医学統計学演習（看護）：資料 2

芳賀昭弘 *

1　復習

確率変数 x に対する、平均 μ、標準偏差 σ の正規分布の値は $\mathrm{dnorm}(x, \mu, \sigma)$ で得ることができます。例えば平均が 50 で標準偏差が 10 の正規分布において、$x = 60$ の正規分布の値は dnorm(60,50,10) とすることで得られます（R のコンソール上で 0.02419707 という値が返ってきます)。次のようにすると図示することも可能です。

```
> x = seq(20,80,by=0.2)
> plot(x, dnorm(x,50,10), type="l", col="blue")
```

1 行目で 20 から 80 までの 0.2 刻みで与えた値を x に入れ、2 行目で対応する正規分布の値を求めてプロットしています。また、上と同じグラフ化は $\mathrm{curve}(\mathrm{dnorm}(x, 50, 10)$,20,80) とするだけでも可能です。

練習問題：上記の正規分布の下側と上側それぞれ 2.5%（合わせて 5%）を与える x を求めよ。

ここで取り上げた正規分布は統計学で非常に重要な役割を担っています。その土台を築いているのが以下に示す中心極限定理です。

2　中心極限定理

中心極限定理：

1. 標本抽出を繰り返したとき、"一回のサンプルの平均値"が示す分布の平均は母集団の平均に一致する。
2. 標本抽出を繰り返したとき、"サンプルの平均値"の標準偏差は'母集団の標準偏差 $/\sqrt{n}$に一致する。ここで n は抽出するときのサンプルの数（標本数・サンプルサイズ）である。
3. 母集団の分布が正規分布なら、標本抽出を繰り返したときの "サンプルの平均値"の分布も正規分布になる。サンプルサイズが十分に大きいとき（$n \sim 30$ 程度)、母集団の分布の形によらず、標本抽出を繰り返したときの "サンプルの平均値"の分布は正規分布になる。

中心極限定理を使った例題をやってみましょう。

* Electronic address: haga@tokushima-u.ac.jp

例題１：日本人の成人男性の $1\mu l$ あたりの白血球数の平均値が 6300 で標準偏差が 1700 の正規分布をしているとします。日本人の成人男性からランダムに 16 名のデータを取り出して $1\mu l$ あたりの白血球数の平均を求める、ということを繰り返すと、その平均値の分布はどのようになると期待されるでしょうか？ R でその分布を図示してみましょう。

（答え）中心極限定理に従えば、平均 6300 で標準偏差が $425(=1700/\sqrt{16})$ の正規分布となる（はず）。

例題２：0 から 10000 番まで番号がついた紙がある。そのうち 100 枚を選んでその番号の平均値を求める、という作業を繰り返すとその平均値の分布はどのようになるでしょうか？ R でその分布を図示してみましょう。

（答え）中心極限定理に従えば、平均 5000 で標準偏差が $288.7(=2887/\sqrt{100})$ となります。

例題 2 の母集団は一様分布をしていますが、それでも「“サンプルの平均値”の平均」の分布は、定理 3 の通り正規分布になるはずです。

上記の項目は、統計解析を行うことを前提に何かの研究を実施する上で都合の良い性質を与えてくれます。すなわち、どの程度の個数の標本を抽出すれば、母集団の性質を推定したり異なる母集団同士を比べたりすることができるのか、ということを事前に検討をつけることができます。これは「研究計画」を立てるのに役立ちます。もちろん、「研究計画」が立てられるような研究ばかりではなく、標本の個数が限られてしまうケースも多くありますが、そのような限定された場合でも推定能力を示した上で、何かしらの推論を行うことが可能です。それが数学の力であり、統計解析の真髄です。

それでは、この中心極限定理が本当に成り立つかどうか、例題 1 と 2 に対してコンピュータで実験してみたいと思います。

3　乱数を使った標本抽出と中心極限定理

0 から 10000 の番号が書かれている札から適当に札 100 枚選ぶという例題 2 の状況を想像してください。これは、一様で離散的な確率変数 $x \in \{0, 1, 2, \cdots, 10000\}$ の中から 100 枚選ぶということであり、**一様分布からのランダム・サンプリング**と言います。一様分布ではないある確率分布に従う変数を適当に抽出する場合はどうなるでしょう？ “適当”に変数を選ぶ、といっても、0 から 10000 の札から 100 枚選ぶというのとは状況が異なります。0 から 10000 の番号が書かれている札でも、前半の若い番号が選ばれやすいという状況の場合はもはや一様分布からのランダム・サンプリングではないでしょう。例えば 0 から 5000 まで 5001 から 10000 よりも 10 倍の確率で選ばれるような場合、0 から 5000 の札をそれぞれ 10 枚用意して全部で 50010（0 から 5000 まで 10 枚）+5000（5001 から 10000 まで 1 枚）枚の中から 100 枚を抽出することで所望の分布からのランダム・サンプリングが実現できるでしょう。このように任意の確率分布から “適当”に変数を選ぶということは、**その確率分布に則った重みを付けて**変数を選ぶということです。正規分布の例で言えば、ピーク付近の x が最も選ばれる確率が高く、その辺りの x を選ぶということになります。ですが、正規分布の裾野（端っこ）も、確率は低くても選ばれないということではありません。それこそ何千回、何万回というサンプリングを繰り返したら、端っこの変数も選ばれるべきでしょう。このようなサンプリングを実現しているのがコンピュータを使ったモンテカルロ・サンプリングです。この具体的なサンプリング方法は、現在でも研究の対象となるくらい奥が深いので、これ以上述べません（興味のある人はぜひ検索してみてくださいね）。ここでは、R で用意されている確率分布サンプリング関数を使い、中心極限定理の言うところのサンプルを繰り返

して得られた平均値の分布が正規分布になることを確かめてみたいと思います。以下のスクリプトを作成しましょう。

```
SampMean = numeric(length=10000)
for(i in 1:10000){
sample = rnorm(16,6300,1700)
SampMean[i] = mean(sample)
}
hist(SampMean,probability=T,xlim=c(5000,8000))
```

（上記の説明）

1 行目：「SampMean」という名前の 10000 の数値データを格納できる配列を作る。

2 行目：「for」は以下に続くの内容を繰り返す。この場合、10000 回繰り返す（i in i:10000）。

3 行目：「rnorm(n, μ, σ)」は平均 μ、標準偏差 σ の正規分布からランダムに n 個の値を抽出する。

4 行目：抽出された 16 個の数値の平均を求め、SampMean に格納する。

5 行目：3-4 行目を繰り返す。

6 行目：10000 個の平均値のヒストグラムを表示する。

このスクリプトのファイル名を MonteCarlo1.R とします。次にスクリプトを実行してみよう。

```
> source(“MonteCarlo1.R”)
```

釣鐘型の分布ができます。これは平均が 6300、標準偏差 $1700/\sqrt{16}$ の正規分布とほぼ一致することが以下を実行することで確かめられます（中心極限定理の２）。

```
> par(new=T)
> curve(dnorm(x,6300,1700/sqrt(16)), 5000,8000)
```

4 演習

1. runif 関数は一様分布からランダムに値を抽出する関数である。サイコロを模擬する場合、ceiling(runif$(n, 0, 6)$) と書く。n はサイコロを振る回数である。ceiling 関数は小数点を切り上げる関数である。さて、サイコロを 10 回, 100 回, 10000 回振った結果をヒストグラムで表示せよ。
   （ヒント：下線には何が入るでしょう？）

   ```
   > hist(ceiling(_____),breaks=seq(0.5,6.5,1))
   ```

   （なお、上記を理解するのために、以下も実行してみてください）

   ```
   > runif(10,0,6)
   > ceiling(runif(10,0,6))
   ```

   何をやっているか、理解できましたか？

2. 0 から 10000 番までの札がある。番号の平均値と不変標準偏差はいくらか。

（ヒント：平均値を出す関数は mean、不変標準偏差を出す関するは sd です）
> x <- seq(0,10000,1)
> _____(x)
> _____(x)

3. 0 から 10000 番までの札がある。そこから 100 枚ランダムに選択したときの平均値を 10000 回計算した結果の分布をヒストグラムで表示せよ。なお、0 から 10000 までの一様分布からランダムに 100 個の値を抽出する関数は runif($n = 100, 0, 10000$) である。→MonteCarlo1.R のスクリプトファイルをコピーして別の名前で保存し、rnorm(n=10, mean=50,sd=10) を適切に修正してみてください。また、最後の hist(SampMean,xlim=c(40,60)) の部分は hist(SampMean,xlim=c(4000,6000)) にしてください（横軸の表示を変える）。

4. 平均 5000、標準偏差 288.7 の正規分布を表示し、上のヒストグラムと一致していることを確かめよ。上記で使った分布は一様分布なのに、なぜ正規分布と一致するのでしょう？（ヒント：中心極限定理と、0 から 10000 までの一様分布の標準偏差が 2887 である）

5. 試行回数 100, 表の出る確率 0.5 の二項分布からランダムに 10 個の値を抽出した場合、表=1、裏=0 としてその 10 個の平均値の分布をヒストグラムで表示せよ（→MonteCarlo1.R のスクリプトファイルをコピーして別の名前で保存し、rnorm(n=10, mean=50,sd=10) を適切に修正してみてください。なお、この場合の二項分布からランダムに値を抽出する関数は rbinom($n = 10, 100, 0.5$) です）。同時に、平均が 50、標準偏差が $\sqrt{100 * 0.25/10}$ である正規分布をプロットせよ。両者は重なりますか？

いつものように manaba のレポートで解答ください。3, 4, 5 についてはスクリプトを貼り付けてください。