■ 相関のノンパラメトリックなど　◆ 北野 利一 ◆　2005 年 11月 29日

● いわゆる相関係数とよばれるものは，ピアソンの積率相関係数である．これに対し，２つの順位相関係数が知られている．

１）スピアマンの順位相関係数　～　両者の順位を対象に，ピアソンの積率相関係数を算出するもの．

２）ケンドールの順位相関係数　～　２変量の順位についての全ての組み合わせの内，一方の順位に対して，他方が大きい場合と，その逆の場合に分け，それらの個数の差を全ての組み合わせの数で除したもの．したがって，混乱の度合いをみている（完全な秩序のある場合に，完全相関．無秩序は無相関）．

例１： NHK 放送世論調査所による全国県民意識調査（1978年）での「お好きな花の順番」は，

```
male   <- c("Cherry", "Chrysanth", "Rose", "Plum", "Lily", "Tulip", "Carnation", "Camellia")
female <- c("Chrysanth", "Rose", "Cherry", "Lily", "Plum", "Carnation", "Tulip", "Camellia")
```

であった（東京大学教養学部統計学教室編，1991，pp.55）．この結果を以下のように表にすることができる．

```
nhk <- data.frame(matrix(c(1:8, match(male, female)), nc=2, dimnames=list(male, c("M","F"))))
```

```
> t(nhk)
  Cherry Chrysanth Rose Plum Lily Tulip Carnation Camellia
M      1         2    3    4    5     6         7        8
F      3         1    2    5    4     7         6        8
```

この時，順位相関係数は次のように計算できる（手計算による確認も行うこと）．

```
cor.test(nhk$M, nhk$F, meth="spearman") # = cor.test(nhk$M, nhk$F, meth="pearson") in this case
cor.test(nhk$M, nhk$F, meth="kendall")
```

順位相関係数は，いずれも定義が異なるので，数値にあまり意味を考えても仕方がない．
むしろ，無相関かどうかの議論に意味があると考えよ（そのためか，R にて，検定用の cor.test により，順位相関として指定することが可能であるが，単なる算出用の cor は積率相関のみで，順位相関を指定することはできない．これは，個人的な憶測だろうか？ R の設計思想の一端がみえる好例といえる）
また，スピアマンとケンドールのいずれか一方に優劣があるというものではない．視点が異なる．

順位相関係数は，上述のようなカテゴリカルデータのみに適用を限定されるわけではない．むしろ，量的データに適用されるべきものである．どのようなものであれ，量的データであれば，積率相関係数は形式的に算出は可能である．しかし，意味があるかどうかは別の問題である．つまり，積率相関係数の標本値により，無相関を検定するためには，２次元正規分布を母分布とした標本であることが前提となる．したがって，そのような前提に無理があるときは，順位相関係数により無相関性を検定しなければならない．

例２：英国 Dover 港と Harwich 港における年最大潮位記録（Stephenson，2004）

```
library(evd); data(sealevel)
sealevel[complete.cases(sealevel),] -> sl
```

```
op <- par(mfrow=c(1,2)); qqnorm(sl$dov); qqline(sl$dov, col="red")
qqnorm(sl$har); qqline(sl$har, col="red"); par(op)
```

いずれも，正規分布に従う変動とはいえないことが確認できる（資料の前提から明らかであるが）．

```
plot(sl); abline(v=quantile(sl$dov, exp(-1)), h=quantile(sl$har, exp(-1)), lty=2)
cor.test(sl$dov, sl$har, meth="s")
```

２次元極値分布に対して，Pickands dependence 関数を用いて，スピアマンおよびケンドールの順位相関係数と一般的に結びつける関係式が知られている（Beirlant et al., 2004 および別紙）．

例３：相関のある２元の極値分布に従う乱数の発生

```
rbvevd(1000, dep=.3, model="log") -> xy # do it again, changing the value of dep
cor.test(xy[,1], xy[,2], meth="sp")

hurlimann.log <- function(w, alpha) 1/(1 + (w^(1/alpha) + (1-w)^(1/alpha))^alpha)^2
12*integrate(hurlimann.log, 0,1, alpha=.3)$value - 3
```

● 相関という概念に対して，注意すべきことが幾つかある（以下２つは，回帰の問題にも同様にいえる）．

◆ みかけ上の相関　・・・＞　偏相関係数　（で，解決！）

例４：徒競走で負ければ負けるほど，収入は多くなる？（永田，1996）

```
mikake <- data.frame(footrace=c(7.7, 8.2, 8.5, 7.8, 8.0, 7.8, 7.7, 8.2, 8.5, 8.1,
                                8.4, 7.7, 7.9, 8.3, 8.2, 7.9, 7.8, 8.4, 7.8, 7.7),
                     income=c(342, 923, 985, 581, 627, 388, 290, 860, 787, 654,
                              788, 334, 412, 915, 648, 761, 589, 946, 477, 412))

> cor(mikake); plot(mikake)
          footrace    income
footrace 1.0000000 0.8780669
income   0.8780669 1.0000000
```

上のデータは，２０歳から５０歳代の男性会社員に対して，５０ｍ走の記録（footrace）と年収（income）を調べたものである．相関係数は，0.88 と非常に高い！給料と徒競走の能力に何か秘められた関係があるのだろうか？（まさか，これをまともに信じる人はいまいが，．．．）

もちろん，給料と徒競走の記録の背後に，年齢（age）という第三の変数が隠されていることに，この場合は容易に気づく．

```
mikake <- cbind(mikake,
          data.frame(age=c(23, 43, 50, 35, 33, 25, 20, 44, 48, 37,
                           39, 22, 29, 46, 43, 33, 30, 47, 28, 25)))

cor(mikake); pairs(mikake) # OK = plot(mikake)
```

footrace と age，および income と age の相関係数はいずれも非常に高い！footrace と income には，本来，相関関係が無いにも関わらず，いずれも age に操られているために，“みかけ上の相関”が現れたという次第である．そこで，いずれも age に操られない本来の footrace と income の相関を知りたいが，それには，次式で表される偏相関係数 r.xy.z を用いる．

```
r.xy.z <- function(r.xy, r.xz, r.yz) (r.xy - r.xz * r.yz)/sqrt(1 - r.xz^2)/sqrt(1 - r.yz^2)
```

ここで，footrace，income および age の３者の互いの相関係数について，

```
cor.xy <- cor(mikake$footrace, mikake$income)
cor.xz <- cor(mikake$footrace, mikake$age)
cor.yz <- cor(mikake$income,   mikake$age)
```

と置くことにより，給料と徒競走の記録の相関は，以下のとおり非常に小さいことがわかる．

```
> r.xy.z(r.xy=cor.xy, r.xz=cor.xz, r.yz=cor.yz)
[1] -0.05382888
```

ところで，上述の R での計算は美しくない．偏相関係数の公式の導出（別紙参照）を考えれば，より自然に計算が可能である（偏相関係数だけを計算したい，という状況は通常あり得ない）．

```
par(mfrow=c(1,2)) -> opar
plot(mikake[,c("age", "footrace")]);
lm(footrace ~ age, data=mikake) -> xz.lm; abline(xz.lm, col="red")
segments(mikake$age, mikake$footrace, mikake$age, predict(xz.lm), lty=3)
plot(mikake[,c("age", "income")]);
lm(income   ~ age, data=mikake) -> yz.lm; abline(yz.lm, col="red")
segments(mikake$age, mikake$income,   mikake$age, predict(yz.lm), lty=3)
par(opar)

> cor(xz.lm$residuals, yz.lm$residuals)
[1] -0.05382888

plot(xz.lm$residuals, yz.lm$residuals); abline(h=0, v=0, lty=3)
```

また，Crawley (2002) では，anova を用いた議論をしている（pp.191-192）．

（宿題）背後に 1 変量が隠れた偏相関係数の式の導出を参考に，背後に 2 変量が隠れている場合の偏相関係数の公式を導け．

◇ ケンドールの順位相関係数について，“混乱の度合い”としての意味を保ったまま，偏相関を定義できる．結果として，ピアソンの積率相関係数に対する偏相関の公式と同じ形式となることは興味深い．詳しくは，Kendall (1948) p.103 および Siegel (1956) を参照せよ．

◆ 層別・合併の影響（Scale-dependent correlations）これも，みかけの問題といえる．

例 5 ：植物の生産性と種の多様性の関係（Crawley，pp.188-190）

```
# access:  http://www.bio.ic.ac.uk/research/mjcraw/statcomp/
read.table("MacOSX3:Users:tk:rm171:crawley1:productivity.txt", header=T) -> veg
plot(veg[,-3], xlab="Productivity", ylab="Species Richness")

cor.test(veg[[1]], veg[[2]], meth="k") # cf. meth="p" and "s"
```

生産性と種の多様性の散布図，および相関係数をみれば，生産性の向上に併せて，種の多様性が増加すること，すなわち，正の相関性が示される．しかし，植生域ごとに層別にみると，．．．

```
points(veg[,-3], pch=as.character(veg[,3]), col=palette()[codes(veg[,3])])
```

明らかに，負の相関性がみられる．このような可視化機能は，R/Splus には多く用意されている．

```
coplot(y ~ x | f, data=veg) # check names(veg)
xyplot(y ~ x | f, data=veg) # library(lattice) required

regr <- function(x,y) {panel.xyplot(x,y)
                       panel.abline(lm(y ~ x))}
xyplot(y ~ x | f, data=veg, panel=regr)
```

その他の相関に関する注意点は，佐伯・松原 (2000) 1.3.3 pp.28-32 を参照せよ．

● 相関は，2 変量（自己と他者）への適用に限らない．1 変量（自分と自分）にも適用可能である．すなわち，時系列データへ適用した自己相関関数（もはや相関“係数”ではなくて，“関数”）である．

例題６：ある女性の血液中のルテイン化ホルモンの１０分間隔での計測値（Diggle, 1990; MASS）

```
data(lh)
embed(lh, 5) -> lhb
pairs(lhb)
pairs(lhb, panel=function(x,y) {points(x,y)
                          abline(lm(y~x), col="blue")},
                  diag.panel=function(x) {par(new=T)
                                hist(x, main="", axes=F, prob=T)
                         lines(density(x))})
library(lattice)
splom(~lhb, panel=function(x,y) {panel.xyplot(x,y)
                           panel.abline(lm(y~x))})

par(mfrow=c(2,1)) -> op; acf(lh); pacf(lh); par(op)

ar(lh)
arima(lh, order = c(1,0,0))
arima(lh, order = c(3,0,0))
arima(lh, order = c(1,0,1))
```

● AR(p) モデルとは，ホワイトノイズ項 Z を用いて，以下に表される自己回帰モデルである．

```
Xt = phi_1 * Xt-1 + ... + phi_p * Xt-p + Zt
```

● MA(q) モデルとは，ホワイトノイズ項 n を用いて，以下に表される移動平均モデルである．

```
Xt = Zt + theta_1 * Zt-1 + ... + theta_q * Zt-q
```

自己回帰モデル AR(1) および移動平均モデル MA(1) の自己相関関数は以下のように得られる．

```
AR(1): acf = phi^abs(h)                              for Xt = phi * Xt-1 + Zt
MA(1): acf = 1                    at h = 0
           = theta/(1 + theta^2) at h = 1
           = 0                    at h > 1 or h < -1  for Xt = Zt + theta * Zt-1
```

なお，MA(q) の自己相関関数は，ラグ q+1 以降で，ゼロとなること，
他方，AR(p) の偏自己相関関数は，ラグ p+1 以降で，ゼロとなることを示すことができる．

● ARMA(p,q) モデルとは，以下に表される自己回帰移動平均モデルである．

```
Xt - phi_1 * Xt-1 - ... - phi_p * Xt-p = Zt + theta_1 * Zt-1 + ... + theta_q * Zt-q
```

例題７：自己回帰移動平均モデルによる時系列のシミュレーション

```
(arima.sim(n = 63, list(ar =0.4522, ma=0.1982, sd=sqrt(0.1923))) -> temp.waves)
plot(temp.waves)
acf(temp.waves)
acf(temp.waves, type="p")
```

◇ 時系列モデルに関しては，様々な視点からの検討と，多くの知識が必要となる．詳しくは別の機会に．

● 連検定（ラン検定と読む）
２つの事象，ＡあるいはＢのうち，いずれかをとる時系列について，事象の連続する長さをランとよぶ．

Ａ Ａ Ｂ Ｂ Ｂ Ｂ Ａ Ｂ Ｂ Ａ Ａ Ｂ Ａ Ｂ Ｂ Ｂ Ａ Ａ Ａ Ａ Ｂ Ａ Ｂ Ｂ Ａ Ｂ Ｂ Ａ　．．．

という時系列に対して，ランは２，４，１，２，２，１，１，３，４，１，１，２，１，２，１，．．．と

なる．このランの個数（上の例では，．．．までにランは１５個）は，Ａの個数をＮ（例では，１３），Ｂの個数Ｍ（例では，１５）により，

平均　＝　２ＮＭ／（Ｎ＋Ｍ）＋１
分散　＝　２ＮＭ（２ＮＭ－Ｎ－Ｍ）／（Ｎ＋Ｍ）＾２／（Ｎ＋Ｍ－１）

の正規分布に従うことが知られている．

例８：円周率の各位の数字について，偶奇がランダムであるか？（松原，1996，pp.125）

```
UNIX> math
 In[1]:= N[Pi, 200]
Out[1]=  3.14159265358979323846264338327950288419716939937510582097494459230781640628620899862803482534211706 79\
           82148086513282306647093844609550582231725359408128481117450284102701938521105559644622948954930382 0\
           22317253594081284811174502841027019385211055596446229489549303820
```

上記は，Mathematica（Wolfram Research）により，円周率を２００桁まで計算したものである．上記の値を以下のようにして，Ｒ に取り込む．

```
as.numeric(unlist(strsplit(paste(
"3.14159265358979323846264338327950288419716939937510582097494459230781640628620899862803482534211706 79",
  "82148086513282306647093844609550582231725359408128481117450284102701938521105559644622948954930382 0",
sep="", collapse=""), split=""))[-2]) -> pi200

pi200cr <- (pi200 %% 2) * 2 - 1
pi200pos <- which((pi200cr[-1] * pi200cr[-200]) < 0)
res.pi <- list(N.even=sum(pi200cr < 0), N.odd=sum(pi200cr > 0), run=c(pi200pos, 200) - c(0, pi200pos))

sum(res.pi$run) # = 200; check!

zscore <- (length(res.pi$run) - (1 + 2*res.pi$N.even*res.pi$N.odd/( res.pi$N.even + res.pi$N.odd)))/
                               sqrt(2*res.pi$N.even*res.pi$N.odd/( res.pi$N.even + res.pi$N.odd)^2*(
                                    2*res.pi$N.even*res.pi$N.odd - res.pi$N.even - res.pi$N.odd)/
                                                                  ( res.pi$N.even + res.pi$N.odd - 1))

pnorm(zscore)*2 # test for both sides
```

（練習）pi2000 （in package: UsingR, Verzani）を用いて，例８と同様のことを試みよ．さらに，＜５ と ＞４の生起のランダム性について検定せよ．また，各位の数字の分布について，一様分布とのカイ自乗適合度検定を行い，ラン検定とカイ自乗適合度検定によるランダム性に対する議論の違いを説明せよ．

▲ トレンドの検定（相関係数の応用）

相関係数を時系列データに対するトレンドの検定に用いることも可能である．

例９：ある地方の冬期の気温が上昇傾向にあるのではないか？という検討（柳川，1981）．
Bhattacharya と Klots は，その地方の湖が毎年11月23日から数えて何日目に凍結したかを表すデータを用いた．以下は，1854年からの１１０年間の１０年単位の平均日数である．

```
frozen <- c(18.0, 19.0, 24.5, 33.5, 26.0, 30.0, 29.0, 23.0, 31.7, 30.5, 22.5)
fyear <- seq(1854, 1954, by=10)
paste(fyear, fyear+9, sep="-") -> names(frozen); frozen
matrix(c(frozen, rank(frozen)), nc=2, dimnames=list(names(frozen), c("mTemp", "Rank")))

plot(frozen, type="S", xlim=c(0,11), axes=F, xlab="year"); axis(2)
lines(0:1, rep(frozen[1],2)); axis(1, 0:11, c(fyear, 1964))

cor.test(1:length(frozen), frozen, meth="k", alt="g")
```

（応用） 各地の有感地震回数（1961-1981年）は，以下のとおりである（渡部ら，1985，表 1.2，p.8）．これらの時系列について，上述の概念や手法を検討せよ（例えば，自己相関関数や偏自己相関関数などを調べ，モデル化することや，トレンドの検定，各地点での相関なども考慮すれば，多変量自己回帰移動平均モデルへ拡張できる）．

```
cat(paste("          ", paste("y", 61:81, sep="", collapse=" ")),                          "\n",
"wakayama     71  74  77  67  71  52  67  71  50  21  32  20  37  32  25  33  97  44  39  54  37\n",
"nemuro       72  54  45  51  37  31  33  59 105  51  29  28 227  42  60  45  42  66  51  63  61\n",
"osaka         5   4   5   2   3   8   2  10   1   2   3   4   5   6   5   1   0   4   5   2   2\n",
"kyoto        14  12  10   7  11  15  12  33   9   8  11   8  11   8   4   5   4   9   6   2   3\n",
"obihiro      18  22  14  15   9  12  16  47  23  26  19  17  52  28  20   8  13  25  15  10  16\n",
"mito         61  48  43  65  68  64  56  52  54  52  61  74  81  66  53  54  67  69  57  53  47\n",
"yokohama     33  31  15  28  28  25  24  27  22  22  28  35  31  31  27  28  24  38  16  38  23\n",
"tokyo        48  43  26  33  21  20  26  29  32  23  34  41  41  32  32  28  30  38  22  31  26\n",
"chyoshi      69  43  30  39  50  39  48  44  26  31  39  40  52  50  49  29  32  45  29  32  27\n",
"fukuoka       4   3   1   4   4   1   0  10   6  10   8   3   4   6   9   5   4   2   6   3   5\n",
"kushiro      72  52  45  40  40  33  50  93  86  32  38  44 130  49  49  36  38  49  50  43  35\n",
"utsunomiya   70  63  41  69  70  69  64  59  78  53  57  60  61  68  56  56  49  72  54  73  47\n",
"hatijyojima   8  11   5   7   2   3   8   8   4   3   8 272 103  48  26   8  10  16   7  34   6\n",
"morioka      39  52  24  43  35  26  24 217  44  46  36  38  36  36  32  22  19  54  25  26  40\n",
"unzen        20  17  10  10   2  37  10  75  82 112  65  72  48  61  18  18  37  21  13  25  13\n",
"aomori       17  17  15  24  10   6   7 124  24  19  23  22  24  21  15  12   9  26  12   8  18\n",
"matsumoto     2   2   6   3  12 106  36  14  14   3   4   4   1   2   1   0   2   7  10   2   1\n",
"nagoya        7  10  14   9   9  15   8   7  15   5   9   6   7   6   2   4   2   7   3   6   9\n",
"aso           9  11  10   6   5   3   4  11   4   7   8   4   4   3  90   8   4   6   5   1   1\n",
"yonago        3   1   3   1   2   2   1   6   1   1   0   3   3   1   1   0   1   6   3   2   0\n",
"hiroo        61  52  40  23  30  30  29  85  37  99  51  44  55  59  45  39  21  42  30  24  30\n",
sep="", file="equakes.txt")

read.table("equakes.txt")
data.frame(t(.Last.value)) -> equa
boxplot(equa)
```

◆ 参考文献：

佐伯　胖・松原　望（2000）：実践としての統計学，東京大学出版会，239p．

東京大学教養学部統計学教室編（1991）：統計学入門，東京大学出版会，307p．

永田　靖（1996）：統計的方法のしくみ，正しく理解するための３０の急所，日科技連，238p．

松原　望（1996）：わかりやすい統計学，丸善，149p．

柳川　堯（1981）：ノンパラメトリック法，培風館，p.255．

渡部　洋・鈴木規夫・山田文康・大塚雄作（1985）：探索的データ解析入門，朝倉書店，188p．

Beirlant, J., Y. Goegebeur, J. Teugels, J. Segers, D. De Waal, C. Ferro (2004):
Statistics of Extremes, Theory and Applications, Wiley, 490p.

Crawley, M. J. (2002): Statistical Computing - An Introduction to Data Analysis using S-Plus, Wiley, 761p.

Diggle, P. J. (1990): Time Series: Biostatistical Introduction, Oxford Univ. Press, 257p.

Kendall, M. G. (1948): Rank correlation methods, London: Griffin.

Siegel, S. (1956): Nonparametric Statistics - For the Behavioral Science, McGraw-Hill.
（藤本　熙 監訳，1958，ノンパラメトリック統計学－行動科学のために－，マグロウヒルブック株式会社）

Stephenson, A. G. (2004) A User's Guide to the evd Package (Version 2.1).
URL http://www.maths.lancs.ac.uk/~stephena/

Verzani, J. (2005): Using R for Introductory Statistics, Chapman & HALL/CRC, 414p.

Venables, W. N. and B. D. Ripley (2002): Modern Applied Statistics with S, (4th ed.), Springer, 495p.
（MASS is one of recommanded packages in R. ）

Wolfram Research: Mathematica 5, http://www.wolfram.com/