10 日目：相関係数（1）

調査系の研究なら，データを集めて，項目別のチェックを済ませたなら，次は尺度の構成，確認に進むことになるでしょう。ここからは，因子分析をしたり，α係数を求めたりして，尺度を確定していきます。

今日は，因子分析に入る一歩手前として，項目間の相関係数を眺めてみたいと思います。

その前に，今回からはサンプルのデータを変更します。ここからしばらくは，ある大学で，リンゴ印の某 PC に関するイメージ調査を行い，性や学年，専攻による差があるのかどうかを分析してみるという順で進んでみたいと思います。そのために架空のデータを作りました。結構面白いものになったと思いますが…

データファイルの方にも記してありますが，変数は，ID，性別，学年，専攻，イメージ項目（b1 から b20）です。今回は「性別」「学年」「専攻」は日本語の変数名にしています。注意してください。２シート目に簡単な説明を入れておきました。

データを読み込んだら，まずは項目のチェックをやってみてください。欠損値も含まれますが，どの変数にいくつの欠損値があるのかも確認しておいてください。

それが終わり，特に各項目に問題はないと判断したとしましょう（正直なところ，ヤバそうな項目は結構ありますが，今回は練習として大目に見てください（苦笑）。もちろん，ヤバそうなものが因子分析の際にどういう影響を及ぼすのかを確認してみるのも一興かと…）。続いて，イメージについて因子分析を行いたいので，その前に項目間相関を見てみます。なお，因子分析の前に項目間相関を眺めてみることは，特に初心者には大事なことだと思います。どのような項目間相関を示すデータ（全体的に，項目間相関が高い／低いのか。項目間相関を見ただけである程度のまとまりが認められるほど，相関の強弱が明確なのか／不明確なのか。他の項目とほとんど相関の無い項目があるのか／無いのか。などなど）を因子分析にかけているのかというイメージを作っておくとよいと思います。

さて，相関係数を計算する命令は単純，cor(x)です。もうわざわざ書かなくても大丈夫だと思いますが，x という名前でデータを読み込んでおいた場合です。

◎ ファイル x にあるすべての変数間の相関を求める

```
cor(x)
```

◎ x の b10，b12 間相関係数を求める

```
cor(x$b10, x$b12)
```

◎ x の 5 行目から 7 行目の変数間の相関係数を求める

```
cor(x[5:7])
```

サンプルデータを使って，cor(x[5:7])を実行してみると以下のような結果が出力されます。

```
> cor(x[5:7])
   b1          b2          b3
b1  1          NA          NA
b2 NA 1.00000000 0.02090073
b3 NA 0.02090073 1.00000000
>
```

b1 に欠損値が含まれるのは確認済みだと思いますが，この簡単な命令だと欠損値がある変数を使って計算した場合は「NA」が返されます。そのため，以下のように命令を加えます。

◎ 5 列目から 7 列目のデータに欠損値を含むサンプルを最初からすべて取り除いてから計算（リストワイズ削除）

```
cor(x[5:7], use="complete.obs")
```

◎ 相関係数を求める 2 変数の組合せごとに欠損値を含むサンプルを取り除いて計算（ペアワイズ削除）

```
cor(x[5:7], use="pairwise.complete.obs")
```

これまたやってみて結果を比較すると一目瞭然なのですが，下の図のように相関係数は若干異なります。どちらがいいかはケース・バイ・ケースですが，レポートや論文に仕上げる場合，後者（いわゆるペアワイズで除く場合）を使ったなら，それぞれで n が変わるので，その数を添えなければなりません。

```
> cor(x[5:7], use="complete.obs")
          b1        b2        b3
b1 1.0000000 0.2934777 0.0947988
b2 0.2934777 1.0000000 0.0240964
b3 0.0947988 0.0240964 1.0000000
> cor(x[5:7], use="pairwise.complete.obs")
          b1         b2         b3
b1 1.0000000 0.29347767 0.09479880
b2 0.2934777 1.00000000 0.02090073
b3 0.0947988 0.02090073 1.00000000
```

この cor(x)は，何も指定しなければピアソンの積率相関係数を計算してくれます。ヘルプをみれば，cor(x, method="pearson")という形式であり，method="pearson"がデフォルト（method=を省略した場合，これが使われる）になっていることがわかります。他

には，"kendall"と"spearman"が使えます。

さて，今回は b1 から b20 までの変数の相関係数を求めるのですが，列番号を使って，cor(x[5:24], use="complete.obs")とやっても問題はありません。しかし以後のことも考えると，６日目にやった，「変数を抽出して，新しい名前でまとめておく」という作業をし，そのファイル名で分析対象を指定するというやり方をしておく方が便利な気がします。つまり…

```
label_b <-
 c("b1","b2","b3","b4","b5","b6","b7","b8","b9","b10","b11",
"b12","b13","b14","b15","b16","b17","b18","b19","b20")
xb <- x[label_b]
```

それから…

```
cor(xb, use="complete.obs")
```

これでｂの 20 項目間の相関マトリックスが出力されます。ちなみに，デフォルトだと小数点以下が多く表示されます。この後エクセルに移して整理してから眺めるなら，このままでも問題はないのですが，丸めておきたければ…

```
round(cor(xb, use="complete.obs"),3)
```

これで小数点以下３桁で表示してくれます。

結果が出たら，エクセルにコピペして整理し，全体を眺めてみてください。先にも触れたように，ポイントは，全体的に項目間相関が高い／低いのか。項目間相関を見ただけである程度のまとまりが認められるほど，相関の強弱が明確なのか／不明確なのか。他の項目とほとんど相関の無い項目があるのか／無いのか，などです。特徴をメモしておくと，実際に因子分析をした際に確認ができるかもしれません。

これで 10 日目は終了です。明日は，いよいよ因子分析をやります。