4,7 の F 乱数を 10000 個生成
> mean(x) # nu2/(nu2-2)=1.2 に近い (大数の法則)
[1] 1.194231

> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=50, col="gray", border="white",
+       main=paste0("F 分布 (自由度", nu1, ", ", nu2, ")")) # ヒストグラム (密度表示)
> curve(df(x, df1=nu1, df2=nu2), add=TRUE,
+        col="red", lwd=2) # 理論上の確率密度関数
> legend(7, 0.6, legend=c("観測値", "理論値"),
+         col=c("gray", "red"), lwd=3) # 凡例を作成
```

(dist-F.r)

Y_1 を自由度 k_1 の χ^2 分布に従う確率変数, Y_2 を自由度 k_2 の χ^2 分布に従う確率変数とし, Y_1, Y_2 は独立であるとする. このとき, 確率変数

$$\frac{Y_1/k_1}{Y_2/k_2} \quad (2.35)$$

は自由度 k_1, k_2 の F 分布に従うことが知られている.

[Figure 2.18 を参照]

```
> ### F 分布の特徴付け: カイ二乗分布を利用して生成
> set.seed(22222) # 亂数の初期値を指定
> k1 <- 20
> k2 <- 10
> y1 <- rchisq(10000, df=k1) # 自由度 20 のカイ二乗分布に従う乱数
> y2 <- rchisq(10000, df=k2) # 自由度 10 のカイ二乗分布に従う乱数
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist((y1/k1)/(y2/k2), freq=FALSE, breaks=50,
+       col="lightblue", border="white",
+       main=bquote(paste(frac(Y[1]/k[1], Y[2]/k[2]), " (",
+                           k[1]==.(k1), ", ", k[2]==.(k2), ")")))) # ヒストグラム (密度表示)
> curve(df(x, df1=k1, df2=k2), add=TRUE,
+        col="red", lwd=2) # 理論上の確率密度関数
> legend(7, 0.6, legend=c("観測値", "理論値"),
+         col=c("lightblue", "red"), lwd=3) # 凡例を作成
```

(dist-F-norm.r)

演習 2.10. t 分布と F 分布の関係を調べてみよ.

R にはここで紹介した以外にも数多くの確率分布を発生させる乱数が実装されている. 詳細は `help(Distributions)` を参照してほしい.

2 確率分布

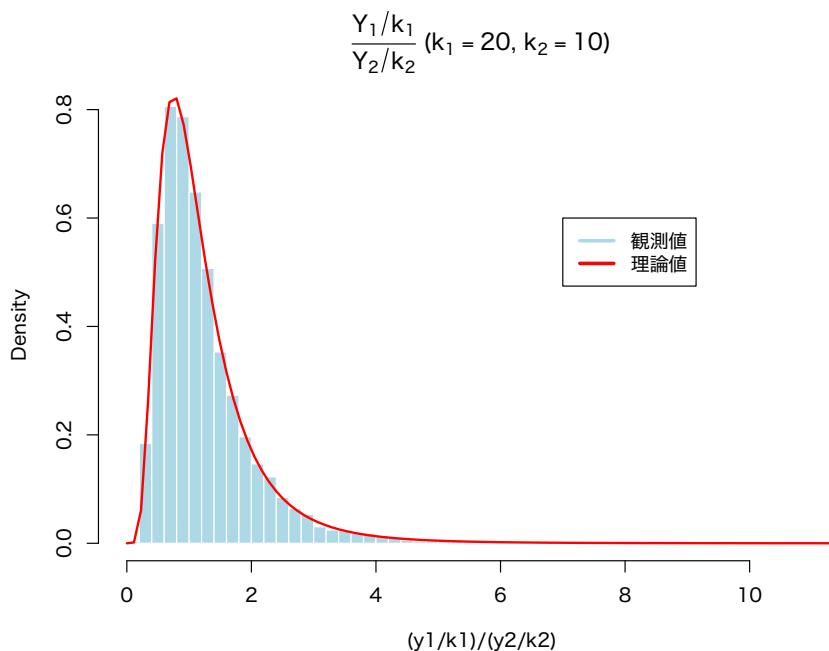


Figure 2.18: F 分布と正規分布の関係

2.4 補遺

2.4.1 参考文献

- [1] 福島正俊. **確率論 (第 5 版)**. 東京: 裳華房, 2006.
- [2] U. リゲス (石田基広訳). **R の基礎とプログラミング技法**. 東京: 丸善出版, 2012.
- [3] 竹村彰通. **統計 (第 2 版)**. 東京: 共立出版, 2007.
- [4] 東京大学教養学部統計学教室. **統計学入門**. 東京: 東京大学出版会, 1991.
- [5] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.

3 基礎的な記述統計量とデータの集約

記述統計量とはデータを簡潔に要約して表すための統計値のことである。要約統計量、基本統計量とも言われる。ヒストグラム（あるいは密度関数）や箱ひげ図などのグラフと併用して、その集団全体の特徴を表す重要な指標となる。本章では、比較的良く用いられる統計量を、その背景となるモーメント、順序、分布という考え方に基づいて分類する。

3.1 モーメントに基づく統計量

3.1.1 平均・分散・標準偏差

n 個のデータ X_1, X_2, \dots, X_n が与えられたとき、それらを代表する値として (標本) 平均

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + X_2 + \dots + X_n}{n} \quad (3.1)$$

が頻繁に利用される。また、データのばらつき具合の指標として (標本) 分散

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n} \quad (3.2)$$

およびその平方根である (標本) 標準偏差が広く利用されている。

5-6 章で述べた通り、統計学では、一つ一つのデータ X_i をある確率変数の実現値とみなすことで、データの背後にある現象に対する統計解析を行う。確率変数 X_1, X_2, \dots, X_n が同分布であれば、適切な次数のモーメントの存在を仮定した下で、 X_i の共通の平均 μ および分散 σ^2 を考えることができる。さらに、 X_1, X_2, \dots, X_n が独立であれば、大数の強法則より、標本平均・標本分散・標本標準偏差はそれぞれ $n \rightarrow \infty$ のとき確率 1 で平均 μ ・分散 σ^2 ・標準偏差 σ に収束する。これは、標本平均・標本分散・標本標準偏差のそれぞれを、対象とする集団の「真の」平均・分散・標準偏差の推定量と考えた場合に、これらの推定量がサンプル数 n が十分大きいときに性質の良い推定量であるという根拠の 1 つを与える。このような性質を推定量の (強) 一致性と呼び、一致性をもつ推定量を (強) 一致推定量と呼ぶ。

一致性はサンプル数が十分大きい場合に推定量がまともであることの 1 つの根拠を与えるが、サンプル数が小さい場合の推定量の性質については何も語っていない。そのような場合の推定量の良さに関する性質の 1 つとして不偏性がある。一般にパラメータ θ の推定量 $\hat{\theta}$ が不偏であるとは、 $\hat{\theta}$ の平均が θ 、すなわち

$$E[\hat{\theta}] = \theta \quad (3.3)$$

が成り立つことをいう。標本平均は μ の不偏推定量である。すなわち

$$E[\bar{X}] = \mu \quad (3.4)$$

が成り立つ。一方で、標本分散は σ^2 の不偏推定量ではない。実際

$$E[S^2] = \frac{n-1}{n} \sigma^2 \quad (3.5)$$

3 基礎的な記述統計量とデータの集約

である。この式は、標本分散は平均的には真の分散を過小推定する傾向にあることを意味する。このバイアスを補正するには、標本分散に $n/(n-1)$ をかけてやれば良い。すなわち

$$s^2 := \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3.6)$$

は σ^2 の不偏推定量となる。わざわざ不偏性を持たない S^2 を σ^2 の推定量として使う理由は通常ないので、標本分散という場合には s^2 のことを指す場合もあるが、バイアス補正をしていることを強調するために**不偏分散**と呼ぶ場合もある。Rには不偏分散を計算するための関数として `var()` が用意されている。同様に、標本標準偏差という場合は、通常は不偏分散の平方根 s を指し、Rでは関数 `sd()` で計算できる。ただし、一般に s は標準偏差 σ の不偏推定量ではない。

```
> #### 標本分散が平均的には過小推定となることの確認
> set.seed(123) # 亂数のシード値を指定
> sample.var <- function(n){ # n 個の標準正規乱数の標本分散を計算する関数
+   x <- rnorm(n)
+   return(mean((x - mean(x))^2))
+ }
> ## Monte-Carlo 実験
> n <- 10 # サンプル数
> mc <- 10000 # 実験回数
> v <- replicate(mc, sample.var(n)) # sample.var(n) を mc 回実行して結果を記録
> head(v) # 実験結果をいくつか見てみる
[1] 0.8187336 0.9698367 0.7797652 0.2502430 1.0546610 0.6601483
> mean(v) # 標本分散推定量の平均 (n-1)/n=0.9 に近い (真の分散 1 を過小推定している)
[1] 0.9009275
> mean((n/(n-1))*v) # バイアス修正すると 1 に近くなる (不偏となる)
[1] 1.001031
```

(summary-unbiased.r)

複数のデータを同時に分析する場合、単位や基準を揃えた方が扱いやすい。このような目的でよく使われる方法に、データの**標準化**がある。データ X_1, X_2, \dots, X_n の標準化は

$$Z_i = \frac{X_i - \bar{X}}{s} \quad (i = 1, 2, \dots, n) \quad (3.7)$$

で定義される¹。定義から明らかのように、 Z_1, Z_2, \dots, Z_n の標本平均は 0、不偏分散は 1 となる。言い換えると、平均は 0、分散は 1 となるようにデータを一次変換したものが標準化である。標準化は**標準得点**あるいは**Zスコア**とも呼ばれる。一方で、教育学や心理学では、データを標本平均が 50、標準偏差 (不偏分散の平方根) が 10 となるように一次変換したもの

$$T_i = 10Z_i + 50 \quad (i = 1, \dots, n) \quad (3.8)$$

を使う場合が多い。これを**偏差值得点**あるいは**T得点**と呼ぶ。

¹ s の代わりに S で割って定義する文献もある。

3 基礎的な記述統計量とデータの集約

```

> ### データの標準化（気候データによる例）
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis") # データの読み込み
> dat <- subset(kikou, select=-c(月, 日)) # 月日は計算対象から削除
> head(dat)

 気温 降水量 日射量 風速
1 7.5      0 11.80 2.6
2 7.3      0 11.59 1.9
3 9.3      0 10.77 1.4
4 9.2      0 11.19 1.6
5 10.9     0 10.57 1.8
6 8.9      0  4.54 1.9

> dat.std <- scale(dat) # 各変数ごとに標準化
> head(dat.std)

 気温    降水量    日射量    風速
1 -1.1682298 -0.3583779 -0.1233405 -0.2189065
2 -1.1942766 -0.3583779 -0.1527083 -1.0461017
3 -0.9338081 -0.3583779 -0.2673829 -1.6369554
4 -0.9468315 -0.3583779 -0.2086471 -1.4006139
5 -0.7254333 -0.3583779 -0.2953523 -1.1642724
6 -0.9859018 -0.3583779 -1.1386296 -1.0461017

> colMeans(dat.std) # 各変数の平均が0であることの確認
 気温    降水量    日射量    風速
-1.460959e-16 1.178285e-17 -6.278182e-17 -9.858538e-18

> apply(dat.std, 2, "sd") # 各変数の標準偏差が1であることの確認
 気温 降水量 日射量 風速
1       1       1       1

>

```

(summary-scale.r)

3.1.2 歪度と尖度

中心極限定理が示すように、正規分布は確率分布のうち最も基本的なものと考えられる。正規分布は平均と分散を決めるに完全に決定されるから、正規分布に従うデータを考える際には標本平均と標本分散（不偏分散）を考えれば十分である。しかし、現実には正規分布では捉えきれない特徴をもつデータに遭遇することがしばしばある。そのようなデータを考える場合の最初のアプローチとして、正規分布からのずれを調べるということがしばしば行われる。そのための統計量として代表的なものに歪度と尖度がある。

X を平均 μ 、分散 σ^2 をもつ確率変数とする。 X が3次のモーメントをもつとき

$$\frac{E[(X - \mu)^3]}{\sigma^3} \quad (3.9)$$

を**歪度**と呼ぶ。歪度は分布の非対称性を表す統計量で、正の場合分布の右の裾の方が重く、負の場合分布の左の裾の方が重いと考えられる。左右に対称的な分布の歪度は0であり、従って正規分布の歪度は0である。正の歪度をもつ分布としては、例えばガンマ分布 $\Gamma(\nu, \alpha)$ があり、その歪度は $2/\sqrt{\nu}$ で与えられる。

3 基礎的な記述統計量とデータの集約

一方で、 X が 4 次のモーメントをもつとき

$$\frac{E[(X - \mu)^4]}{\sigma^4} \quad (3.10)$$

を**尖度**と呼ぶ。尖度は平均の周囲の分布の尖り具合を表す統計量だと考えられる。正規分布の場合 3 であるため、正規分布との比較のため上の定義から 3 を引いた量

$$\frac{E[(X - \mu)^4]}{\sigma^4} - 3 \quad (3.11)$$

のことを尖度と呼ぶ文献も多いが、後者を前者と区別するために**超過尖度**と呼ぶ場合もある。超過尖度が正となる分布は正規分布よりも平均の周囲の分布が尖っており、負となる分布は丸みを帯びていると考えられる。前者の場合、平均まわりの密度が分布の裾の方にまわっていることが多いため、正規分布より裾が重いと解釈されることが多い。正の超過尖度をもつ分布としては、例えば自由度 $\nu > 4$ をもつ t 分布 $t(\nu)$ があり、その超過尖度は $6/(\nu - 4)$ で与えられる ($\nu \leq 4$ のときは $t(\nu)$ は 4 次モーメントをもたない)。また、ガンマ分布 $\Gamma(\nu, \alpha)$ は超過尖度 $6/\nu$ をもつ。

観測データ X_1, X_2, \dots, X_n から歪度と尖度を推定するには、それらの標本バージョンを考えればよい。すなわち、歪度の推定量としては**標本歪度**

$$\frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3}{s^3} \quad (3.12)$$

を考えればよく、尖度の推定量としては**標本尖度**

$$\frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4}{s^4} \quad (3.13)$$

を考えればよい。標本歪度・標本尖度を直接計算するための関数は R の基本パッケージには含まれていないため、自作するかパッケージを利用する。例えば、パッケージ e1071 には標本歪度を計算するための関数 `skewness()` および標本尖度を計算するための関数 `kurtosis()` が実装されている。後者は上の定義の標本尖度から 3 を引いたもの、すなわち標本超過尖度を計算することに注意されたい。

なお、標本歪度・標本尖度の値は標本平均・分散に比べてばらつきが大きくなる傾向があるため、サンプル数が少ない場合の計算結果の解釈には注意を要する。

[Figure 3.1 を参照]

```
> ### 歪度と尖度の計算
> if(require(e1071)) { # パッケージの読み込み
+   print("package e1071 is loaded")
+ } else { # 無ければインストールしてから読み込む
+   install.packages("e1071")
+   require(e1071)
+   print("package e1071 is installed/loaded")
+ }

[1] "package e1071 is loaded"

> set.seed(123)
> ## 正規分布による例
> x <- rnorm(10000) # 標準正規乱数を 10000 個発生
> skewness(x) # 歪度: 0 に近い

[1] 0.008197965
```

3 基礎的な記述統計量とデータの集約

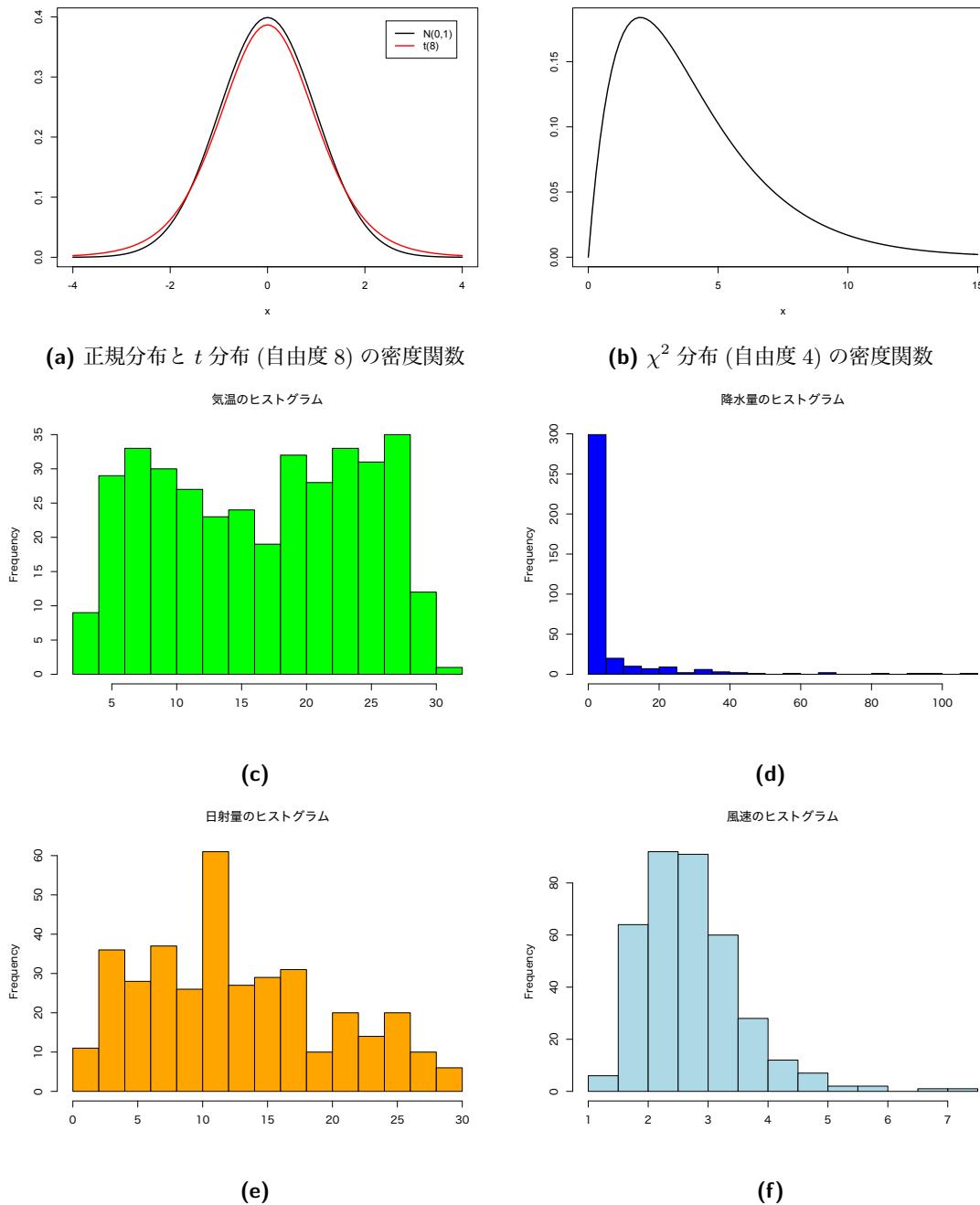


Figure 3.1: 歪度・尖度の計算の例

```
> kurtosis(x) # (超過) 尖度: 0 に近い
[1] 0.01073845

> ## t 分布による例
> y <- rt(10000, df=8) # 自由度 8 の t 乱数を 10000 個発生
> skewness(y) # 歪度: 0 に近い
[1] 0.01196181
```

3 基礎的な記述統計量とデータの集約

```

> kurtosis(y) # (超過) 尖度 : 6/(8-4)=1.5 に近い
[1] 1.445932

> ## グラフで確認してみる
> plot(dnorm, -4 ,4 , lwd=2, ylab="") # 標準正規密度
> curve(dt(x, df = 8), add=TRUE, lwd=2, col="red") # 自由度 8 の t 分布の密度
> legend("topright", inset=1/20, legend=c("N(0,1)", "t(8)"),
+         lwd=2, col=c("black", "red")) # 凡例の追加
> z <- rchisq(10000, df=4) # 自由度 4 のカイ二乗乱数を 10000 個発生
> mean(z) # 平均 : 4(自由度) に近い

[1] 4.013924

> skewness(z) # 歪度 : sqrt(8/4)=1.414... に近い
[1] 1.415016

> kurtosis(z) # (超過) 尖度 : 12/4=3 に近い
[1] 2.973832

> ## グラフで確認してみる
> curve(dchisq(x,df=4), 0, 15, lwd=2, ylab="") # 自由度 4 のカイ二乗分布の密度
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis") # データの読み込み
> dat <- subset(kikou, select=-c(月, 日)) # 月日は計算対象から削除
> apply(dat, 2, "skewness") # 歪度

    気温      降水量      日射量      風速
-0.0508104  4.4538316  0.4333552  1.3670592

> apply(dat, 2, "kurtosis") # (超過) 尖度

    気温      降水量      日射量      風速
-1.3089167 23.4751756 -0.6806749  3.5644397

> # ヒストグラムによる確認
> par(family = "HiraginoSans-W4") # 日本語フォントの指定
> myCol <- c("green", "blue", "orange", "lightblue") # 色を用意
> for(i in 1:4){
+   hist(dat[, i], col=myCol[i], breaks=20, xlab="",
+         main=paste0(names(dat)[i], "のヒストグラム"))
+ }

```

(summary-skewkurt.r)

演習 3.1. 正規分布からサンプルされたデータから計算された標本歪度・標本尖度の不偏性について調べてみよ。

3.1.3 相関と共分散

複数のデータが与えられた場合、それらのデータの間の関係性を知りたい場合が頻繁に生じる。そのような目的のための最も基本的な記述統計量に **(標本)相関** があり、これは 2 種類のデータ間の比例関係の大きさを計測する。2 種類のデータ x_1, x_2, \dots, x_n および y_1, y_2, \dots, y_n に対して、それらの相関は

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.14)$$

3 基礎的な記述統計量とデータの集約

で定義される。ここに、 \bar{x} および \bar{y} はそれぞれ x_1, x_2, \dots, x_n および y_1, y_2, \dots, y_n の標本平均である。相関は -1 以上 1 以下の値をとり、 1 に近いほど正の比例関係が強く、 -1 に近いほど負の比例関係が強いことになる。なお、(3.14) の分子の統計量を $n-1$ で割ったものは**(標本)共分散**と呼ばれる。相関および共分散はそれぞれ関数 `cor()` および関数 `cov()` で計算できる。2種類のデータ `x` および `y` が与えられたとき、それらの相関は `cor(x, y)` で計算できる。一方、`x` がデータフレームのとき、`cor(x)` の (i, j) 成分が `x` の i 列と j 列の間の相関であるような行列を計算する。これを**相関行列**と言う。共分散についても同様である。

```
> ### 相関の計算
> ## sleep データによる例
> ## 2種類の睡眠薬の効果の個人差に相関はあるか?
> x <- subset(sleep, group==1, extra)
> y <- subset(sleep, group==2, extra)
> cor(x, y)

      extra
extra 0.7951702

> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> cor(kikou[, -c(1:2)]) # 相関行列

   気温    降水量    日射量    風速
気温  1.00000000  0.08575259  0.3314077  0.14291472
降水量 0.08575259  1.00000000 -0.3607787  0.07572892
日射量  0.33140773 -0.36077872  1.0000000  0.31826364
風速   0.14291472  0.07572892  0.3182636  1.00000000

                                         (summary-cor.r)
```

演習 3.2. R の組込データセット `state.x77` について、列ごとの標本平均、標準偏差、標本歪度、標本尖度を求めよ。また、相関行列も求めよ。

3.2 順序に基づく統計量

データの順序にもとづく記述統計量もよく利用される。例えば、 X_1, \dots, X_n の**最大値**は関数 `max()` で、**最小値**は関数 `min()` でそれぞれ計算できる。データを

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \quad (3.15)$$

のように昇順に並べ替えた際に中央の位置にくる値を**メディアン**もしくは**中央値**と呼ぶ。 n が奇数の場合メディアンは $X_{((n+1)/2)}$ であり、 n が偶数の場合は $(X_{(n/2)} + X_{(n/2+1)})/2$ である。メディアンは関数 `median()` で計算できる。

メディアンは平均と同様データを代表する値だと考えられるが、平均と比較して、計算結果がデータに含まれる異常な値（**外れ値**と呼ばれる）の影響を受けにくい。

メディアンの一般化として、 $\alpha \in [0, 1]$ に対して、その点以下のデータの個数が全体の $100\alpha\%$ になるような点を $100\alpha\%$ **分位点**と呼ぶ。特に 25% 分位点および 75% 分位点をそれぞれ**第1四分位点**、**第3四分位点**と呼ぶ。**第2四分位点**は 50% 分位点となるが、これはメディアンのことである。ベクトル `x` の $100\alpha\%$ 分位点は `quantile(x, alpha)` で計算できる。分位点は一意的には定まらず、いくつかの計算方式があるので、詳しくは `help(quantile)` を参照してほしい。

3 基礎的な記述統計量とデータの集約

```
> #### 順序に基づく統計量の計算
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> x <- kikou$気温
> mean(x) # 平均
[1] 16.47022

> median(x) # メディアン
[1] 17.15

> quantile(x) # 最小・最大値および四分位点
   0%    25%    50%    75%   100%
2.800  9.425 17.150 23.075 31.900

> quantile(x, 0.5) # メディアンと同じ
50%
17.15

> quantile(x, c(0.25, 0.75)) # 第1四分位点および第3四分位点
  25%    75%
9.425 23.075

> fivenum(x) # 五数要約 (quantile(x) と同様, 表示が異なる)
[1] 2.80 9.40 17.15 23.10 31.90

> summary(x) # 五数要約に平均を加えて集約
   Min. 1st Qu. Median Mean 3rd Qu. Max.
2.800  9.425 17.150 16.470 23.075 31.900

> ## 分位点の計算方式の違いのため, quantileの結果と多少異なることに注意
> y <- c(x, 1000) # データに外れ値を加えてみる
> mean(y) # mean(x) と大きく異なる (外れ値の影響を受けやすい)
[1] 19.15014

> median(y) # median(x) とあまり変わらない (外れ値に頑健)
[1] 17.2

> ## データフレームに対しても同様の計算が可能
> summary(kikou[, -c(1, 2)]) # 五数要約と平均を計算
   気温      降水量      日射量      風速
Min. : 2.800  Min. : 0.000  Min. : 1.11  Min. : 1.200
1st Qu.: 9.425  1st Qu.: 0.000  1st Qu.: 7.16  1st Qu.: 2.200
Median :17.150  Median : 0.000  Median :11.69  Median : 2.700
Mean   :16.470  Mean   : 4.861  Mean   :12.68  Mean   : 2.785
3rd Qu.:23.075  3rd Qu.: 2.000  3rd Qu.:17.58  3rd Qu.: 3.200
Max.  :31.900  Max.  :106.500  Max.  :29.87  Max.  : 7.200
```

(summary-order.r)

確率分布に対しても分位点が定義され, 推定や検定において重要な役割を果たす. $0 < \alpha < 1$ とする. 連続分布の $100\alpha\%$ 分位点は, その分布に従う確率変数を X としたとき, 不等式

$$P(X \leq x) \geq \alpha \quad (3.16)$$

3 基礎的な記述統計量とデータの集約

を満たす実数 x のうち最小のものとして定義される.² そのような実数は常に存在し、それを q_α とすると、

$$P(X \leq q_\alpha) = \alpha \quad (3.17)$$

が成り立つ。 X_1, X_2, \dots, X_n が独立同分布な確率変数の列のとき、 X_1, X_2, \dots, X_n の $100\alpha\%$ 分位点は $n \rightarrow \infty$ のとき X_i の従う分布の $100\alpha\%$ 分位点の一致推定量となることが知られている。

確率分布の分位点は、その分布の省略形が xxx の場合 (6.2.2 節参照)、関数 `qxxx()` で計算できる。例えば、平均 `mu`、標準偏差 `sigma` の正規分布の $100\alpha\%$ 分位点は、

```
qnorm(alpha, mean=mu, sd=sigma)
```

計算できる。

```
> ### 分位点の性質
> ## データから計算された分位点が
> ## 確率分布の分位点の一一致推定量となることの確認
> set.seed(123)
> x <- rnorm(1000) # 標準正規乱数を 1000 個発生
> alpha <- c(0.25, 0.5, 0.75) # 計算する分位点の位置
> quantile(x, probs=alpha) # データの分位点 (推定値)

      25%          50%          75%
-0.628324243  0.009209639  0.664601867

> qnorm(alpha) # 確率分布の分位点 (理論値)
[1] -0.6744898  0.0000000  0.6744898

                                         (summary-quantile.r)
```

順序に基づいてデータのばらつきを測るための記述統計量もいくつか存在する。そのようなものとして、最大値と最小値の差である**範囲**がある。範囲は外れ値の影響を大きく受けるので、第3四分位点と第1四分位点の差である**四分位範囲**もよく使われる。また、データ X_1, X_2, \dots, X_n のメディアンを m としたとき、 $|X_1 - m|, |X_2 - m|, \dots, |X_n - m|$ のメディアンを**メディアン絶対偏差**と呼ぶ。

```
> ### 範囲に関する統計量
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> dat <- subset(kikou, select=-c(月, 日)) # 月日は計算対象から削除
> range(dat$気温) # 最小値と最大値を計算
[1] 2.8 31.9

> diff(range(dat$気温)) # 範囲を計算
[1] 29.1

> apply(dat, 2, function(x) diff(range(x))) # 変数ごとの範囲
```

² より一般に、確率分布の $100\alpha\%$ 分位点は、その分布に従う確率変数を X としたとき、不等式

$$P(X \leq x) \geq \alpha$$

を満たす実数 x の下限として定義される (そのような実数が存在しない場合は ∞ とする)。

3 基礎的な記述統計量とデータの集約

```
気温 降水量 日射量 風速
29.10 106.50 28.76 6.00

> apply(dat, 2, "IQR") # 変数ごとの四分位範囲
気温 降水量 日射量 風速
13.6500 2.0000 10.4225 1.0000

> apply(dat, 2, "mad", constant = 1) # 変数ごとのメディアン絶対偏差
気温 降水量 日射量 風速
6.75 0.00 5.28 0.50

> apply(dat, 2, "sd") # 変数ごとの標準偏差
気温 降水量 日射量 風速
7.6784711 13.5629323 7.1506723 0.8462331

                                         (summary-range.r)
```

演習 3.3. 順序に基づく記述統計量について調べてみよう。

1. 正規乱数から計算された四分位範囲・メディアン絶対偏差と標準偏差の間の関係を調べてみよ。
2. 順序に基づいて複数のデータの間の関係性を要約するための記述統計量について調べてみよ。

3.3 頻度に基づく統計量

データの中で最も頻度が高く現れる値を、モードもしくは最頻値と呼ぶ。モードはデータが有限個の値を取る場合に特に有効であるが、データが連續で無限に多くの値を取ることができる場合には注意が必要である。連続なデータの場合でも有限個の観測データに対してモードは定義できるが、偶々観測値として現れた値なので、その意味はよく考えなくてはならない。必要に応じて、例えば区分的に集計するなどの工夫をすることもある。

```
> ### 最頻値の計算
> ## モードを計算するための関数の作成
> myMode <- function(x){ # 注意:"mode"という関数は既にある
+   return(names(which.max(table(x))))
+ }
> ## シミュレーションによる例
> set.seed(123)
> x <- rpois(1000, lambda = 5) # 強度 5 の Poisson 乱数を 1000 個発生
> myMode(x) # モードの計算
[1] "4"

> table(x) # 度数分布表の確認

x
 0   1   2   3   4   5   6   7   8   9   10  11  12  14 
 7  29  89 135 189 167 155 100  62  35  20   9   2   1 

> ## モードは数値以外のデータ（質的データ）にも適用可能
> ## 東京都 2017 年の日別の風に関するデータ tokyo_wind.csv による例
```

3 基礎的な記述統計量とデータの集約

```
> ## 気象庁のホームページより取得して整理したもの
> ## http://www.data.jma.go.jp/gmd/risk/obsdl/index.php
> ## 東京都の2017年の各日の最多風向(16方位)と風速(m/s)を記録
> wind <- read.csv("tokyo_wind.csv", fileEncoding="utf8")
> head(wind)

年 月 日 最多風向 平均風速 最大風速 最大瞬間風速
1 2017 1 1 北北西 2.0 4.0 5.6
2 2017 1 2 北西 1.4 3.4 5.2
3 2017 1 3 北北西 2.2 5.4 10.7
4 2017 1 4 北北西 2.6 5.9 12.3
5 2017 1 5 北西 4.9 9.0 16.3
6 2017 1 6 北北西 2.3 4.6 10.6

> myMode(wind$最多風向) # モードの計算
[1] "北北西"

> table(wind$最多風向) # 度数分布表の確認
   西 西西北 東 東南東 東北東 南 南西 南東 南南西 南南東 北
1 10 6 1 14 45 2 18 8 53 15
北西 北東 北北西 北北東
75 17 86 14

> ## modeest package を使うことも可能
> ## modeest::mfv 離散データ
> ## modeest::mlv 連続データ

                                         (summary-mode.r)
```

3.4 補遺

3.4.1 参考文献

- [1] 東京大学教養学部統計学教室. **統計学入門**. 東京: 東京大学出版会, 1991.
- [2] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.

4 推定

分析対象の観測データに基づいて統計解析を行う際、統計学では観測データをある確率変数の実現値の集合と考えてモデル化する。このとき、確率変数の従う分布のもつなんらかの特性量(平均や分散など)を評価したり、分布そのものを決定することが統計解析の目標の1つとなるが、この作業を一般に**推定**と呼ぶ。本章の目的は、統計学で広く利用されている代表的な推定方法を説明することである。

以下本章では、観測データが独立同分布な確率変数列 X_1, X_2, \dots, X_n がモデル化されている状況を考える。この場合、 X_1, \dots, X_n が従う共通の分布 \mathcal{L} に関する推定を行うことが目標となるが、この分布としてすべての分布を考察対象とすると、対象とする範囲が広くなりすぎて、サンプル数 n が十分大きくない限り応用上意味のある結論を導き出すことが困難になる。そのため、以下では確率分布 \mathcal{L} を特徴づけるなんらかのパラメータ θ 、例えば \mathcal{L} の平均、分散、歪度、尖度などを考察対象とする。

4.1 点推定

\mathcal{L} に含まれるあるパラメータ θ を X_1, \dots, X_n のある関数

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) \quad (4.1)$$

で推定することを、**点推定**と呼ぶ。このとき、 $\hat{\theta}$ は θ の**推定量**と呼ばれる。点推定については、すでに前章で記述統計量との関連で議論した。そこでみたように、多くの記述統計量は、適当なモデル化の下で分布の性質を記述するなんらかのパラメータの推定量とみなせる。例えば、 \mathcal{L} の平均 μ を標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ によって推定することが点推定であり、 \bar{X} は μ の推定量となる。

一般に、1つのパラメータに対して推定量は無数に存在するため、使うべき推定量を決定するためには推定量の良さを評価する基準が必要である。前章で議論したように、そのような基準として代表的なものに、**不偏性**と**一致性**がある: $\hat{\theta}$ が θ の不偏推定量であるとは、 $\hat{\theta}$ の平均が θ 、すなわち

$$E[\hat{\theta}] = \theta \quad (4.2)$$

が成り立つことをいい、 $\hat{\theta}$ が θ の(強)一致推定量であるとは、 $n \rightarrow \infty$ のとき $\hat{\theta}$ が θ に収束する確率が 1 であることをいう。

シンプルな状況では、1つのパラメータに対して複数の不偏推定量が存在する場合も起こりうる。例えば、 \mathcal{L} の平均 μ の不偏推定量としては、標本平均 \bar{X} 以外にも、例えば X_1 (1番目の観測データそのもの)が考えられる。より“自然な”推定量の例を挙げると、 \mathcal{L} が直線 $x = \mu$ に関して線対称な密度をもつ連続分布であったならば、 X_1, \dots, X_n のメディアンも μ の不偏推定量である。このような場合、不偏推定量の中から使うべきものを決定するための基準を設ける必要がある。自然な基準として、推定値のばらつき(分散)が最も小さいものを選ぶという基準が考えられる。すなわち、パラメータ θ の不偏推定量 $\hat{\theta}$ の中で、任意の不偏推定量 $\hat{\theta}'$ に対して

$$\text{Var}[\hat{\theta}] \leq \text{Var}[\hat{\theta}'] \quad (4.3)$$

を満たすようなものを選ぶということである。このような推定量 $\hat{\theta}$ を**一様最小分散不偏推定量**と呼ぶ。一様最小分散不偏推定量を見出す方法の1つとして、次の結果が利用できる: \mathcal{L} は1つの(1次元)パラメータ θ を含む連続分布であるとし、その確率密度関数 $f_\theta(x)$ は θ に

4 推定

関して偏微分可能であるとする。このとき、緩やかな仮定の下で、 θ の任意の不偏推定量 $\hat{\theta}$ に対して以下の不等式が成り立つ：

$$\text{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta)}. \quad (4.4)$$

ただし

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta} \log f_{\theta}(x) \right)^2 f_{\theta}(x) dx \quad (4.5)$$

である。不等式 (4.4) は **Cramér-Rao の不等式** と呼ばれ、その下界 $1/(nI(\theta))$ は **Cramér-Rao 下界** と呼ばれる。また、 $I(\theta)$ は **Fisher 情報量** と呼ばれる。Cramér-Rao の不等式より、もし θ の不偏推定量 $\hat{\theta}$ で分散が Cramér-Rao 下界 $1/(nI(\theta))$ に一致するものが存在すれば、それは一様最小分散不偏推定量となる。

例題 4.1 (正規分布の平均の推定). \mathcal{L} が平均 μ 、分散 σ^2 の正規分布でモデル化されているとする。このとき、平均パラメータ μ に関する Fisher 情報量は

$$I(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{(x - \mu)^2}{\sigma^4} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sigma^2} \quad (4.6)$$

となるから、Cramér-Rao 下界は σ^2/n となる。従って、標本平均 \bar{X} の分散は Cramér-Rao 下界に一致するので、 \bar{X} は μ の一様最小分散不偏推定量である。

```
> ### 標本平均が一様最小分散不偏推定量であることの確認
> set.seed(123) # 亂数のシード値を指定
> mu <- 10
> sigma <- 5
> ## 平均値の推定を行う関数(標本平均, x1, メディアン)
> mean.est <- function(n, mu, sigma){
+   x <- rnorm(n, mean=mu, sd=sigma)
+   return(c(xbar=mean(x), x1=x[1], med=median(x)))
+ }
> ## Monte-Carlo 実験
> n <- 10 # サンプル数
> mc <- 30000 # 実験回数
> mu.hats <- as.data.frame(t( # mc 行 x 3 種の推定量
+   replicate(mc, mean.est(n, mu, sigma))
+ ))
> head(mu.hats) # 実験結果の表示

  xbar      x1     med
1 10.373128 7.197622 9.600827
2 11.043110 16.120409 11.901463
3  7.877206  4.660881  6.615174
4 11.610223 12.132321 12.450955
5  9.956422  6.526465  8.472995
6 11.108430 11.266593 10.849490

> apply(mu.hats, 2, mean) # 推定値の平均 (colMeans(mu.hats) も可)

  xbar      x1     med
10.00933 10.01986 10.01210

> apply(mu.hats, 2, var) # 推定値の分散

  xbar      x1     med
2.502021 25.157491  3.457196
```

```
> sigma^2/n          # Cramer-Rao 下界
[1] 2.5

```

(est-umvue.r)

4.2 最尤法

興味あるパラメータが、平均や分散といった記述統計量と自然に関連づけられるパラメータではない場合、推定量の構成が自明ではないことがある。このような場合でも、確率分布 \mathcal{L} に対するモデルがいくつかのパラメータ $\theta_1, \theta_2, \dots, \theta_p$ を除いて特定されている状況であれば、一般的に適用可能な $\theta_1, \theta_2, \dots, \theta_p$ の推定量の構成方法がいくつか知られている。ここでは代表的な方法として最尤法を説明する。

まず \mathcal{L} が離散分布の場合を考え、その確率関数を $f_{\theta}(x)$ と書くことにする（確率関数のパラメータ $\theta := (\theta_1, \dots, \theta_p)$ への依存を明示するために添え字 θ をついている）。このとき、パラメータ θ を一つ定めれば、観測値として $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ が得られる理論上の確率を

$$\prod_{i=1}^n f_{\theta}(x_i) = f_{\theta}(x_1) \cdot f_{\theta}(x_2) \cdots f_{\theta}(x_n) \quad (4.7)$$

で求めることができる。実際に得られている観測データは X_1, X_2, \dots, X_n であるから、パラメータ θ に対してそのような観測データが得られる理論上の確率は、

$$L(\theta) := \prod_{i=1}^n f_{\theta}(X_i) \quad (4.8)$$

で与えられる。 $L(\theta)$ は観測データ X_1, X_2, \dots, X_n が現れるのにパラメータ θ の値がどの程度尤もらしいか測る尺度と解釈でき、 θ の尤度と呼ばれる。 $L(\theta)$ を θ の関数とみなしたもののが尤度関数と呼ぶ。最尤法は観測データに対して「最も尤もらしい」パラメータを θ の推定量として採用する方法である。すなわち、尤度関数 $L(\theta)$ を最大化するパラメータ値 $\hat{\theta}$ を θ の推定量とする：

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta). \quad (4.9)$$

ここで、 Θ は尤度関数の定義域である。上の式は以下のようにも書かれる：

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta). \quad (4.10)$$

$\hat{\theta}$ は**最尤推定量**と呼ばれる。

なお、尤度関数は積の形をしていて扱いにくいので、和の形に直すために対数を取ることが多い：

$$\ell(\theta) := \log L(\theta) = \sum_{i=1}^n \log f_{\theta}(X_i). \quad (4.11)$$

対数関数は狭義増加であるから、 $\ell(\theta)$ の最大化と $L(\theta)$ の最大化は同義である。 $\ell(\theta)$ は**対数尤度関数**と呼ばれる。

例題 4.2 (Poisson 分布の最尤推定)。 \mathcal{L} がパラメータ $\lambda > 0$ の Poisson 分布としてモデル化されている場合を考える。このとき、未知パラメータは λ であり、対数尤度関数は

$$\ell(\lambda) = \sum_{i=1}^n \log \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} = \sum_{i=1}^n (X_i \log \lambda - \log X_i!) - n\lambda \quad (4.12)$$

4 推定

で与えられる。いま、少なくとも 1 つの i について $X_i > 0$ であると仮定する。このとき、 $\ell(\lambda)$ を微分すると、

$$\ell'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^n X_i - n, \quad \ell''(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^n X_i < 0 \quad (4.13)$$

を得る。従って方程式 $\ell'(\lambda) = 0$ の解が $\ell(\lambda)$ を最大化するから、 $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i$ が λ の最尤推定量である。

\mathcal{L} が連続分布の場合は、確率関数の代わりに確率密度関数を用いて尤度を計算する。

例題 4.3 (指数分布の最尤推定)。 \mathcal{L} がパラメータ $\lambda > 0$ の指数分布としてモデル化されている場合を考える。このとき、未知パラメータは λ であり、対数尤度関数は

$$\ell(\lambda) = \sum_{i=1}^n \log \lambda e^{-\lambda X_i} = n \log \lambda - \lambda \sum_{i=1}^n X_i \quad (4.14)$$

で与えられる。 $\ell(\lambda)$ を微分すると、

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n X_i, \quad \ell''(\lambda) = -\frac{n}{\lambda^2} < 0 \quad (4.15)$$

を得る。従って方程式 $\ell'(\lambda) = 0$ の解が $\ell(\lambda)$ を最大化するから、 $\hat{\lambda} = (\frac{1}{n} \sum_{i=1}^n X_i)^{-1}$ が λ の最尤推定量である。

例題 4.4 (ガンマ分布の最尤推定)。 \mathcal{L} がパラメータ $\nu, \alpha > 0$ のガンマ分布としてモデル化されている場合を考える。このとき、未知パラメータは ν, α であり、対数尤度関数は

$$\ell(\nu, \alpha) = \sum_{i=1}^n \log \frac{\alpha^\nu}{\Gamma(\nu)} X_i^{\nu-1} e^{-\alpha X_i} \quad (4.16)$$

$$= n\nu \log \alpha - n \log \Gamma(\nu) + \sum_{i=1}^n \{(\nu-1) \log X_i - \alpha X_i\} \quad (4.17)$$

で与えられる。 $\ell(\nu, \alpha)$ を最大化するような ν, α は解析的には求まらないため、実際の計算では数値的に求めることになる(以下の実行例を参照)。数値計算の際に対数尤度関数の勾配(偏導関数からなるベクトル)があると便利なので計算しておく:

$$\frac{\partial \ell}{\partial \nu}(\nu, \alpha) = n \log \alpha - n \psi(\nu) + \sum_{i=1}^n \log X_i, \quad \frac{\partial \ell}{\partial \alpha}(\nu, \alpha) = \frac{n\nu}{\alpha} - \sum_{i=1}^n X_i \quad (4.18)$$

ここに、 $\psi(\nu) = \frac{d}{d\nu} \log \Gamma(\nu)$ であり、ディガンマ関数と呼ばれる(R では関数 `digamma()` で計算できる)。

広い範囲の確率分布に対して、最尤推定量は一致性を持つことが知られている。

[Figure 4.1 を参照]

```
> ### ガンマ分布の最尤推定
> require(stats4) # 関数 mle を利用するため
> ## 観測データから最尤推定量を計算する関数の作成
> ## x: 観測データ, nu0: nu の初期値, alpha0: alpha の初期値
```

4 推定

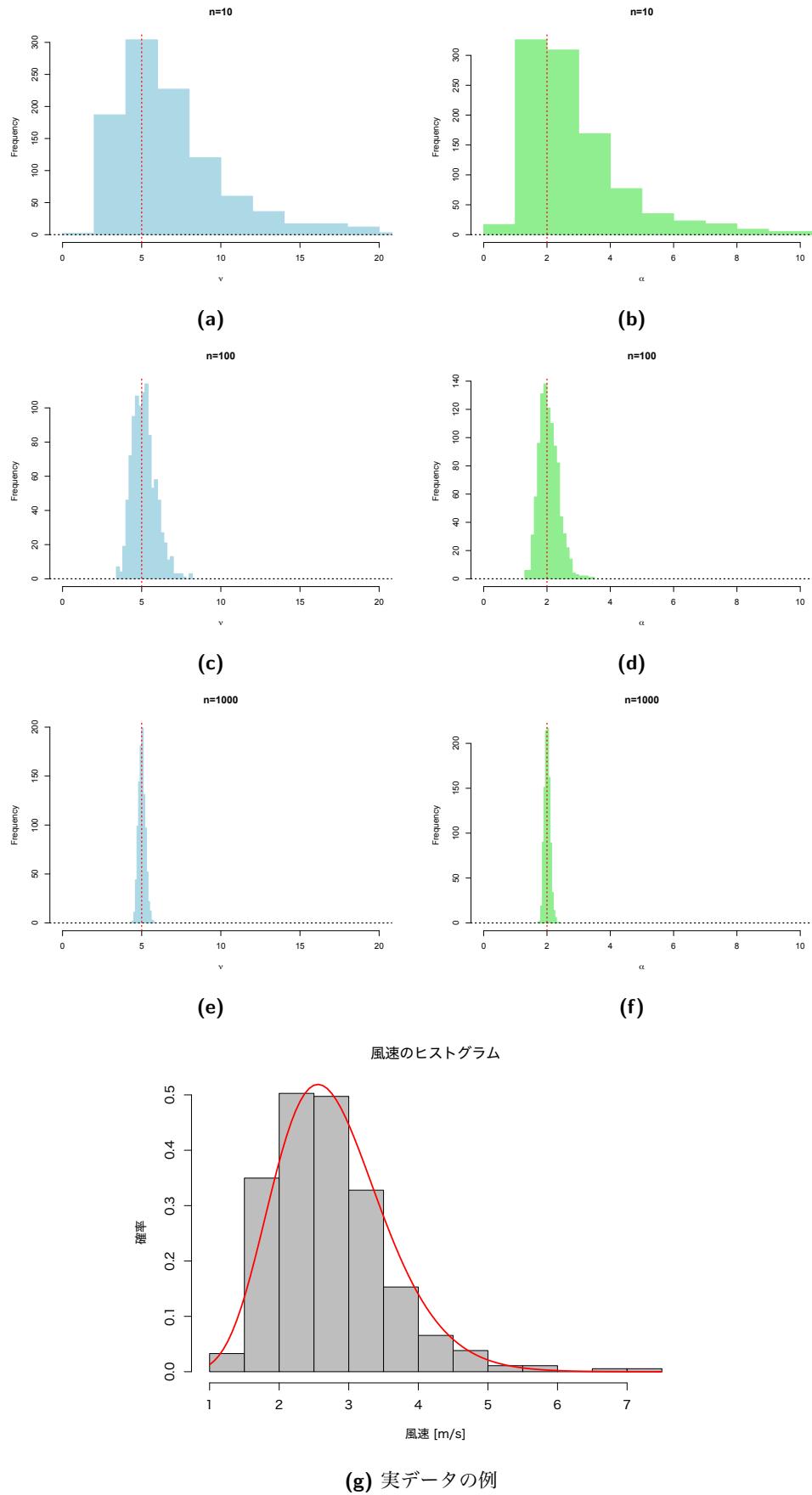


Figure 4.1: 最尤推定量の計算の例

4 推定

```

> ## 数値的最適化のためには尤度関数を最初に評価する初期値が必要
> mle.gamma <- function(x, nu0=1, alpha0=1, verbose=FALSE){
+   n <- length(x)
+   # 負の対数尤度関数を定義 (最小化を考えるため)
+   # suppressWarnings で定義域外で評価された際の警告を表示させない
+   ll <- function(nu, alpha)
+     suppressWarnings(-sum(dgamma(x, nu, alpha, log=TRUE)))
+   # 最尤推定
+   est <- mle(minuslogl=ll, # 負の対数尤度関数
+               start=list(nu=nu0, alpha=alpha0), # 初期値
+               method="BFGS", nobs=n) # 最適化方法は選択可能
+   if(verbose) {
+     return(est) # verbose=TRUEなら mle の結果を全て返す
+   } else {
+     return(coef(est)) # 推定値のみ返す
+   }
+ }
> ## シミュレーションによる一致性の検証
> set.seed(123) # 亂数のシード値の指定
> nu <- 5
> alpha <- 2
> n <- 100 # データ数
> x <- rgamma(n, shape=nu, rate=alpha)
> mle.gamma(x)

      nu      alpha
6.486969 2.730202

> ## 複数回行なった場合のヒストグラムの描画
> mc <- 1000 # 各実験の繰り返し回数
> for(n in c(10, 100, 1000)){ # データ数を変えて実験
+   ## Monte-Carlo 実験
+   theta <- data.frame(t( # 推定値の data.frame
+     replicate(mc, mle.gamma(rgamma(n, nu, alpha))))
+   ))
+   ## 結果をヒストグラムで表示
+   hist(theta$nu, breaks=20, col="lightblue", border="lightblue",
+         xlim=c(0, 20), main=paste0("n=",n), xlab=expression(nu))
+   abline(h=0, lwd=2, lty="dotted")
+   abline(v=nu, col="red", lwd=2, lty="dotted")
+   hist(theta$alpha, breaks=20, col="lightgreen", border="lightgreen",
+         xlim=c(0, 10), main=paste0("n=",n), xlab=expression(alpha))
+   abline(h=0, lwd=2, lty="dotted")
+   abline(v=alpha, col="red", lwd=2, lty="dotted")
+ }
> ## 実データへの適用
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> x <- kikou$風速 # 風速のデータにガンマ分布をあてはめてみる
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> hist(x, freq=FALSE, breaks=20, col="gray", # ヒストグラム(確率値)の描画
+       main="風速のヒストグラム", xlab="風速 [m/s]", ylab="確率")
> (theta <- mle.gamma(x)) # 最尤推定

      nu      alpha
12.23397 4.39242

> curve(dgamma(x, theta[1], theta[2]), # あてはめたガンマ分布の密度関数
+        add=TRUE, col="red", lwd=2)

```

(est-mle.r)

4 推定

パッケージ `stats4` は R に標準で含まれているが、これを使わないので最尤推定を行う関数を作成するには以下のようにすれば良い。

```
> #### optim を用いた最尤推定量の関数
> #### mle のかわりに optim を用いると細かな最適化の設定が行える
> mle.gamma <- function(x, nu0=1, alpha0=1){
+   n <- length(x) # データ数 (尤度関数で用いる)
+   ## 対数尤度関数のマイナスを計算する関数 (optim は最小化を行うため)
+   ## suppressWarnings で定義域外で評価された際の警告を表示させない
+   f <- function(theta) # theta=(nu,alpha)
+     suppressWarnings(-sum(log(dgamma(x, theta[1], theta[2]))))
+   ## f の勾配 (偏微分からなるベクトル) を計算する関数
+   ## 指定しなくても大丈夫だが、指定した方が計算が速いことが多い
+   gr <- function(theta){ # theta=(nu,alpha)
+     ## nu に関する偏微分
+     gr.nu <- -n*log(theta[2])+n*digamma(theta[1])-sum(log(x))
+     ## alpha に関する偏微分
+     gr.alpha <- -n*theta[1]/theta[2]+sum(x)
+     return(c(gr.nu, gr.alpha))
+   }
+   ## 関数 f の最小化
+   opt <- optim(c(nu0, alpha0), # パラメータの初期値
+                 fn=f, gr=gr, # 関数 fn とその勾配 gr
+                 method="BFGS") # BFGS は準 Newton 法の一つ
+   theta <- opt$par # 推定値
+   names(theta) <- c("nu", "alpha")
+   return(theta)
+ }
```

(est-mle-optim.r)

演習 4.5. 最尤法について調べてみよう。

1. 確率分布 \mathcal{L} が成功確率 p の幾何分布でモデル化される場合の、パラメータ p に対する最尤推定量を求めよ。
2. 確率分布 \mathcal{L} が平均 μ , 分散 σ^2 の正規分布でモデル化される場合の、パラメータ μ, σ^2 に対する最尤推定量を求めよ。

4.3 区間推定

未知パラメータ θ を推定量 $\hat{\theta}$ で点推定した場合、通常推定値は真のパラメータ値とは異なるため、推定誤差が必ず存在する。そのため、推定結果の定量的な評価には、推定誤差の評価が重要となる。統計学では、ある値 $\alpha \in (0, 1)$ を固定したとき、

$$P(l \leq \hat{\theta} - \theta \leq u) \geq 1 - \alpha \quad (4.19)$$

が成り立つような l, u を観測データから推定することで推定誤差の評価を試みる。上の式の意味するところは、「誤差 $\hat{\theta} - \theta$ が区間 $[l, u]$ の外側にある確率が α 以下」ということである。上の式を変形すると、

$$P(\hat{\theta} - u \leq \theta \leq \hat{\theta} - l) \geq 1 - \alpha \quad (4.20)$$

となる。従って、ここで行っているのは、パラメータ θ が含まれているような確率が $1 - \alpha$ 以上となるような区間 $[\hat{\theta} - u, \hat{\theta} - l]$ を推定することだと言い換えられる。このように、未知

4 推定

パラメータが含まれている確率があらかじめ決められたある値以上となるような区間を推定することを**区間推定**と呼ぶ。

より一般には、未知パラメータ θ とある値 $\alpha \in (0, 1)$ に対して、

$$P(L \leq \theta \leq U) \geq 1 - \alpha \quad (4.21)$$

が成り立つような確率変数 L, U を観測データから求めることになる。このとき、区間 $[L, U]$ を $100(1-\alpha)\%$ 信頼区間、 L を $100(1-\alpha)\%$ 下側信頼限界、 U を $100(1-\alpha)\%$ 上側信頼限界、 $1-\alpha$ を信頼係数とそれぞれ呼ぶ。慣習として $\alpha = 0.01, 0.05, 0.1$ とすることが多い。

信頼区間は幅が狭いほど真のパラメータが取りうる値の範囲を限定することになるため、推定精度が良いといえる。一方で、信頼区間 $[L, U]$ の幅が狭いほど確率 $P(L \leq \theta \leq U)$ は小さくなるため、最も推定精度の良い $100(1-\alpha)\%$ 信頼区間 $[L, U]$ は、

$$P(L \leq \theta \leq U) = 1 - \alpha \quad (4.22)$$

を満たす。そのため、実行可能である限り、 $100(1-\alpha)\%$ 信頼区間 $[L, U]$ の構成では上の式を満たすように L, U を決定する。

4.4 正規母集団の区間推定

この節では、 X_i の従う分布が平均 μ 、分散 σ^2 の正規分布の場合に、 μ および分散 σ^2 の区間推定をする方法を説明する。

4.4.1 分散既知における平均の区間推定

はじめに、分散 σ^2 がすでにわかっている場合に平均 μ の区間推定をする方法を説明する。これには次の結果を用いる：

命題 4.1. Z_1, Z_2, \dots, Z_k を独立な確率変数列とし、各 $i = 1, 2, \dots, k$ に対して Z_i は平均 μ_i 、分散 σ_i^2 の正規分布に従うとする。このとき、 a_0 を実数、 a_1, \dots, a_k を k 個の 0 でない実数とすると、 $a_0 + \sum_{i=1}^k a_i Z_i$ は平均 $a_0 + \sum_{i=1}^k a_i \mu_i$ 、分散 $\sum_{i=1}^k a_i^2 \sigma_i^2$ の正規分布に従う。

上の命題を

$$k = n, \mu_i = \mu, \sigma_i^2 = \sigma^2, a_0 = 0, a_i = 1/n \quad (i = 1, \dots, n) \quad (4.23)$$

として適用すると、標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ は平均 μ 、分散 σ^2/n の正規分布に従うことがわかる。再び命題 4.1 を $k = 1$ 、 $\mu_1 = \mu$ 、 $\sigma_1^2 = \sigma^2/n$ 、 $a_0 = -\sqrt{n}\mu/\sigma$ 、 $a_1 = \sqrt{n}/\sigma$ として適用すると、確率変数

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \quad (4.24)$$

が標準正規分布に従うことがわかる。従って、 $\alpha \in (0, 1)$ を定めたとき、 $z_{1-\alpha/2}$ を標準正規分布の $1-\alpha/2$ 分位点とすれば、

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{1-\alpha/2}\right) = 1 - \alpha \quad (4.25)$$

が成り立つことがわかる。実際、 ϕ を標準正規分布の確率密度関数とすると、

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{1-\alpha/2}\right) = \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \phi(x) dx \quad (4.26)$$

$$= \int_{-\infty}^{z_{1-\alpha/2}} \phi(x) dx - \int_{-\infty}^{-z_{1-\alpha/2}} \phi(x) dx \quad (4.27)$$

4 推定

となるが、ここで定義より

$$\int_{-\infty}^{z_{1-\alpha/2}} \phi(x)dx = 1 - \alpha/2 \quad (4.28)$$

であり、また

$$\int_{-\infty}^{-z_{1-\alpha/2}} \phi(x)dx = - \int_{\infty}^{z_{1-\alpha/2}} \phi(-y)dy \quad (y = -x \text{ と置換}) \quad (4.29)$$

$$= \int_{z_{1-\alpha/2}}^{\infty} \phi(y)dy \quad (\phi \text{ は偶関数}) \quad (4.30)$$

$$= \int_{-\infty}^{\infty} \phi(y)dy - \int_{-\infty}^{z_{1-\alpha/2}} \phi(y)dy \quad (4.31)$$

$$= 1 - (1 - \alpha/2) = \alpha/2 \quad (4.32)$$

(全確率 $\int_{-\infty}^{\infty} \phi(y)dy = 1$ に注意) であるから、

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{1-\alpha/2}\right) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha \quad (4.33)$$

となる。カッコ内を μ について解くと、

$$P\left(\bar{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right) = 1 - \alpha \quad (4.34)$$

が得られるので、 σ が既知であれば、区間

$$[\bar{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}, \bar{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}] \quad (4.35)$$

が平均 μ の $100(1 - \alpha)\%$ 信頼区間を与える。

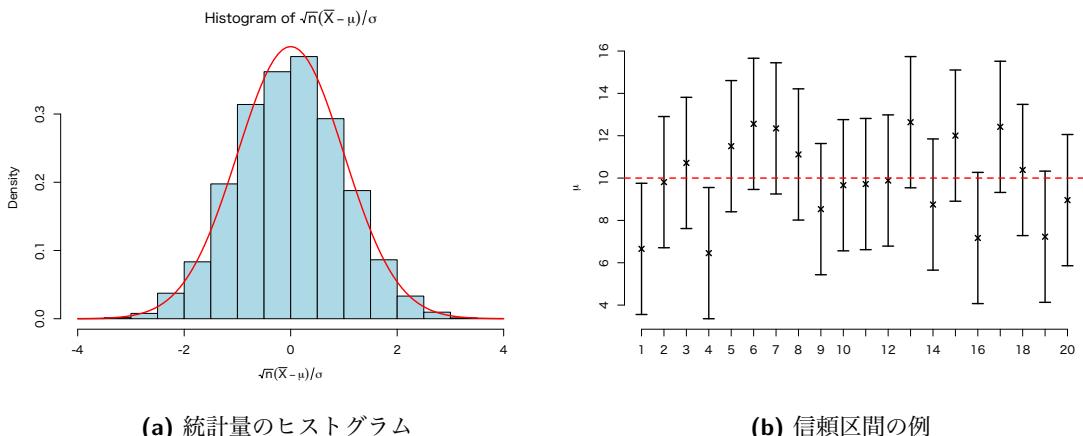


Figure 4.2: 分散既知における平均の区間推定の例

[Figure 4.2 を参照]

```
> ### 分散既知における平均の区間推定
> ## 正規標本から標本平均と平均の 100(1-alpha)%信頼区間の上端・下端を
> ## 計算する関数の作成
> myest <- function(x, sigma, alpha=0.05){
```

```

+   n <- length(x) # データ数
+   xbar <- mean(x) # 標本平均
+   ui <- xbar + qnorm(1-alpha/2)*sigma/sqrt(n) # 上側信頼限界
+   li <- xbar - qnorm(1-alpha/2)*sigma/sqrt(n) # 下側信頼限界
+   return(c(xbar=xbar, ui=ui, li=li))
+ }
> ## Monte-Carlo 実験
> set.seed(111) # 亂数のシード値の指定
> mu <- 10      # 平均
> sigma <- 5    # 標準偏差
> n <- 10       # データ数
> mc <- 10000 # シミュレーション回数
> result <- as.data.frame(t(
+   replicate(mc, myest(rnorm(n, mean=mu, sd=sigma), sigma)))
+ ))
> ## sqrt(n)*(xbar-mu)/sigma が標準正規分布に従うことの確認
> hist(sqrt(n)*(result$xbar-mu)/sigma, freq=FALSE,
+       col="lightblue", xlab=expression(sqrt(n)*(bar(X)-mu)/sigma),
+       main=expression(paste("Histogram of ", sqrt(n)*(bar(X)-mu)/sigma)))
> curve(dnorm, add=TRUE, col="red", lwd=2)
> ## 真の平均 mu が 95%信頼区間に含まれている割合が約 95%であることの確認
> with(result, sum((mu<=ui & mu>=li))/mc

[1] 0.9502

> ## はじめの 20 個の信頼区間の可視化 (plotrix パッケージを利用)
> if(require(plotrix)) { # パッケージの読み込み
+   print("package plotrix is loaded")
+ } else { # 無ければインストールしてから読み込む
+   install.packages("plotrix")
+   require(plotrix)
+   print("package plotrix is installed/loaded")
+ }

[1] "package plotrix is loaded"

> with(result, # 信頼区間と標本平均の図示
+       plotCI(1:20, xbar[1:20], ui=ui[1:20], li=li[1:20],
+               pch=4, axes=FALSE, xlab="", ylab=expression(mu), lwd=2))
> axis(1, 1:20, 1:20) # x 軸の追加
> axis(2) # y 軸の追加
> abline(h=mu, col="red", lty="dashed", lwd=2) # 真の平均の図示

```

(est-ci-mean.r)

4.4.2 分散未知における平均の区間推定

分散 σ^2 が既知であることは稀である。 σ^2 が未知の場合、不偏分散 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ で代用するのが自然な考え方である。 s^2 の分布については次の結果が知られている ($n \geq 2$ とする):

命題 4.2. X_1, X_2, \dots, X_n は独立同分布な確率変数列で、平均 μ 、分散 σ^2 の正規分布に従うとする。このとき、 \bar{X} と s^2 は独立であり、確率変数 $(n-1)s^2/\sigma^2$ は自由度 $n-1$ の χ^2 分布に従う。

4 推定

上の命題と $\sqrt{n}(\bar{X} - \mu)/\sigma$ が標準正規分布に従うことから、確率変数

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} \quad (4.36)$$

は自由度 $n-1$ の t 分布に従うことがわかる（2.3.5 節参照）。従って、 $\alpha \in (0, 1)$ を定めたとき、 $t_{1-\alpha/2}(n-1)$ を自由度 $n-1$ の t 分布の $1-\alpha/2$ 分位点とすれば、上と同様の議論によって、

$$P\left(-t_{1-\alpha/2}(n-1) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{s} \leq t_{1-\alpha/2}(n-1)\right) = 1 - \alpha \quad (4.37)$$

が成り立つことがわかる。カッコ内を μ について解くことで、

$$[\bar{X} - t_{1-\alpha/2}(n-1) \cdot s/\sqrt{n}, \bar{X} + t_{1-\alpha/2}(n-1) \cdot s/\sqrt{n}] \quad (4.38)$$

が平均 μ の $100(1-\alpha)\%$ 信頼区間を与えることがわかる。

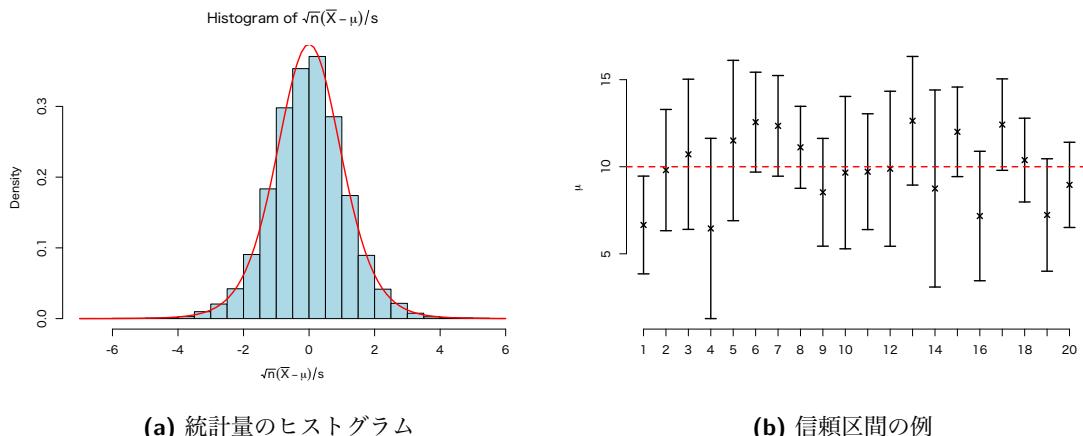


Figure 4.3: 分散未知における平均の区間推定の例

[Figure 4.3 を参照]

```
> #### 分散未知における平均の区間推定
> ## 正規標本から標本平均・標準偏差と平均の 100(1-alpha)% 信頼区間の
> ## 上端・下端を計算する関数の作成
> myest <- function(x, alpha=0.05){
+   n <- length(x) # データ数
+   xbar <- mean(x) # 標本平均
+   s <- sd(x) # 標本標準偏差 (不偏分散の平方根)
+   ui <- xbar + qt(1-alpha/2, df=n-1)*s/sqrt(n) # 上側信頼限界
+   li <- xbar - qt(1-alpha/2, df=n-1)*s/sqrt(n) # 下側信頼限界
+   return(c(xbar=xbar, s=s, ui=ui, li=li))
+ }
> ## Monte-Carlo 実験
> set.seed(111) # 亂数のシード値の設定
> mu <- 10 # 平均
> sigma <- 5 # 標準偏差
> n <- 10 # データ数
```

```

> mc <- 10000 # シミュレーション回数
> result <- as.data.frame(t(
+   replicate(mc, myest(rnorm(n, mean=mu, sd=sigma))))
+ ))
> ## sqrt(n)*(xbar-mu)/s が自由度 n-1 の t 分布に従うことの確認
> hist(sqrt(n)*(result$xbar-mu)/result$s, freq=FALSE,
+       col="lightblue", breaks=20,
+       xlab=expression(sqrt(n)*(bar(X)-mu)/s),
+       main=expression(paste("Histogram of ", sqrt(n)*(bar(X)-mu)/s)))
> curve(dt(x, df=9), add=TRUE, col="red", lwd=2)
> ## 真の平均 mu が 95%信頼区間に含まれている割合が約 95%であることの確認
> with(result, sum((mu<=ui & mu>=li))/mc
[1] 0.95

> ## はじめの 20 個の信頼区間の可視化
> require(plotrix)
> with(result, # 信頼区間と標本平均の図示
+   plotCI(1:20, xbar[1:20], ui=ui[1:20], li=li[1:20],
+           pch=4, axes=FALSE, xlab="", ylab=expression(mu), lwd=2))
> axis(1, 1:20, 1:20) # x 軸の追加
> axis(2) # y 軸の追加
> abline(h=mu, col="red", lty="dashed", lwd=2) # 真の平均の図示

```

(est-ci-mean-unkonwn.r)

4.4.3 分散の区間推定

分散 σ^2 の区間推定には命題 4.2 を利用する。すなわち、 $(n-1)s^2/\sigma^2$ が自由度 $n-1$ の χ^2 分布に従うので、 $\chi_{\alpha/2}^2(n-1)$, $\chi_{1-\alpha/2}^2(n-1)$ をそれぞれ自由度 $n-1$ の χ^2 分布の $\alpha/2, 1-\alpha/2$ 分位点とすれば、

$$P(\chi_{\alpha/2}^2(n-1) \leq (n-1)s^2/\sigma^2 \leq \chi_{1-\alpha/2}^2(n-1)) = 1-\alpha \quad (4.39)$$

が成り立つ。実際、自由度 $n-1$ の χ^2 分布の確率密度関数を f とすれば、

$$P(\chi_{\alpha/2}^2(n-1) \leq (n-1)s^2/\sigma^2 \leq \chi_{1-\alpha/2}^2(n-1)) = \int_{\chi_{\alpha/2}^2(n-1)}^{\chi_{1-\alpha/2}^2(n-1)} f(x)dx \quad (4.40)$$

$$= \int_{-\infty}^{\chi_{1-\alpha/2}^2(n-1)} f(x)dx - \int_{-\infty}^{\chi_{\alpha/2}^2(n-1)} f(x)dx = 1-\alpha/2 - \alpha/2 = 1-\alpha \quad (4.41)$$

となる。 (4.39) の左辺のカッコ内を σ^2 について解くと、

$$P\left((n-1)s^2/\chi_{1-\alpha/2}^2(n-1) \leq \sigma^2 \leq (n-1)s^2/\chi_{\alpha/2}^2(n-1)\right) = 1-\alpha \quad (4.42)$$

が得られるので、

$$\left[(n-1)s^2/\chi_{1-\alpha/2}^2(n-1), (n-1)s^2/\chi_{\alpha/2}^2(n-1)\right] \quad (4.43)$$

が σ^2 の $100(1-\alpha)\%$ 信頼区間を与える。

[Figure 4.4 を参照]

4 推定

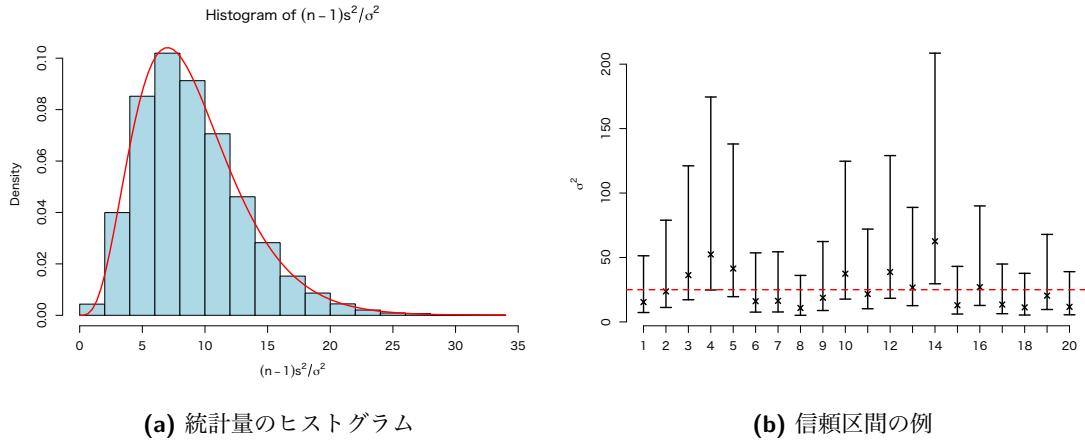


Figure 4.4: 分散の区間推定の例

```

> #### 分散の区間推定
> ## 正規標本から不偏分散と分散の 100(1-alpha)%信頼区間の
> ## 上端・下端を計算する関数の作成
> myest <- function(x, alpha=0.05){
+   n <- length(x)
+   s2 <- var(x) # 不偏分散
+   ui <- (n-1) * s2/qchisq(alpha/2, df=n - 1) # 上側信頼限界
+   li <- (n-1) * s2/qchisq(1-alpha/2, df=n - 1) # 下側信頼限界
+   return(c(s2=s2, ui=ui, li=li))
+ }
> ## Monte-Carlo 実験
> set.seed(111) # 亂数のシード値の設定
> mu <- 10      # 平均
> sigma <- 5    # 標準偏差
> n <- 10       # データ数
> mc <- 10000   # シミュレーション回数
> result <- as.data.frame(t(
+   replicate(mc, myest(rnorm(n, mean=mu, sd=sigma))))
+ ))
> ## (n-1)*s^2/sigma^2が自由度 n-1 のカイ二乗分布に従うことの確認
> hist((n-1)*result$s2/sigma^2, freq=FALSE, col="lightblue",
+       breaks=20, xlab=expression((n-1)*s^2/sigma^2),
+       main=expression(paste("Histogram of ", (n-1)*s^2/sigma^2)))
> curve(dchisq(x, df=9), add=TRUE, col="red", lwd=2)
> ## 真の分散 sigma^2 が 95%信頼区間に含まれている割合が約 95%であることの確認
> with(result, sum((sigma^2<=ui & sigma^2>=li))/mc
[1] 0.9528

> ## はじめの 20 個の信頼区間の可視化
> require(plotrix)
> with(result, # 信頼区間と不偏分散の図示
+   plotCI(1:20, s2[1:20], ui=ui[1:20], li=li[1:20],
+          pch=4, axes=FALSE, xlab="", ylab=expression(sigma^2), lwd=2))
> axis(1, 1:20, 1:20) # x 軸の追加
> axis(2) # y 軸の追加
> abline(h=sigma^2, col="red", lty="dashed", lwd=2) # 真の分散の図示

```

(est-ci-var.r)

4.5 漸近正規性による区間推定

前節のように信頼区間を正確に計算できることは稀である。しかし、未知パラメータ θ のある推定量 $\hat{\theta}$ について、推定誤差 $\hat{\theta} - \theta$ の分布がある正規分布で近似できる状況はしばしばある。このような推定量の性質を**漸近正規性**と呼ぶ。漸近正規性をもつ推定量がある場合、推定誤差がある区間に含まれる確率を近似的に求めることができるから、近似的に正しい信頼区間を構成することが可能となる。この節ではそのような方法の具体例を説明する。

4.5.1 平均の信頼区間

以前の章で確認したように、確率分布 \mathcal{L} が 2 次のモーメントを持つば、中心極限定理より、 \mathcal{L} の平均 μ の推定量である標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ は漸近正規性をもつ。より正確に述べると、 \mathcal{L} の標準偏差を σ とすれば、任意の $a \leq b$ に対して、

$$P\left(a \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq b\right) \rightarrow \int_a^b \phi(x) dx \quad (n \rightarrow \infty) \quad (4.44)$$

が成り立つ。ただし ϕ は標準正規分布の確率密度関数である。従って、 $\alpha \in (0, 1)$ を定めたとき、 $z_{1-\alpha/2}$ を標準正規分布の $1-\alpha/2$ 分位点とすれば、

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{1-\alpha/2}\right) \rightarrow 1 - \alpha \quad (n \rightarrow \infty) \quad (4.45)$$

が成り立つ。カッコ内を μ について解くと、

$$P\left(\bar{X} - z_{1-\alpha/2} \cdot \sigma / \sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \sigma / \sqrt{n}\right) \rightarrow 1 - \alpha \quad (n \rightarrow \infty) \quad (4.46)$$

が得られるので、 σ が既知であれば、

$$[\bar{X} - z_{1-\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X} + z_{1-\alpha/2} \cdot \sigma / \sqrt{n}] \quad (4.47)$$

はサンプル数 n が十分大きい場合に近似的に正しい平均 μ の $100(1-\alpha)\%$ 信頼区間を与える。通常は σ は未知であるが、近似 (4.44) は σ をその一致推定量 $\hat{\sigma}$ で置き換えてそのまま成立することが知られている。従って、上の式で σ を $\hat{\sigma}$ で置き換えたもの

$$[\bar{X} - z_{1-\alpha/2} \cdot \hat{\sigma} / \sqrt{n}, \bar{X} + z_{1-\alpha/2} \cdot \hat{\sigma} / \sqrt{n}] \quad (4.48)$$

も、サンプル数 n が十分大きい場合に近似的に正しい平均 μ の $100(1-\alpha)\%$ 信頼区間を与える。 $\hat{\sigma}$ としては例えば不偏分散の平方根

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (4.49)$$

を使うことができる。

[Figure 4.5 を参照]

```
> ### 一様分布の平均の信頼区間
> ## (0,1) 上の一様乱数から標本平均・標準偏差と平均の 100(1-alpha)% 信頼区間の
> ## 上端・下端を計算する関数の作成
> myest <- function(x, alpha=0.05){
+   n <- length(x)
```

4 推定

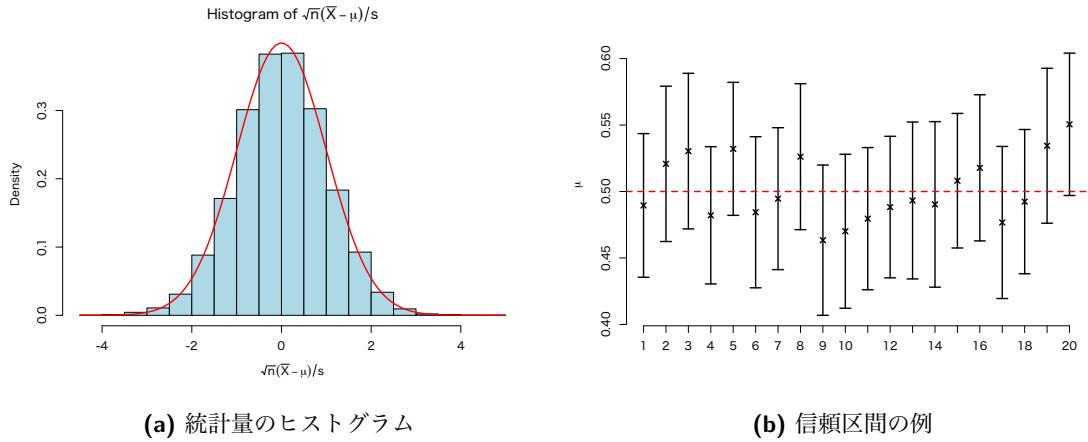


Figure 4.5: 漸近正規性による平均の区間推定の例

```

+   xbar <- mean(x) # 標本平均
+   s <- sd(x) # 標準偏差 (不偏分散の平方根)
+   ui <- xbar + qnorm(1-alpha/2)*s/sqrt(n) # 上側信頼限界
+   li <- xbar - qnorm(1-alpha/2)*s/sqrt(n) # 下側信頼限界
+   return(c(xbar=xbar, s=s, ui=ui, li=li))
+ }
> ## Monte-Carlo 実験
> set.seed(111) # 亂数のシード値の設定
> n <- 100      # データ数
> mc <- 10000   # シミュレーション回数
> result <- as.data.frame(t(
+   replicate(mc, myest(runif(n)))
+ ))
> ## sqrt(n)*(xbar-mu)/s の分布が標準正規分布で近似できることの確認
> mu <- 0.5 # 真の平均
> hist(sqrt(n)*(result$xbar-mu)/result$s, freq=FALSE,
+       col="lightblue", breaks=20,
+       xlab=expression(sqrt(n)*(bar(X)-mu)/s),
+       main=expression(paste("Histogram of ", sqrt(n)*(bar(X)-mu)/s)))
> curve(dnorm, add=TRUE, col="red", lwd=2)
> ## 真の平均 mu が 95%信頼区間に含まれている割合が約 95%であることの確認
> with(result,sum((mu<=ui & mu>=li))/mc
[1] 0.9484

> ## はじめの 20 個の信頼区間の可視化
> require(plotrix)
> with(result, # 信頼区間と標本平均の図示
+   plotCI(1:20, xbar[1:20], ui=ui[1:20], li=li[1:20],
+           pch=4, axes=FALSE, xlab="", ylab=expression(mu), lwd=2))
> axis(1, 1:20, 1:20) # x 軸の追加
> axis(2) # y 軸の追加
> abline(h=mu, col="red", lty="dashed", lwd=2) # 真の平均の図示

```

(est-ci-mean-asymp.r)

4.5.2 最尤法による区間推定

広い範囲の確率分布に対して、最尤推定量は漸近正規性をもつことが知られている。より正確に述べるために、確率分布 \mathcal{L} は p 個の未知パラメータ $\boldsymbol{\theta} := (\theta_1, \dots, \theta_p)$ を含む確率分布で

4 推定

モデル化されているとして、その対数尤度関数を $\ell(\boldsymbol{\theta})$ 、最尤推定量を $\hat{\boldsymbol{\theta}}$ とする。また、 $\ell(\boldsymbol{\theta})$ は C^2 級 (2 階微分可能) であると仮定し、 p 次正方行列 $I(\boldsymbol{\theta})$ を

$$I(\boldsymbol{\theta}) = \left(\frac{1}{n} E \left[\left(\frac{\partial \ell}{\partial \theta_i}(\boldsymbol{\theta}) \right) \left(\frac{\partial \ell}{\partial \theta_j}(\boldsymbol{\theta}) \right) \right] \right)_{1 \leq i, j \leq p} \quad (4.50)$$

で定義する (期待値は定義できると仮定する)。さらに、真のパラメータ値を $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_p^*)$ とし、 $I(\boldsymbol{\theta}^*)$ は正則であると仮定する。このとき、緩やかな仮定の下で、各 $j = 1, \dots, p$ について、 v_j を行列 $I(\boldsymbol{\theta}^*)^{-1}$ の第 j 対角成分とすると、推定誤差 $\hat{\theta}_j - \theta_j^*$ の分布は平均 0、分散 v_j/n の正規分布で近似できていることが知られている。より正確には、任意の $a \leq b$ に対して、

$$P \left(a \leq \frac{\sqrt{n}(\hat{\theta}_j - \theta_j^*)}{\sqrt{v_j}} \leq b \right) \rightarrow \int_a^b \phi(x) dx \quad (n \rightarrow \infty) \quad (4.51)$$

が成り立つ。従って、上と同様の議論によって、

$$[\hat{\theta}_j - z_{1-\alpha/2} \cdot v_j / \sqrt{n}, \hat{\theta}_j + z_{1-\alpha/2} \cdot v_j / \sqrt{n}] \quad (4.52)$$

はサンプル数 n が十分大きい場合に近似的に正しいパラメータ θ_j の $100(1-\alpha)\%$ 信頼区間を与える。一般には v_j は未知パラメータ $\boldsymbol{\theta}^*$ に依存するため、 $I(\hat{\boldsymbol{\theta}})^{-1}$ の第 j 対角成分で置き換える。この場合でも近似式 (4.51) は成り立つため、上の議論は引き続き正当化される。

行列 $I(\boldsymbol{\theta}^*)$ は **Fisher 情報行列** と呼ばれる。特に $p = 1$ の場合は 4.1 節で述べた Fisher 情報量と同じものとなる。多くの場合に次の等式が成り立つ:

$$I(\boldsymbol{\theta}^*) = \left(-\frac{1}{n} E \left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}^*) \right] \right)_{1 \leq i, j \leq p}. \quad (4.53)$$

のことから、 $I(\boldsymbol{\theta}^*)$ を解析的に求めるのが困難な場合、

$$\left(-\frac{1}{n} \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}^*) \right)_{1 \leq i, j \leq p} \quad (4.54)$$

で代用することが多い。これは、上の行列の各成分は独立同分布な確率変数の平均で表されるため、適当な仮定の下で大数の強法則より $I(\boldsymbol{\theta}^*)$ の一致推定量となるからである。関数 `optim()` で対数尤度関数のマイナスを最小化する場合、オプション `hessian` を `TRUE` に指定することで、上の行列に n をかけたものを数値計算した値を計算するようにできる (詳細は `help(optim)` を参照すること)。

例題 4.6 (Poisson 分布の Fisher 情報量). 例 4.2 より、パラメータ λ の Poisson 分布の Fisher 情報量は $I(\lambda) = \lambda^{-1}$ で与えられる。

例題 4.7 (指数分布の Fisher 情報量). 例 4.3 より、パラメータ λ の指数分布の Fisher 情報量は $I(\lambda) = \lambda^{-2}$ で与えられる。

例題 4.8 (ガンマ分布の Fisher 情報行列). 例 4.4 より、パラメータ ν, α のガンマ分布の Fisher 情報行列は

$$I(\nu, \alpha) = \begin{pmatrix} \psi'(\nu) & -\alpha^{-1} \\ -\alpha^{-1} & \nu \alpha^{-2} \end{pmatrix} \quad (4.55)$$

で与えられる。ここで、ディガンマ関数の導関数 $\psi'(\nu)$ はトリガンマ関数と呼ばれ、R では関数 `trigamma()` で計算できる。Fisher 情報行列の逆行列は、

$$I(\nu, \alpha)^{-1} = \frac{1}{\nu \psi'(\nu) - 1} \begin{pmatrix} \nu & \alpha \\ \alpha & \alpha^2 \psi'(\nu) \end{pmatrix} \quad (4.56)$$

となる。

4 推定

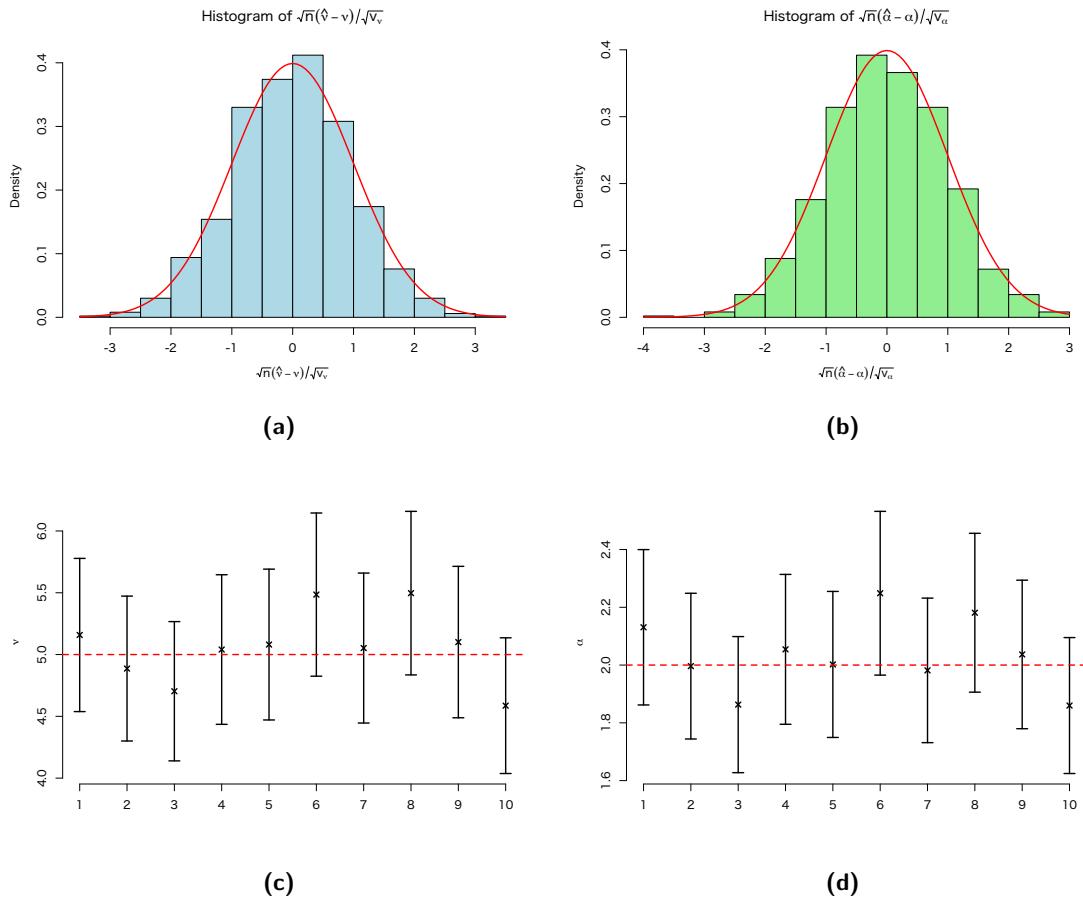


Figure 4.6: 最尤推定による区間推定の例

[Figure 4.6 を参照]

```
> #### 最尤法によるガンマ分布のパラメーターの区間推定
> ## ガンマ乱数から標本平均・標準偏差と平均の 100(1-conf)%信頼区間の
> ## 上端・下端を計算する関数の作成
> ## 以下注意：ガンマ関数のパラメータ alpha と区別して信頼係数は conf を使用
> require(stats4) # 関数 mle を利用するため
> mle.gamma <- function(x, nu0=1, alpha0=1){ # 最尤推定で用いた定義を簡略化
+   n <- length(x)
+   ll <- function(nu, alpha)
+     suppressWarnings(-sum(dgamma(x, nu, alpha, log=TRUE)))
+   est <- mle(minuslogl=ll, # 負の対数尤度関数
+             start=list(nu=nu0, alpha=alpha0), # 初期値
+             method="BFGS", nobs=n)
+   return(coef(est))
+ }
> ci.gamma <- function(theta, n, conf=0.05){ # theta=(nu, alpha)
+   ## nu の最尤推定量の漸近分散
+   v1hat <- theta[1]/(theta[1]*trigamma(theta[1])-1)
+   ## alpha の最尤推定量の漸近分散
+   v2hat <- theta[2]^2*trigamma(theta[1])/
+     (theta[1]*trigamma(theta[1])-1)
+   ui1 <- theta[1] + qnorm(1-conf/2)*sqrt(v1hat/n) # nu の上側信頼限界
+   li1 <- theta[1] - qnorm(1-conf/2)*sqrt(v1hat/n) # nu の下側信頼限界
+   ui2 <- theta[2] + qnorm(1-conf/2)*sqrt(v2hat/n) # alpha の上側信頼限界
```

4 推定

```

+     li2 <- theta[2] - qnorm(1-conf/2)*sqrt(v2hat/n) # alpha の下側信頼限界
+     return(c(hat=theta[1], hat=theta[2], v=v1hat, v=v2hat,
+             ui=ui1, li=li1, ui=ui2, li=li2))
+ }
> myest <- function(x, conf=0.05){
+   theta <- mle.gamma(x) # 最尤推定量
+   return(ci.gamma(theta,length(x),conf))
+ }
> ## Monte-Carlo 実験
> set.seed(123) # 亂数のシード値の設定
> nu <- 5
> alpha <- 2
> n <- 500 # データ数
> mc <- 1000 # シミュレーション回数
> result <- as.data.frame(t(
+   replicate(mc, myest(rgamma(n, shape=nu, rate=alpha))))
+ ))
> ## 漸近正規性の確認
> hist(sqrt(n)*(result$hat.nu-nu)/sqrt(result$v.nu), freq=FALSE,
+       col="lightblue", breaks=20,
+       xlab=expression(sqrt(n)*(hat(nu)-nu)/sqrt(v[nu])),
+       main=expression(paste("Histogram of ",
+                             sqrt(n)*(hat(nu)-nu)/sqrt(v[nu]))))
> curve(dnorm, add=TRUE, col="red", lwd=2)
> hist(sqrt(n)*(result$hat.alpha-alpha)/sqrt(result$v.alpha), freq=FALSE,
+       col="lightgreen", breaks=20,
+       xlab=expression(sqrt(n)*(hat(alpha)-alpha)/sqrt(v[alpha])),
+       main=expression(paste("Histogram of ",
+                             sqrt(n)*(hat(alpha)-alpha)/sqrt(v[alpha]))))
> curve(dnorm, add=TRUE, col="red", lwd=2)
> ## 真のパラメーターが 95%信頼区間に含まれている割合が約 95%であることを確認
> with(result, sum((nu<=ui.nu & nu>=li.nu))/mc
[1] 0.956
> with(result, sum((alpha<=ui.alpha & alpha>=li.alpha))/mc
[1] 0.954
> ## はじめの 10 個の信頼区間の可視化
> require(plotrix)
> with(result, # 信頼区間と最尤推定値の図示
+       plotCI(1:10, hat.nu[1:10], ui=ui.nu[1:10], li=li.nu[1:10],
+               pch=4, axes=FALSE, xlab="", ylab=expression(nu), lwd=2))
> axis(1, 1:10, 1:10) # x 軸の追加
> axis(2) # y 軸の追加
> abline(h=nu, col="red", lty="dashed", lwd=2) # 真のパラメータの図示
> with(result, # 信頼区間と最尤推定値の図示
+       plotCI(1:10, hat.alpha[1:10], ui=ui.alpha[1:10], li=li.alpha[1:10],
+               pch=4, axes=FALSE, xlab="", ylab=expression(alpha), lwd=2))
> axis(1, 1:10, 1:10) # x 軸の追加
> axis(2) # y 軸の追加
> abline(h=alpha, col="red", lty="dashed", lwd=2) # 真のパラメータの図示
> ## 実データへの適用
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> x <- kikou$風速 # 風速のデータにガンマ分布をあてはめてみる
> (theta <- mle.gamma(x)) # 最尤推定
      nu      alpha
12.23397  4.39242
> ci <- ci.gamma(theta,length(x),conf=0.05)
> c(ci["li.nu"], ci["ui.nu"]) # nu の 95%信頼区間

```

```
    li.nu      ui.nu
10.48509 13.98285
> c(ci["li.alpha"], ci["ui.alpha"]) # alpha の 95%信頼区間
li.alpha ui.alpha
3.751466 5.033374
(est-ci-mle.r)
```

演習 4.9. 最尤推定量の漸近正規性について調べてみよう.

1. 上の実行例において, 最尤推定量の漸近分散を関数 `optim()` のオプション `hessian` を `TRUE` と指定して計算する方法について調べてみよ.
2. Poisson 分布や指数分布の場合の最尤推定量の漸近正規性について, シミュレーションで確認せよ.

4.6 補遺

4.6.1 参考文献

- [1] 東京大学教養学部統計学教室. **統計学入門**. 東京: 東京大学出版会, 1991.
- [2] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.

5 検定

仮説が正しいかどうかをデータにもとづいて判断する統計的な方法を**統計的仮説検定**あるいは単に**検定**という。推定と大きく異なるのは、母集団の分布に対して何らかの**仮説**を考えるところにある。基本的な手続きとしては、適当な統計量に対して仮説が正しいときの標本分布を調べ、データから計算した統計量の値が仮説に従う母集団から得られたと考えるに十分高い確率かどうかに基づいて仮説が正しいか否かを判断する。この判断に用いられる統計量を**検定統計量**と呼ぶ。

5.1 統計的検定の考え方

例えば、新しい薬が古い薬より良いことを示したい場合、「新しい薬も古い薬も効能は同じ」という仮説を考え、新しい薬のデータも古い薬のデータも同じ分布から出たと仮定する。このとき2つのデータ集合は、平均値が等しいとか分散が等しいとか、統計的に同じ性質を持っていると考えられるが、これをデータから実際に計算される統計量にもとづき判断する。

5.1.1 帰無仮説と対立仮説

これまで学んだように、統計学ではデータを確率変数の実現値とみなしてモデル化するため、統計量はデータの実現値ごとに異なる値を持ちばらつくが、そのばらつきは統計量の分布の性質によって予想することができる。データから計算された検定統計量の値がこの予想と著しく異なる場合には、最初に立てた仮説がおそらく正しくないと考えるのが妥当であろう。このとき検定のために立てる仮説を**帰無仮説**と呼ぶ。多くの場合「この仮説を捨てて無に帰したい」ことを期待して立てられるため、「帰無」という言葉が使われる。帰無仮説が正しい場合の検定統計量の分布を**帰無分布**と呼ぶ。帰無分布は、帰無仮説の下で検定統計量が取りうる値の範囲を予想するのに必要となる。

検定統計量の分布はデータから計算される検定統計量の値の出現し易さを表しているので、実験を繰り返した場合には分布の中心部分の値を多く観測することになる。一方、帰無仮説が正しくない場合には本来起こり難い分布の裾部分の値が観測されることになる。帰無仮説のもとで起こり易い値と起こり難い値を区別するために、帰無仮説に対して比較したい仮説として考えるのが**対立仮説**である。帰無仮説のもとでは起こり難く、対立仮説のもとでは起こり易い値の領域を決めて、検定統計量がこの領域に入ったら帰無仮説を信じるには根拠が薄いと考えて、「帰無仮説を棄却する」というのが一般的な検定の手続きとなる。逆に検定統計量からは帰無仮説を積極的に棄却することができない場合「帰無仮説を受容する」という。この帰無仮説を棄却するために決める領域を**棄却域**と言う。

通常、棄却域の大きさは帰無仮説が正しいときに検定統計量が棄却域に入る確率を用いて定め、この確率を**有意水準**と言う。

有意水準は、帰無仮説が正しいにも係わらず帰無仮説を棄却する確率なので、誤りを起こす確率であることに注意しよう。この誤りを**第一種過誤**という。一方、対立仮説が正しいにも係わらず帰無仮説を受容してしまう誤りもありうる。この誤りを**第二種過誤**という。どちらの誤りも小さいほど良いが、棄却域をどのように取るかを考えるときには、一般には第一種過誤が起きる確率の上限(有意水準)を定めた上で、できるだけ第二種過誤の起きる確率を小さくするような戦略が取られる。第一種過誤が起きる確率を**サイズ**、第二種過誤が起きない確率を**検出力**と呼ぶ。上で述べたことは、サイズを有意水準以下に抑えた上で、可能な限り検出力を大きくするように棄却域をとるのが一般的な戦略であると言えられる。

5.1.2 片側検定と両側検定

前節で述べた第一種過誤と第二種過誤のバランスを考えると、帰無仮説のみによって棄却域が決まるのではなく、対立仮説の立て方によって棄却域の形は変わり得ることがわかる。

例えば、先の薬の例であれば帰無仮説として「新しい薬も古い薬も効能は同じ」、対立仮説として「新しい薬の方が古い薬より効能が高い」を考えるのが自然であろう。また、標本から計算する効能に関する統計量が大きいほど効能が高いとしよう。このとき、対立仮説が正しければ、統計量は帰無仮説のもとで期待される値より大きくなるであろう。この場合、値の大きな領域に棄却域を適切な大きさで取ることになる。一方、同様の薬の比較であっても、新しい薬がコストを下げたもので効能が変わらないことを確認したい場合には、帰無仮説が正しいことを期待して対立仮説として「新しい薬と古い薬の効能が異なる」を考えることになる。この場合には、もし対立仮説が正しいとすれば値が大きい場合も小さい場合も考えられるので、棄却域としては値の大きな領域と小さな領域の両方を考えるのが自然である。

棄却域が片側に寄った前者の例を**片側検定**、棄却域が両側にある後者の例を**両側検定**と呼ぶ。検定を考える際には、このように帰無仮説と対立仮説の関係を明瞭にしておくことも重要である。

5.1.3 *p* 値

上記のように最初に有意水準を決めて棄却域を定める場合もあるが、データから計算された検定統計量の値に対して、その値が棄却域に含まれるような有意水準の最小値(厳密には下限)に基づいて検定を行う場合もある。この値のことを*p* 値という。この場合、検定の*p* 値が有意水準未満のときに帰無仮説を棄却することとなる。

母集団が正規分布の場合、その平均や分散に関しては正規分布、*t* 分布、 χ^2 分布、*F* 分布などが検定統計量の分布として現れることをこれまでに学んだ。これらの理論的に求められる分布に基づいて棄却域の設定、*p* 値の計算を行えばよい。また、例えば最尤推定量のように、サンプル数が十分に大きければ推定量の分布が正規分布で十分良く近似できる場合もある。このときも正規分布にもとづいた検定を近似的に用いることができる。

5.2 正規母集団に対する検定(1標本)

この節では、観測データ X_1, X_2, \dots, X_n が平均 μ 、分散 σ^2 の正規分布に従う独立同分布な確率変数列としてモデル化されている場合に、 μ および分散 σ^2 に対する検定を行う方法を説明する。

5.2.1 平均の検定

まず、 μ_0 を既知の定数として、平均 μ が μ_0 であるか否かを検定する問題を考える。検定の用語を使って述べると、帰無仮説を $\mu = \mu_0$ 、対立仮説を $\mu \neq \mu_0$ とする検定を考える。これはしばしば次の記号で表される(H_0 が帰無仮説、 H_1 が対立仮説を表す):

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0. \quad (5.1)$$

この仮説に対する検定は標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ が μ_0 からどの程度離れているかを検証することで行われる。より具体的には、 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ を不偏分散とし、検定統計量として

$$t = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \quad (5.2)$$

を考える。仮に帰無仮説 H_0 が正しいとすると、 t は自由度 $n-1$ の *t* 分布に従う(4.4.2 節参照)。従って、 $\alpha \in (0, 1)$ に対して、自由度 $n-1$ の *t* 分布の $1-\alpha/2$ 分位点を $t_{1-\alpha/2}(n-1)$ と

5 檢定

すれば、 H_0 の下では

$$P(|t| > t_{1-\alpha/2}(n-1)) = \alpha \quad (5.3)$$

が成り立つ(4.4.2節参照)。以上より、有意水準を α とする場合、棄却域を

$$(-\infty, -t_{1-\alpha/2}(n-1)) \cup (t_{1-\alpha/2}(n-1), \infty) \quad (5.4)$$

と設定すれば、第一種過誤の上限が α となる。具体的な検定の手順としては、データから検定統計量 t の値を計算し、

$$|t| > t_{1-\alpha/2}(n-1) \quad (5.5)$$

であった場合には帰無仮説を棄却する。もしくは前節で述べたように、検定の p 値を計算して、 p 値が α 未満の場合に帰無仮説を棄却するという手順をとってもよい。いまの場合の検定の p 値は、 $f(x)$ を自由度 $n-1$ の t 分布の確率密度関数として、

$$2 \int_{|t|}^{\infty} f(x) dx \quad (5.6)$$

によって与えられる。

$$|t| > t_{1-\alpha/2}(n-1) \Leftrightarrow \int_{-\infty}^{|t|} f(x) dx > 1 - \alpha/2 \Leftrightarrow 2 \int_{|t|}^{\infty} f(x) dx < \alpha \quad (5.7)$$

が成り立つから、検定統計量が棄却域に入ることと p 値が有意水準未満となることは同じ意味である。(5.6)に現れる積分は関数 `pt()` のオプション `df` に自由度を、オプション `lower.tail` に `FALSE` を指定することで計算できるが、いまの場合は上の検定を実行するための関数 `t.test()` が p 値も計算してくれる。

なお、この検定のように、帰無分布が t 分布となるような検定を **t 検定** と呼ぶ。また、上の検定は **Student の t 検定** と呼ばれることがある。

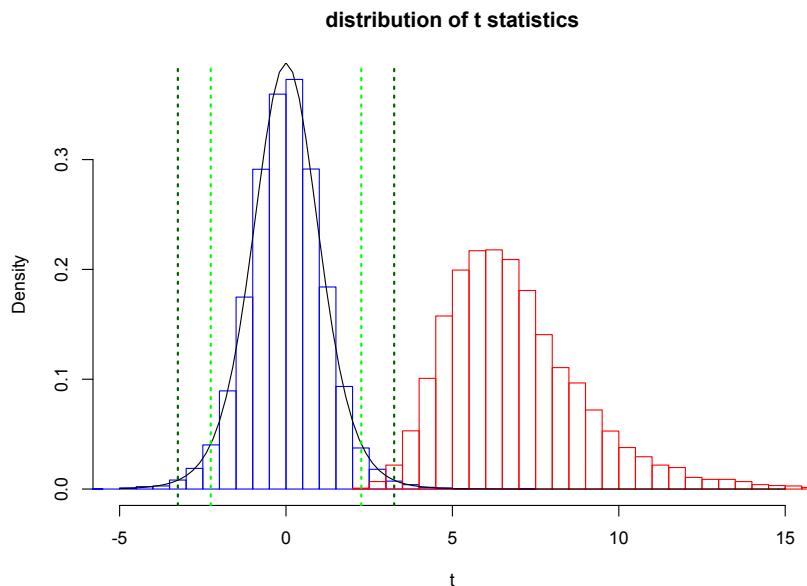


Figure 5.1: 平均値の検定の例

5 検定

[Figure 5.1 を参照]

```
> ### 平均値の検定 (Student の t 検定)
> set.seed(123) # 亂数のシード値の設定
> mu <- 10    # 平均
> sigma <- 5  # 標準偏差
> n <- 10     # データ数
> x <- rnorm(n, mean=mu, sd=sigma) # 正規乱数の生成
> t.test(x, mu=mu) # mu0=mu として検定を実行 (帰無仮説は正しい)

One Sample t-test

data: x
t = 0.24742, df = 9, p-value = 0.8101
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 6.961648 13.784608
sample estimates:
mean of x
10.37313

> t.test(x)          # mu0=0 として検定を実行 (帰無仮説は誤り)

One Sample t-test

data: x
t = 6.8784, df = 9, p-value = 7.239e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 6.961648 13.784608
sample estimates:
mean of x
10.37313

> ## 実験を繰り返した場合の検定の棄却率の確認
> mytest <- function(x, mu0=0){ # 検定統計量の値と p 値を計算する関数
+   res <- t.test(x, mu=mu0) # t 検定の実行
+   t.val <- res$statistic  # 検定統計量の値
+   p.val <- res$p.value    # p 値
+   return(c(t.val, p=p.val))
+ }
> ## Monte-Carlo 実験
> mc <- 10000 # 実験回数
> ## 帰無仮説が正しい場合 (mu0=mu として実験)
> res1 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(n, mean=mu, sd=sigma), mu0=mu)))
+ ))
> head(res1)

      t         p
1 0.63552401 0.54090693
2 -1.44237189 0.18307871
3 1.93132875 0.08549017
4 -0.02546003 0.98024361
5 0.81853782 0.43418722
6 0.41566425 0.68739126

> mean(res1$p < 0.05) # 有意水準 5%で棄却された実験の割合
[1] 0.0485

> mean(res1$p < 0.01) # 有意水準 1%で棄却された実験の割合
```

5 檢定

```
[1] 0.01
> ## 帰無仮説が誤りの場合 (mu0=0 として実験)
> res2 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(n, mean=mu, sd=sigma))))
+ ))
> mean(res2$p < 0.05) # 有意水準 5%で棄却された実験の割合
[1] 0.9999
> mean(res2$p < 0.01) # 有意水準 1%で棄却された実験の割合
[1] 0.9926
> ## 検定統計量のヒストグラム
> hist(res1$t, freq=FALSE, xlim=c(-5, 15), breaks=20, border="blue",
+       main="distribution of t statistics", xlab="t")
> hist(res2$t, freq=FALSE, add=TRUE, breaks=30, border="red")
> ## 帰無分布の理論曲線
> curve(dt(x, df=9), add=TRUE)
> ## 棄却域の可視化 (有意水準 5%)
> abline(v=c(-qt(0.975, df=9), qt(0.975, df=9)),
+          lty=3, lwd=2, col="green")
> ## 棄却域の可視化 (有意水準 1%)
> abline(v=c(-qt(0.995, df=9), qt(0.995, df=9)),
+          lty=3, lwd=2, col="darkgreen")

```

(test-mean.r)

前節で述べたように、対立仮説の立て方によって棄却域の形は変わりうる。例えば、前節で述べた例に対応して、観測データ X_1, X_2, \dots, X_n が新しい薬の効能を確認するためにその薬を n 人の被験者に投与した際の治験結果であったとする。(例えばその薬が睡眠薬であれば、データは睡眠時間の伸び具合に対応するであろう)。このとき、母集団分布の平均 μ は新薬の「真の」、もしくは「平均的な」効能に対応するから、 μ が古い薬の効能 μ_0 と比較して大きいと言えるのかどうかを考えるのが自然である。すなわち、帰無仮説として $\mu = \mu_0$ 、対立仮説として $\mu > \mu_0$ を設定した検定

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0. \quad (5.8)$$

を考えるのが自然である。この場合、帰無仮説は検定 (5.1) と同一なので、検定統計量としても同一のもの t が利用出来る(帰無分布が計算可能であるため)。他方、対立仮説の下では検定統計量 t の値が正の方向に大きくなると期待される。従って、棄却域としては、 c をある正の数として、“ $t > c$ ” という形のものを考えるのが自然である(すなわち片側検定となる)。いま、 $\alpha \in (0, 1)$ に対して、自由度 $n-1$ の t 分布の $1-\alpha$ 分位点を $t_{1-\alpha}(n-1)$ とすれば、 H_0 の下で

$$P(t > t_{1-\alpha}(n-1)) = \alpha \quad (5.9)$$

が成り立つ。以上より、棄却域を

$$(t_{1-\alpha}(n-1), \infty) \quad (5.10)$$

と設定すれば、第一種過誤の上限が α となる。具体的な検定の手順としては、データから検定統計量 t の値を計算し、

$$t > t_{1-\alpha}(n-1) \quad (5.11)$$

であった場合には帰無仮説を棄却する。もしくは、 $f(x)$ を自由度 $n-1$ の t 分布の確率密度関数として、 p 値

$$\int_t^\infty f(x)dx \quad (5.12)$$

5 検定

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等である。

反対向きの対立仮説 $\mu < \mu_0$ を考えた検定

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0. \quad (5.13)$$

の場合は、自由度 $n-1$ の t 分布の α 分位点を $t_\alpha(n-1)$ として、

$$t < t_\alpha(n-1) \quad (5.14)$$

であった場合に帰無仮説を棄却すれば良い。これは、 p 値

$$\int_{-\infty}^t f(x)dx \quad (5.15)$$

が α 未満であった場合に帰無仮説を棄却するということと同じである。

```
> ##### 平均値の検定における両側・片側検定の検出力の比較
> set.seed(123) # 亂数のシード値の設定
> mytest <- function(x, mu0=0){ # 各種 p 値を計算する関数
+   p1 <- t.test(x, mu=mu0)$p.value # 両側検定の p 値
+   p2 <- t.test(x, mu=mu0, alternative="greater")$p.value # 右側検定の p 値
+   p3 <- t.test(x, mu=mu0, alternative="less")$p.value # 左側検定の p 値
+   return(c(two=p1, right=p2, left=p3))
+ }
> ## Monte-Carlo 実験
> mu <- 0.5 # 平均
> sigma <- 1 # 標準偏差
> n <- 5 # データ数
> mc <- 1000 # 実験回数
> ## 帰無仮説を mu0=0 として実験 (帰無仮説が誤りで棄却して欲しい場合)
> result <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(n, mean=mu, sd=sigma)))
+ ))
> alpha <- 0.05 # 有意水準
> colMeans(result<alpha) # 各検定 (両側, 右側, 左側) の棄却率=検出力

  two right left
0.134 0.244 0.004

> ##### データセット sleep による例 (睡眠薬の効果)
> plot(extra ~ group, data = sleep)
> (x <- subset(sleep, group==1, extra, drop=TRUE)) # group 1 の睡眠時間の伸び
[1] 0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
> (y <- subset(sleep, group==2, extra, drop=TRUE)) # group 2 の睡眠時間の伸び
[1] 1.9  0.8  1.1  0.1 -0.1  4.4  5.5  1.6  4.6  3.4
> t.test(x, mu=0, alternative="greater")

One Sample t-test

data: x
t = 1.3257, df = 9, p-value = 0.1088
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
-0.2870553      Inf
sample estimates:
mean of x
0.75
```

5 檢定

```
> t.test(y, mu=0, alternative="greater")
One Sample t-test

data: y
t = 3.6799, df = 9, p-value = 0.002538
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
1.169334      Inf
sample estimates:
mean of x
2.33

(test-region.r)
```

演習 5.1. 平均の検定について調べてみよう.

1. 適当な正規分布を設定し, 検定のための統計量の標本分布をシミュレーションにより求めなさい.
2. 帰無仮説が正しい状況および正しくない状況を設定し, 正しく受容される確率および正しく棄却される確率をシミュレーションにより調べなさい.
3. 実際のデータについて適当な仮説を設定して, Student の t 検定を実行してみよ.

5.2.2 分散の検定

次に, σ_0^2 を既知の定数として, 分散 σ^2 が σ_0^2 であるか否かを検定する問題を考える. 検定の用語を使って述べると, 帰無仮説を $\sigma^2 = \sigma_0^2$, 対立仮説を $\sigma^2 \neq \sigma_0^2$ とする検定

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 \neq \sigma_0^2 \quad (5.16)$$

を考える. ここで $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ を不偏分散とする. 仮に帰無仮説 H_0 が正しいとすると, 統計量

$$\chi^2 = (n-1)s^2 / \sigma_0^2 \quad (5.17)$$

は帰無仮説 H_0 の下で自由度 $n-1$ の χ^2 分布に従う (4.4.2 節, 命題 4.2 参照). 従って, $\alpha \in (0, 1)$ に対して, 自由度 $n-1$ の χ^2 分布の $\alpha/2, 1-\alpha/2$ 分位点をそれぞれ $\chi_{\alpha/2}^2(n-1), \chi_{1-\alpha/2}^2(n-1)$ とすれば, H_0 の下では

$$P(\chi^2 < \chi_{\alpha/2}^2(n-1) \text{ または } \chi^2 > \chi_{1-\alpha/2}^2(n-1)) = \alpha \quad (5.18)$$

が成り立つ (4.4.3 節参照). 以上より, 有意水準を α とする場合, 棄却域を

$$(-\infty, \chi_{\alpha/2}^2(n-1)) \cup (\chi_{1-\alpha/2}^2(n-1), \infty) \quad (5.19)$$

と設定すれば, 第一種過誤の上限が α となる. 具体的な検定の手順としては, データから検定統計量 χ^2 の値を計算し,

$$\chi^2 < \chi_{\alpha/2}^2(n-1) \text{ または } \chi^2 > \chi_{1-\alpha/2}^2(n-1) \quad (5.20)$$

であった場合には帰無仮説を棄却する. もしくは, この場合の p 値は, 自由度 $n-1$ の χ^2 分布の確率密度関数を $f(x)$ とすると

$$2 \min \left\{ \int_0^{\chi^2} f(x) dx, \int_{\chi^2}^{\infty} f(x) dx \right\} \quad (5.21)$$

5 検定

で与えられるので、この値が α 未満の場合に帰無仮説を棄却するというのと同じである。

対立仮説が片側の場合

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 > \sigma_0^2 \quad (5.22)$$

を考えたときも、前の小節と同様の議論によって検定を構成できる。すなわち、自由度 $n-1$ の χ^2 分布の $1-\alpha$ 分位点を $\chi_{1-\alpha}^2(n-1)$ として、

$$\chi^2 > \chi_{1-\alpha}^2(n-1) \quad (5.23)$$

であった場合に帰無仮説を棄却すれば良い。検定の p 値は

$$\int_{\chi^2}^{\infty} f(x) dx \quad (5.24)$$

で与えられるので、この値が α 未満の場合に帰無仮説を棄却するとしても同じである。左側対立仮説 $H_1 : \sigma^2 < \sigma_0^2$ の場合も同様なので詳細は省略する。

なお、この検定のように、帰無分布が χ^2 分布となるような検定を χ^2 検定と呼ぶ。

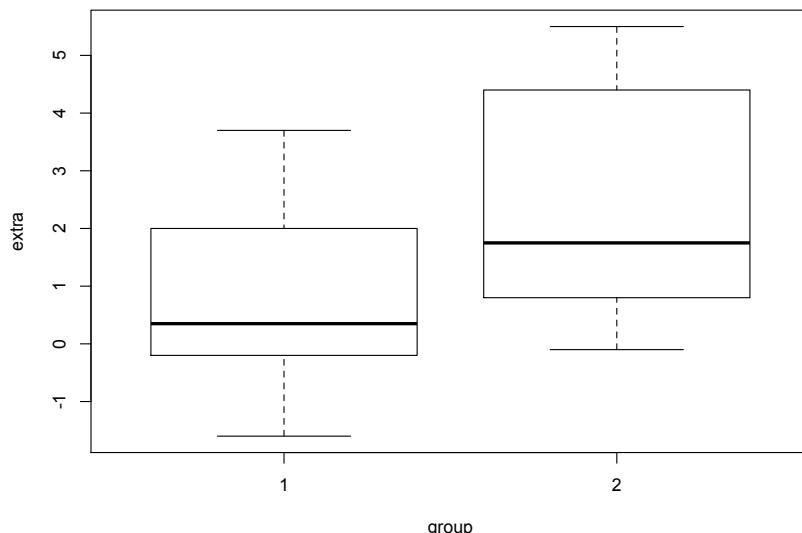


Figure 5.2: 分散の検定の例

[Figure 5.2 を参照]

```
> ### 分散の検定
> set.seed(123) # 亂数のシード値の設定
> mu <- 5      # 真の平均
> sigma <- 3    # 真の分散
> n <- 10      # データ数
> x <- rnorm(n, mean=mu, sd=sigma) # 正規乱数の生成
> ## 帰無仮説が正しい場合 (両側検定)
> sigma0 <- sigma
> chi2 <- (n-1)*var(x)/sigma0^2 # 検定統計量
> p0 <- pchisq(chi2, df=n-1)
> 2*min(p0, 1-p0) # p 値
```

5 検定

```
[1] 0.9692336

> ## 帰無仮説が誤っている場合 (両側検定)
> sigma0 <- 2
> chi2 <- (n-1)*var(x)/sigma0^2 # 検定統計量
> p0 <- pchisq(chi2, df=n-1)
> 2*min(p0, 1-p0) # p 値

[1] 0.06117309

> ## 実験を繰り返した場合の検定の棄却率の確認
> mytest <- function(x, sigma0){ # 検定統計量の値と p 値を計算する関数
+   chi2.val <- (n-1)*var(x)/sigma0^2 # 検定統計量
+   p0 <- pchisq(chi2.val, df=n-1)
+   p.val <- 2*min(p0, 1-p0) # p 値
+   return(c(chisq=chi2.val, p=p.val))
+ }
> ## Monte-Carlo 実験
> mc <- 10000 # 実験回数
> ## 帰無仮説が正しい場合 (sigma0=sigma として実験)
> res1 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(n, mean=mu, sd=sigma), sigma0=sigma)))
+ ))
> mean(res1$p < 0.05) # 有意水準 5%で棄却された実験の割合

[1] 0.0518

> mean(res1$p < 0.01) # 有意水準 1%で棄却された実験の割合

[1] 0.0094

> ## 帰無仮説が誤りの場合 (sigma0=2 として実験)
> res2 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(n, mean=mu, sd=sigma), sigma0=2)))
+ ))
> mean(res2$p < 0.05) # 有意水準 5%で棄却された実験の割合

[1] 0.4884

> mean(res2$p < 0.01) # 有意水準 1%で棄却された実験の割合

[1] 0.3171

> ## 検定統計量のヒストグラム
> hist(res1$chisq, freq=FALSE, xlim=c(0, 50), breaks=20, border="blue",
+       main=expression(paste("distribution of ", chi^2, " statistics")),
+       xlab=expression(chi^2))
> hist(res2$chisq, freq=FALSE, add=TRUE, breaks=30, border="red")
> ## 帰無分布の理論曲線
> curve(dchisq(x, df=9), add=TRUE)
> ## 棄却域の可視化 (有意水準 5%)
> abline(v=c(qchisq(0.025, df=9), qchisq(0.975, df=9)),
+          lty=3, lwd=2, col="green")
> ## 棄却域の可視化 (有意水準 1%)
> abline(v=c(qchisq(0.005, df=9), qchisq(0.995, df=9)),
+          lty=3, lwd=2, col="darkgreen")

(test-var.r)
```

演習 5.2. 分散の検定について調べてみよう。

1. 適当な正規分布を設定し、検定のための統計量の標本分布をシミュレーションにより

5 検定

求めなさい.

2. 帰無仮説が正しい状況および正しくない状況を設定し, 正しく受容される確率および正しく棄却される確率をシミュレーションにより調べなさい.
3. 実際のデータについて適当な仮説を設定して, 分散の χ^2 検定を実行してみよ.

5.3 正規母集団に対する検定 (2 標本)

この節では, 2 種類の観測データとして, X_1, X_2, \dots, X_m および Y_1, Y_2, \dots, Y_n が与えられている状況で, 両者の平均や分散が一致するかどうかを検定する問題を考える. 以下では次の 3 つの条件が満たされていると仮定する:

1. $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ は独立な確率変数列である.
2. X_1, X_2, \dots, X_m は同分布であり, 平均 μ_1 , 分散 σ_1^2 の正規分布に従う.
3. Y_1, Y_2, \dots, Y_n は同分布であり, 平均 μ_2 , 分散 σ_2^2 の正規分布に従う.

5.3.1 平均の差の検定 (等分散の場合)

2 種類のデータの平均が等しいか否かを検定する問題

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2 \quad (5.25)$$

を考える. まず, 2 種類のデータの分散が一致することがあらかじめわかっている場合, すなわち $\sigma_1^2 = \sigma_2^2 =: \sigma^2$ であることがわかっている場合について考察する. この場合, それぞれのデータの標本平均を $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ とすると, 命題 8.1 より $\bar{X} - \bar{Y}$ は平均 $\mu_1 - \mu_2$, 分散 $\sigma^2(\frac{1}{m} + \frac{1}{n})$ の正規分布に従うことがわかる. さらに,

$$s^2 = \frac{1}{m+n-2} \left\{ \sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 \right\} \quad (5.26)$$

とおくと, $(m+n-2)s^2/\sigma^2$ は自由度 $m+n-2$ の χ^2 分布に従い, かつ $\bar{X} - \bar{Y}, (m+n-2)s^2/\sigma^2$ は独立となることが知られている. 以上より, 検定統計量として

$$t = \frac{\bar{X} - \bar{Y}}{s \sqrt{\frac{1}{m} + \frac{1}{n}}} \left(= \frac{\frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{\frac{(m+n-2)s^2/\sigma^2}{m+n-2}}} \right) \quad (5.27)$$

を考えると, H_0 の下で t は自由度 $m+n-2$ の t 分布に従う. 従って, $\alpha \in (0, 1)$ に対して, 自由度 $m+n-2$ の t 分布の $1-\alpha/2$ 分位点を $t_{1-\alpha/2}(m+n-2)$ とすれば, H_0 の下では

$$P(|t| > t_{1-\alpha/2}(m+n-2)) = \alpha \quad (5.28)$$

が成り立つ. 以上より, 有意水準を α とする場合, 棄却域を

$$(-\infty, -t_{1-\alpha/2}(m+n-2)) \cup (t_{1-\alpha/2}(m+n-2), \infty) \quad (5.29)$$

と設定すれば, 第一種過誤の上限が α となる. 具体的な検定の手順としては, データから検定統計量 t の値を計算し,

$$|t| > t_{1-\alpha/2}(m+n-2) \quad (5.30)$$

であった場合には帰無仮説を棄却する.

p 値の計算方法や片側対立仮説の場合への対応方法は, 9.2.1 節と類推の議論となるため, ここでは省略する.

5 檢定

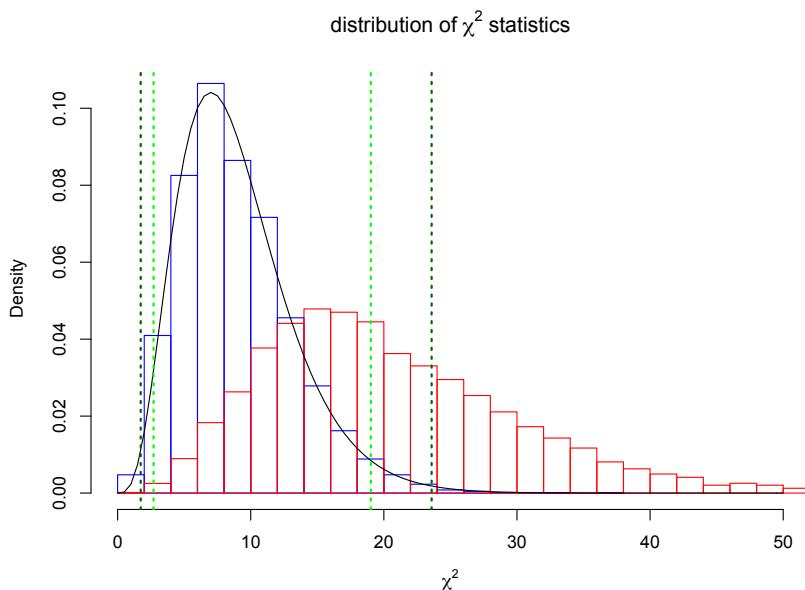


Figure 5.3: 等分散の平均値の差の検定の例

[Figure 5.3 を参照]

```
> ### 平均の差の検定（等分散の場合）
> set.seed(123) # 亂数のシード値の設定
> mu1 <- 5      # X/Yの平均
> mu2 <- 7.5    # Yの平均
> sigma <- 1    # 分散
> m <- 8        # Xのデータ数
> n <- 12       # Yのデータ数
> ## 帰無仮説が正しい場合（平均が等しい）
> x <- rnorm(m, mean=mu1, sd=sigma)
> y <- rnorm(n, mean=mu1, sd=sigma)
> t.test(x, y, var.equal=TRUE) # 等分散を仮定して t 検定を実行
```

Two Sample t-test

```
data: x and y
t = 0.34174, df = 18, p-value = 0.7365
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.7998173  1.1105593
sample estimates:
mean of x mean of y
5.234846 5.079475
```

```
> ## 帰無仮説が誤りの場合（平均が異なる）
> x <- rnorm(m, mean=mu1, sd=sigma)
> y <- rnorm(n, mean=mu2, sd=sigma)
> t.test(x, y, var.equal=TRUE) # 等分散を仮定して t 検定を実行
```

Two Sample t-test

```
data: x and y
t = -9.8499, df = 18, p-value = 1.127e-08
alternative hypothesis: true difference in means is not equal to 0
```

```

95 percent confidence interval:
-4.031983 -2.614354
sample estimates:
mean of x mean of y
4.454842 7.778010

> ## 実験を繰り返した場合の検定の棄却率の確認
> mytest <- function(x, y){
+   res <- t.test(x, y, var.equal=TRUE) # 等分散を仮定して t 検定を実行
+   t.val <- res$statistic # 検定統計量の値
+   p.val <- res$p.value # p 値
+   return(c(t.val, p=p.val))
+ }
> ## Monte-Carlo 実験
> mc <- 10000 # 実験回数
> ## 帰無仮説が正しい場合 (X と Y の平均は等しい)
> res1 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(m, mean=mu1, sd=sigma),
+                      rnorm(n, mean=mu1, sd=sigma))))
+ ))
> mean(res1$p < 0.05) # 有意水準 5%で棄却された実験の割合
[1] 0.0517

> mean(res1$p < 0.01) # 有意水準 1%で棄却された実験の割合
[1] 0.0098

> ## 帰無仮説が誤りの場合 (X と Y の平均は異なる)
> res2 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(m, mean=mu1, sd=sigma),
+                      rnorm(n, mean=mu2, sd=sigma))))
+ ))
> mean(res2$p < 0.05) # 有意水準 5%で棄却された実験の割合
[1] 0.9995

> mean(res2$p < 0.01) # 有意水準 1%で棄却された実験の割合
[1] 0.9906

> ## 検定統計量のヒストグラム
> hist(res1$t, freq=FALSE, xlim=c(-12, 5), breaks=20, border="blue",
+       main="distribution of t statistics", xlab="t")
> hist(res2$t, freq=FALSE, add=TRUE, breaks=20, border="red")
> ## 帰無分布の理論曲線
> curve(dt(x, df=m+n-2), add=TRUE)
> ## 棄却域の可視化 (有意水準 5%)
> abline(v=c(-qt(0.975, df=m+n-2), qt(0.975, df=m+n-2)),
+          lty=3, lwd=2, col="green")
> ## 棄却域の可視化 (有意水準 1%)
> abline(v=c(-qt(0.995, df=m+n-2), qt(0.995, df=m+n-2)),
+          lty=3, lwd=2, col="darkgreen")

```

(test-diff.r)

5.3.2 平均の差の検定 (一般の場合)

前小節に引き続いだ検定問題 (5.25) について考察するが、今度は σ_1^2, σ_2^2 が一致するかどうかはわからない状況を考える。この場合の検定問題 (5.25) は **Behrens-Fisher 問題**として知られており、正確かつ適切な検定を導出することは難しいことが知られている。そのため

5 檢定

め、通常は以下で説明する **Welch の近似法** (*Satterthwaite の近似法*と呼ばれることがある) と呼ばれる近似解が用いられる。

Welch の近似法は、一般に 2 つの独立な確率変数 Z, W があり、それぞれ自由度 k_1, k_2 の χ^2 分布に従うとき、 Z, W の線形結合 $aZ + bW$ (a, b は正の実数) で表される確率変数の分布を、正の実数 c, ν をうまく選んで、 $c\chi_\nu^2$ という形の確率変数の分布で近似する方法である。ただし、 χ_ν^2 は自由度 ν の χ^2 分布に従う確率変数である。パラメータ c, ν は $aZ + bW, c\chi_\nu^2$ の平均・分散が互いに一致するように選ぶ。すなわち、 c, ν は次の連立方程式の解となる：

$$\begin{cases} E[aZ + bW] = E[c\chi_\nu^2], \\ \text{Var}[aZ + bW] = \text{Var}[c\chi_\nu^2]. \end{cases} \quad (5.31)$$

この方程式を解くと、以下の解を得る：

$$c = \frac{a^2 k_1 + b^2 k_2}{a k_1 + b k_2}, \quad \nu = \frac{(a k_1 + b k_2)^2}{a^2 k_1 + b^2 k_2}. \quad (5.32)$$

Welch の近似法を適用することで、検定問題 (5.25) に対して以下のような近似解が得られる。まず、 X_1, \dots, X_m の不偏分散を s_1^2 、 Y_1, \dots, Y_n の不偏分散を s_2^2 とする：

$$s_1^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2, \quad s_2^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2. \quad (5.33)$$

このとき、 $\bar{X} - \bar{Y}, s_1^2, s_2^2$ は独立となることが知られている。また、命題 8.2 より $(m-1)s_1^2/\sigma_1^2, (n-1)s_2^2/\sigma_2^2$ はそれぞれ自由度 $m-1, n-1$ の χ^2 分布に従う。よって、確率変数 $s_1^2/m + s_2^2/n$ に Welch の近似法を適用すると、 $a = \sigma_1^2/m(m-1), b = \sigma_2^2/n(n-1), k_1 = m-1, k_2 = n-1$ であるから、

$$c = \frac{\frac{(\sigma_1^2/m)^2}{m-1} + \frac{(\sigma_2^2/n)^2}{n-1}}{\sigma_1^2/m + \sigma_2^2/n}, \quad \nu = \frac{(\sigma_1^2/m + \sigma_2^2/n)^2}{\frac{(\sigma_1^2/m)^2}{m-1} + \frac{(\sigma_2^2/n)^2}{n-1}} \quad (5.34)$$

を得る。よって、検定統計量

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_1^2/m + s_2^2/n}} \quad (5.35)$$

を考えると、 t の分布は $(\bar{X} - \bar{Y})/\sqrt{c\chi_\nu^2}$ で近似できる。ただし、 $\bar{X} - \bar{Y}, s_1^2/m + s_2^2/n$ は独立であるから、 $\bar{X} - \bar{Y}, \chi_\nu^2$ も独立である。いま、命題 8.1 より $((\bar{X} - \bar{Y}) - (\mu_1 - \mu_2))/\sqrt{\sigma_1^2/m + \sigma_2^2/n}$ は標準正規分布に従うので、 H_0 の下で

$$\frac{\bar{X} - \bar{Y}}{\sqrt{c\chi_\nu^2}} \left(= \frac{(\bar{X} - \bar{Y})/\sqrt{\sigma_1^2/m + \sigma_2^2/n}}{\sqrt{\frac{c}{\sigma_1^2/m + \sigma_2^2/n}\chi_\nu^2}} \right) \quad (5.36)$$

は自由度 ν の t 分布に従う ($\frac{c}{\sigma_1^2/m + \sigma_2^2/n} = \nu^{-1}$ に注意)。従って、元々の検定統計量 t の帰無分布は自由度 ν の t 分布で近似できることになる。 ν は未知の分散 σ_1^2, σ_2^2 を含むので、実際の応用ではこれらを不偏推定量 s_1^2, s_2^2 で代用して、次式で与えられる自由度 $\hat{\nu}$ を用いる：

$$\hat{\nu} = \frac{(s_1^2/m + s_2^2/n)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}. \quad (5.37)$$

5 検定

具体的な検定の手順は以下の通りである。有意水準を $\alpha \in (0, 1)$ とする場合、自由度 $\hat{\nu}$ の t 分布の $1-\alpha/2$ 分位点を $t_{1-\alpha/2}(\hat{\nu})$ として、棄却域を

$$(-\infty, -t_{1-\alpha/2}(\hat{\nu})) \cup (t_{1-\alpha/2}(\hat{\nu}), \infty) \quad (5.38)$$

と設定する。従って、データから検定統計量 t の値を計算し、

$$|t| > t_{1-\alpha/2}(\hat{\nu}) \quad (5.39)$$

であった場合には帰無仮説を棄却することになる。この検定は **Welch の t 検定** と呼ばれることがある。

前の小節と同様に、 p 値の計算方法や片側対立仮説の場合への対応方法は 9.2.1 節と類推の議論となるため、ここでは省略する。

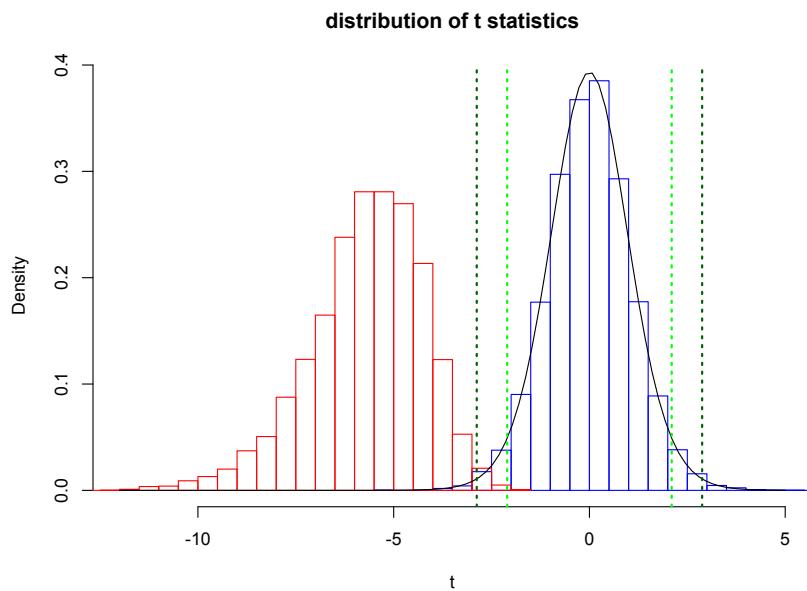


Figure 5.4: 一般の平均値の差の検定の例

[Figure 5.4 を参照]

```
> ### 平均の差の検定（一般の場合； Welch の  $t$  検定）
> set.seed(123) # 亂数のシード値の設定
> mu1 <- 5      # X/Y の平均
> mu2 <- 7.5    # Y の平均
> sigma1 <- 1   # X の分散
> sigma2 <- 2   # Y の分散
> m <- 8        # X のデータ数
> n <- 12       # Y のデータ数
> ## 帰無仮説が正しい場合（平均が等しい）
> x <- rnorm(m, mean=mu1, sd=sigma1)
> y <- rnorm(n, mean=mu1, sd=sigma2)
> t.test(x, y) # Welch の  $t$  検定を実行
```

Welch Two Sample t-test

data: x and y

5 検定

```
t = 0.11283, df = 17.166, p-value = 0.9115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.342201  1.493993
sample estimates:
mean of x mean of y
5.234846 5.158951

> ## 帰無仮説が誤りの場合 (平均が異なる)
> x <- rnorm(m, mean=mu1, sd=sigma1)
> y <- rnorm(n, mean=mu2, sd=sigma2)
> t.test(x, y) # Welch の t 検定を実行

Welch Two Sample t-test

data: x and y
t = -7.2931, df = 17.637, p-value = 1.008e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.640098 -2.562260
sample estimates:
mean of x mean of y
4.454842 8.056021

> ## 実験を繰り返した場合の検定の棄却率の確認
> mytest <- function(x, y){
+   res <- t.test(x, y)      # Welch の t 検定を実行
+   t.val <- res$statistic # 検定統計量の値
+   p.val <- res$p.value   # p 値
+   return(c(t.val, p=p.val))
+ }
> ## Monte-Carlo 実験
> mc <- 10000 # 実験回数
> ## 帰無仮説が正しい場合 (X と Y の平均は等しい)
> res1 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(m, mean=mu1, sd=sigma1),
+                      rnorm(n, mean=mu1, sd=sigma2))))
+ ))
> mean(res1$p < 0.05) # 有意水準 5%で棄却された実験の割合

[1] 0.0522

> mean(res1$p < 0.01) # 有意水準 1%で棄却された実験の割合

[1] 0.0104

> ## 帰無仮説が誤りの場合 (X と Y の平均は異なる)
> res2 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(m, mean=mu1, sd=sigma1),
+                      rnorm(n, mean=mu2, sd=sigma2))))
+ ))
> mean(res2$p < 0.05) # 有意水準 5%で棄却された実験の割合

[1] 0.9317

> mean(res2$p < 0.01) # 有意水準 1%で棄却された実験の割合

[1] 0.7621

> ## 検定統計量のヒストグラム
> hist(res1$t, freq=FALSE, xlim=c(-12, 5), breaks=20, border="blue",
+       main="distribution of t statistics", xlab="t")
> hist(res2$t, freq=FALSE, add=TRUE, breaks=20, border="red")
```

```

> ## 帰無分布の理論曲線(自由度は理論値を使う)
> a <- sigma1^2/m
> b <- sigma2^2/n
> nu <- (a+b)^2/(a^2/(m-1)+b^2/(n-1)) # 自由度の理論値
> curve(dt(x, df=nu), add=TRUE)
> ## 棄却域の可視化(有意水準5%, 自由度は理論値を使う)
> abline(v=c(-qt(0.975, df=nu), qt(0.975, df=nu)),
+           lty=3, lwd=2, col="green")
> ## 棄却域の可視化(有意水準1%, 自由度は理論値を使う)
> abline(v=c(-qt(0.995, df=nu), qt(0.995, df=nu)),
+           lty=3, lwd=2, col="darkgreen")
> ### 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> x <- subset(kikou, 月==1, select=気温, drop=TRUE)
> y <- subset(kikou, 月==2, select=気温, drop=TRUE)
> z <- subset(kikou, 月==8, select=気温, drop=TRUE)
> t.test(x, y, alternative="less") # 2月より1月の方が平均気温は低いか?

Welch Two Sample t-test

data: x and y
t = -1.719, df = 50.436, p-value = 0.04587
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -0.02893813
sample estimates:
mean of x mean of y
6.080645 7.227586

> t.test(x, z, alternative="less") # 8月より1月の方が平均気温は低いか?

Welch Two Sample t-test

data: x and z
t = -43.147, df = 57.404, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -20.22041
sample estimates:
mean of x mean of y
6.080645 27.116129

```

(test-welch.r)

演習 5.3. 2標本の場合の平均の差の検定について調べてみよう。

1. Welch の近似法の近似精度をシミュレーションで確認せよ。
2. 適当な正規分布を設定し、検定のための統計量の標本分布をシミュレーションにより求めなさい。
3. 帰無仮説が正しい状況および正しくない状況を設定し、正しく受容される確率および正しく棄却される確率をシミュレーションにより調べなさい。
4. 実際のデータについて適当な仮説を設定して、Welch の t 検定を実行してみよ。

5.3.3 平均の差の検定(対応がある場合)

2種類のデータを考える場合、2つのデータ間に自然な対応を考えることができることがある。例えば、2種類の薬の効能を比較するために、 n 人の被験者にそれぞれの薬を投与したと

5 検定

する。このとき、各 $i = 1, \dots, n$ について、 i 番目の被験者にそれぞれの薬を投与した場合の治験結果を X_i, Y_i とした場合、 X_i と Y_i には「同一の被験者に対する治験結果」という意味で対応がある。このような場合、仮説検定(5.25)の代わりに、「対応がある観測値の差の平均が0か否か」という仮説検定を考えることができる。すなわち、 $Z_i = X_i - Y_i$ ($i = 1, \dots, n$) として、 Z_1, \dots, Z_n の平均が0か否かを 5.2.1 節の方法で検定すれば良い。

一般に対応のある二標本間の平均の差の検定では、上のように対応に関する情報を利用した検定の方が検出力が優れているため、対応がある場合はその情報を利用することが推奨される。

```
> ### 平均の差の検定（対応がある場合）
> ## データセット sleep による例（睡眠薬による睡眠時間の伸び）
> (x <- subset(sleep, group==1, extra, drop=TRUE)) # group 1 の睡眠時間の伸び
[1] 0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
> (y <- subset(sleep, group==2, extra, drop=TRUE)) # group 2 の睡眠時間の伸び
[1] 1.9  0.8  1.1  0.1 -0.1  4.4  5.5  1.6  4.6  3.4
> t.test(x, y) # 対応を考慮しない t 検定 (Welch の t 検定)
    Welch Two Sample t-test

data: x and y
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.3654832  0.2054832
sample estimates:
mean of x mean of y
0.75      2.33

> t.test(x, y, paired=TRUE) # 対応を考慮する t 検定
    Paired t-test

data: x and y
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58

> t.test(x-y) # 対応を考慮する場合と同じ結果
    One Sample t-test

data: x - y
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of x
-1.58

> plot(x,y,xlim=range(x,y),ylim=range(x,y),
+       main="scatter plot of sleep data")
> abline(a=0,b=1,col="red",lwd=2)
```

(test-paired.r)

5 検定

演習 5.4. 対応がある場合の平均の検定について調べてみよう.

1. 対応を考慮した場合としない場合における t 検定の検出力をシミュレーションで確認してみよ.
2. 実際のデータについて適当な仮説を設定して、対応がある場合の平均の差の t 検定を実行してみよ.

5.3.4 分散の比の検定

最後に、2種類のデータの分散が等しいか否かを検定する問題

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \quad (5.40)$$

を考える。 X_1, \dots, X_m の不偏分散を s_1^2 , Y_1, \dots, Y_n の不偏分散を s_2^2 とする。 s_1^2, s_2^2 は独立であり、また命題 8.2 より $(m-1)s_1^2/\sigma_1^2, (n-1)s_2^2/\sigma_2^2$ はそれぞれ自由度 $m-1, n-1$ の χ^2 分布に従う。従って、検定統計量として

$$F = s_1^2/s_2^2 \quad (5.41)$$

を考えると、 H_0 の下で F は自由度 $m-1, n-1$ の F 分布に従う(2.3.6節参照)。よって、 $\alpha \in (0, 1)$ に対して、自由度 $m-1, n-1$ の F 分布の $\alpha/2, 1-\alpha/2$ 分位点をそれぞれ $F_{\alpha/2}(m-1, n-1), F_{1-\alpha/2}(m-1, n-1)$ とすれば、 H_0 の下では

$$P(F < F_{\alpha/2}(m-1, n-1) \text{ または } F > F_{1-\alpha/2}(m-1, n-1)) = \alpha \quad (5.42)$$

が成り立つ。以上より、有意水準を α とする場合、棄却域を

$$(-\infty, F_{\alpha/2}(m-1, n-1)) \cup (F_{1-\alpha/2}(m-1, n-1), \infty) \quad (5.43)$$

と設定すれば、第一種過誤の上限が α となる。具体的な検定の手順としては、データから検定統計量 F の値を計算し、

$$F < F_{\alpha/2}(m-1, n-1) \text{ または } F > F_{1-\alpha/2}(m-1, n-1) \quad (5.44)$$

であった場合には帰無仮説を棄却する。 p 値の計算方法や片側対立仮説の場合への対応方法は 9.2.2 節と類推の議論となるため、ここでは省略する。

なお、この検定のように、帰無分布が F 分布となるような検定を **F 検定** と呼ぶ。

[Figure 5.5 を参照]

```
> ### 分散の比の検定 (F 検定)
> set.seed(123) # 亂数のシード値の設定
> mu1 <- 5      # X/Y の平均
> mu2 <- 8      # Y の平均
> sigma1 <- 4   # X/Y の分散
> sigma2 <- 2   # Y の分散
> m <- 15       # X のデータ数
> n <- 20       # Y のデータ数
> ## 帰無仮説が正しい場合 (分散が等しい)
> x <- rnorm(m, mean=mu1, sd=sigma1)
> y <- rnorm(n, mean=mu2, sd=sigma1)
> var.test(x, y)
```

5 検定

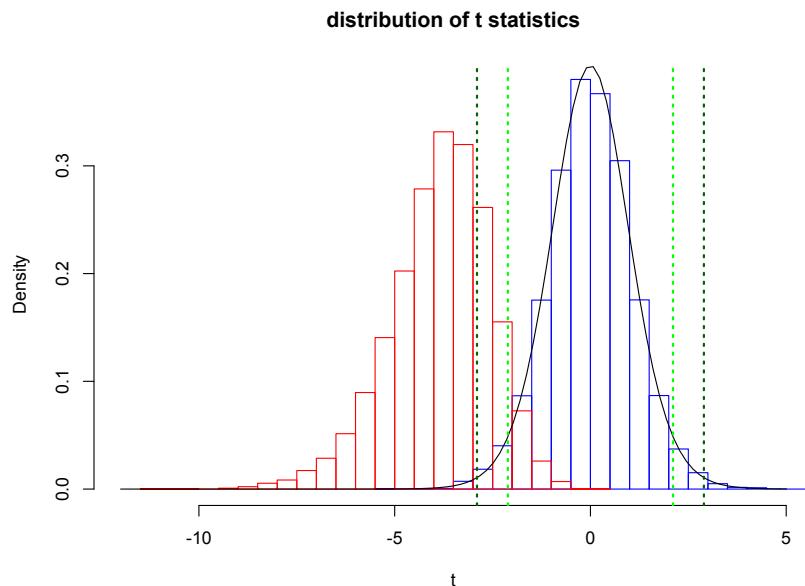


Figure 5.5: 分散の比の検定の例

```
F test to compare two variances

data: x and y
F = 0.67572, num df = 14, denom df = 19, p-value = 0.4594
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.255286 1.933058
sample estimates:
ratio of variances
0.6757238

> ## 帰無仮説が誤りの場合
> x <- rnorm(m, mean=mu1, sd=sigma1)
> y <- rnorm(n, mean=mu2, sd=sigma2)
> var.test(x, y)

F test to compare two variances

data: x and y
F = 4.3154, num df = 14, denom df = 19, p-value = 0.003739
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
1.630325 12.345021
sample estimates:
ratio of variances
4.315352

> ## 実験を繰り返した場合の検定の棄却率の確認
> mytest <- function(x, y){
+   res <- var.test(x, y)
+   f.val <- res$statistic # 検定統計量の値
+   p.val <- res$p.value # p 値
+   return(c(f.val, p=p.val))
+ }
> ## Monte-Carlo 実験
```

5 檢定

```
> mc <- 10000 # 実験回数
> ## 帰無仮説が正しい場合 (X と Y の分散は等しい)
> res1 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(m, mean=mu1, sd=sigma1),
+   rnorm(n, mean=mu2, sd=sigma1))))
+ ))
> mean(res1$p < 0.05) # 有意水準 5%で棄却された実験の割合
[1] 0.0504

> mean(res1$p < 0.01) # 有意水準 1%で棄却された実験の割合
[1] 0.0113

> ## 帰無仮説が誤りの場合 (X と Y の分散は異なる)
> res2 <- as.data.frame(t(
+   replicate(mc, mytest(rnorm(m, mean=mu1, sd=sigma1),
+   rnorm(n, mean=mu2, sd=sigma2))))
+ ))
> mean(res2$p < 0.05) # 有意水準 5%で棄却された実験の割合
[1] 0.7793

> mean(res2$p < 0.01) # 有意水準 1%で棄却された実験の割合
[1] 0.5617

> ## 検定統計量のヒストグラム
> hist(res1$F, freq=FALSE, xlim=c(0, 15), breaks=20, border="blue",
+       main="distribution of F statistics", xlab="F")
> hist(res2$F, freq=FALSE, add=TRUE, border="red", breaks=40)
> ## 帰無分布の理論曲線
> curve(df(x, df1=m-1, df2=n-1), add=TRUE)
> ## 棄却域の可視化 (有意水準 5%)
> abline(v=c(qf(0.025, df1=m-1, df2=n-1), qf(0.975, df1=m-1, df2=n-1)),
+          lty=3, lwd=2, col="green")
> ## 棄却域の可視化 (有意水準 1%)
> abline(v=c(qf(0.005, df1=m-1, df2=n-1), qf(0.995, df1=m-1, df2=n-1)),
+          lty=3, lwd=2, col="darkgreen")
> ### 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> x <- subset(kikou, 月==7, select=気温, drop=TRUE)
> y <- subset(kikou, 月==8, select=気温, drop=TRUE)
> z <- subset(kikou, 月==3, select=気温, drop=TRUE)
> var.test(x, y) # 7月と8月の気温のばらつき具合は異なるか?

  F test to compare two variances

data: x and y
F = 1.706, num df = 30, denom df = 30, p-value = 0.1492
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.822587 3.538146
sample estimates:
ratio of variances
 1.705999

> var.test(z, y, alternative="greater") # 8月より3月の気温のばらつきは大きいか?

  F test to compare two variances

data: z and y
```

5 検定

```
F = 4.0031, num df = 30, denom df = 30, p-value = 0.0001408
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 2.174555      Inf
sample estimates:
ratio of variances
        4.003077
```

(test-ratio.r)

演習 5.5. 2 標本の場合の分散の比の検定について調べてみよう。

1. 適当な正規分布を設定し、検定のための統計量の標本分布をシミュレーションにより求めなさい。
2. 帰無仮説が正しい状況および正しくない状況を設定し、正しく受容される確率および正しく棄却される確率をシミュレーションにより調べなさい。
3. 実際のデータについて適当な仮説を設定して、分散の比の検定を実行してみよ。

5.4 補遺

5.4.1 参考文献

- [1] 東京大学教養学部統計学教室. **統計学入門**. 東京: 東京大学出版会, 1991.
- [2] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.

6 分散分析

前章の平均の差の検定の節では、2つのグループ間で平均の差があるか否かを検定する方法を学習した。分散分析とは、大雑把にいうと、2つ以上のグループ間で平均の差があるか否かを検定する方法である。例えば、ある小売店について「売上高は月によって差があるか」という仮説を検定したり、また、ある銘柄の株価について「収益率は曜日によって差があるか」という仮説を検定するのに分散分析は有用である。

分散分析の基本的な考え方は、データの変動からグループ間での変動と観測誤差のみに起因する変動を抽出し、両者を比較することである。もしグループ間で平均に差がなければ、グループ間での変動は観測誤差のみに起因する変動と自由度を除いて本質的な差がないはずである。逆にグループ間で平均に差があれば、前者はその分だけ変動が増して後者より大きくなるはずなので、両者の比較によって目的の検定が実行できる。従って、分散分析は「分散の分析」というよりむしろ「データの変動の分析」であるといえる。

6.1 一元配置

この節ではグループ分けが1種類の場合を考え、 p 個のグループ A_1, A_2, \dots, A_p があるとする。なお、統計学では、グループ分けのことを因子と呼び、因子内の各グループのことを水準と呼ぶことが多いため、以下これらの用語を用いることにする。

各 $i = 1, 2, \dots, p$ について n_i 個の観測データ $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ が与えられている状況を考える。例えば、 A_1, A_2, \dots, A_p が月に対応し、 $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ が i 月の各日における売上高に対応していると考えれば良い。観測データは以下のモデルに従うと仮定する：

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, \dots, p; j = 1, \dots, n_i). \quad (6.1)$$

ここで、 μ_i は定数であり、水準 A_i における観測データの平均値を表す。また、 ε_{ij} は観測にともなう不確定性を表す確率変数であり、 $\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \dots, \varepsilon_{p1}, \dots, \varepsilon_{pn_p}$ は独立同分布で平均 0、分散 σ^2 の正規分布に従うと仮定する。水準 A_1, A_2, \dots, A_p の間の平均値に差があるか否かを検定する問題は、以下のように定式化できる：

$$H_0 : \mu_1 = \dots = \mu_p \quad \text{vs} \quad H_1 : \text{ある } i, j \text{ に対して } \mu_i \neq \mu_j.$$

冒頭で述べたように、分散分析ではデータの変動から因子間での変動と観測誤差のみに起因する変動を抽出し、両者を比較することで検定を構成する。まず、データ全体の標本平均 $\bar{Y}_{..}$ および水準 A_i における標本平均 $\bar{Y}_{i..}$ を以下で定義する：

$$\begin{aligned} \bar{Y}_{..} &= \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} Y_{ij}, \\ \bar{Y}_{i..} &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad (i = 1, \dots, p). \end{aligned}$$

ただし、 $n := \sum_{i=1}^p n_i$ は全サンプル数を表す。次に、データ全体の変動 SS_T 、各水準内での

6 分散分析

データの変動(の合計) SS_W , 水準間でのデータの変動 SS_B を以下で定義する:

$$\begin{aligned} SS_T &= \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2, \\ SS_W &= \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2, \\ SS_B &= \sum_{i=1}^p \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = \sum_{i=1}^p n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2. \end{aligned}$$

SS_W を**級内変動**, SS_B を**級間変動**と呼ぶ。いまの設定では, 級内変動 SS_W は観測誤差にのみ起因して生じる。仮に帰無仮説 H_0 が正しければ, 水準内でのデータの変動・水準間でのデータの変動とともに観測誤差のみが原因で生じるはずなので, 自由度を除けば本質的な違いはないはずである。逆に, 対立仮説 H_1 が正しければ, 水準間でのデータの変動は観測誤差のみならず, 水準間での平均値 μ_1, \dots, μ_p の異質性にも影響されるはずなので, SS_B は SS_W より本質的に大きくなるはずである。3つの変動の間には以下の分解

$$SS_T = SS_B + SS_W \quad (6.2)$$

が成り立ち, SS_T, SS_B, SS_W のうち2つがわかれば, 残り1つも計算できる。また, 帰無仮説の下で $SS_W/(n-p)$, $SS_B/(p-1)$ はともに σ^2 の不偏推定量となることが示せる。従つて, 検定統計量として

$$F = \frac{SS_B/(p-1)}{SS_W/(n-p)}$$

を考えるのが自然である。対立仮説の下では F は大きな値をとるはずなので, この検定は右片側検定となる。帰無仮説の下で次の事実が成り立つことが知られている: SS_B, SS_W は独立であり, SS_B は自由度 $p-1$ の χ^2 分布に従い, SS_W は自由度 $n-p$ の χ^2 分布に従う。従つて, 帰無仮説の下で F は自由度 $p-1, n-p$ の F 分布に従う(2.3.6節参照)。よって, $\alpha \in (0, 1)$ に対して, 自由度 $p-1, n-p$ の F 分布の $1-\alpha$ 分位点を $F_{1-\alpha}(p-1, n-p)$ とすれば, H_0 の下では

$$P(F > F_{1-\alpha}(p-1, n-p)) = \alpha$$

が成り立つ。以上より, 有意水準を α とする場合, 棄却域を

$$(F_{1-\alpha}(p-1, n-p), \infty)$$

と設定すれば, 第一種過誤の上限が α となる。具体的な検定の手順としては, データから検定統計量 F の値を計算し,

$$F > F_{1-\alpha}(p-1, n-p)$$

であった場合には帰無仮説を棄却する。もしくは, $f(x)$ を自由度 $p-1, n-p$ の F 分布の確率密度関数として, p 値

$$\int_F^\infty f(x) dx$$

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等である。

分散分析の結果は表6.1のように表形式にまとめることが多い。このような表を**分散分析表**と呼ぶ。なお, 全変動の欄は他の2つから計算できるので省略されることが多い。

モデル(6.1)では各水準の効果をその水準における平均値で表していたが, 因子A全体の平均効果を μ で表して, 平均 μ を基準とした各水準 A_i の相対的な効果 α_i で表すことも可能である。すなわち,

$$\mu_i = \mu + \alpha_i, \quad \mu = \frac{1}{n} \sum_{i=1}^p n_i \mu_i$$

6 分散分析

Table 6.1: 分散分析表 (一元配置の場合)

	自由度	平方和	平均平方和	F 値	p 値
級間	$p-1$	SS_B	$SS_B/(p-1)$	F	$\int_F^\infty f(x)dx$
級内	$n-p$	SS_W	$SS_W/(n-p)$		
全変動	$n-1$	SS_T			

とする。このとき、

$$\sum_{i=1}^p n_i \alpha_i = 0$$

であるから、帰無仮説 H_0 は

$$\alpha_1 = \dots = \alpha_p = 0$$

と同等となる。

R には分散分析を実行するための関数 `aov()` が用意されている。

[Figure 6.1 を参照]

```
> ### 一元配置分散分析
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> ## 月ごとの気温に差があるか否かを分散分析
> kikou$月 <- as.factor(kikou$月) # 因子扱いするために月を factor に変換
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> boxplot(気温 ~ 月, data=kikou, # 箱ひげ図で可視化
+           col="lavender", main="月ごとの気温")
> (result <- aov(気温 ~ 月, data=kikou)) # 月ごとの気温差に関する分散分析

Call:
aov(formula = 気温 ~ 月, data = kikou)

Terms:
月 Residuals
Sum of Squares 19134.42 2385.59
Deg. of Freedom 11 354

Residual standard error: 2.59595
Estimated effects may be unbalanced

> summary(result) # 分散分析表の表示(棄却される)

Df Sum Sq Mean Sq F value Pr(>F)
月 11 19134 1739.5 258.1 <2e-16 ***
Residuals 354 2386 6.7
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

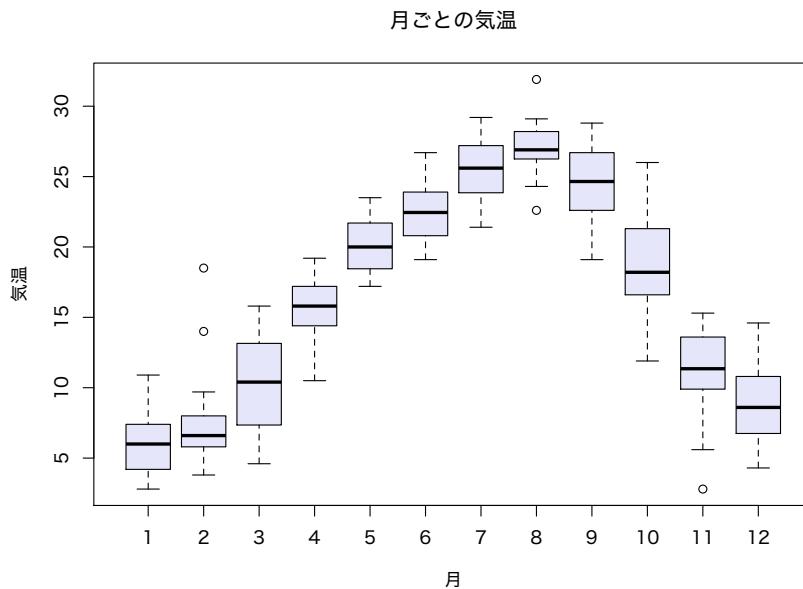
> model.tables(result, type="means") # 水準(月)ごとの平均値

Tables of means
Grand mean

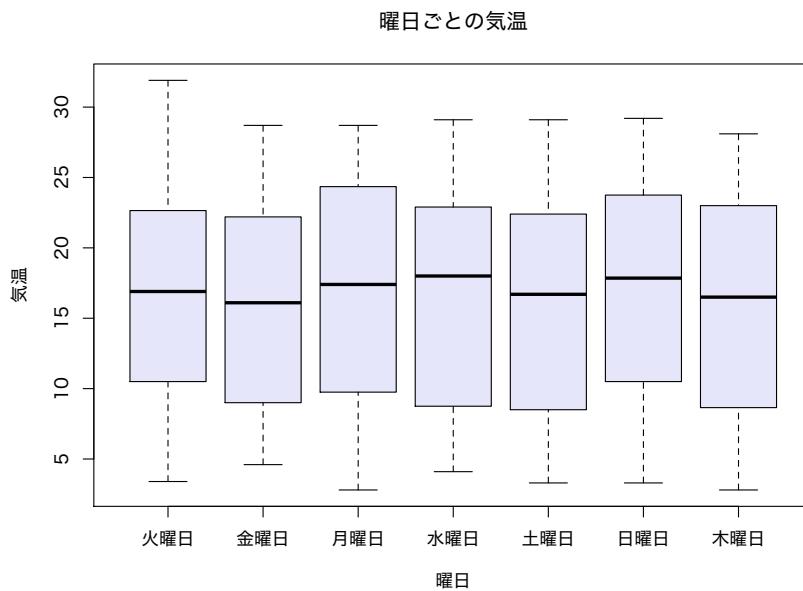
16.47022

月
1 2 3 4 5 6 7 8 9 10 11 12
6.081 7.228 10.14 15.45 20.16 22.35 25.37 27.12 24.4 18.72 11.41 8.865
rep 31.000 29.000 31.00 30.00 31.00 30.00 31.00 31.00 30.0 31.00 30.00 31.000
```

6 分散分析



(a) 月ごとの分析



(b) 曜日ごとの分析

Figure 6.1: 一元配置分散分析の例

```
> model.tables(result, type="effects") # 水準(月)ごとの効果
```

```
Tables of effects
```

月

	1	2	3	4	5	6	7	8	9	10	11
rep	-10.39	-9.243	-6.328	-1.024	3.691	5.883	8.904	10.65	7.93	2.252	-5.064
rep	31.00	29.000	31.000	30.000	31.000	30.000	31.000	31.00	30.00	31.000	30.000
	12										

6 分散分析

```
-7.606
rep 31.000

> ## 検定のみ実行する場合
> oneway.test(気温 ~ 月, data=kikou, var.equal=TRUE) # 等分散での検定

One-way analysis of means

data: 気温 and 月
F = 258.12, num df = 11, denom df = 354, p-value < 2.2e-16

> oneway.test(気温 ~ 月, data=kikou) # Welch の近似法による検定

One-way analysis of means (not assuming equal variances)

data: 気温 and 月
F = 320.99, num df = 11.00, denom df = 139.09, p-value < 2.2e-16

> ## 曜日ごとの気温に差があるか否かを分散分析
> days <- as.Date(paste(2016, kikou$月, kikou$日, sep="-")) # 日付オブジェクト
> youbi <- weekdays(days) # 各日付の曜日を計算
> kikou2 <- cbind(kikou, 曜日=as.factor(youbi)) # 曜日因子を追加したデータセット
> boxplot(気温 ~ 曜日, data=kikou2, # 箱ひげ図で可視化
+           col="lavender", main="曜日ごとの気温")
> (result <- aov(気温 ~ 曜日, data=kikou2)) # 曜日ごとの気温差に関する分散分析

Call:
aov(formula = 気温 ~ 曜日, data = kikou2)

Terms:
          曜日 Residuals
Sum of Squares   34.959 21485.047
Deg. of Freedom      6        359

Residual standard error: 7.73608
Estimated effects may be unbalanced

> summary(result) # 分散分析表の表示 (棄却されない)

Df Sum Sq Mean Sq F value Pr(>F)
曜日       6     35    5.83   0.097  0.997
Residuals 359  21485    59.85

> model.tables(result, type="means") # 水準 (曜日) ごとの平均値

Tables of means
Grand mean

16.47022

曜日
火曜日 金曜日 月曜日 水曜日 土曜日 日曜日 木曜日
16.47 16.03 16.52 16.43 16.32 17.14 16.4
rep 52.00 53.00 52.00 52.00 53.00 52.00 52.0

> model.tables(result, type="effects") # 水準 (曜日) ごとの効果

Tables of effects

曜日
火曜日 金曜日 月曜日 水曜日 土曜日 日曜日 木曜日
-0.0009878 -0.4381 0.04709 -0.03945 -0.1532 0.6663 -0.07022
rep 52.000000 53.0000 52.00000 52.00000 53.0000 52.00000 52.00000
```

(anova-oneway.r)

演習 6.1. 一元配置分散分析について調べてみよう.

1. 式 (6.2) の分解を確認せよ.
2. 上の実行例において, 気温の代わりに日射量もしくは風速とした場合に結果がどうなるか観察せよ.
3. 適当な正規分布を設定し, 検定のための統計量の標本分布をシミュレーションにより求めなさい.
4. 帰無仮説が正しい状況および正しくない状況を設定し, 正しく受容される確率および正しく棄却される確率をシミュレーションにより調べなさい.
5. 実際のデータについて適用な仮説を設定して, 一元配置分散分析を実行してみよ.

6.2 二元配置

前節では因子が 1 種類の場合について考えたが, 本節では因子が 2 種類ある場合を考え, 一方の因子の水準間の平均値に差があるか否かを検定する問題を考える. このとき, もう一方の因子の水準間で平均値に差があるかは問わない.

例えば, いくつかの薬の効能を比較するために何人かの被験者にそれぞれの薬を投与して治験結果を集めた場合, 観測データには「薬の種類」と「被験者番号」という 2 種類の因子が設定される. この場合, 検証したいのは「薬の種類」という因子の水準間で効能に差があるかということだが, 薬の効き目には個人差があると考えられるため, 同一の薬に対して被験者間で効能に差があることは許容したい. この場合, 前節の方法では因子内での効能は一定でなければならないと仮定するため, そのままでは適用できない. しかし, 同一被験者に対しては, 薬の効能の上乗せ/下乗せ分は薬によらず一定であると考え, 「被験者番号」という因子を考慮(コントロール)することで, 薬の効能の違いについて検証できると考えられる.

2 種類の因子 A, B があるとし, 因子 A には a 個の水準 A_1, \dots, A_a があり, 因子 B には b 個の水準 B_1, \dots, B_b があるとする. 因子 A, B の水準がそれぞれ A_i, B_j であるようなデータの観測値が Y_{ij} で与えられているとし, 以下のモデルに従うとする:

$$Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, \dots, a; j = 1, \dots, b). \quad (6.3)$$

ここで, α_i, β_j はともに定数であり, それぞれ因子 A, B の水準 A_i, B_j における効果を表す. ε_{ij} は確率変数であり, $\varepsilon_{11}, \dots, \varepsilon_{1b}, \dots, \varepsilon_{a1}, \dots, \varepsilon_{ab}$ は独立同分布で平均 0, 分散 σ^2 の正規分布に従うと仮定する. 上の例でいうと, 因子 A が「薬の種類」, 因子 B が「被験者番号」に対応し, α_i は薬 A_i の効能を, β_j は被験者 B_j 固有の薬の効きやすさに対応すると考えられる. そして, 薬の効能に差があるか否かという検定は, 因子 A の水準間の効果に差があるか否かを検定する問題となる. 因子 A の水準間の効果に差があるか否かの検定は以下のように定式化できる:

$$H_0: \alpha_1 = \dots = \alpha_a \quad \text{vs} \quad H_1: \text{ある } i_1, i_2 \text{ に対して } \alpha_{i_1} \neq \alpha_{i_2}.$$

前節と同様に, データの変動から因子間での変動と観測誤差のみに起因する変動を抽出し, 両者を比較することで検定を構成する. まず, データ全体の標本平均 $\bar{Y}_{..}$, 因子 A の水準 A_i における標本平均 $\bar{Y}_{i..}$, および因子 B の水準 B_j における標本平均 $\bar{Y}_{..j}$ をそれぞれ以下

6 分散分析

で定義する:

$$\begin{aligned}\bar{Y}_{..} &= \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b Y_{ij}, \\ \bar{Y}_{i\cdot} &= \frac{1}{b} \sum_{j=1}^b Y_{ij} \quad (i = 1, \dots, a), \\ \bar{Y}_{\cdot j} &= \frac{1}{a} \sum_{i=1}^a Y_{ij} \quad (j = 1, \dots, b).\end{aligned}$$

次に、データ全体の変動 SS_T 、因子 A 内でのデータの変動 SS_A 、および因子 B 内でのデータの変動 SS_B を以下で定義する:

$$\begin{aligned}SS_T &= \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2, \\ SS_A &= b \sum_{i=1}^a (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2, \\ SS_B &= a \sum_{j=1}^b (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2.\end{aligned}$$

SS_A を行間変動、 SS_B を列間変動と呼ぶ。仮に帰無仮説 H_0 が正しければ、因子 A 内でのデータの変動 SS_A は観測誤差のみが原因で生じるはずなので、 SS_A を観測誤差による変動と比較するのが自然である。観測誤差による変動は次の統計量で計算できる:

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2.$$

実際、 $\bar{Y}_{i\cdot}, \bar{Y}_{\cdot j}, \bar{Y}_{..}$ はそれぞれ $\alpha_i + \frac{1}{b} \sum_{j=1}^b \beta_j, \frac{1}{a} \sum_{i=1}^a \alpha_i + \beta_j, \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b (\alpha_i + \beta_j) = \frac{1}{a} \sum_{i=1}^a \alpha_i + \frac{1}{b} \sum_{j=1}^b \beta_j$ の推定量とみなせるため、 $Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..}$ は観測誤差 ε_{ij} に対応するものと考えられる。 SS_E は誤差変動と呼ばれる。このとき、全変動、行間変動、列間変動、誤差変動の間には以下の分解が成り立つ:

$$SS_T = SS_E + SS_A + SS_B. \quad (6.4)$$

特に、 SS_T, SS_E, SS_A, SS_B のうち 3 つがわかれば、残り 1 つも計算できる。

帰無仮説 H_0 が正しければ、変動 SS_A, SS_E はともに観測誤差のみが原因で生じるはずなので、自由度を除けば本質的な違いはないはずである。逆に、対立仮説 H_1 が正しければ、因子 A 内でのデータの変動は観測誤差のみならず、因子 A 内の水準間での効果 $\alpha_1, \dots, \alpha_a$ の異質性にも影響されるはずなので、 SS_A は SS_E より本質的に大きくなるはずである。数学的には、帰無仮説の下で、 $SS_A/(a-1), SS_E/\{(a-1)(b-1)\}$ はともに σ^2 の不偏推定量となることが示せる。従って、検定統計量として

$$F_A = \frac{SS_A/(a-1)}{SS_E/\{(a-1)(b-1)\}}$$

を考えるのが自然である。対立仮説の下では F_A は大きな値をとるはずなので、この検定は右片側検定となる。帰無仮説の下で次の事実が成り立つことが知られている： SS_A, SS_E は独立であり、 SS_A は自由度 $a-1$ の χ^2 分布に従い、 SS_E は自由度 $(a-1)(b-1)$ の χ^2 分布

6 分散分析

に従う。従って、帰無仮説の下で F_A は自由度 $a-1, (a-1)(b-1)$ の F 分布に従う (2.3.6 節参照)。よって、 $\alpha \in (0, 1)$ に対して、自由度 $a-1, (a-1)(b-1)$ の F 分布の $1-\alpha$ 分位点を $F_{1-\alpha}(a-1, (a-1)(b-1))$ とすれば、 H_0 の下では

$$P(F_A > F_{1-\alpha}(a-1, (a-1)(b-1))) = \alpha$$

が成り立つ。以上より、有意水準を α とする場合、棄却域を

$$(F_{1-\alpha}(a-1, (a-1)(b-1)), \infty)$$

と設定すれば、第一種過誤の上限が α となる。具体的な検定の手順としては、データから検定統計量 F_A の値を計算し、

$$F_A > F_{1-\alpha}(a-1, (a-1)(b-1))$$

であった場合には帰無仮説を棄却する。もしくは、 $f_{a-1, (a-1)(b-1)}(x)$ を自由度 $a-1, (a-1)(b-1)$ の F 分布の確率密度関数として、 p 値

$$\int_{F_A}^{\infty} f_{a-1, (a-1)(b-1)}(x) dx$$

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等である。

因子 A ではなく因子 B の水準間の平均の差に関心がある場合、すなわち検定

$$H_0 : \beta_1 = \cdots = \beta_b \quad \text{vs} \quad H_1 : \text{ある } j_1, j_2 \text{ に対して } \beta_{j_1} \neq \beta_{j_2}$$

に興味がある場合は、行間変動 SS_A の代わりに列間変動 SS_B を考えればよい。この場合、検定統計量は

$$F_B = \frac{SS_B/(b-1)}{SS_E/\{(a-1)(b-1)\}}$$

となり、帰無分布は自由度 $b-1, (a-1)(b-1)$ の F 分布となる。

一元配置の場合と同様に、分散分析の結果は表 6.2 のような分散分析表にまとめることが多い。

Table 6.2: 分散分析表 (二元配置の場合)

	自由度	平方和	平均平方和	F 値	p 値
因子 A	$a-1$	SS_A	$SS_A/(a-1)$	F_A	$\int_{F_A}^{\infty} f_{a-1, (a-1)(b-1)}(x) dx$
因子 B	$b-1$	SS_B	$SS_B/(b-1)$	F_B	$\int_{F_B}^{\infty} f_{b-1, (a-1)(b-1)}(x) dx$
誤差	$(a-1)(b-1)$	SS_E	$SS_E/\{(a-1)(b-1)\}$		

前節と同様に、モデル (6.3) において各水準の効果を全体の平均 μ^* に対する相対効果で表すことも可能である。実際、 $\bar{\alpha} = \frac{1}{a} \sum_{i=1}^a \alpha_i$, $\bar{\beta} = \frac{1}{b} \sum_{j=1}^b \beta_j$ とし、

$$\mu^* = \bar{\alpha} + \bar{\beta}, \quad \alpha_i^* = \alpha_i - \bar{\alpha}, \quad \beta_j^* = \beta_j - \bar{\beta}$$

とおけば、 α_i^*, β_j^* はそれぞれ水準 A_i, B_j の相対効果に対応し、モデル (6.3) は

$$Y_{ij} = \mu^* + \alpha_i^* + \beta_j^* + \varepsilon_{ij} \quad (i = 1, \dots, a; j = 1, \dots, b)$$

と書き直せる。このとき

$$\sum_{i=1}^a \alpha_i^* = \sum_{j=1}^b \beta_j^* = 0$$

6 分散分析

であるから、帰無仮説 H_0 は、因子 A について考える場合は

$$\alpha_1^* = \cdots = \alpha_a^* = 0$$

と同等となり、因子 B について考える場合は

$$\beta_1^* = \cdots = \beta_b^* = 0$$

と同等となる。

```
> ### 二元配置分散分析
> ## sleep データによる例
> ## 2種類の睡眠薬の効能は異なるか? (有意水準 5%では棄却できない)
> oneway.test(extra ~ group, data = sleep, var.equal = TRUE)

One-way analysis of means

data: extra and group
F = 3.4626, num df = 1, denom df = 18, p-value = 0.07919

> ## 薬の効き目の個人差をコントロールするために、薬の種類と
> ## 被験者 ID という 2つの因子を考慮した分散分析を実行
> (result <- aov(extra ~ group + ID, data = sleep))

Call:
aov(formula = extra ~ group + ID, data = sleep)

Terms:
group           ID   Residuals
Sum of Squares 12.482 58.078    6.808
Deg. of Freedom     1         9         9

Residual standard error: 0.8697381
Estimated effects may be unbalanced

> summary(result) # 分散分析表の表示 (棄却される)

Df Sum Sq Mean Sq F value Pr(>F)
group      1 12.482 12.482 16.501 0.00283 **
ID         9 58.08   6.453   8.531 0.00190 **
Residuals  9   6.81   0.756
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ## この例では因子 group は 2つの水準しか持たないため、検定統計量
> ## F は「対応のある t 検定」の検定統計量の二乗と同一のものとなり、
> ## 検定結果はその場合と同じになる
> t.test(extra ~ group, data = sleep, paired = TRUE)

Paired t-test

data: extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

```
> model.tables(result, type = "means") # 水準ごとの効果
Tables of means
Grand mean
1.54

group
group
  1     2
0.75 2.33

ID
ID
  1     2     3     4     5     6     7     8     9     10
1.30 -0.40  0.45 -0.55 -0.10  3.90  4.60  1.20  2.30  2.70

> model.tables(result, type = "effects") # 水準ごとの相対効果
Tables of effects

group
group
  1     2
-0.79  0.79

ID
ID
  1     2     3     4     5     6     7     8     9     10
-0.24 -1.94 -1.09 -2.09 -1.64  2.36  3.06 -0.34  0.76  1.16
```

(anova-twoway.r)

演習 6.2. 二元配置分散分析について調べてみよう.

1. 式 (6.4) の分解を確認せよ.
2. 適当な正規分布を設定し, 検定のための統計量の標本分布をシミュレーションにより求めなさい.
3. 帰無仮説が正しい状況および正しくない状況を設定し, 正しく受容される確率および正しく棄却される確率をシミュレーションにより調べなさい.
4. 実際のデータについて適用な仮説を設定して, 二元配置分散分析を実行してみよ.

6.3 補遺

6.3.1 参考文献

- [1] 竹内啓. **数理統計学**. 東京: 東洋経済, 1963.
- [2] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.

7 回帰分析

回帰分析とは、ある変量やデータを別の変量・データを用いて説明・予測するためのモデル（回帰モデル）を構築することを目的とする分析法である。回帰分析においては、説明される側の変量・データは**目的変数**・**被説明変数**・**従属変数**・**応答変数**などと呼ばれ、説明する側の変量・データは**説明変数**・**独立変数**・**共変量**などと呼ばれる。目的変数・説明変数ともに複数個あってもよいが、目的変数については変数ごとにそれぞれ回帰モデルを構築すればよいので、通常は1つの場合を考える。説明変数については、1つの場合を**单回帰**、2つ以上の場合を**重回帰**として区別することが多い。この講義では单回帰のみ扱う。

7.1 回帰モデル

以下では、説明変数を X 、目的変数を Y で表すことにする。 Y を X で説明するための関係式は、一般にはある関数 $f(x)$ を使って、

$$Y = f(X) \quad (7.1)$$

と書ける。しかし、このモデルでは一般的すぎて分析に不向きのため、通常は f の関数形に何らかの制約を課す。最も広く利用されているのは、 $f(x)$ として一次関数のみ考えるというものである。すなわち、ある定数 α, β が存在して、

$$f(x) = \alpha + \beta x$$

と書ける場合のみを分析対象とする。この場合(7.1)式は

$$Y = \alpha + \beta X \quad (7.2)$$

となる。モデル(7.2)を分析対象とする回帰分析を**線形回帰**と呼び、 f としてより一般的な関数形を許す回帰分析を**非線形回帰**と呼ぶ。この講義では線形回帰分析を取り扱う。モデル(7.2)において、 α は**定数項**、 β は X の**回帰係数**と呼ばれる。

なお、非線形な関係であっても、データに適切な変数変換（二乗する、対数をとるなど）を施すことで線形な関係に変換可能な場合や、線形な関係で近似できる場合がよくあることに注意しておく。

7.2 回帰係数の点推定

モデル(7.2)は未知のパラメータ α, β を含むから、これらを観測データから推定する必要がある。この節ではこの問題について議論する。

7.2.1 観測データ

n 個の個体について説明変数と目的変数の組 (X, Y) を観測して得られたデータ $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ が与えられているとする。実際のデータには観測誤差のようなランダムな変動が含まれていると考えられるから、モデル(7.2)が観測データに対してもそのまま成立するとは考えづらい

7 回帰分析

い. そのため, データのランダムな変動を表す項を $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ として, 以下の形の確率モデルを分析することを考える:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, \dots, n. \quad (7.3)$$

$\epsilon_1, \dots, \epsilon_n$ は誤差項もしくは攪乱項と呼ばれる. 以下の分析では次の仮定をおく:

1. データ X_1, \dots, X_n は確率変数ではなく確定値であり, また一定値ではない. すなわち, $X_1 = \dots = X_n$ ではない.
2. 誤差項 $\epsilon_1, \dots, \epsilon_n$ は独立同分布な確率変数列であり, 平均 0, 分散 σ^2 である.

7.2.2 最小二乗法

回帰モデルの推定には通常最小二乗法が用いられる. 最小二乗法の考え方は以下の通りである. パラメータの組 (α, β) を 1 つ決めたとき, 回帰モデルでは説明できない目的変数の変動は,

$$e_i(\alpha, \beta) = y_i - (\alpha + \beta X_i), \quad i = 1, \dots, n$$

で与えられる. これらの変動 $e_1(\alpha, \beta), \dots, e_n(\alpha, \beta)$ はいずれも絶対値が小さいほど当てはまりがよいと考えられる. そこで, 最小二乗法では, $e_1(\alpha, \beta), \dots, e_n(\alpha, \beta)$ の平方和

$$S(\alpha, \beta) := \sum_{i=1}^n e_i(\alpha, \beta)^2 = \sum_{i=1}^n \{Y_i - (\alpha + \beta X_i)\}^2$$

を最小にするようにパラメータ (α, β) を決定する. $S(\alpha, \beta)$ は残差平方和と呼ばれ, $S(\alpha, \beta)$ を最小にするパラメータの組 (α, β) は最小二乗推定量と呼ばれる. 最小二乗推定量はしばしば記号 $(\hat{\alpha}, \hat{\beta})$ で表される.

最小二乗推定量は具体的に求めることができる. 実際, 最小二乗推定量はもし存在すれば次の連立方程式の解とならなければならない:

$$\begin{cases} \frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^n \{Y_i - (\alpha + \beta X_i)\} = 0, \\ \frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^n \{Y_i - (\alpha + \beta X_i)\} X_i = 0. \end{cases} \quad (7.4)$$

この式を整理して, α, β に関する連立一次方程式

$$\begin{cases} n\alpha + \beta \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i, \\ \alpha \sum_{i=1}^n X_i + \beta \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \end{cases}$$

を得る. これは正規方程式と呼ばれる. この連立一次方程式を解くと以下の解を得る:

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}, \quad (7.5)$$

ただし

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

R では線形回帰分析を実行するための関数 `lm()` が用意されている. モデル (7.3) において, 説明変数 X および目的変数 Y の観測データに対応するベクトルがそれぞれ `x` および `y` で与えられているとする. このとき, モデル (7.3) の回帰係数の推定は,

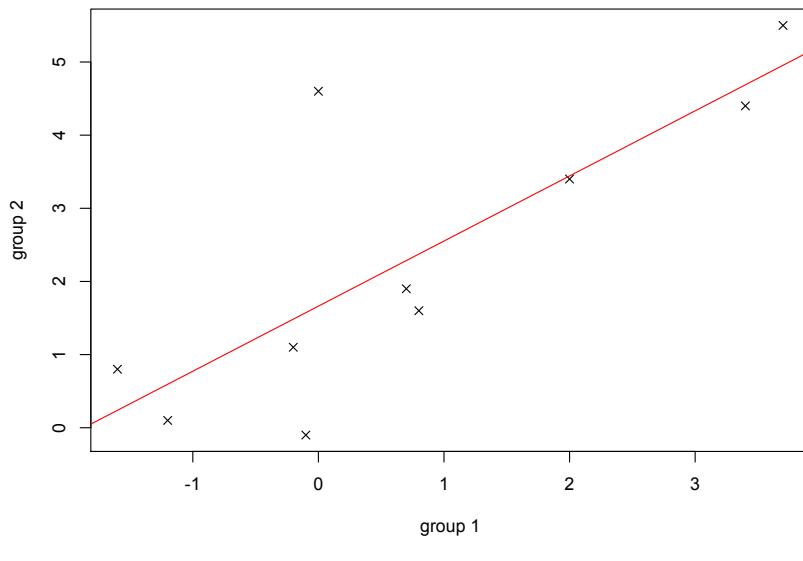
7 回帰分析

`lm(y ~ x)`

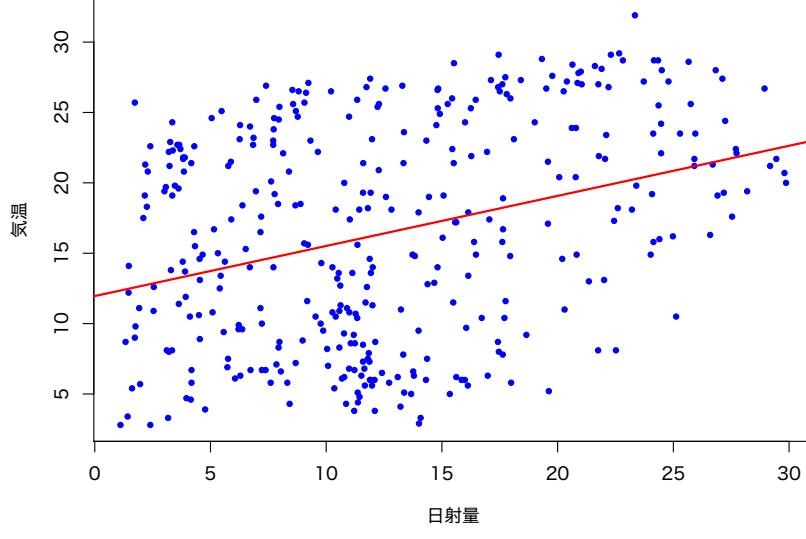
で実行できる。また、実際のデータを使って解析する際は、データセットの一部の変数を目的変数および説明変数として回帰分析をすることが多い。そのような場合、データセットに対応するデータフレームを `dat` とすれば、以下のコマンドで回帰係数の推定を実行できる：

`lm(Y の変数名 ~ X の変数名, data = dat)`

ここで、`dat` は列が各変数に対応するような形式になっている必要がある。



(a)



(b)

Figure 7.1: 回帰分析の例

[Figure 7.1 を参照]

```
> ### 線形回帰分析（単回帰）
> ## データセット sleep による例
> x <- subset(sleep, group == 1, extra, drop=TRUE)
> y <- subset(sleep, group == 2, extra, drop=TRUE)
> ## 線形回帰分析の実行（モデルの作成）
> (myModel <- lm(y ~ x))

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
1.6625      0.8899

> coef(myModel) # 推定されたパラメータ値

(Intercept)          x
1.6625378  0.8899497

> ## 最小二乗推定量の式にもとづく計算（結果の比較）
> (beta.hat <- cov(x, y)/var(x))

[1] 0.8899497

> (alpha.hat <- mean(y) - beta.hat * mean(x))

[1] 1.662538

> ## データの散布図と回帰直線の図示
> plot(x, y, xlab="group 1", ylab="group 2", pch=4)
> abline(reg=myModel, col="red")
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> (myModel <- lm(気温 ~ 日射量, data=kikou)) # 気温を日射量で説明

Call:
lm(formula = 気温 ~ 日射量, data = kikou)

Coefficients:
(Intercept) 日射量
11.9571     0.3559

> ## データの散布図と回帰直線の図示
> par(family="HiraginoSans-W4") # 日本語フォントの指定
> plot(気温 ~ 日射量, data=kikou, pch=20, col="blue")
> abline(reg=myModel, col="red", lwd=2)
> confint(myModel)

              2.5 %    97.5 %
(Intercept) 10.4372617 13.4769302
日射量        0.2514435  0.4602951

(reg-simple.r)
```

7.2.3 Gauss-Markov の定理

最小二乗推定量は以下の性質をもつことが確認できる：

7 回帰分析

1. $\hat{\alpha}, \hat{\beta}$ は不偏推定量である:

$$E(\hat{\alpha}) = \alpha, \quad E(\hat{\beta}) = \beta.$$

2. $\hat{\alpha}, \hat{\beta}$ は Y_1, \dots, Y_n の線形和で表される。すなわち、 (X_1, \dots, X_n) に依存するかもしれない定数 $a_1, \dots, a_n, b_1, \dots, b_n$ が存在して、

$$\hat{\alpha} = \sum_{i=1}^n a_i Y_i, \quad \hat{\beta} = \sum_{i=1}^n b_i Y_i$$

が成り立つ。

3. $\hat{\alpha}, \hat{\beta}$ の分散は次式で与えられる:

$$\text{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}, \quad \text{Var}(\hat{\beta}) = \frac{\sigma^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}.$$

実は、最小二乗推定量は上の性質 1 および 2 を満たすもののうちで、分散が最小のものであるということが知られている。すなわち、次の定理が成り立つ:

定理 7.1 (Gauss-Markov の定理). α, β の推定量 $\tilde{\alpha}, \tilde{\beta}$ が以下の 2 条件を満たすとする:

1. $\tilde{\alpha}, \tilde{\beta}$ は不偏推定量である:

$$E(\tilde{\alpha}) = \alpha, \quad E(\tilde{\beta}) = \beta.$$

2. $\tilde{\alpha}, \tilde{\beta}$ は Y_1, \dots, Y_n の線形和で表される。すなわち、 (X_1, \dots, X_n) に依存するかもしれない定数 $a_1, \dots, a_n, b_1, \dots, b_n$ が存在して、

$$\tilde{\alpha} = \sum_{i=1}^n a_i Y_i, \quad \tilde{\beta} = \sum_{i=1}^n b_i Y_i$$

が成り立つ。

このとき、次の不等式が成り立つ:

$$\text{Var}(\tilde{\alpha}) \geq \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}, \quad \text{Var}(\tilde{\beta}) \geq \frac{\sigma^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}.$$

演習 7.1. 最小二乗推定量について調べてみよう。

1. 正規方程式の解が (7.5) 式で与えられることを実際に確認してみよ。
2. (7.5) 式で与えられる $(\hat{\alpha}, \hat{\beta})$ が実際に $S(\alpha, \beta)$ を最小化していることを確認してみよ。

7.3 回帰係数の区間推定

この節ではパラメータ α, β の区間推定について議論する。そのために、誤差項に関して以下の仮定を追加する:

1. ϵ_i は正規分布に従う。

7 回帰分析

上の仮定と命題 4.1 より、 $\hat{\alpha}, \hat{\beta}$ もそれぞれ正規分布に従うことがわかり、平均と分散は

$$E(\hat{\alpha}) = \alpha, \quad E(\hat{\beta}) = \beta,$$

$$\text{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}, \quad \text{Var}(\hat{\beta}) = \frac{\sigma^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

で与えられる。従って、もし σ^2 が既知であれば、4.4.1 節と同様の議論によって α, β の信頼区間をそれぞれ構成できる。一般には σ^2 は既知でないため、データから推定する必要がある。 σ^2 が ϵ_i に共通の分散であったことと、 ϵ_i の平均は 0 であること、および

$$\epsilon_i = Y_i - (\alpha + \beta X_i) \quad (i = 1, \dots, n)$$

と書き直せることに注意すれば、

$$\hat{\epsilon}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i), \quad i = 1, \dots, n$$

と定義して、 $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ の二乗の平均 $\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2$ を σ^2 の推定量として考えるのが自然なようと思える。 $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ は**残差**と呼ばれ、以下を満たす((7.4) 式より従う):

$$\sum_{i=1}^n \hat{\epsilon}_i = 0, \quad \sum_{i=1}^n \hat{\epsilon}_i X_i = 0. \quad (7.6)$$

実際

$$E[\hat{\epsilon}_i^2] = \frac{n-2}{n} \sigma^2 \quad (i = 1, \dots, n)$$

となるため、 $\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2$ を σ^2 の不偏推定量となるように補正した以下の推定量が利用される:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2.$$

従って、 $\hat{\alpha}, \hat{\beta}$ の分散の推定量として

$$s.e.(\hat{\alpha})^2 := \frac{\hat{\sigma}^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}, \quad s.e.(\hat{\beta})^2 := \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

を考えるのが自然である。これらの推定量の平方根をとつて得られる $\hat{\alpha}, \hat{\beta}$ の標準偏差の推定量 $s.e.(\hat{\alpha}), s.e.(\hat{\beta})$ をそれぞれ $\hat{\alpha}, \hat{\beta}$ の**標準誤差**と呼ぶ。

以上の準備の下、 α, β の信頼区間を構成する方法を説明する。まず、 $(n-2)s.e.(\hat{\alpha})^2 / \text{Var}[\hat{\alpha}]$ は $\hat{\alpha}$ と独立で、かつ自由度 $n-2$ の χ^2 分布に従うことが知られている。従って、

$$\frac{\hat{\alpha} - \alpha}{s.e.(\hat{\alpha})} = \frac{\frac{\hat{\alpha} - \alpha}{\sqrt{\text{Var}[\hat{\alpha}]}}}{\sqrt{\frac{(n-2)s.e.(\hat{\alpha})^2}{\text{Var}[\hat{\alpha}]} / (n-2)}}$$

は自由度 $n-2$ の t 分布に従うことがわかる(2.3.5 節参照)。以上より、 $\gamma \in (0, 1)$ に対して、

$$[\hat{\alpha} - t_{1-\gamma/2}(n-2) \cdot s.e.(\hat{\alpha}), \hat{\alpha} + t_{1-\gamma/2}(n-2) \cdot s.e.(\hat{\alpha})]$$

は α の $100(1-\gamma)\%$ 信頼区間を与えることがわかる。ただし $t_{1-\gamma/2}(n-2)$ は自由度 $n-2$ の t 分布の $1-\gamma/2$ 分位点を表す。

7 回帰分析

β の信頼区間の構成も同様の議論ができる。 $(n-2)s.e.(\hat{\beta})^2 / \text{Var}[\hat{\beta}]$ は $\hat{\beta}$ と独立で、かつ自由度 $n-2$ の χ^2 分布に従うことが知られているので、

$$\frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} = \frac{\frac{\hat{\beta} - \beta}{\sqrt{\text{Var}[\hat{\beta}]}}}{\sqrt{\frac{(n-2)s.e.(\hat{\beta})^2}{\text{Var}[\hat{\beta}]} / (n-2)}}$$

は自由度 $n-2$ の t 分布に従うことがわかる(2.3.5節参照)。以上より、 $\gamma \in (0, 1)$ に対して、

$$[\hat{\beta} - t_{1-\gamma/2}(n-2) \cdot s.e.(\hat{\beta}), \hat{\beta} + t_{1-\gamma/2}(n-2) \cdot s.e.(\hat{\beta})]$$

は β の $100(1-\gamma)\%$ 信頼区間を与えることがわかる。

```
> ### 線形回帰の信頼区間
> ## データセット sleep による例
> x <- subset(sleep, group == 1, extra, drop=TRUE)
> y <- subset(sleep, group == 2, extra, drop=TRUE)
> (myModel <- lm(y ~ x)) # 線形回帰分析の実行

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
1.6625        0.8899

> confint(myModel)           # 95%信頼区間(標準値)

              2.5 %    97.5 %
(Intercept) 0.6358501 2.689225
x            0.3366380 1.443261

> confint(myModel, level=0.99) # 99%信頼区間

              0.5 %    99.5 %
(Intercept) 0.16863991 3.156436
x            0.08484491 1.695054

(reg-ci.r)
```

演習 7.2. 最小二乗推定量の性質について調べてみよう。

1. (7.6) 式が成立することを実際に確認してみよ。
2. $(\hat{\alpha} - \alpha)/s.e.(\hat{\alpha})$, $(\hat{\beta} - \beta)/s.e.(\hat{\beta})$ がそれぞれ自由度 $n-2$ の t 分布に従うことをシミュレーションで確認してみよ。

7.4 回帰係数の有意性の検定

回帰分析において、説明変数 X が目的変数 Y を説明・予測するのに本当に役立っているか検証することは重要である。線形回帰モデル(7.3)においてこれを検証するには、検定問題

$$H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0 \tag{7.7}$$

7 回帰分析

を考えればよい。この検定は β の**有意性の検定**と呼ばれ、帰無仮説 H_0 が有意水準 γ で棄却されるとき、 β は有意水準 γ で**有意である**といわれる。この節では、前節に引き続き 1 を仮定した下で、上の検定を実行する方法を説明する。

前節で述べたことから、帰無仮説 H_0 の下で、統計量

$$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

は自由度 $n-2$ の t 分布に従う。一方、対立仮説 H_1 が正しければ、 $\hat{\beta}$ は 0 でない値 β に近い値を取ることが期待されるから、 $|t|$ は 0 から離れた値を取ることが予想される。以上より、有意水準を $\gamma \in (0, 1)$ とする場合、検定 (7.7) は次の手順で実行できる：データから検定統計量 t の値を計算し、

$$|t| > t_{1-\gamma/2}(n-2)$$

であった場合には帰無仮説を棄却する。もしくは、検定の p 値

$$2 \int_{|t|}^{\infty} f(x) dx \quad (7.8)$$

が γ 未満の場合に帰無仮説を棄却するとしても同等である。ここに、 $f(x)$ は自由度 $n-2$ の t 分布の確率密度関数を表す。なお、検定統計量の値 t を $\hat{\beta}$ の t 値と呼び、検定の p 値 (7.8) を $\hat{\beta}$ の p 値と呼ぶ。

定数項 α についても同様の方法で検定を実行することが可能であるが、詳細は省略する。

```
> #### 回帰係数の検定
> ## 気候データによる例
> kikou <- read.csv("kikou2016.csv", fileEncoding="sjis")
> ## 気温を日射量で説明するモデルの作成
> (myModel1 <- lm(気温 ~ 日射量, data=kikou))

Call:
lm(formula = 気温 ~ 日射量, data = kikou)

Coefficients:
(Intercept)      日射量
           11.9571        0.3559

> summary(myModel1) # パラメータの推定値・標準誤差・t 値・p 値などを表示

Call:
lm(formula = 気温 ~ 日射量, data = kikou)

Residuals:
    Min      1Q   Median      3Q      Max 
-14.0428 -6.4502 -0.2706  7.2320  13.1273 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.9571     0.7729  15.471 < 2e-16 ***
日射量       0.3559     0.0531   6.702 7.86e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.254 on 364 degrees of freedom
Multiple R-squared:  0.1098,    Adjusted R-squared:  0.1074 
F-statistic: 44.91 on 1 and 364 DF,  p-value: 7.863e-11
```

```

> ## 日射量の回帰係数の p 値は非常に小さいので、気温の説明に有用であるといえそう
> ##
> ## 気温を降水量で説明するモデルの作成
> (myModel2 <- lm(気温 ~ 降水量, data=kikou))

Call:
lm(formula = 気温 ~ 降水量, data = kikou)

Coefficients:
(Intercept)      降水量
16.23425        0.04855

> summary(myModel2) # パラメータの推定値・標準誤差・t 値・p 値などを表示

Call:
lm(formula = 気温 ~ 降水量, data = kikou)

Residuals:
    Min      1Q   Median      3Q      Max 
-16.6869 -6.9685  0.1832  6.6494 15.6658 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.23425   0.42543  38.159 <2e-16 ***
降水量       0.04855   0.02956   1.642   0.101    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.661 on 364 degrees of freedom
Multiple R-squared:  0.007354, Adjusted R-squared:  0.004626 
F-statistic: 2.697 on 1 and 364 DF,  p-value: 0.1014

> ## 降水量の回帰係数の p 値は小さくないので、有意であるとはいえない

```

(reg-test.r)

演習 7.3. 回帰係数の有意性の検定について、そのサイズおよび検出力をシミュレーションによって計算してみよ。

7.5 決定係数

前節で議論した回帰係数の有意性の検定では、説明変数 X が目的変数 Y の説明・予測に役立つかどうかを検証することはできたが、実際に X が Y の変動をどの程度説明できているかということについては何も述べていない。このことを評価する指標として**決定係数**(あるいは**寄与率**)がある。決定係数は次式で定義される:

$$R^2 := \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

ただし、

$$\hat{Y}_i := \hat{\alpha} + \hat{\beta}X_i \quad (i = 1, \dots, n)$$

であり、 $\hat{Y}_1, \dots, \hat{Y}_n$ は**あてはめ値**または**予測値**と呼ばれる。 $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ ($i = 1, \dots, n$) が成り立つことに注意すると、(7.6) 式より

$$\frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \bar{Y}$$

7 回帰分析

が成り立つ。この式より、 R^2 の分子・分母はそれぞれあてはめ値・目的変数の（標本平均まわりでの）変動に対応しており、従って回帰モデルが目的変数の変動を何割程度説明できているかを測る評価指標であると解釈できる。従って R^2 が大きいほど回帰式の説明力が高いと解釈される。

R^2 は以下のように書き直すことも可能である:

$$R^2 = \left\{ \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}} \right\}^2. \quad (7.9)$$

すなわち、 R^2 は目的変数の観測データとあてはめ値の相関の二乗に等しく、回帰モデルによるあてはめが目的変数にどの程度連動しているかを測る指標であるとも解釈できる。さらに、等式 $\hat{Y}_i - \bar{Y} = \hat{\beta}(X_i - \bar{X})$ を使うことで、上の式は

$$R^2 = \left\{ \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2} \cdot \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right\}^2$$

とも書ける。すなわち、 R^2 は説明変数と目的変数の観測データの間の相関の二乗にも等しくなっている。

(7.6) 式を使うことで、 R^2 のさらなる別表示として次式を得る：

$$R^2 = 1 - \frac{\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2}. \quad (7.10)$$

この式において、右辺第2項の分子、分母はそれぞれ確率変数 ϵ_i 、 Y_i の分散の標本分散による推定値ともみなせる。この観点から考えると、推定量としては不偏なものを用了方がよいと考えられる。そこで、標本分散を対応する不偏推定量で置き換えた以下のような評価指標を考えられる：

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

これを自由度調整済み決定係数（または自由度調整済み寄与率）と呼ぶ。

```

> ### 決定係数の計算
> ## データセット sleepによる例
> x <- subset(sleep, group == 1, extra, drop=TRUE)
> y <- subset(sleep, group == 2, extra, drop=TRUE)
> ## 線形回帰モデルの作成
> myModel <- lm(y ~ x)
> (mySum <- summary(myModel)) # 決定係数と自由度調整済み決定係数は下から二行目

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6735 -0.4673 -0.3365  0.3979  2.9375 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.6623    0.1654   3.988  0.0011 ** 
x            0.3393    0.1654   2.078  0.0493 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


```

```
(Intercept) 1.6625      0.4452     3.734  0.00575 ** 
x           0.8899      0.2399     3.709  0.00596 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.288 on 8 degrees of freedom
Multiple R-squared:  0.6323,    Adjusted R-squared:  0.5863 
F-statistic: 13.76 on 1 and 8 DF,  p-value: 0.005965

> coef(mySum) # パラメータの推定値・標準誤差・t値・p値
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.6625378 0.4452237 3.734163 0.005753296
x           0.8899497 0.2399439 3.708990 0.005964996

> mySum$r.squared      # 決定係数
[1] 0.6322957
> mySum$adj.r.squared # 自由度調整済み決定係数
[1] 0.5863326
> ## 決定係数の様々な計算方法の確認
> ybar <- mean(y) # 目的変数の標本平均
> yhat <- fitted(myModel) # あてはめ値
> ehat <- resid(myModel) # 残差
> n <- length(y) # データ数
> sum((yhat-ybar)^2)/sum((y-ybar)^2) # もともとの定義式
[1] 0.6322957
> cor(yhat, y)^2 # あてはめ値と目的変数の相関の二乗
[1] 0.6322957
> 1-mean(ehat^2)/mean((y-ybar)^2) # 残差による計算
[1] 0.6322957
> 1-(sum(ehat^2)/(n-2))/(sum((y-ybar)^2)/(n-1)) # 自由度調整済み決定係数
[1] 0.5863326
```

(reg-rsquared.r)

演習 7.4. 決定係数について調べてみよう.

1. (7.9) 式が成り立つことを確認してみよ.
2. (7.10) 式が成り立つことを確認してみよ.

7.6 補遺

7.6.1 参考文献

- [1] 東京大学教養学部統計学教室. **統計学入門**. 東京: 東京大学出版会, 1991.
- [2] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.

7.6.2 回帰分析の例

パッケージ MASS に付属するデータセット `Animals` を用いて、回帰分析を行った例を以下に示す。

[Figure 7.2 を参照]

```
> #### 線形回帰分析(単回帰)の例
> #### - Brain and Body Weights for 28 Species
>
> ## データの読み込み ("MASS::Animals"を用いる)
> require(MASS) # パッケージの読み込み
> data(Animals)
> ## データの内容を確認
> help(Animals) # 内容の詳細を表示
> str(Animals) # データの構造を表示

'data.frame': 28 obs. of 2 variables:
 $ body : num 1.35 465 36.33 27.66 1.04 ...
 $ brain: num 8.1 423 119.5 115 5.5 ...

> print(Animals) # データの表示

      body   brain
Mountain beaver 1.350    8.1
Cow             465.000  423.0
Grey wolf       36.330   119.5
Goat            27.660   115.0
Guinea pig      1.040    5.5
Dipliodocus    11700.000 50.0
Asian elephant  2547.000 4603.0
Donkey          187.100   419.0
Horse           521.000   655.0
Potar monkey    10.000   115.0
Cat              3.300    25.6
Giraffe          529.000   680.0
Gorilla          207.000   406.0
Human            62.000   1320.0
African elephant 6654.000 5712.0
Triceratops     9400.000  70.0
Rhesus monkey    6.800    179.0
Kangaroo         35.000   56.0
Golden hamster   0.120    1.0
Mouse            0.023    0.4
Rabbit           2.500    12.1
Sheep            55.500   175.0
Jaguar           100.000   157.0
Chimpanzee      52.160   440.0
Rat              0.280    1.9
Brachiosaurus   87000.000 154.5
Mole             0.122    3.0
Pig              192.000   180.0

> ## データのプロット (normal plot)
> plot(Animals, ann=FALSE) # タイトルやラベルを付けない (ann=FALSE)
> title(main="Brain and Body Weights (normal plot)",
+       xlab="body [kg]", ylab="brain [g]")
> ## データのプロット (log-log plot)
> plot(Animals, log="xy", ann=FALSE)
> title(main="Brain and Body Weights (log-log plot)",
+       xlab="body [kg]", ylab="brain [g]")
```

7 回帰分析

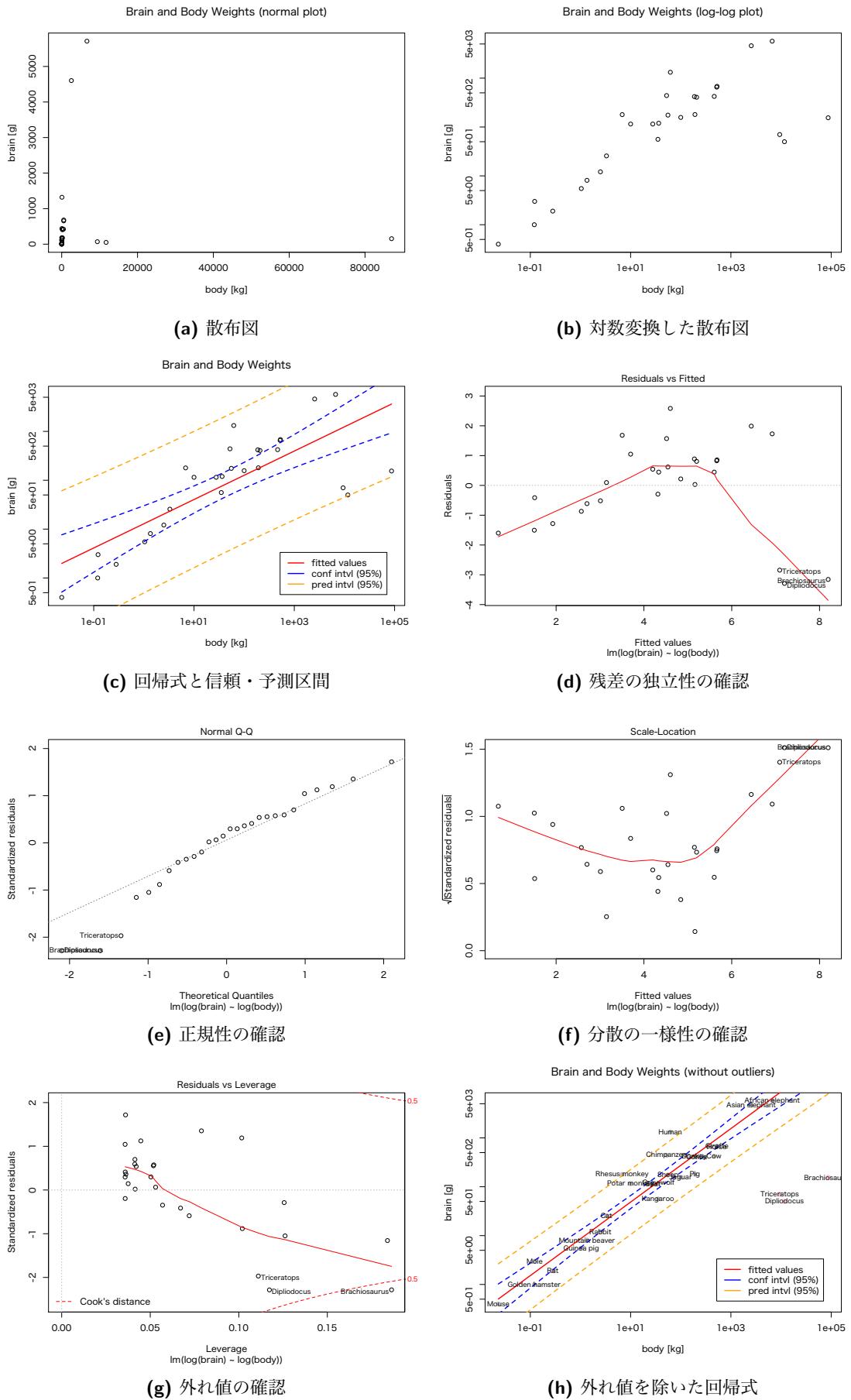


Figure 7.2: 回帰分析の例

```

> ## 回帰分析 (単回帰)
> model <- lm(log(brain) ~ log(body), data=Animals)
> summary(model) # 分析結果のまとめを表示

Call:
lm(formula = log(brain) ~ log(body), data = Animals)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2890 -0.6763  0.3316  0.8646  2.5835 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.55490   0.41314   6.184 1.53e-06 ***
log(body)    0.49599   0.07817   6.345 1.02e-06 ***  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.532 on 26 degrees of freedom
Multiple R-squared:  0.6076,    Adjusted R-squared:  0.5925 
F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06

> ## 回帰式および信頼区間・予測区間の表示
> plot(Animals, log="xy", ann=FALSE)
> ## 区間推定
> r.x <- with(Animals, range(log(body))) # x軸の範囲を取得
> x <- seq(r.x[1], r.x[2], length=50)
> y <- predict(model, newdata=data.frame(body=exp(x)),
+               interval="confidence")
> yp <- predict(model, newdata=data.frame(body=exp(x)),
+               interval="prediction")
> ## グラフ表示 (lty:スタイル, lwd:太さ, col:色)
> matlines(exp(x), exp(y), # 信頼区間
+           lty=c("solid", "dashed", "dashed"), lwd=2,
+           col=c("red", "blue", "blue"))
> matlines(exp(x), exp(yp[, c("lwr", "upr")]), # 予測区間
+           lty=c("dashed", "dashed"), lwd=2,
+           col=c("orange", "orange"))
> title(main="Brain and Body Weights", # タイトル
+       xlab="body [kg]", ylab="brain [g]")
> legend("bottomright", inset=.05, # 凡例の作成
+        legend=c("fitted values", "conf intvl (95%)", "pred intvl (95%)"),
+        col=c("red", "blue", "orange"), lty=1, lwd=2)
> ## 診断プロット
> plot(model)
> ## 外れ値を除いた回帰分析
> idx <- c(6, 16, 26) # 外れ値の index
> model <- lm(log(brain) ~ log(body), data=Animals, subset=-idx)
> summary(model)

Call:
lm(formula = log(brain) ~ log(body), data = Animals, subset = -idx)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9125 -0.4752 -0.1557  0.1940  1.9303 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.15041   0.20060   10.72 2.03e-10 ***
log(body)    0.75226   0.04572   16.45 3.24e-14 ***

```

7 回帰分析

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7258 on 23 degrees of freedom
Multiple R-squared:  0.9217,    Adjusted R-squared:  0.9183
F-statistic: 270.7 on 1 and 23 DF,  p-value: 3.243e-14

> ## 回帰式および信頼区間・予測区間の表示
> plot(Animals,log="xy",ann=FALSE,col="gray",pch=19)
> points(Animals[idx,],col="pink",pch=19)
> ## 区間推定
> r.x <- with(Animals,range(log(body)))
> x <- seq(r.x[1],r.x[2],length=50)
> y <- predict(model,newdata=data.frame(body=exp(x)),
+               interval="confidence")
> yp <- predict(model,newdata=data.frame(body=exp(x)),
+               interval="prediction")
> ## グラフ表示
> matlines(exp(x),exp(y),
+            lty=c("solid","dashed","dashed"),lwd=2,
+            col=c("red","blue","blue"))
> matlines(exp(x),exp(yp[,c("lwr","upr")]),
+            lty=c("dashed","dashed"),lwd=2,
+            col=c("orange","orange"))
> title(main="Brain and Body Weights (without outliers)",
+       xlab="body [kg]", ylab="brain [g]")
> legend("bottomright",inset=.05,
+        legend=c("fitted values","conf intvl (95%)","pred intvl (95%)"),
+        col=c("red","blue","orange"),lty=1,lwd=2)
> ## 種の名前をグラフ上に表示
> text(Animals,labels=rownames(Animals),cex=0.75)
```

(reg-example.r)