April 23, 2008 V1R3 A
April 28, 2010 V1R5 B
May 10, 2011 V2R0

情報リテラシー演習

S 言語と R 言語 第 3 回
大数の法則と中心極限定理

w.yamamoto

1 確率シミュレーション (続き)

理論については、講義での解説を待たねばならないが、今回は乱数シミュレーションを用いて、大数の法則と中心極限定理についての実験を行ってもらう。

2 離散確率事象のモデル

表が上になる確率、裏が上になる確率とも 1/2 の平等なコインを投げるシミュレーションを行う。

2.1 二項分布 (コイン投げ)

S 言語では、r で始まる関数は、特定の確率分布に従う疑似乱数を発生させる。コイン投げは、1 回ごとの試行はベルヌーイ試行と呼ばれる。そして n 回の試行のうちの表が出る回数 k は、各試行で表が出る確率を p として、二項分布 $B(n, p)$ に従う、と言う。

2.2 多項分布 (賽子投げ)

6 面の賽子を考える。おのおのの面が上になる確率はすべて 1/6 とする。

```
> rmultinom(n=1,size=1,prob=c(1/6,1/6,1/6,1/6,1/6,1/6))
```

を 1 回実行すると、さいころを一回だけ振る確率シミュレーションを 1 セット、行うことができる。

```
> rmultinom(n=1,size=10,prob=c(1/6,1/6,1/6,1/6,1/6,1/6))
```

とすると、10 回振って出た目の頻度を得る確率シミュレーションを 1 セット、行う。

```
> rmultinom(n=10,size=5,prob=c(1/6,1/6,1/6,1/6,1/6,1/6))
```

とすると、5 回振って出た目を得る確率シミュレーションを 10 セット、行う。

2.3 ポアソン分布 (ポアソン到着)

開店時刻から閉店時刻までの間を通じて、10 分あたり平均して 3 人の到着が見込まれるお店で、10 分刻みで何人到着するか、という確率シミュレーションをする。開店時刻を 10 時、閉店時刻を 8 時とすると、営業時間は 10 時間であり、10 分刻みの時間単位では 60 単位のシミュレーションを行うことになる。

```
> rpois(n=60,lambda=3)
```

たとえば一時間あたりの来客数に変換すにのは、

```
> X <- rpois(n=60,lambda=3)
> X.sum.6 <- c(0)
> for ( i in c(1:10) ) {
+   tmp <- 0
+   for ( j in c(1:6) ) {
+      tmp <- tmp + X[6*(i-1)+(j-1)+1]
+   }
+   X.sum.6 <- append(X.sum.6,tmp)
+ }
> X.sum.6 <- X.sum.6[2:11]
apply(matrix(rpois(n=60,lambda=3),ncol=6), 1, "sum")
```

と発生させたデータを、60 分ずつ合計していく。これを S 言語風に書くのであれば

```
> apply(matrix(rpois(n=60,lambda=3),ncol=6), 1, "sum")
```

となるが、この科目では、特にこの書き方は強要しない。

3 連続確率事象のモデル

3.1 正規分布

正規分布 $N\left(50, 10^2\right)$ に従う疑似乱数を発生させるには、

```
> x <- qnorm(runif(1000,0,1),50,10)
```

これら二つのコードで、使用するメモリ領域の大きさが異なるが、小規模な計算では、どちらを用いるかは、趣味の問題に過ぎない。当面は理解しやすい方も用いることを薦めるが、前者の方法ならば一様疑似乱数が保存されるという利点はある。

他にも累積分布関数の逆関数を与えられれば、任意の確率分布に従う疑似乱数を得ることができるが、主な確率分布については表 3.1 にある通り、既に用意されている。離散分布の場合には、上のように簡単には変換できず、工夫が必要になるため、予め用意されていると便利である。実際 R では、正規分布 $N\left(50, 10^2\right)$ に従う疑似乱数の生成は

```
> x <- rnorm(1000,50,10)
```

と簡単に指定でき、`x <- qnorm(runif(1000,0,1),50,10)` と比べると記述が若干短くなる。

表 1: 主な分布と乱数を発生させる関数

確率分布	関数の記述
一様分布	`runif(n= , min= , max= )`
正規分布	`rnorm(n= , mean= , sd= )`
t 分布	`rt(n= , df= )`
ガンマ分布	`rgamma(n= , shape= , scale= )`
χ^2 分布	`rchisq(n= , df= , ncp= )`
ベータ分布	`rbeta(n= , shape1= , shape2= )`
F 分布	`rf(n= , df1= , df2= )`
指数分布	`rexp(n= , rate= )`
ロジスティック	`rlogis(n= , location= , scale= )`
対数正規分布	`rlnorm(n= , meanlog= , sdlog= )`
ワイブル分布	`rweibull(n= , shape= , scale= )`
コーシー分布	`rcauchy(n= , location= , scale= )`
二項分布	`rbinom(n= , size= , prob= )`
幾何分布	`rgeom(n= , prob= )`
負の二項分布	`rnbinom(n= , size= , prob= )` or `rnbinom(n= , size= , mu= )`
ポアソン分布	`rpois(n= , lambda= )`
超幾何分布	`rhyper(nn= , m= , n= , k= )`
多項分布	`rmultinom(n= , size= , prob= )`

4 R における疑似乱数の生成

R のデフォルトは `Mersenne-Twister` という日本人が発明した一様疑似乱数で、周期が $2^{19937}-1$ であり、623 次元に均等に分布することが証明されている。詳細は

`http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/mt.html`

を参照されたい。

```
RNGkind();
    # 疑似乱数の発生方法を得る
RNGkind(kind="一様乱数の発生アルゴリズム");
    # 疑似乱数の発生方法を設定する
```

他に `Wichmann-Hill`, `Marsaglia-Multicarry`, `Super-Duper`, `Knuth-TAOCP`, `Knuth-TAOCP-2002` が搭載されていて、指定可能である。

まず Mersenne-Twister は 0 から $2^{32}-1$ までの整数値を一様な確率で生成する乱数発生器、と見なせる。これから得る値を $2^{32}-1$ で除算すると、区間 $[0,1]$ 上の一様分布に従う疑似乱数を得る。これを $\tilde{U}$ と置く。

R では Mersenne-Twister に直接触れる必要はなく、`runif()` という関数を用いて、$\tilde{U}$ を発生させることができる。

```
> runif(n=1,min=0,max=1)
```

これで、0 から 1 までの小数が一つ、生成される。一つ発生させただけでは、一様分布にも見えないであろうと、

とりあえず 1000 個の疑似乱数を発生させる。

```
> runif(n=1000,min=0,max=1)
```

今度は多すぎて、やはり見えない。

```
mean(実数ベクトル);
    # 算術平均を計算する
sd(実数ベクトル);
    # 標準偏差を計算する
summary(実数ベクトル);
    # 最大値、最小値、四分位点、算術平均を計算する
```

そこで、平均や標準偏差、最大値、最小値、中央値、四分位などを計算してみる。平均や標準偏差を計算してみると、どうだろうか。

```
> u <- runif(1000,0,1)    # n=や min=や max=は省略可能
> mean(u)       # 算術平均を計算する関数
> sd(u)         # 標準偏差を計算する関数
> summary(u)    # 実数ベクトルに対して平均、最大、最小、四分位点を計算する
```

一様分布であれば、平均は 1/2、標準偏差は $1/\sqrt{12} = 0.2887$ に近いことが期待される。また最大値は 1 に、最小値は 0 にそれぞれ近く、1/4 位点は 0.25、中央値は 0.5、3/4 位点は 0.75 にそれぞれ近いはずである。

Q1 以上のことを、何度か繰り返して、各数字のばらつきを眺めて考察せよ。また `n=10000` とするとどうか、についても検討せよ。(ソースコードの方は、数字を出力するだけで良い。)

一様分布に従う乱数から、各種連続分布に従うのは原理は簡単である。次に、ある連続分布の累積分布関数 $F(x)$ とその逆関数 $F^{-1}(u)$ の計算方法が既に手元にあるとする。このとき、その連続分布に従う疑似乱数 $\tilde{X}$ は、

$$\tilde{X} \quad = \quad F^{-1}\left(\tilde{U}\right)$$

と $\tilde{U}$ を変換することで得られる。

例として、平均が 50 で標準偏差が 10 の正規分布 $N\left(50, 10^2\right)$ に従う疑似乱数を、一様疑似乱数を変換して得ることを試す。

```
> u <- runif(1,0,1)
> qnorm(u,50,10)
```

正規分布の累積分布関数の逆関数は `qnorm(u , mean=50, sd=10)` という関数を用いる。ここでも、疑似乱数を一つ生成した程度では、確率分布の確認も難しかろうと、1000 個の値を発生させてみる。

```
> u <- runif(1000,0,1)
> x <- qnorm(u,50,10)
```

`u` が実数でも実数ベクトルでも、`qnorm()` は正しく、正規分布の累積分布関数の逆関数で変換するところに、オブジェクト指向が見え隠れする。

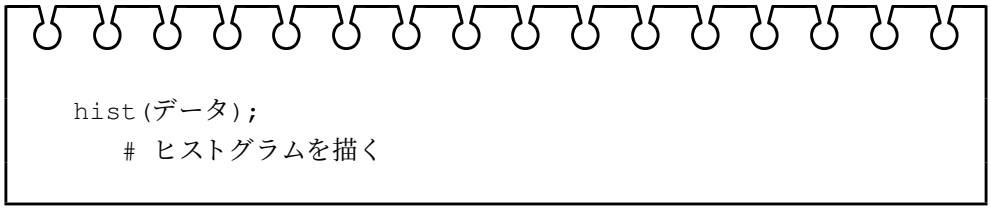
平均や標準偏差などを計算させてみると、平均は 50 に、標準偏差は 10 に、それぞれ近くなっているだろうか。また四分位点から、分布が対称であると判断できるだろうか。

```
> mean(x)
> sd(x)
> summary(x)
```

今度はヒストグラムも描いてみる。

```
> hist(x)
```

`hist()` は、データのベクトルを与えると、そのヒストグラムを描画する関数である。これを用いると、なんとか正規分布に見えそうな、対称かつ単峰で裾が軽いヒストグラムが描かれることを予想している。



5 大数の法則

一般の確率分布 F について、その累積分布関数を $F(x)$、その分布に独立に従う確率変数列を $X_1, X_2, X_3, \ldots$ と記す。また確率分布の期待値 $\mu = \int xdF(x)$ が有限と仮定する。このとき、大数の法則、は次のことを主張する。

> **大数の法則** n 個の確率変数の標本平均 $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ が、$n \to \infty$ とともに μ に収束する。

このことを、シミュレーションで眺めてみよう。

まず、正規分布 $N\left(50, 10^2\right)$ に従う疑似乱数を 10 個発生させ、その平均を計算する。

```
> x <- rnorm(10,50,10)
> mean(x)
```

このことを、1000 回繰り返してみると、何が分かるだろうか。

```
> x <- mean(rnorm(10,50,10))
> for ( i in c(2:1000) ) {
+   x <- append(x, mean(rnorm(10,50,10)))
+ }
> hist(x)
```

元々が平均が 50 で、標準偏差が 10 だったので、確率統計学で教わる通り、$n = 10$ 個の平均は同じく 50 に近く、標準偏差は $10/sqrt10$ に近いことが期待される。上の計算は、繰り返し文を用いずに `apply()` という関数を用いて

```
> X.10 <- matrix(rnorm(10000,50,10),ncol=10)
> x.10 <- apply(X.10,1,mean)
> hist(x.10)
```

としてもよい。`apply(X,1,mean)` は、行列オブジェクト `X` の各行ごとに `mean()` を実行せよ、という意味にな

る。apply(X,2,mean) も実行して、違いを確認してみることを薦める。(前者が行平均で、後者が列平均を得る)

```
apply(行列オブジェクト, 次元番号, 関数);
    # 行または列に関数を適用する
```

更に $n = 100$、$n = 1000$、$n = 10000$ と疑似乱数を発生させて平均を計算することを 1000 回ずつ繰り返し、ヒストグラムを描いてみる。

```
> X.100   <- matrix(rnorm(100000,50,10),ncol=100)
> X.1000  <- matrix(rnorm(1000000,50,10),ncol=1000)
> X.10000 <- matrix(rnorm(10000000,50,10),ncol=10000)
> x.100 <- apply(X.100,1,mean)
> x.1000 <- apply(X.1000,1,mean)
> x.10000 <- apply(X.10000,1,mean)
```

ヒストグラムを描く前に、par(mfrow=c(2,2)) と実行してから、最初のヒストグラムを描き直し、更に残りの三枚も描く。

```
> par(mfrow=c(2,2))
> hist(x.10)
> hist(x.100)
> hist(x.1000)
> hist(x.10000)
```

Q2 平均や標準偏差等も計算した上で、これらのグラフとデータは、大数の法則を支持するかどうか、検討しなさい。(ソースコードの方は、数字を出力するだけで良い。)

Q3 密度関数が連続かつ期待値に関して対称な正規分布では面白くないので、期待値が 5 のポアソン分布 (lambda=5) についても同じことを試みよ。また、元のデータのヒストグラムが対称でないことも、確認せよ。(ソースコードの方は、ヒストグラムも表示して良い。)

6 中心極限定理

中心極限定理には、いくつかのバリエーションがあるが、まずは簡単なものを引用しておく。

中心極限定理 $X_1, X_2, \ldots$ を独立に同一の分布に従う確率変数とし、それらの期待値は $\mu = \mathrm{E}[X_i]$、分散を $\sigma^2 = \mathrm{Var}[X_i]$ と記す。μ と σ^2 がともに有限の値であると仮定すると、任意の実数 $-\infty < x < \infty$ に対して

$$P\left[\sqrt{n}\left(\overline{X}_n - \mu\right) < x\right] \to \Phi\left(\frac{x}{\sigma}\right), \quad (n \to \infty)$$

が成り立つ。ただし Φ は標準正規分布 $N(0,1)$ の累積分布関数 $\int_{-\infty}^{x} e^{-t^2/2}/\sqrt{2\pi}dt$ である。

平たく言えば「標本平均から期待値を引いたものを $\sqrt{n}$ 倍した統計量の従う確率分布は、$n \to \infty$ につれて、分散が σ^2 の正規分布に近づく」となる。

前節で紹介した正規分布の例では、**Q3** で生成した平均から期待値 50 を引いて、それぞれ $\sqrt{n}$ 倍すれば良い。それを R で実行するコードは下記のようになる。

```
> par(mfrow=c(2,2))
> hist(sqrt(10)*(x.10-50))
> hist(sqrt(100)*(x.100-50))
> hist(sqrt(1000)*(x.1000-50))
> hist(sqrt(10000)*(x.10000-50))
```

元々が、正規分布に従うものとして生成した疑似乱数なので、これらはいずれも分散が 100(標準偏差が 10) の正規分布に従うことは明らかである。そこで、正規分布以外についても、中心極限定理を確認してみる、という課題を二つ課す。

Q4 ポアソン分布の場合にも、中心極限定理を確認してみなさい。

Q5 他に幾何分布についても、中心極限定理を確認してみなさい。

7 おわりに

グラフが沢山、レポートに貼り付くことが予想される。うまく大きさを調整して、読みやすいようにレイアウトを工夫することを、期待する。

参考文献

1. 「準数計算法/乱数」D. E. Knuth 著, 渋谷政昭 訳, サイエンス社, 1981.
2. 「乱数」 伏見正則 著, UP 応用数学選書 12, 東京大学出版会, ISBN 4-13-064072-0, 1989.
3. 「モンテカルロ法の金融工学への応用」 湯前祥二・鈴木輝好 共著, シリーズ現代金融工学 6, 朝倉書店, ISBN 4-254-27506-4, 2000.
4. 「統計」竹村彰通 著, 共立講座 21 世紀の数学 14, 共立出版, ISBN 4-320-01566-5, 1997.
5. 「工学のためのデータサイエンス入門」 間瀬茂・神保雅一・鎌倉稔成・金藤浩司 共著, 工学のための数学 EKM-3, 数理工学社, ISBN 4-901683-12-8, 2004.