サンプルの相関は母集団の相関をうまく代表しているのか？　（1）

たとえば，学生数が 10000 人のある大学があったとします。これが母集団です。この母集団においては，「自分の大学が好きですか」と「授業には積極的に参加していますか」という質問に対する回答の相関係数は，r=.57 だとします。

では，この大学の学生の中からランダムに 74 名を抽出して，「自分の大学が好きですか」と，「授業には積極的に参加していますか」という質問をしたとすると，このサンプルにおける相関係数はどの程度だと思いますか？

R を使ってこのような場合のシミュレーションをすることができます。

任意の相関係数を持つ変量を発生させるには，MASS パッケージに入っている mvrnorm 関数を使います。命令は以下のようになります。

```
library(MASS)
x <- matrix(c(1, 0.57, 0.57, 1), ncol=2)
data <- mvrnorm(n= 10000, mu= c(0, 0), Sigma= x, empirical= TRUE)
```

最初は MASS パッケージの呼び出しです。

２行目は，以下のような相関マトリックスを x として認識させている部分です。これは母集団における相関マトリックスですね。1, 0.57, 0.57, 1 という数字の列を，２列（ncol=2）で折り返すようにする命令で，以下のようなマトリックスを認識させています。

	好き	積極参加
好き	1.00	0.57
積極参加	0.57	1.00

３行目がデータを発生させる命令です。mvrnorm の後のカッコの中は，n=で発生させるデータ数，mu=で作成されるデータの平均値，Sigma=で相関マトリックス，empirical=で誤差を指定します。この場合，n= 10000 として 10000 人分のデータを発生させます。mu=の後は c(0, 0)となっていますが，これは２つの変数それぞれの平均値を 0 とするという命令になります。もし平均値を 1 と 10 にしたければ，c(1, 10)とすればよいわけです。Sigma=の後は，２行目で作成した相関マトリックスを指定します。empirical=の後は，TRUE とすると指示通りのデータを作成してくれます。FALES とすれば，ほんのちょっとのズレがあるデータが作成されます。そしてこれを data という名前にしているということです。

ここで

```
head(data)
```

と入力してそれを見ると，以下のようになっているはずです。しかし，これを読んで同じ命令をしたとしても，以下の結果は同じではないでしょう。データはランダムに発生させるので，同じ命令でも違うデータになります。

```
> library(MASS)
> x <- matrix(c(1, 0.57, 0.57, 1), ncol=2)
> data <- mvrnorm(n = 10000, mu = c(0, 0), Sigma =x, empirical = TRUE)
> head(data)
            [,1]        [,2]
[1,] -0.32254914  0.30779944
[2,] -1.56107505 -0.02560525
[3,]  1.49608310  0.77718623
[4,] -0.28796966 -0.88894165
[5,] -0.08284855  0.65749954
[6,] -0.85731526  0.16321179
>
```

さて，ここで作成されたデータはマトリックス型のデータ形式になっているので，これをデータフレーム形式になおして，d.data という名前にしておきます。また変数名はあった方がわかりやすいでしょうから，以下のようにして１列目と２列目に「好き」と「積極参加」という名前を付けておきます。

```
d.data <- data.frame(data)
colnames(d.data) <- c("好き", "積極関与")
```

すると，こんな感じになります。

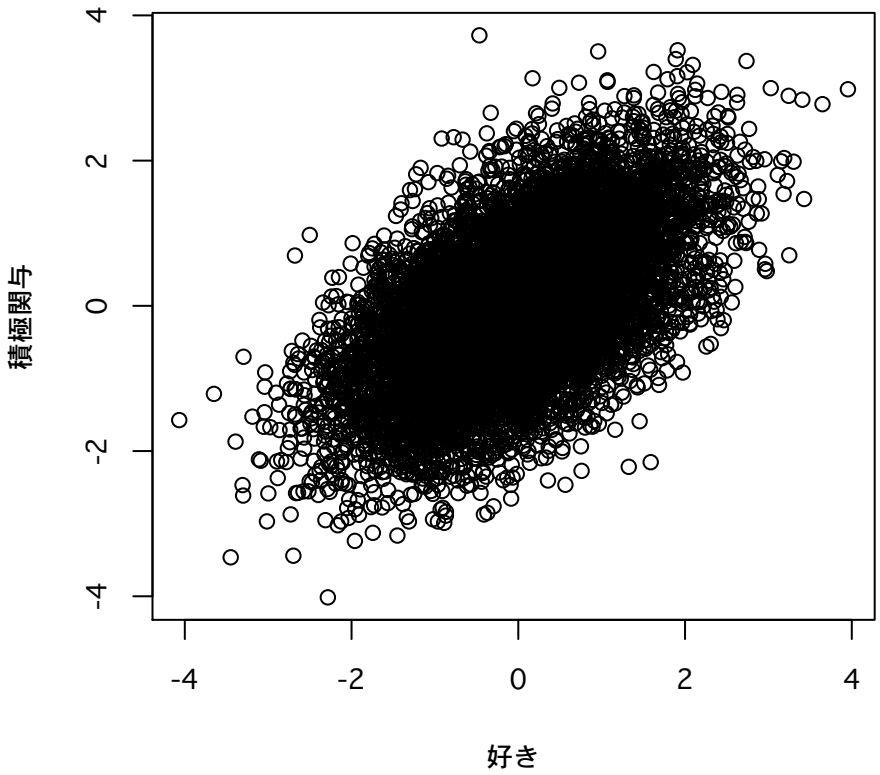
```
> d.data <- data.frame(data)
> colnames(d.data) <- c("好き", "積極関与")
> head(d.data)
         好き     積極関与
1 -0.32254914  0.30779944
2 -1.56107505 -0.02560525
3  1.49608310  0.77718623
4 -0.28796966 -0.88894165
5 -0.08284855  0.65749954
6 -0.85731526  0.16321179
>
```

これで 10000 人分のデータができました。平均値や標準偏差等，２変数の散布図の作成，そして相関係数を算出してみてください。

たとえば今回のデータを describe で計算させると，平均値が０，標準偏差が１になっていることが確認できます。作成されたデータは標準正規分布ということです。先の mvrnorm で，empirical=を FALSE にすると，ほんの少し平均値や標準偏差がズレます。一度確認してみてください。

```
> library(psych)
> describe(d.data)
          var     n mean sd median trimmed  mad   min  max range  skew kurtosis   se
好き        1 10000    0  1   0.02       0 1.00 -4.07 3.95  8.02 -0.01     0.02 0.01
積極関与    2 10000    0  1   0.00       0 0.98 -4.01 3.73  7.74  0.04     0.04 0.01
> 
```

また散布図を作成してみると以下のようになりました。



そして，corr.test で相関係数を確認してみると…

```
> corr.test(d.data)
Call:corr.test(x = d.data)
Correlation matrix
          好き 積極関与
好き      1.00      0.57
積極関与 0.57      1.00
Sample Size
           好き 積極関与
好き      10000     10000
積極関与 10000     10000
Probability values (Entries above the diagonal are adjusted for multiple tests.)
          好き 積極関与
好き         0         0
積極関与     0         0
```

ちゃんと両者の間の相関係数は .57 になっていることが確認できます。サンプル数も 10000 であることが確認できます。もちろん有意な相関です。

では，この母集団から 74 名をランダムに取り出したデータを作成してみます。やり方は以下のようです。

```
qq1 <- c(1:10000)
qq2 <- sample(qq1, 74, replace= FALSE)
sub.d.data <- subset(d.data[qq2, ])
```

１行目は１から 10000 までのナンバーリストを作って，qq1 に入れています。２行目がランダムサンプリングの指示で，１から 10000 までのナンバーが入った qq1 から，74 個を取り出して，qq2 に入れろという命令です。なお，replace=は重複を許すかどうかで，許さないなら FALSE にしておきます。そして３行目が，d.data から qq2 の番号の行だけをとりだして，sub.d.data というデータのサブセットを作りなさいという命令です。

これで sub.d.data という名前で，74 個をランダムサンプリングしたデータができました。これについて，また平均値や標準偏差等，２変数の散布図の作成，そして相関係数を算出してみてください。

たいていは，.57 前後の相関係数が得られると思います。

そして，極めてまれに，かなり離れた相関係数が得られます。では，サンプリングを 1000

回繰り返した場合，最低でどの程度，最高でどの程度の相関係数が得られるでしょうか？

この人間がやるととんでもなく面倒な作業を PC はやってくれます。

命令は以下のようで OK でしょう。（他にもやり方はあるでしょうが…）

```
box0 <- rep(NA, 1000)
box <- matrix(box0, ncol=1)
qq1 <- c(1:10000)
for(m in 1:1000) {
        qq2 <- sample(qq1, 74, replace = FALSE)
        sub.d.data <- subset(d.data[qq2, ])
        box[m,1] <- cor(sub.d.data$好き, sub.d.data$積極関与)
        }
```

１，２行目で計算された相関係数を収納する入れ物を準備しています，1000 行１列に NA が並ぶ box という名前の入れ物です。

それ以下は，先のサンプリングのやりかたを 1000 回繰り返すものです。順番にたどっていけば理解は難しくないでしょう，cor(sub.d.data$好き，sub.d.data$積極関与)を計算して，その結果を box に入れさせています。

つまり，box には，1000 回分の相関係数が並ぶということなのです。上をやった後， box の中身を確認してみてください。

では，この 1000 回のサンプリングの結果，相関係数はどの程度の値をとるでしょうか。box の最小値，最大値，平均や標準偏差，そしてヒストグラムを作って確認してみてください。

やってみたところ，平均は .57，標準偏差は 0.08，最小値は .27，最大値は .79 でした（もちろんこれは唯一の答えではありません。それぞれの場合で，少しずつ違ってきます）。ヒストグラムも描いてみました。

つまり，概ね 70%弱くらいは .49 から .65 くらいの間（正規分布ではないので，単純に ±１SD で判断するわけにはいきませんが）におさまるようです。しかし，そこからさらに離れるケースも 30%強くらいはあるわけです。

我々が手にできるのは，サンプルの相関係数です。母集団の相関係数は，通常はわかりま

せん。サンプルの相関係数の近くにあるだろうなということが推測できるだけです。R の `cor.test` は 95%の信頼区間も計算してくれますが，このような情報も加味しながら母集団の様相を推測する必要があります。

ちなみに，母集団の相関係数と，無相関検定はまったく関係がありません。無相関検定は，サンプルの相関係数が 0 か否かを問題にしているだけですから。