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？　（2）

サンプル数が増えれば増えるほど，有意な結果が出やすいということは知っていると思います。では，サンプルが多ければ多いほど，母集団の相関がより正確に推定できるでしょうか？　今回はこれを確認してみます。

データは前回と同じとしておきましょう。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。

抽出する人数を変えながら，サンプルにおける相関係数と母集団の相関の関係を見てみます。

前回と同様にして，まずは母集団のデータを作成します。

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

とりあえず，両者の間の相関係数は .57 になっているか，サンプル数が 10000 であるかを確認しておきましょう。

では，この母集団からｎ名をランダムに取り出したデータを作成し，得られる相関係数の最大値や最小値，ヒストグラムの様子をみてみます。やり方は前回のものをそのまま使って，サンプリングの行の 74 を任意の数字に変えてやれば OK です。

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

何だか間違いそうだという場合は，最初に抽出する n を指定するように変更することもできます。たとえば

```
n <- 20                                                    #←ここで n を指定
box0 <- rep(NA, 1000)
box <- matrix(box0, ncol=1)
qq1 <- c(1:10000)
for(m in 1:1000) {
        qq2 <- sample(qq1, n, replace = FALSE)     #←ここを n に
        sub.d.data <- subset(d.data[qq2, ])
        box[m,1] <- cor(sub.d.data$好き, sub.d.data$積極関与)
        }
```

n の数はいろいろと試してみて欲しいのですが，たとえば，25，50，75，100，150，200，300，500，1000 くらいをやってみましょう。

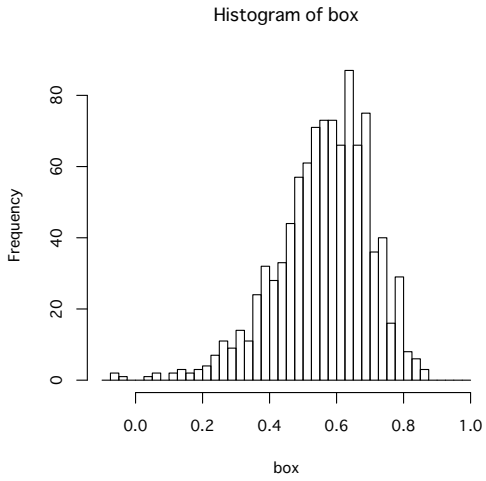
試しにやってみて，主な指標をまとめたら以下のようになりました。ヒストグラムもあわせて示しておきます。

サンプル数	平均	標準偏差	最小値	最大値	レンジ
25	0.56	0.14	−0.07	0.87	0.94
50	0.57	0.10	0.24	0.82	0.58
75	0.57	0.08	0.28	0.78	0.50
100	0.57	0.07	0.33	0.73	0.40
150	0.57	0.06	0.30	0.71	0.41
200	0.57	0.05	0.41	0.70	0.29
300	0.57	0.04	0.42	0.68	0.26
500	0.57	0.03	0.47	0.65	0.18
1000	0.57	0.02	0.51	0.64	0.13

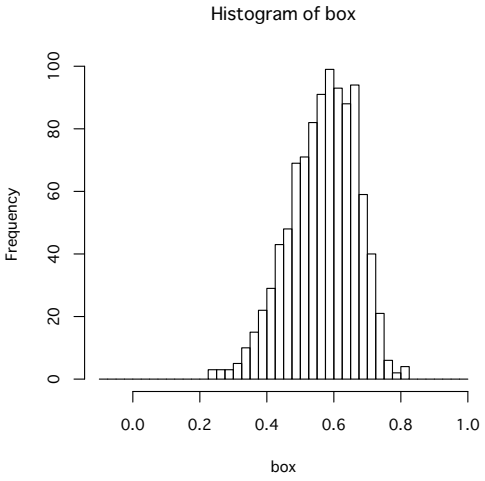
サンプル数が 25 だと，なんと負の相関まで出てきました。レンジも 0.94 とやたらと大きいです。確かにその平均は母集団の平均とほぼ一致するのですが，これでは１回の調査から母集団について話をするのは無理と言わざるを得ません。

サンプル数が増えれば，徐々にレンジも小さくなり，母集団の相関の周りに集中してくることがわかります。ヒストグラムもあわせて考えると，やはり 200 あたりは欲しいなという感じがします。調査においては，ちょっとがんばれば現実的な数字ではないでしょうか。これくらいのサンプル数があれば，そこで得られた相関係数はかなり母集団の値に近いと言えるでしょう。

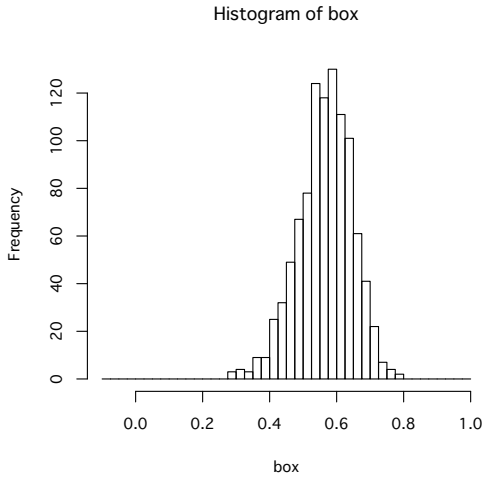
n=25



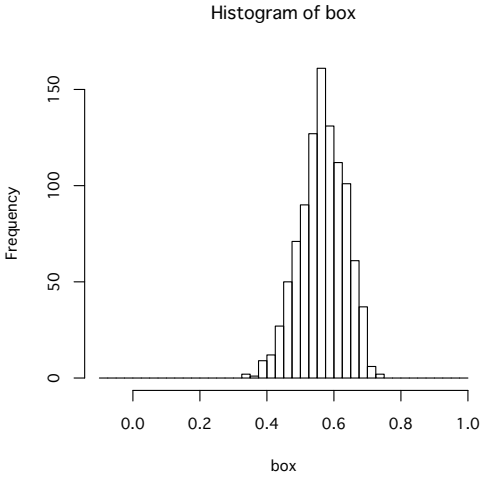
n=50



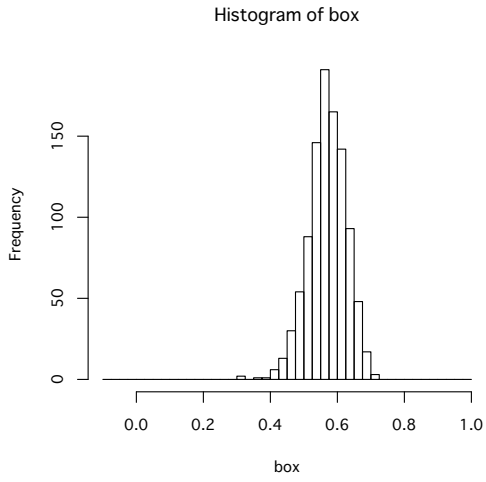
n=75



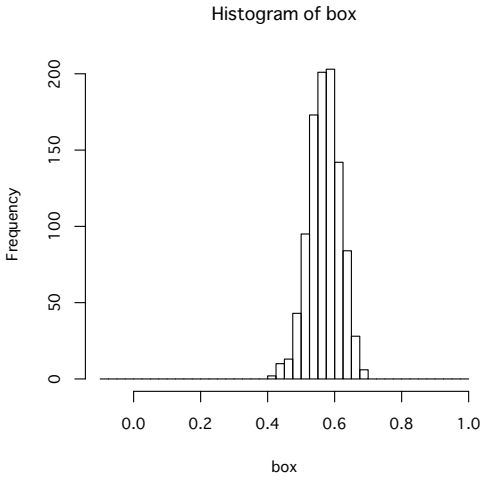
n=100



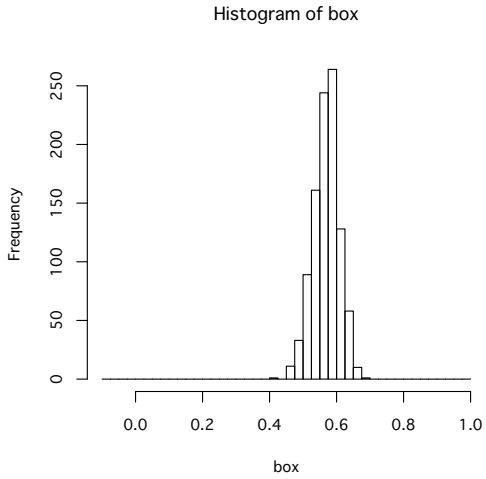
n=150



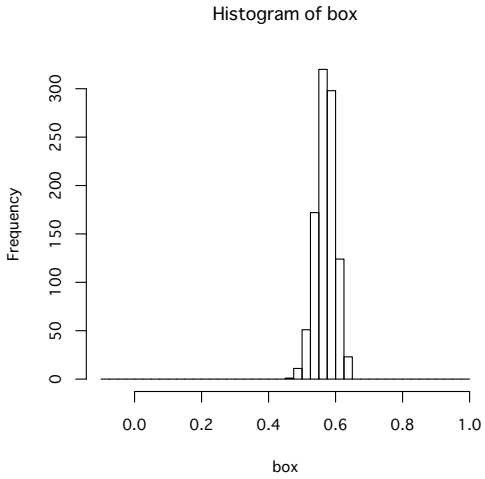
n=200



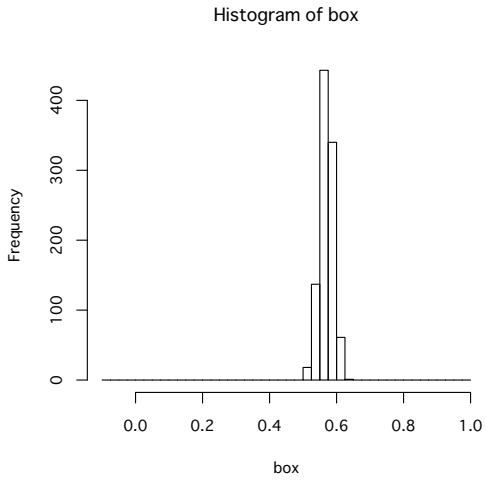
n=300



n=500



n=1000



なお，今回はヒストグラムの横軸をすべて同じにしてあります。これは以下のようにすればできます。

```
hist(box, breaks=seq(-0.1,1,0.025))
```

seq のカッコの中は，最小，最大，間隔の３つを指定します。もし指定した幅を越える値があった場合，

'x' の一部分が数えられていません。多分 'breaks' が 'x' の範囲全体をカバーしていません

というようなエラーが出ます。その場合は，最大，最小の値を見直してください。

さて，以上のようにnをいちいち指定してやるのもよいのですが，一気にやってみたいとも思います。これが自動的にできれば，もっと詳細に検討できるはずです。ですので，少し工夫してみました。

nを10から10ずつ増やして，1000まで順にやります。ひとつのnについてサンプリングを1000回，nを10，20，30…と増やして1000まで100回やります。ちょっと時間はかかりますがやってくれます。

```
x <-  c(1:100)*10
box0 <- rep(NA, 1000)
box.a <- matrix(box0, ncol=1)
box00 <- rep(NA, 400)
box.b <- matrix(box00, ncol=4)
qq1 <- c(1:10000)
for(n in x) {
       for(m in 1:1000)  {
              qq2 <- sample(qq1, n, 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[n/10,1] <- describe(box.a)$sd
       box.b[n/10,2] <- describe(box.a)$min
       box.b[n/10,3] <- describe(box.a)$max
       box.b[n/10,4] <- describe(box.a)$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")
```

何をやっているかは，これを読み解いていただければよいのですが，for のカッコ内には数字の列（ベクトル）も指定できます。１行目で，10，20，30…1000という数字の列xを作成しています。そしてfor(n in x)で，それを順にやらせます。

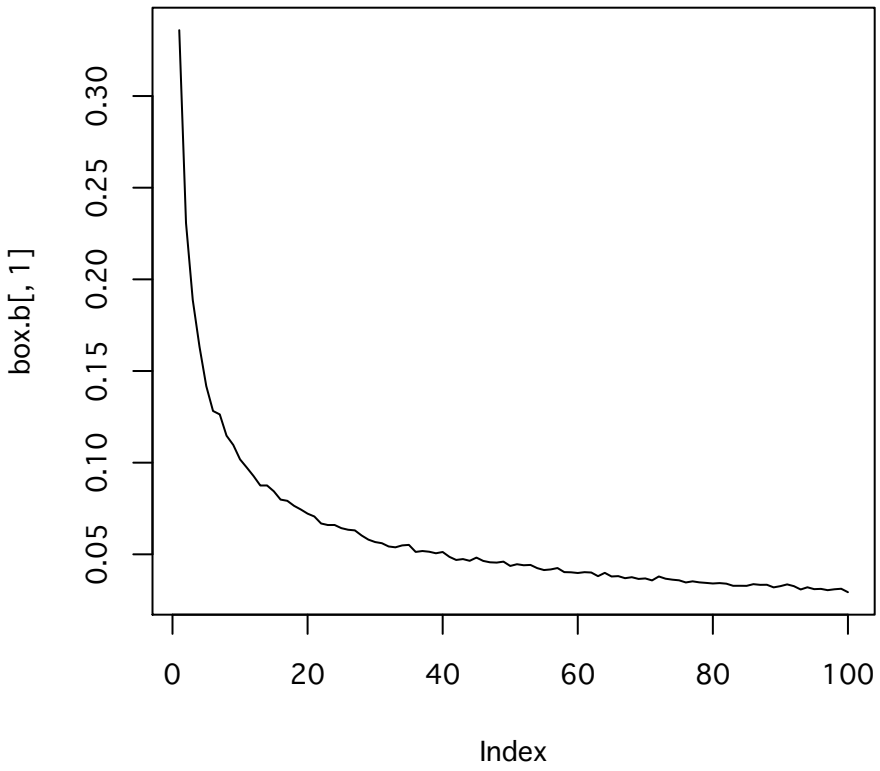
結果を入れるboxは２つ用意しています。それぞれのnで1000回のサンプリングの結果を入れるbox.aと，100種類のnについての結果を入れるbox.bです。box.bにはdescribe

で計算した結果の標準偏差（sd），最小値（min），最大値(max)，レンジ(range)を入れています。

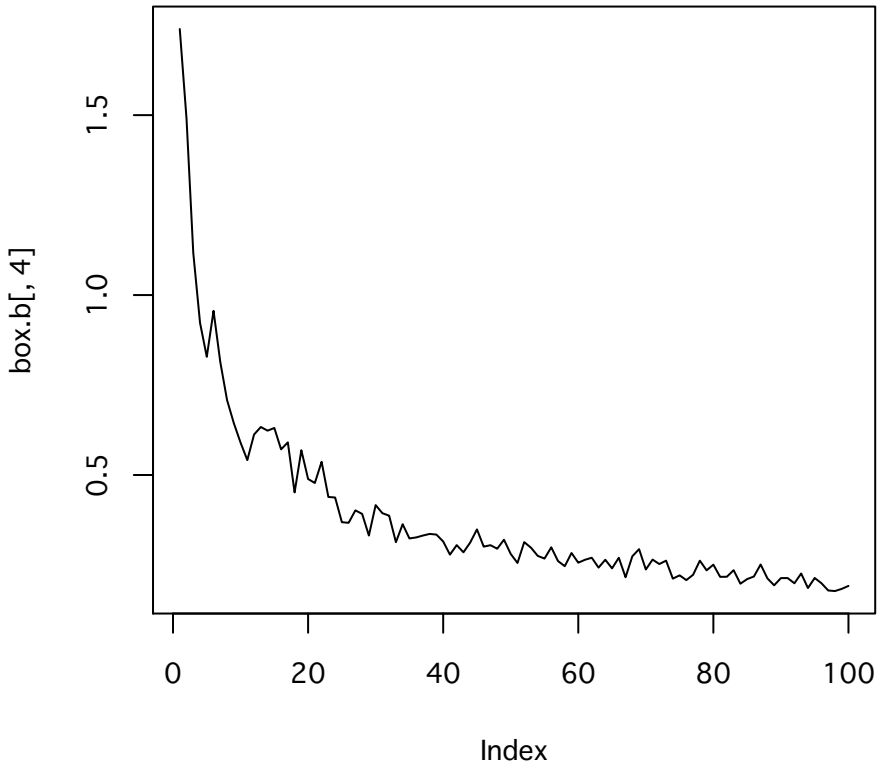
その 100 回分の結果を plot でグラフにします。なお，type="l"は折れ線プロットをさせるためのものです。他にもプロットの種類はありますので，必要なら調べてください。

４つのグラフを描かせるのですが,そのうちの標準偏差とレンジを以下に示しておきます。

次は，標準偏差の変化の様子です。横軸はサンプル数ですが 10 倍します。



こちらがレンジの変化の様子です。



これを見ると， 300 を越えてくるとあまり大差がないようです。やはり 200 程度を目指すべきなのでしょうか…。サンプル数を決める参考になるのではないでしょうか。