■ 相関と比率　◆ 北野 利一 ◆　2005 年 10月 11日

● 量的変数の相関係数

例１：相関係数を生み出した歴史的データ：Pearson and Lee (1903) ＝ 解説およびデータ取得は，
Weisberg (2005) による（http://www.stat.umn.edu/alr/Links/alr3data.zip）.

```
read.table("MacOSX3:Users:tk:envstat:heights.txt", header=T) -> pear

plot(pear) # Fig. 1
cor(pear)
pear[1:11,]
dim(pear)
names(pear)

attach(pear); cor(Mheight, Dheight); detach(pear)
with(pear, cor(Mheight, Dheight)) # cf.
```

例２：歴史的データ２：F. Galton のスィートピーの種子の大きさ
（東京大学教養学部統計学教室編，1991，表 1.3，p.6 を参照）

親：　15　16　17　18　19　20　21　（× 1/100 インチ）
子：　15.3 16.0 15.6 16.3 16.0 17.3 17.5　（× 1/100 インチ）

```
> x <- scan()
1: 15   16   17   18   19   20   21
8: ### just hit return ###
Read 7 items
> y <- scan() # do the similar actions

cor(x,y)

# cf. 1
z <- data.frame(x,y)
cor(z)
# cf. 2 : check it by your own hands
mean.x <- mean(x)
mean.y <- mean(y)
sum((x - mean.x)*(y - mean.y))/sqrt(sum((x - mean.x)^2)*sum((y - mean.y)^2)) # DEFINITION
```

● カテゴリカル・データの相関係数は？

例３：火山噴火と凶作年（高橋，1989，表 6，p.49）

		火山噴火		
		有	無	和
大凶作	有	5	27	32
	無	12	242	254
	和	17	269	286

この場合，火山噴火と凶作年に【相関】があるか？否か？ということが問題となる.

```
c(1700, 1739, 1755, 1783, 1815, 1822, 1835, 1875, 1883, 1886, 1902, 1907,
  1912, 1932, 1956, 1980, 1982) -> eruption

c(1705, 1707, 1720, 1737, 1747, 1749, 1753, 1755, 1757, 1767, 1782, 1783,
  1786, 1825, 1832, 1833, 1835, 1836, 1837, 1838, 1860, 1866, 1869, 1884,
  1897, 1902, 1905, 1931, 1934, 1941, 1945, 1980) -> lean
```

```
(intersect(eruption, lean) -> a.tak)
(setdiff(eruption, lean)   -> c.tak)
(setdiff(lean, eruption)   -> b.tak)
(union(lean, eruption)     -> d.tak)

matrix(unlist(lapply(list(a.tak, c.tak, b.tak, d.tak), length))*c(1,1,1,-1) + c(0,0,0,286),
       nr=2, dimnames=list("lean"=c("Yes","No"),
                           "eruption"=c("Yes","No"))) -> tak
addmargins(tak)
```

例４：量的変数を分割したデータ（例えば，上記の pear を平均より大／小に分けたもの）しか与えられない時，どのように，相関をみればよいか？

```
(mean(pear) -> mpear)
(sum(pear[[1]] <=mpear[1] & pear[[2]] > mpear[2]) -> a.pear)
(sum(pear[[1]] > mpear[1] & pear[[2]] > mpear[2]) -> b.pear)
(sum(pear[[1]] <=mpear[1] & pear[[2]] <=mpear[2]) -> c.pear)
(sum(pear[[1]] > mpear[1] & pear[[2]] <=mpear[2]) -> d.pear)
(matrix(c(a.pear, b.pear, c.pear, d.pear),
        2,2, byrow=T, dimnames=list(c("Dtall","Dshort"),c("Mshort","Mtall"))) -> tab.pear)

> addmargins(tab.pear)
       Mshort Mtall  Sum
Dtall     212   450  662
Dshort    478   235  713
Sum       690   685 1375

abline(v=mpear[1], h=mpear[2], col="red") # on Fig. 1
```

▲ 唐突かもしれないが，１次元の比率テストから議論をはじめる．

例５：『クイズ１００人にききました』と，１０００人にきいた場合との違い．

```
rbinom(1, 100, prob=0.3)                  # ask 100 pearsons
rbinom(200, 100, prob=0.3)/100 -> rat1    # and the 200 worlds
hist(rat1)

rbinom(200, 1000, prob=0.3)/1000 -> rat2   # ask 1000 pearsons
hist(rat1, xlim=c(0,1), ylim=c(0, 120)) -> rat1.hist
hist(rat2, add=TRUE, br=rat1.hist$br, dens=10, col="red")
abline(v=0.3, col="blue")
```

例６：貧困率 11.3 ％（２０００年），11.7 ％（２００１年）に対して，0.4 ％の上昇は有為か？ただし，２００１年は，５万世帯を調査対象としている（出典：Verzani，2005，Example 8.3，pp.219-220）．

（以下は，貧困率 11.3 ％を仮に母比率とし，標本比率 11.7 ％が確率的に起こりやすいか，否かを議論）

```
50000 * 0.117 # = 5850
binom.test(5850, 50000, p=0.113) # Exact binomial test
prop.test(5850, 50000, p=0.113)  # 1-sample proportions test with continuity correction
```

▽ binom.test を median に対する検定に用いることもできる（sign test）．
（例４の続きとして，Mheight において，平均値は，中央値とみなせるか？という議論）

```
binom.test(sum(pear$Mheight <= mpear[1]), length(pear$Mheight)) # p = 0.5 for median
```

◎ ２項分布は，正規分布で近似できる．

```
rbinom(200, 50000, p=.113) -> poverty
hist(poverty, xlim=c(5300, 6000))
approx <- function(x) dnorm(x, mean=50000*.113, sd=sqrt(50000*.113*(1-.113)))
hist(poverty, xlim=c(5300, 6000), prob=TRUE)
curve(approx, 5300, 6000, col="red", add=TRUE)
```

２項分布の平均および分散は，mean = n * p0，var = n * p0 * (1 - p0) であることと，分散の一般的な性質（Var(a * X) = a^2 * Var(X)；a = 定数，X = 確率変数）から，比率は以下のように規格化できる．

```
   (value - mean)/sqrt(var) = (p - p0)/sqrt(p0*(1-p0)/n); p = value/n
```

```
prop.test(5850, 50000, p=0.113, correct=F) # without continuity correction

((5850/50000 - .113)/sqrt(.113*(1 - .113)/50000) -> zval)
pnorm(zval, low=F)*2

prop.test(5850, 50000, p=0.113, alt="greater")
```

● ２次元の比率テスト

例７：（出典：Verzani，2005，Example 8.8，pp.234-235）【例６のつづき】
２００２年は，６万世帯を調査対象とし，貧困率は，12.1 %であった．
２００１年の貧困率と比較せよ（ここでは，２つの標本比率の差に注目する）．

```
60000 * .121 # = 7260
prop.test(c(5850, 7260), c(50000,60000), alt="less")

pp <- (50000 *.117 + 60000 *.121)/110000
(.117 - .113)/sqrt(pp*(1-pp)*(1/50000 + 1/60000)) -> zval; pnorm(zval, low=F)
prop.test(c(5850, 7260), c(50000,60000), alt="less", cor=F)
```

比率の差 p1 - p2 は，帰無仮説のもとで，平均がゼロ，分散が，

```
    p1 * (1 - p1)/n1 + p2 * (1 - p2)/n2 = p * (1 - p) * (1/n1 + 1/n2)
```

となる正規分布に従うものと扱うことができる（p = p1 = p2）．

例３についても，以下のように実行できる．

```
prop.test(tak[,1], apply(tak, 1, sum))
prop.test(tak[,1], apply(tak, 1, sum), correct=F)
prop.test(tak[,1], apply(tak, 1, sum), alt="gr") # it should be! # Why?
```

● 独立性の検定（or カイ自乗適合検定

```
  X-squared = N * (a * d - b * c)^2/n1/n2/m1/m2, in the 2 by 2 contingency table:
                                  sum
                      a       b     n1
                      c       d     n2
               sum    m1      m2     N  )
```

```
chisq.test(tak)
chisq.test(tak, correct=F)
```

１）「２次元の比率テスト」と「独立性の検定」の同等性（別紙参照）
つまり，prop.test(tak[,1], apply(tak, 1, sum)) と chisq.test(tak) は，ともに同じ出力結果を与えている．
（同様に，prop.test(tak[,1], apply(tak, 1, sum), correct=F) と chisq.test(tak, correct=F) も同じ結果）

```
X-squared = 6.0402, df = 1, p-value = 0.01398
```

２） sqrt(X-squared/N) は，相関係数を表す（別紙参照）．

```
(242 * 5 - 12 * 27)^2/32/254/17/269*286   # = X-squared; confirm it in the result by chisq.test/prop.test

> (242 * 5 - 12 * 27)/sqrt(32*254*17*269) # cor. value for 2x2 contingency table
[1] 0.1453253
```

したがって，火山噴火と大凶作の相関係数は，0.15 であるといえる．両者の関係に相関なしといえないことは， chisq.test などの p値からわかるが，その相関の度合いは強いものではない．
なお，渡部（1999）での「相関係数」についての説明（p.49）は正しいだろうか？

３）厳密な検定（Fisher の正確な検定）

```
fisher.test(tak)
```

（考え方：周辺度数を固定して，（？）に入る組み合わせの頻度を考え，p値を算出するのである．

		火山噴火		
		有	無	和
大凶作	有	（？）	（？）	32
	無	（？）	（？）	254
	和	17	269	286 ）

● 同じ２×２分割表でも，次の例は，独立性の検定は適用できないことに注意．
例８：２種類の検査法を比べる場合（McNemar 検定）
患者 166 人に対して，血清 IgE 抗体の定量検査を RAST 法とスクラッチ法と比較した．検査法の感度の差があるといえるか（古川・丹後 例題 7.6 を参照）．

```
matrix(c(85, 18, 2, 51), nr=2, dimnames=list("RAST"=c("+","-"),
                                        "SCRATCH"=c("+","-"))) -> tab29

> addmargins(tab29)
     SCRATCH
RAST   +  - Sum
  +   85  2  87
  -   18 51  69
  Sum 103 53 156

mcnemar.test(tab29)
```

■ 相関係数は，本来的に，統計量である（例２での DEFINITION に見るとおり）．従って，記述統計的に用いられやすい．推測統計的な観点での議論（相関係数の標本分布，無相関の検定）や，偏相関や順位相関などの概念については，後述に再び取り扱う．

◆ 参考文献：

高橋浩一郎 (1989)：デタラメを科学する - カオスの世界 -，丸善，248p．
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．
古川俊之・丹後俊郎 (1993)：新版 医学への統計学，朝倉書店，317p．
渡部 洋 (1999)：ベイズ統計学入門，福村出版，249p．
Verzani, J. (2005): Using R for Introductory Statistics, Chapman & HALL/CRC, 414p.
Weisberg, S. (2005): Applied Linear Regression (3rd ed.), Wiley Interscience, 310p.