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？　（3）

今回は，母集団の相関が異なる時，サンプルの相関がどのように影響を受けるかを見てみたいと思います。たとえば，母集団での相関が .80 であった時と，.20 であった時でサンプルの相関係数はどのように違ってくるかということです。

？？と思われるかもしれませんが，実は母集団における相関の高低で，サンプルの相関係数の信頼区間は変わってくるのです。

今回もとりあえず 10000 人の母集団を考えます。そして，前回の確認で「これくらいかな」というサンプル数だった 200 人を採用します。母集団の相関係数を，0，.2，.4，.6，.8 と変えてみると，サンプルにおける相関係数はどのように変化するかを確認してみましょう。

前回も使ったデータの発生部分を使い，matrix の部分を変えることで簡単にシミュレーションができます。やってみてください。

```
library(MASS)
x <- matrix(c(1, 0.57, 0.57, 1), ncol=2)          #←ここを変える
data <- mvrnorm(n= 10000, mu= c(0, 0), Sigma= x, empirical= TRUE)
d.data <- data.frame(data)
colnames(d.data) <- c("好き", "積極関与")

box0 <- rep(NA, 1000)
box <- matrix(box0, ncol=1)
qq1 <- c(1:10000)
for(m in 1:1000)  {
        qq2 <- sample(qq1, 200, replace = FALSE)   #←ここは 200 で
        sub.d.data <- subset(d.data[qq2, ])
        box[m,1] <- cor(sub.d.data$好き, sub.d.data$積極関与)
        }
```

やってみて，主な指標をまとめたら以下のようになりました。レンジは .40 の時が一番大きくなっていますが，標準偏差は徐々に小さくなっています。最大値や最小値を見てみると，おそらく 1000 回のサンプリングの中に，かなり小さい相関が得られたケースがあるためではないかと思われます。

母集団の相関	平均	標準偏差	最小値	最大値	レンジ
.00	0.01	0.07	−0.18	0.21	0.39
.20	0.20	0.07	0.00	0.39	0.39
.40	0.40	0.06	0.12	0.57	0.44
.60	0.60	0.05	0.46	0.72	0.26
.80	0.80	0.03	0.71	0.86	0.16

これを見ると，母集団の相関が高いほど，サンプルの相関のブレが小さくなる傾向があることがわかるでしょう。つまり，母集団における相関が高い場合には，サンプルの相関もそれに比較的近い値にあることが多いのですが，母集団における相関が低いと，サンプルの相関はバラつきやすくなります。

もちろん母集団における相関はわからないので，これでは一概に 200 人くらいのデータがあればなんとか…という判断ができません。しかし，サンプル数が多くなればそのバラつき具合が小さくなるという性質もあります。

ということは，仮説を立てた時点で母集団における相関を予測し，もしそれほど高くないだろうと考えられるのならば多めのサンプル数を確保することが必要になるでしょう。

さて，こちらについても　母集団の相関をいちいち指定するのではなく，一気にやってみたいと思います。

母集団の相関を 0 から 0.025 ずつ増やして，1 まで順にやります。前回のスクリプトをもとにして，少し修正してやればできますので考えてみてください。なお，こちらも計算にちょっと時間がかかります。

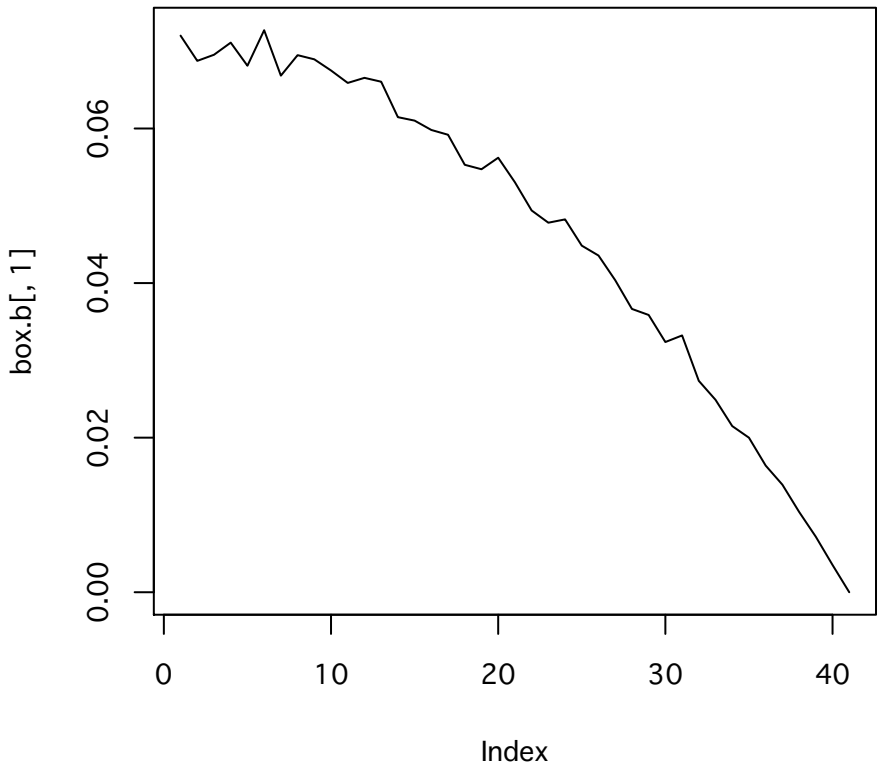
```
rr <-  c(0:40)/40
box0 <- rep(NA, 1000)
box.a <- matrix(box0, ncol=1)
box00 <- rep(NA, 164)
box.b <- matrix(box00, ncol=4)
qq1 <- c(1:10000)
for(r in rr) {
        x <- matrix(c(1, r, r, 1), ncol=2)
        data <- mvrnorm(n= 10000, mu= c(0, 0), Sigma= x, empirical= FALSE)
        d.data <- data.frame(data)
        colnames(d.data) <- c("好き", "積極関与")
        for(m in 1:1000)  {
                qq2 <- sample(qq1, 200, replace = FALSE)
```

```
            sub.d.data <- subset(d.data[qq2, ])
            box.a[m,1] <- cor(sub.d.data$好き, sub.d.data$積極関与)
            }
        box.b[1+40*r,1] <- describe(box.a)$sd
        box.b[1+40*r,2] <- describe(box.a)$min
        box.b[1+40*r,3] <- describe(box.a)$max
        box.b[1+40*r,4] <- describe(box.a)$range
        }
plot(box.b[,1], type="l")
plot(box.b[,2], type="l")
plot(box.b[,3], type="l")
plot(box.b[,4], type="l")
```

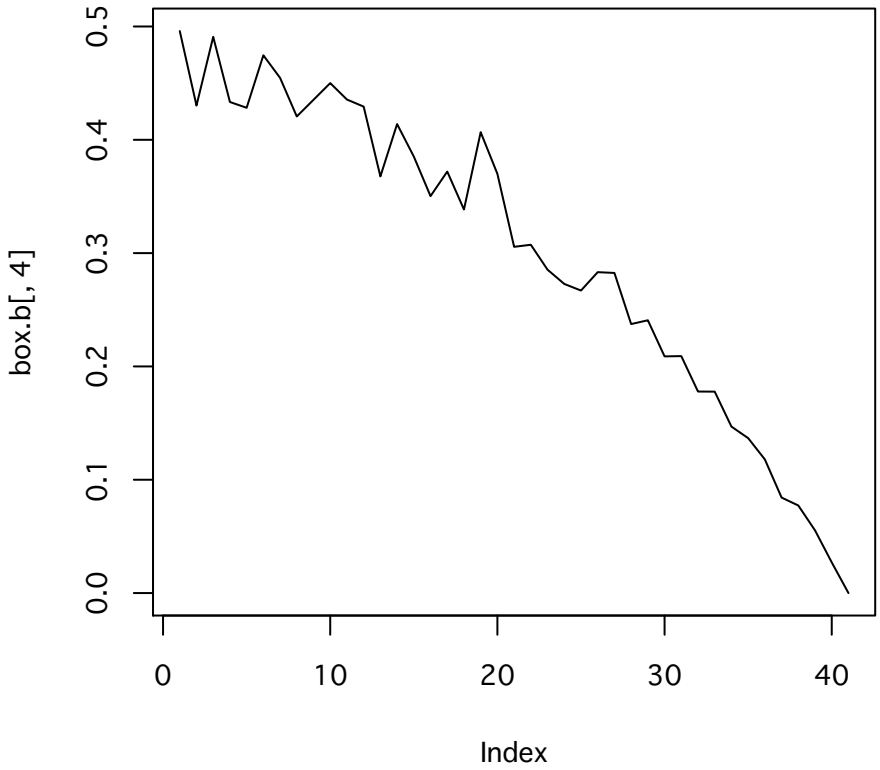
今回は matrix の部分を変えます。１行目で，0，0.025，0.050，0.075…1 という数字の列 rr を作成しています。そして for(r in rr)で，それを順にやらせています。また結果を入れる box.b は 41 行分必要です。このあたりがちょっと違うところでしょうか。

４つのグラフのうちの標準偏差とレンジの様子を以下に示しておきます。

まず，標準偏差の変化の様子です。横軸は母集団の相関係数です。読みにくいのですが 10 が .25，20 が .50，最後の 41 が 1.00 になります。



こちらがレンジの変化の様子です。



これを見ると，母集団における相関係数とサンプルの相関係数のブレは直線的ではないことがわかります。母集団における相関係数が低い場合は，変化量は少ないですが，高くなるとブレが小さくなるようです。母集団における相関係数が.40 くらいより小さい場合には，ブレが気になる感じがします。