

クレジット:

UTokyo Online Education 統計データ解析 I 2017 小池祐太

ライセンス:

利用者は、本講義資料を、教育的な目的に限ってページ単位で利用することができます。特に記載のない限り、本講義資料はページ単位でクリエイティブ・コモンズ 表示-非営利-改変禁止 ライセンスの下に提供されています。

<http://creativecommons.org/licenses/by-nc-nd/4.0/>

本講義資料内には、東京大学が第三者より許諾を得て利用している画像等や、各種ライセンスによって提供されている画像等が含まれています。個々の画像等を本講義資料から切り離して利用することはできません。個々の画像等の利用については、それぞれの権利者の定めるところに従ってください。



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5章 シミュレーションと極限定理

データ解析の際に、分析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。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、分析対象としたい集団に「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」が含まれてしまうため、そもそも集団全体のデータの入手は不可能である。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実行し、その結果から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は難しくなることは直感的には理解できると思う。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ってしまうと、得られたデータから計算された平均気温は明らかに真に知りたい値より高くなってしまったため、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と思う。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。

5.1. 乱数

乱数とはランダムに生成された数列のことである。もちろん、コンピューターでは完全にランダムに数字を発生させることは不可能なため、それらの乱数は厳密には**擬似乱数**である。¹ 特に、数値シミュレーションを行う上では、それが再現可能であることが要請されるため、発生される乱数も再現可能である必要がある。R ではこれを実行するために、乱数の初期値を指定するための関数 `set.seed()` が用意されている(同一の初期値から生成される乱数は同一のものとなる)。

ここでは基本的な乱数として、ランダムサンプリング、二項乱数および一様乱数を考える。ランダムサンプリングは、その名の通り、与えられた集合の要素をランダムに抽出することで発生する乱数のことである。二項乱数は、「確率 p で表が出るコインを n 回投げた際の表が出る回数」に対応する乱数である。従って p と n によって乱数の発生が変わるため、それを明示するために「確率 p に対する次数 n の二項乱数」とも呼ぶ。一様乱数は、ある決まった区間 (a, b) ($a < b$) に含まれる数字からランダムに発生する乱数のことである。² 従って区間 (a, b) によって乱数の発生が変わるため、それを明示するために「区間 (a, b) 上の一様乱数」とも呼ぶ。

¹R では擬似乱数を発生させるための方法として Mersenne ツイスターがデフォルトでは用いられている。 `help(Random)` 参照。

² (a, b) は a より大きく b より小さい実数全体からなる集合を表す。

ランダムサンプリングは関数 `sample()` で実行できる。二項乱数および一様乱数はそれぞれ関数 `rbinom()` および `runif()` で発生させられる。

```
> ## 関数 sample の使い方
> x <- 1:10 # 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(x, 5) # x から 5 つの要素を重複なしでランダムに抽出
[1] 3 8 4 7 6
> sample(x, 10) # x の要素のランダムな並べ替えとなる
[1] 1 5 8 4 3 9 2 6 7 10
> sample(x, 5, replace = TRUE) # x から 5 つの要素を重複ありでランダムに抽出
[1] 9 3 1 4 10
> sample(1:6, 10, replace = TRUE) # 「サイコロを 10 回振って出た目を記録する実験」の再現
[1] 6 5 4 6 4 5 4 4 2 1
> sample(1:6, 10, prob = 6:1, replace = TRUE) # 出る目の確率に偏りがある場合
[1] 6 5 3 4 1 2 4 1 2 1
> ## 関数 rbinom の使い方
> rbinom(10, size = 4, prob = 0.5) # 確率 0.5 に対する次数 4 の二項乱数を 10 個発生
[1] 1 2 2 2 1 1 1 2 1 3
> rbinom(20, size = 4, prob = 0.2) # 個数を 20, 確率を 0.2 に変更
[1] 0 1 1 0 1 0 0 1 2 0 1 0 0 0 1 1 1 1 1 1
> ## 関数 runif の使い方
> runif(5, min = -1, max = 2) # 区間 (-1, 2) 上の一様乱数を 5 個発生
[1] 1.2634255 0.8876634 1.1305472 -0.9981257 0.4259497
> runif(5) # 何も指定しないと区間 (0, 1) を指定したことになる
[1] 0.2201189 0.3798165 0.6127710 0.3517979 0.1111354
> ## 関数 set.seed について
> set.seed(1) # 乱数の初期値を seed=1 で指定
> runif(5)
[1] 0.2655087 0.3721239 0.5728534 0.9082078 0.2016819
> set.seed(2) # 乱数の初期値を seed=2 で指定
> runif(5) # seed=1 の場合と異なる結果となる
[1] 0.1848823 0.7023740 0.5733263 0.1680519 0.9438393
> set.seed(1) # 乱数の初期値を seed=1 で指定
> runif(5) # 初めの seed=1 の結果と同じ
[1] 0.2655087 0.3721239 0.5728534 0.9082078 0.2016819
```

(sample.r)

R には他にも様々な種類の確率分布に従う乱数が実装されているが、それらについては次章で詳しく説明する。

5.2. 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。

5.2.1. 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。³

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 $P(X = a)$ で表すことにする。

³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。

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。

5.2.2. 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

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

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

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

で定義する。例えば、整数 p に対して

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

である。

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
のばらつき具合を定量化した指標

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

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
ため、単位をあわせるために、分散の平方根

$$\sqrt{\text{Var}[X]}$$

を考えることもよくある。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:

$$(5.1) \quad \text{Var}[X] = E[X^2] - (E[X])^2$$

演習 5.1. (5.1) 式の成立を確認せよ ($\sum_{i=1}^N P(X = x_i) = 1$ に注意)。

5.2.3. 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、 \dots 、 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

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

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。

定義 5.1 (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

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

が成り立つことをいう。

直感的には、 X_1, X_2, \dots, X_n が独立であるというのは、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 5.2 (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

$$P(X_1 = x) = P(X_2 = x) = \dots = P(X_n = x)$$

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)。

5.3. 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、その平均値は集団全体の平均値に近い値であると考えられることは、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無いコインを繰り返し投げ続けると、投げた回数に対する表が出た回数の割合は理論上の平均値 $\frac{1}{2}$ に近づいて行くことも、経験則として知っていると思う。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 5.3 (無限列の場合の独立性と同分布性). (a) 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。

(b) 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。

(c) 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 5.1 (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

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

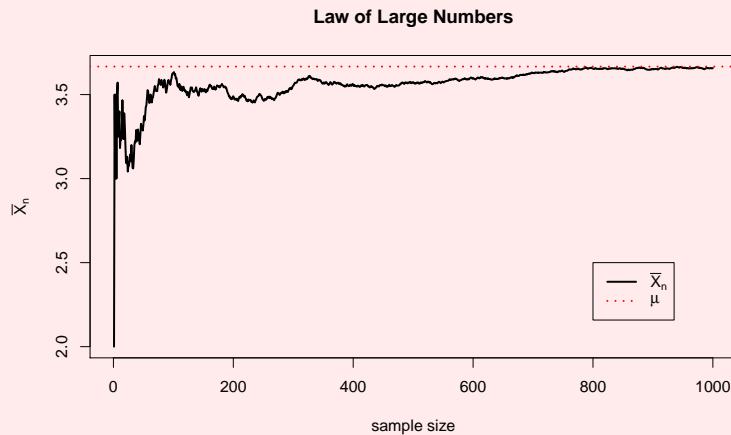
が $n \rightarrow \infty$ のとき μ に収束する確率は 1 である (このことを、「 $\frac{1}{n} \sum_{i=1}^n X_i$ は $n \rightarrow \infty$ のとき μ に**概収束する**」という)。

```
> ### 大数の法則
> ## 偶数が出る確率が奇数が出る確率の 2 倍あるようなサイコロを
> ## 振ったときに出る目の平均値を実験で確かめる
> set.seed(123) # 乱数の初期値の指定
> ## 1 回の実験で、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の確認
> omega <- 1:6 # 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
> 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 # 各実験回数ごとの標本平均の計算
> plot(xbar, type = "l", xlab = "sample size",
```

```

+   ylab = expression(bar(X)[n]), lwd = 2,
+   main = "Law of Large Numbers") # 標本平均の推移のプロット
> abline(mu, 0, lty = "dotted", col = "red", lwd = 2) # 理論平均に対応する点線
> legend(800,2.5,legend = c(expression(bar(X)[n]), expression(mu)),
+       col = c("black", "red"), lty = c("solid", "dotted"), lwd = 2) # 凡例の追加

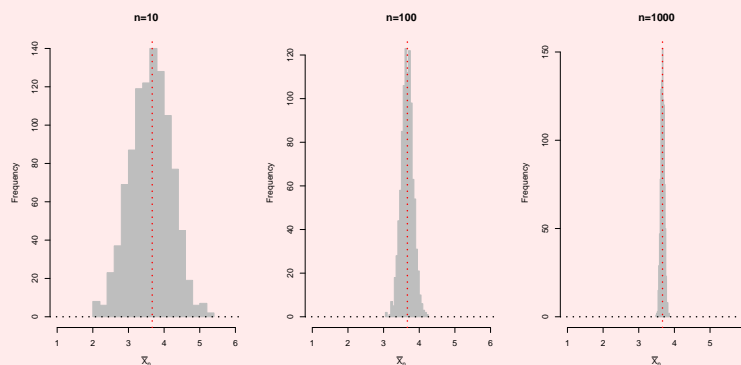
```



```

> ## 複数回実験を行ったとき、サンプル数が大きければ、標本平均は実験によらず
> ## ほとんどの場合に理論平均に近い値を取ることを確認
> 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 # 各実験の繰り返し回数
> nsample <- c(10, 100, 1000) # 実験するサンプル数
> op <- par(mfrow = c(1, 3)) # 画面を3分割
> for(n in nsample){
+   xbar <- replicate(MC, mymean(n)) # 実験をMC回実行して、標本平均を記録
+   hist(xbar, breaks = 20, xlim = c(1,6),
+        col = "gray", border = "gray",
+        main = paste0("n=",n), xlab = expression(bar(X)[n]))
+   abline(h = 0, lwd = 2, lty = "dotted")
+   abline(v = mu, col = "red", lwd = 2, lty = "dotted")
+ }
> par(op) # プロット環境の初期化

```



(LLN.r)

演習 5.2. 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。

5.4. 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

$$(5.2) \quad P(\alpha \leq \bar{X}_n - \mu \leq \beta)$$

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 (5.2) 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 5.2 (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

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

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

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

は定積分

$$(5.3) \quad \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

によって近似できる。積分 (5.3) 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 (5.3) 自体は関数 `pnorm()` を用いてコマンド `pnorm(b) - pnorm(a)` で計算できる。

注意 5.1. 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 (5.3) 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値の分布の仕方を**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 たち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の $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ の分布の仕方は標準正規確率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

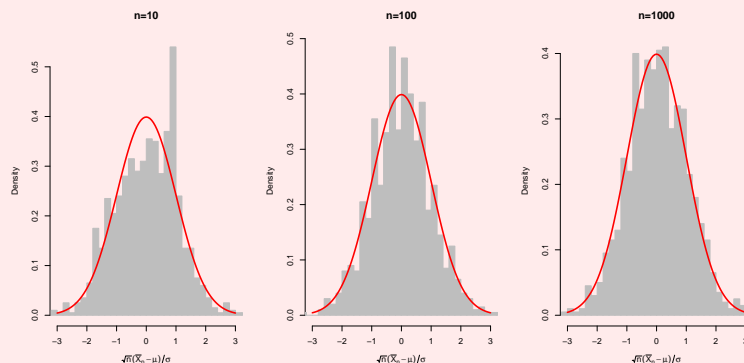
$$\frac{1}{b-a} P\left(a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b\right)$$

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)。従ってビンの幅 $b-a$ 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できる。


```

> ### 中心極限定理
> ## 偶数が出る確率が奇数が出る確率の2倍あるようなサイコロを
> ## 何度か振って出た目の標本平均の分布を確認
> set.seed(123) # 乱数の初期値の指定
> omega <- 1:6 # 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
> (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 # 各実験の繰り返し回数
> nsample <- c(10, 100, 1000) # 実験するサンプル数
> op <- par(mfrow = c(1, 3)) # 画面を3分割
> for(n in nsample){
+   xbar <- replicate(MC, mymean(n)) # 実験をMC回実行して、標本平均を記録
+   hist(sqrt(n) * (xbar - mu)/sigma, breaks = 25, freq = FALSE,
+        main = paste0("n=",n), xlim = c(-3, 3),
+        col = "gray", border = "gray",
+        xlab = expression(sqrt(n) * (bar(X)[n] - mu)/sigma))
+   curve(dnorm, add = TRUE, col = "red", lwd = 2)
+ }
> par(op) # プロット環境の初期化

```



(CLT.r)

演習 5.3. 様々な乱数を用いて、中心極限定理を確認してみなさい。

5.5. 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 5.3 (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、

$$(5.4) \quad \limsup_{n \rightarrow \infty} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log \log n}} = 1 \quad \text{a.s.},$$

$$(5.5) \quad \liminf_{n \rightarrow \infty} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log \log n}} = -1 \quad \text{a.s.}$$

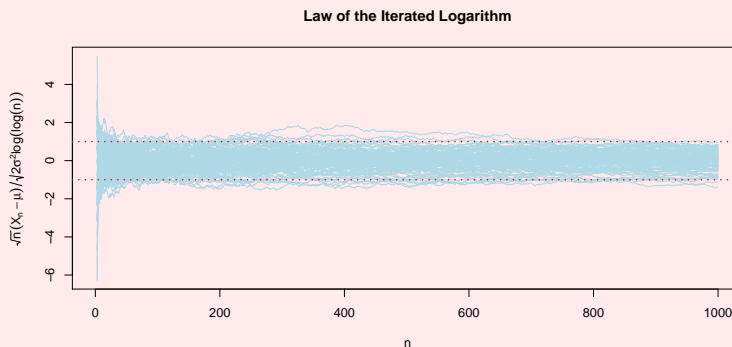
が成り立つ. より一般に, 列

$$\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{2\sigma^2 \log \log n}} \right)_{n=3}^{\infty}$$

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 5.3 の後半の主張は **Hartman-Wintner の定理** として知られている.

```
> ### 重複対数の法則
> ## 偶数が出る確率が奇数が出る確率の 2 倍あるようなサイコロを
> ## 何度か振って出た目の平均に対する重複対数の法則の確認
> set.seed(111)
> omega <- 1:6 # 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 <- 100 # 実験回数
> x <- sample(omega, size = n * MC, prob = p, replace = TRUE) # 乱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
> 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") # プロット
> abline(h = 1, lty = "dotted", col = "red", lwd = 2) # y=1 のプロット
> abline(h = -1, lty = "dotted", col = "red", lwd = 2) # y=-1 のプロット
```



(LIL.r)

演習 5.4. 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.

5.6. 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. 一日に n 個 (非常に多数とする) 生産するとき, 不良品は平均的には $\lambda = np$ 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 5.4 (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:

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

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

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

が成り立つならば, 任意の 0 以上の整数 k に対して

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

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.

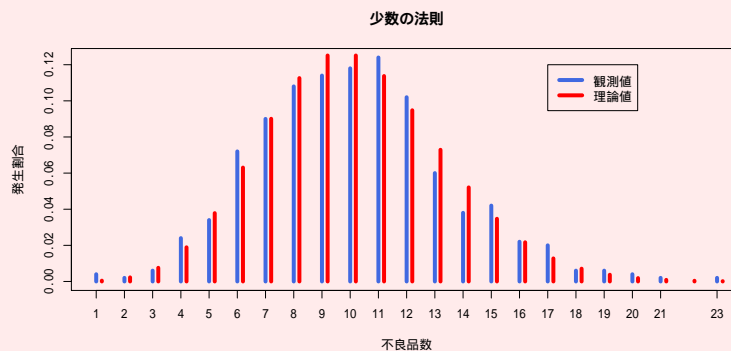
注意 5.2. 詳細は次章で説明するが, 取りうる値が 0 以上の整数全体で, 値が整数 $k \geq 0$ となる確率が

$$(5.6) \quad e^{-\lambda} \frac{\lambda^k}{k!}$$

で与えられる確率変数を強度 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. (5.6) は関数 `dpois()` で計算できる.

```
> ### 少数の法則
> ### 大きさ 1 の 2 項分布を使う
> set.seed(123)
> 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
> (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 22 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
> plot(A/MC, type = "h", lwd = 5, col="royalblue",
+      main = "少数の法則", xlab = "不良品数", ylab = "発生割合")
> lines(min(x):max(x)+0.2, dpois(min(x):max(x), n * p), type = "h",
+      col = "red", lwd = 5) # 理論上の割合を上書き
> legend(17, 0.12, legend = c("観測値", "理論値"),
+      col = c("royalblue", "red"), lwd = 5) # 凡例を作成
```



(LRE.r)

演習 5.5. 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.

5.7. 参考文献

1. P. Billingsley 著「Convergence of probability measures (第2版)」, Wiley (1999年).
2. 福島正俊著「確率論(第5版)」, 裳華房(2006年).
3. 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).
4. 竹村彰通著「統計(第2版)」, 共立出版(2007年).