クレジット:
UTokyo Online Education 統計データ解析 I 2017 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析Ⅰ (平成２９年度)

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

村田　昇 (早稲田大学, 東京大学)
吉田朋広 (東京大学)
小池祐太 (首都大学東京, 東京大学)

第 6 章 確率分布

6.1. 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ような変数のことである. 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, 確率統計学では, 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.

6.2. 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**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確率統計学において重要な概念である. いまの場合, 確率変数の取りうる値が無限個あるかもしれないため, 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

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

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 $x_1, \ldots, x_n$ とし,

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

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 $x_1, x_2, \ldots$ に対して級数 $\sum_{i=1}^{\infty} \varphi(x_i)$ が絶対収束するとき,

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

[1] $\mathcal{X}$ は空集合でないとする.

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. 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

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

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

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

で定義する. 特に, 正の整数 p に対して,

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

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 $\sum_{x\in\mathcal{X}} x^pP(X=x)$ 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, $q\le p$ 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, 平均からのばらつき具合を定量化した指標として, X の**分散**を

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

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 $\sqrt{\mathrm{Var}[X]}$ 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:

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

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. 定義より明らかなように, 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 $X_1, X_2, \ldots, X_n$ が**独立**であるとは, $a_i \le b_i$ $(i=1,\ldots,n)$ なる任意の実数 $a_1, b_1, \ldots, a_n, b_n$ に対して,

$$\begin{aligned}(6.1)\quad & P(a_1\le X_1\le b_1, a_2\le X_2\le b_2, \ldots, a_n\le X_n\le b_n)\\ &= P(a_1\le X_1\le b_1)P(a_2\le X_2\le b_2)\cdots P(a_n\le X_n\le b_n)\end{aligned}$$

が成り立つことをいう. ここに, (6.1) 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, $\ldots$, 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, $X_1, X_2, \ldots, X_n$ が**同分布**であるとは, $a\le b$ なる任意の実数 a, b に対して,

$$P(a\le X_1\le b) = P(a\le X_2\le b) = \cdots = P(a\le X_n\le b)$$

[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.

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.

6.2.1. 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

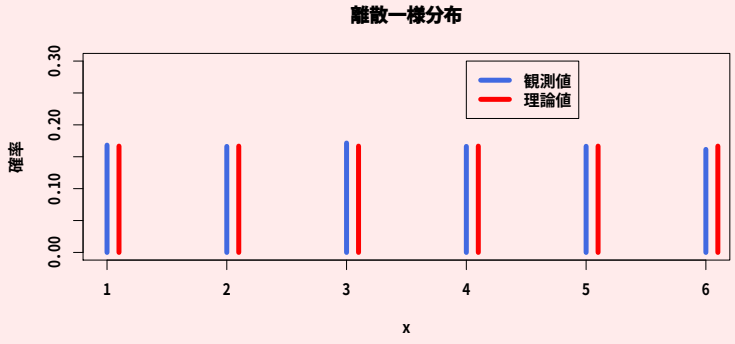
$$f(x) = \frac{1}{n}, \qquad x \in \{x_1, \dots, x_n\}$$

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 $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$, 分散は $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ 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).

```
> 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) # 乱数の初期値を指定
> 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] 2 5 3 6 6 1 4 6 4 3 6 3 5 4 1 6 2 1 2 6
> ## 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] 3.4809
> (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
0.1682 0.1662 0.1715 0.1662 0.1664 0.1615
> 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
```



(sample2.r)

演習 6.1. 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.

6.2.2. 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

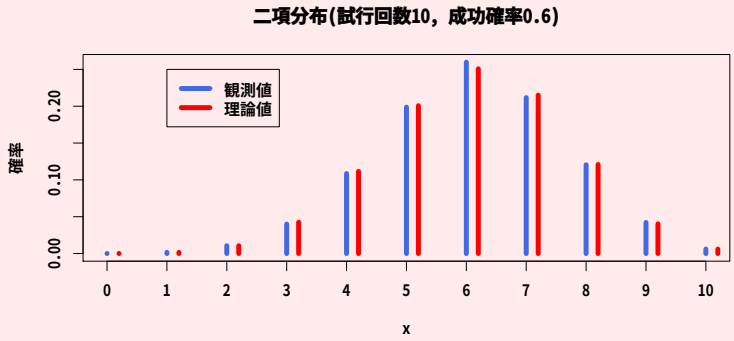
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \dots, n$$

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関数 `dbinom()` で計算できる.

```
> set.seed(123) # 乱数の初期値を指定
> 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
> 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
> 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
```



(rbinom2.r)

演習 6.2. 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.

6.2.3. 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

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

演習 6.2. 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.

6.2.3. 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

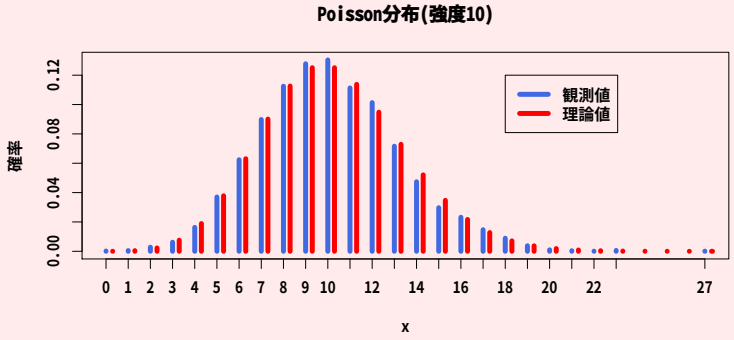
$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \dots$$

で与えられる離散分布をパラメーター 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.

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.

```
> set.seed(12345) # 乱数の初期値を指定
> 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
> (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
> 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
```



(rpois2.r)

演習 6.3. 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.

6.2.4. 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

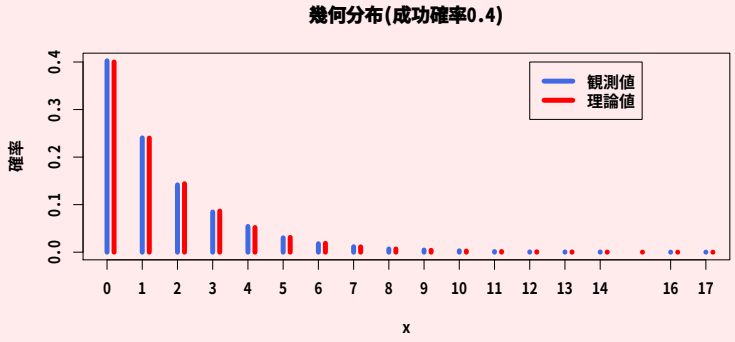
$$f(x) = p(1-p)^x, \qquad x = 0, 1, \ldots$$

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.

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.

```
> set.seed(777) # 乱数の初期値を指定
> 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
> 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
```



(rgeom.r)

演習 6.4. 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.

6.3. 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

が成り立つことをいい, 対応する確率分布を**連続分布**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

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

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

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

で定義する. 特に, 正の整数 p に対して

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

であり, これを 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

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

で定義する. 分散の平方根 $\sqrt{\mathrm{Var}[X]}$ 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:

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

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. 定義より明らかなように, 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, 確率変数たちが 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.

6.3.1. 一様分布. $a < b$ とする. 確率密度関数が

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

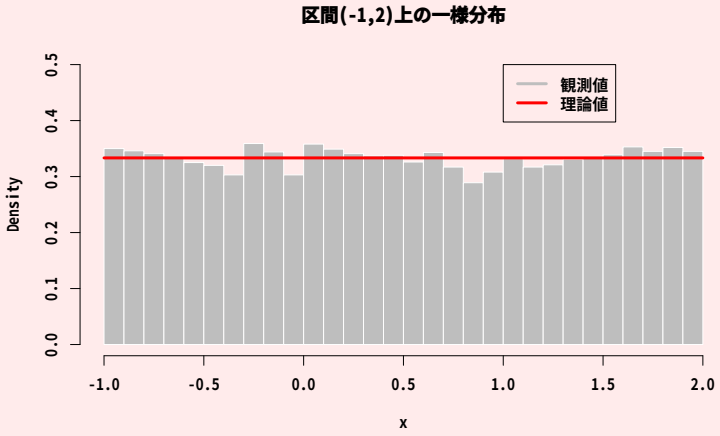
で与えられる連続分布を区間 (a, b) 上の**一様分布**と呼び, 記号 $U(a, b)$ で表す. 平均は $(a + b)/2$, 分散は $(b - a)^2/12$ で与えられる.

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, 連続分布の場合, 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.

[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.

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, 連続分布の場合, 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.

```
> set.seed(1) # 乱数の初期値を指定
> 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
> 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
```



(runif2.r)

演習 6.5. 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.

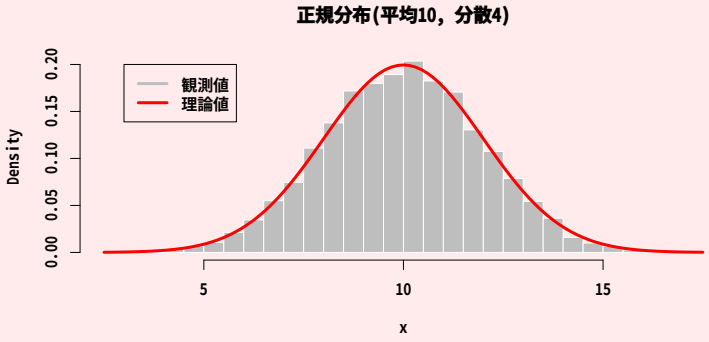
6.3.2. 正規分布. μ を実数, σ を正の実数とする. 確率密度関数が

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

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. 物理実験等の観測誤差の分布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.

```
> ## 統計的性質の確認
> mu <- 10
> sigma <- 2
> x <- rnorm(10000, mean = mu, sd = sigma)
> mean(x) # mu=10 に近い (大数の法則)
[1] 9.986371
> 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
```



(rnorm2.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, $X_1,\ldots,X_n$ 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,

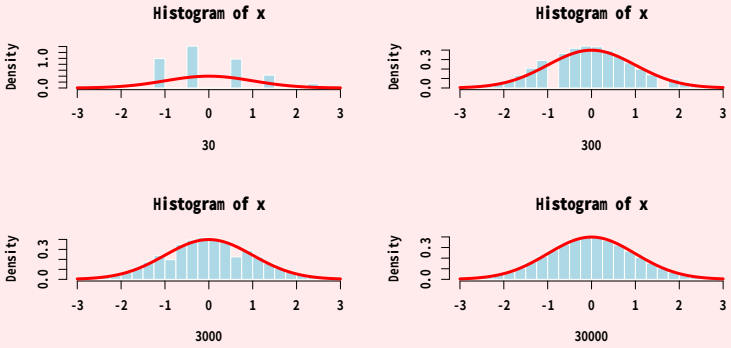
$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i - p)}{\sqrt{p(1-p)}}$$

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.

```
> # 二項分布の極限：離散分布から連続分布へ
> set.seed(123)
> op <- par(mfrow=c(2,2))
> 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
+ }
> par(op)
```

[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 $X_1,\ldots,X_n$ 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.

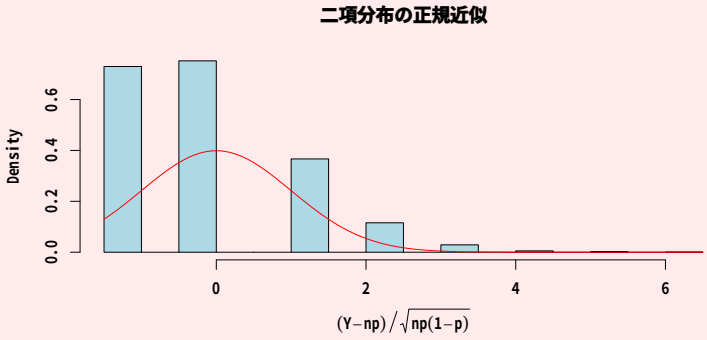
```
+         xlim=c(-3,3),xlab=n,col="lightblue",border = "white")
+    curve(dnorm(x, mean = 0, sd = 1), add = TRUE,
+          col = "red", lwd = 3) # 理論上の確率密度関数
+ }
> par(op)
```



(rbinom-normal2.r)

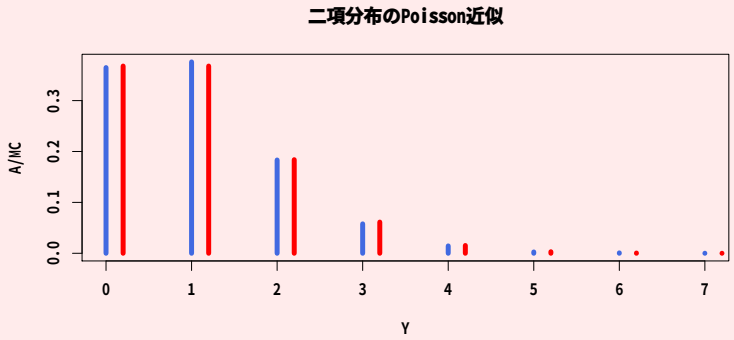
ただし, p が非常に小さい場合, 特に np がそれほど大きくならない程度に p が小さい場合は, $(Y-np)/\sqrt{np(1-p)}$ の分布の正規近似よりも, Y の分布のパラメーター np の Poisson 分布による近似の方が精度がよい (後者は前節で述べた少数の法則の特殊な場合である).

```
> # 二項分布の Poisson 近似が有効な場合
> set.seed(123)
> n <- 1000
> p <- 0.001
> MC <- 10000
> y <- rbinom(MC, n, p)
> ## まず正規近似を試す
> 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
```

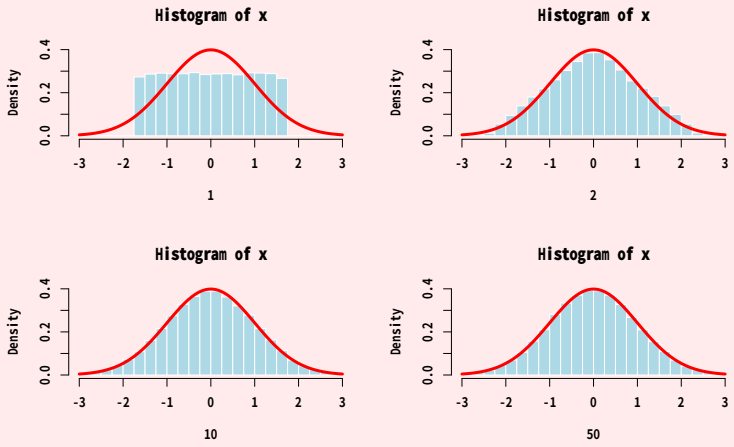
```
> plot(A/MC, type = "h", lwd = 5, col="royalblue",
+      main = "二項分布の Poisson 近似", xlab = expression(Y))
> lines(min(y):max(y)+0.2, dpois(min(y):max(y), n * p), type = "h",
+       col = "red", lwd = 5) # うまくいく
```



(rbinom-poisson.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.

```
> # 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
> op <- par(mfrow=c(2,2))
> 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
+ }
> par(op)
```



(runif-normal.r)

演習 6.6. 正規分布について調べてみよう.

(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，

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

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．

6.3.3. ガンマ分布. ν, α を正の実数とする．確率密度関数が

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

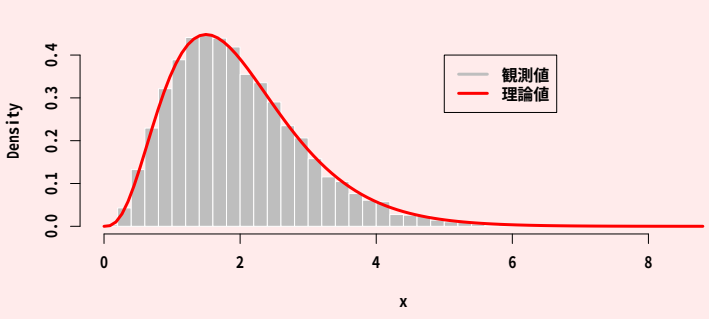
で与えられる連続分布をパラメータ ν, α の**ガンマ分布**と呼び，記号 $\Gamma(\nu,\alpha)$ や $G(\alpha,\nu)$ で表す．ただし，$\Gamma(\nu)$ はガンマ関数

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

を表す．ν, α はそれぞれ**形状パラメーター**, **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．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．

```
> set.seed(123) # 乱数の初期値を指定
> # ガンマ分布に従う乱数
> 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
> 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
> 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
```



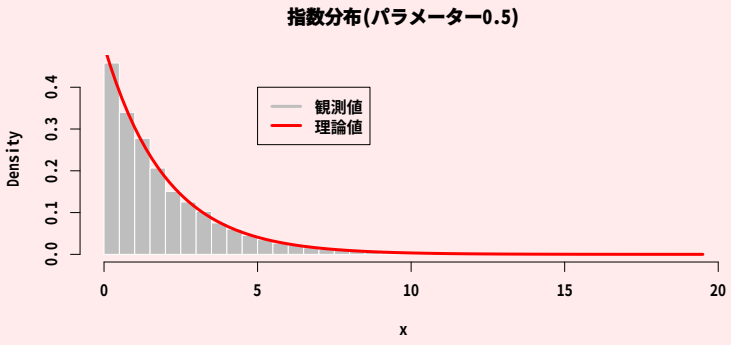
(rgamma2.r)

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 $\mathrm{Exp}(\lambda)$ で表す. λ は**レート**と呼ばれることがある. 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] 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.

```
> set.seed(20) # 乱数の初期値を指定
> 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
> 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
```

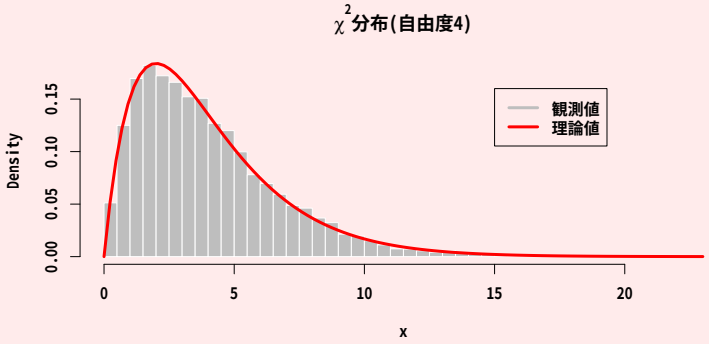


(rexp.r)

```
> set.seed(20) # 乱数の初期値を指定
> 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
> 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)
```

[5] χ^2 は「カイ二乗」と読む.

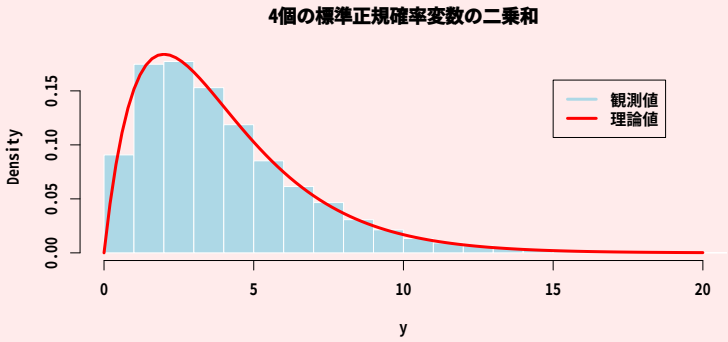
```
> curve(dchisq(x, k), add = TRUE, col = "red", lwd = 3) # 理論上の確率密度関数
> legend(15, 0.16, legend = c("観測値", "理論値"),
+        col = c("gray", "red"), lwd = 3) # 凡例を作成
```



(rchisq.r)

標準正規分布に従う k 個の独立な確率変数の二乗和は自由度 k の χ^2 分布に従うことが知られている.

```
> ## 標準正規確率変数の二乗和
> set.seed(123) # 乱数の初期値を指定
> 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.
> 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
```



(rgamma-chi2.r)

演習 6.7. 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 $\int_{-\infty}^{\infty} f(x)dx = 1$ 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と近くなるか確認せよ.

6.3.4. ベータ分布. α, β を正の実数とする. 確率密度関数が

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

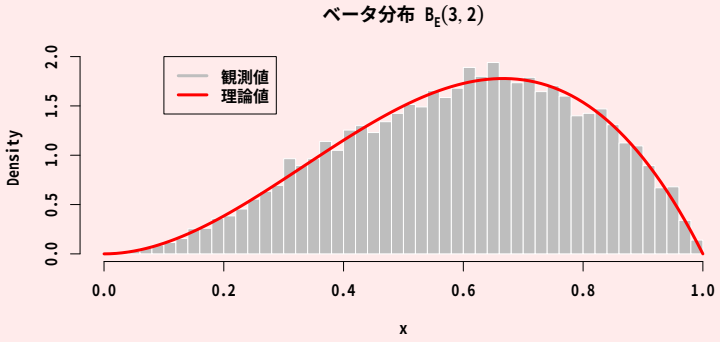
で与えられる連続分布を, パラメーター α, β の**ベータ分布**と呼び, 記号 $B_E(\alpha,\beta)$ で表す. ただし, $B(\alpha,\beta)$ はベータ関数

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

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.

```
> set.seed(123) # 乱数の初期値を指定
> 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
> 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
```



(rbeta.r)

演習 6.8. 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 $\int_{-\infty}^{\infty} f(x)dx = 1$ を満たすことを確認せよ.

6.3.5. t 分布. ν を正の実数とする. 確率密度関数が

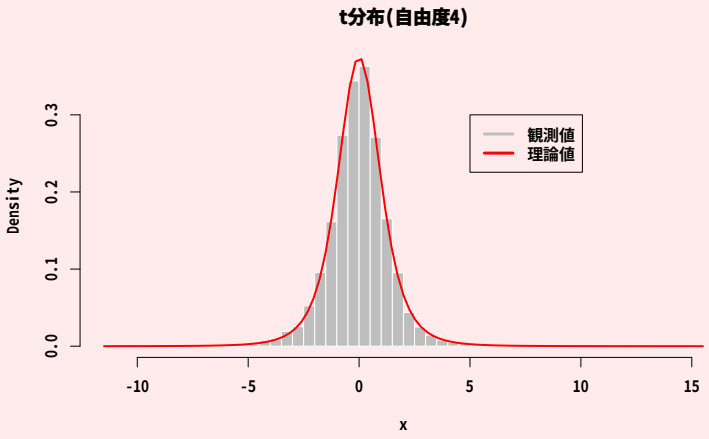
$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

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, 0 で与えられる. 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.

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

```
> set.seed(3) # 乱数の初期値を指定
> 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
> ### 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
> 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)
```



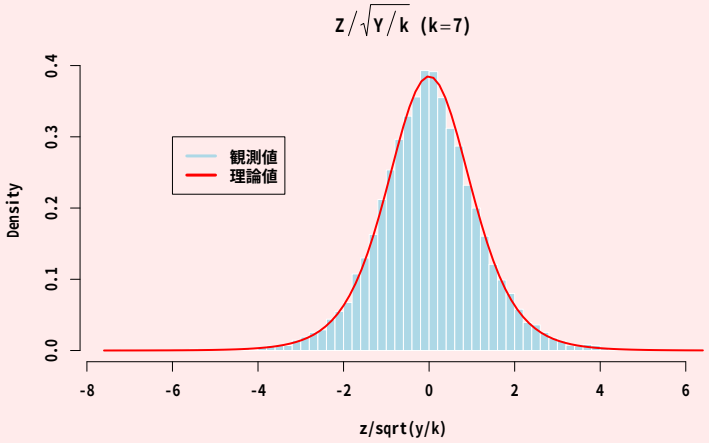
(rt2.r)

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

$$\frac{Z}{\sqrt{Y/k}}$$

は自由度 k の t 分布に従うことが知られている.

```
> ## 正規分布とカイ二乗分布を利用して t 分布を生成
> set.seed(11111) # 乱数の初期値の設定
> 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
> 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
```



(normal-t.r)

演習 6.9. 自由度が非常に大きい t 分布はどのような分布と近くなるか確認せよ.

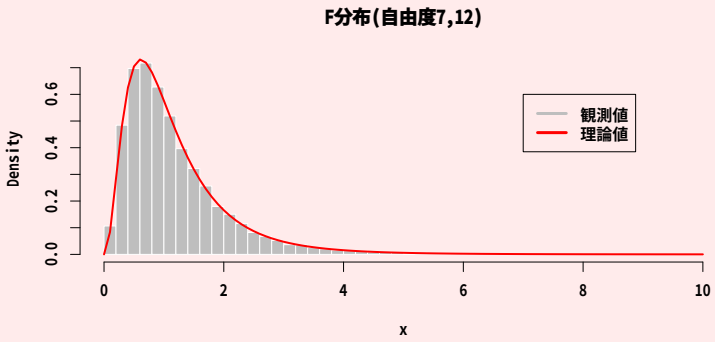
6.3.6. F **分布.** ν_1, ν_2 を正の実数とする. 確率密度関数が

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

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.

```
> set.seed(1) # 乱数の初期値を指定
> 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
> 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
```



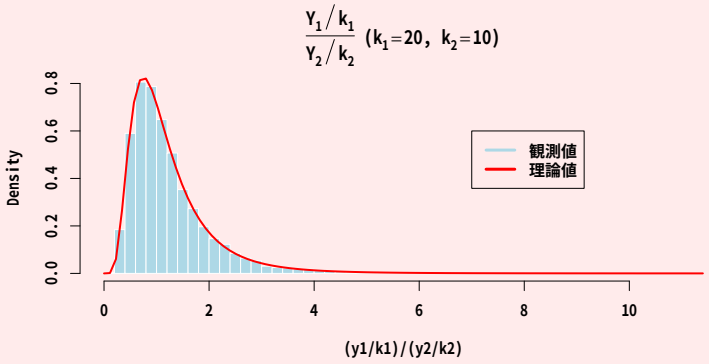
(rf2.r)

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

$$\frac{Y_1/k_1}{Y_2/k_2}$$

は自由度 k_1, k_2 の F 分布に従うことが知られている.

```
> ### カイ二乗分布を利用して F 分布を生成
> set.seed(22222) # 乱数の初期値を指定
> 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
> 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
```



(normal-f.r)

演習 6.10. t 分布と F 分布の関係を調べてみよ.

6.4. その他

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.

6.5. 参考文献

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