■ 正規分布からの標本　◆ 北野 利一 ◆　2005 年 10月 18日

● 独立性の検定は，2x2 に限らず，2xn に拡張できる（さらに，rxc の分割表にも適用可）．つまり，２群の分布について，一方が他方と同一視できるか否かを議論する．しばしば，他方を理論分布とすることが多い．この場合，理論分布との“適合性”を議論するわけである．

例１：母親の足のサイズと帝王切開の関係（Dalgaard，2002，p.133-135 で議論）

```
library(ISwR); data(caesarean); ?caesarean
chisq.test(caesar.shoe) # cf.  prop.test(caesar.shoe[1,], margin.table(caesar.shoe, 2))
```

例２：歴史的なデータ：Feller（1968）（鈴木・山田（1996）例 6.10 より孫引き）第２次世界大戦でのロンドン南部に投下された爆弾数の統計（爆弾はランダムに投下されたのか？それとも，．．．）

```
N.sec <- c(229, 211, 93, 35, 7, 1); names(N.sec) <- c(0:4, "5+")
list(total.N.sections=sum(N.sec), total.N.bombs=sum(c(0:4, 7)*N.sec)) # "5+" is regarded as 7

poi <- c(dpois(0:4, lamb=537/576), ppois(4, lamb=537/576, low=F))

chisq.test(N.sec, p=poi) # with a warning message! WHY?
chisq.test(c(N.sec[-1 * 5:6], sum(N.sec[5:6])), p=c(poi[-1 * 5:6], sum(poi[5:6])))
```

ここで，検定統計量は，

カイ自乗統計量 ＝ （観測頻度　－　期待頻度）＾２／（期待頻度）の総和

である．なお，「観測頻度が期待頻度に適合するか」という立場ではなく，「（想定される）期待頻度が，（現実の）観測頻度に適合しているか」という立場では，以下を用いる．

修正カイ自乗統計量 ＝ （観測頻度　－　期待頻度）＾２／（観測頻度）の総和

```
(EN.sec <- 576 * poi)
list(X2.stat=sum((N.sec - EN.sec)^2/EN.sec), modif.X2.stat=sum((N.sec - EN.sec)^2/N.sec))
```

◎ 歴史的に有名なデータとして，「プロシアの馬蹴られ死亡兵」というものがある．多くの本に掲載されているので，各自，確認してみよ（他にも例題はいろいろあるはず．それらの例も検討せよ）．

なお，期待頻度に対する観測頻度は，多項分布（分割数が２つであれば，２項分布）に従う．上述のように定義されるカイ自乗統計量が，多項分布の尤度比統計量の対数×（－２）に，近似的に等しいことが示される（鈴木・山田（1996），p.243-245 を参照）．よって，尤度比検定の枠組みから，カイ自乗検定が成立することがいえるのである．

◆ 注意：オーバーフィットに，ご用心！
例３：メンデルの法則とその根拠データ： 佐伯・松原（2000）の練習問題 4.4，p.178
（“あてはまり”過ぎ ＝ 帰無仮説の下で，与えられた結果が実現される確率はわずかである）

```
(chisq.test(c(315, 101, 108, 32), p=c(9,3,3,1)/16) -> mendel.chisq)
pchisq(mendel.chisq$stat, df=3)
```

（カイ自乗検定に関する，その他の注意点（ウェルドンのさいころ，シンプソンのパラドックスなど）については，佐伯・松原（2000）p.32-34 を参照）

このように，比較されるべき理論分布は，カテゴリーに分類する必要があるので，はじめから整数値というカテゴリーに分類された離散分布（カウントの分布）に対して適用されることが多い．
連続分布に対しては，区間分割する必要がある．ただし，区間の切り方に注意と工夫がいる．

● 連続分布に対する適合度は，コルモゴロフ・スミルノフ検定を用いることが多い．
例４：２地点で採取された鳥（？）の翼の長さの相違（Crawley，2005，pp.101）

```
setwd("MacOSX3:Users:tk:rm171:crawley2"); read.table("wings.txt", header=T) -> wings
# cf. wings <- read.table(
# url("http://www.bio.ic.ac.uk/research/crawley/statistics/data/wings.txt"), header=T)

table(wings$loc)
(unstack(wings) -> wng)
library(stepfun); plot(ecdf(wng$B), lty=3); plot(ecdf(wng$A), add=T)
legend(15, .95, c("A","B"), lty=c(1,3), pch=1)
hist(wng$B, ylim=c(0,20), main="") -> out
hist(wng$A, bresks=out$br, add=T, dens=10, col="red")
legend(30,20, c("A","B"), lty=1, col=c("red", "black"))

ks.test(wng$A, wng$B) # cf.  t.test(wng$A, wng$B)
```

ところで，例１の話題は，２つの標本に対する分布の一致についての検討といえる．平均という代表値のみに着目して議論するものが，t 検定である．

● ２標本問題：２つの条件の下での実験データの分布を比較する問題．
“ある意味では１標本問題は現実の実験データを，すでに知られている分布と比較する問題として，いわば２標本問題の退化した場合として理解することもできる．そういう意味では２標本問題の方が１標本問題よりかえって基本的であると考えることもできる（竹内・大橋，1981，p.1）”

例５：ダーウィンのデータ（トウモロコシ属）（竹内・大橋，1981，p.4；原出典は，Fisher，1935）

```
zea <- data.frame(cross=c(23 + 4/8,  12,        21,
                          22,        19 + 1/8,  21 + 4/8,
                          22 + 1/8,  20 + 3/8,  18 + 2/8,  21 + 5/8,  23 + 2/8,
                          21,        22 + 1/8,  23,        12),
                  self=c(17 + 3/8,  20 + 3/8,  20,
                         20,        18 + 3/8,  18 + 5/8,
                         18 + 5/8,  15 + 2/8,  16 + 4/8,  18,        16 + 2/8,
                         18,        12 + 6/8,  15 + 4/8,  18),   pot=rep(1:4, c(3,3,5,4)))

boxplot(zea[,-3])
t.test(zea$cross, zea$self, var.eq=T)
```

ここでの議論は，２標本の平均に違いがあるか，否かである．平均という代表値１つのみに着目し，分布の他の特性は，なにも議論しない．“議論しない”というのは，“どうでもいい”といっているのだろうか？むしろ，“議論しない”のは，すでに定まっているからである．つまり，２つの標本は，正規分布から抽出されたものとし，分散は未知であるが等しいという前提のもとで，『平均値も一致していること』を帰無仮説として検定しているのである（言うまでもないが，２標本は，独立に抽出されていることを前提．したがって，対応のある２標本に，適用はできない）．

なお，分散が等しいと仮定しない場合には，Welch による近似的な検定を行う．R の t.test は，デフォルトで，Welch による検定を行うように設定されている．つまり，等分散性を確認してから，いわゆる t 検定を行うべきだという立場である．等分散性の確認には，F 検定を行う（R では，var.test が用意されている）．

例５（つづき）：

```
var.test(zea$cross, zea$self)
t.test(zea$cross, zea$self)
```

◇ F 検定は，エフ分布に基づく．つまり，２標本の分散比は，値 1 （＝帰無仮説）の周辺で，変動することを利用するものである．

標本分散 ＝ （（標本値 － 標本平均）＾２ の総和）／（標本サイズ － １）

であるので，母分散の比が１である時，標本分散の比は，カイ自乗の比に相当する．２標本は，独立に抽出されていることを前提としている．したがって，独立な２変量のカイ自乗分布を変形することにより，エフ分布が導かれる（別紙参照）．

◇ t 検定は，ティ分布に基づく．２標本の平均値の差を標本分散で規格化した量は，次のように変形できる．

（標本平均１ － 標本平均２）／（標本分散／標本サイズ）
＝（（標本平均１ － 母平均）／（ 母分散／標本サイズ）
－ （標本平均２ － 母平均）／（ 母分散／標本サイズ））／（カイ自乗／（標本サイズ －１））

ただし，（カイ自乗）＝（標本サイズ － １）＊（標本分散／母分散）である．これにより，正規分布に従う２変量の差（あるいは，和）の量も正規分布に従うので，標本平均（の差）は，正規分布に従う．
他方，標本分散（を規格化したもの）はカイ自乗分布に従う．さらに，標本平均と標本分散は，独立であること（例えば，鈴木・山田（1996）pp.127-128）から，正規分布とカイ自乗分布を独立に結合させた分布を変形することにより，ティ分布が導かれる（別紙参照）．

「独立」とは，２変量の結合分布が，それぞれの１次元の分布を掛けたものとして表される場合（つまり，一方の頻度が，他方の頻度に全く影響を受けないので，結合確率は，両方の確率の積となる）をいう．
「独立」に類似する概念として「無相関」がある．「無相関」とは，２変量の相関係数＝ゼロを意味する．

● 無相関の検定
例６：標本平均と標本分散に対する無相関の検定

```
ms <- vs <- numeric(100); for (j in 1:100) {dat <- rnorm(1000, mean=1, sd=1)
                           ms[j]  <-  mean(dat)
                           vs[j]  <-   var(dat)}

plot(ms,vs); cor(ms,vs)
cor.test(ms,vs)
```

例７：母相関係数＝ゼロに対する標本相関係数の分布

```
cors <- numeric(1000)
for (j in 1:1000) cors[j] <- cor(rnorm(10, mean=1, sd=1), rnorm(10, mean=1, sd=1))
hist(cors, prob=T)
```

● 標本相関係数の分布
ここでは，あえて式を示さないが，東京大学教養学部統計学教室編（1991）式（1.2）をみよ．
（また，導出については，竹内，1963，pp.135-138 を参照）

```
library(SuppDists)
rr <- seq(-1, 1, length=300)
lines(rr, dPearson(rr, 10, rho=0), col="red") # on the fig. of ex. 7

plot( rr, dPearson(rr, 8, rho=0.8), type="l")
lines(rr, dPearson(rr, 8, rho=0.0), col="blue") # fig. 1.1 in text Tokyo Univ. (1991)

r2 <- seq(-.99, .99, length=300)
plot( atanh(r2), dPearson(r2, 8, rho=0.8) * (1-r2^2), type="l")
lines(atanh(r2), dPearson(r2, 8, rho=0.0) * (1-r2^2), col="blue") # Fisher!
```

● Fisher の z 変換
　標本相関係数 r （および母相関係数 rho）を

$$z = 1/2 * \log((1 + r)/(1 - r))$$

で変換した量 z （母相関係数 rho を r に上式に代入して得られる量を eta とおく）は，近似的に，平均を eta および分散を１／（標本サイズ　－　３）とする正規分布に従うことが知られている（Fisher, 1925）．cor.test での信頼区間は，この知見に基づく（他方，検定は，これとは別の公式による）．上式の逆の関係は，次式で表せる．

$$r = \arctanh(z)$$

● 母相関係数がゼロでない多次元正規分布からのサンプル生成
　パッケージ： MASS あるいは，mvtnorm を用いて，以下のようにできる．

```
(Sigma <- matrix(c(10,3,3,2),2,2))
library(MASS);     var(mvrnorm(n=1000, rep(0, 2), Sigma) -> xy1)
library(mvtnorm);  var(rmvnorm(n=1000, rep(0, 2), Sigma) -> xy2)

xy <- rbind(xy1, xy2)
par(fig=c(0,  .7,   0, .7));           plot(xy)
par(fig=c(0,  .7, .65,  1), new=T);    hist(xy[,1])
par(fig=c(.65, 1,   0, .7), new=T); boxplot(xy[,2])
```

◇ t 検定に関する注意点（１）：対応のある２標本の検定として考えるべき問題もあることに注意せよ．

　例５（さらに，つづき）：Fisher の見解は，ギリギリ有為（であって，確信できるほどの有為ではない）．

```
t.test(zea$cross, zea$self, pair=T)
```

◇ t 検定に関する注意点（２）：
上述のストーリーから明らかなように，母分布を正規分布とみなせない場合には，適用不可．
この場合には，wilcox.test を用いることになる（が，それは別の機会に紹介する予定）．

▲ カイ自乗分布，ティ分布，エフ分布のいずれにも，自由度という概念が含まれている．
自由度は標本のサイズに依存するものである．細かい定義は，しかし，種々の整合をとるために便宜的に決めている節がある．自由度については別の機会に触れる．

◆ 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.
鈴木　武・山田作太郎 (1996): 数理統計学，－ 基礎から学ぶデータ解析 －，内田老鶴圃，406p.
竹内　啓 (1963): 数理統計学，データ解析の方法，東洋経済，373p.
竹内　啓・大橋靖雄 (1981): 統計的推測 － ２標本問題，数学セミナー増刊，日本評論社，192p.
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.
Crawley, M. J. (2005): Statistics, an introduction using R, Wiley, 327p.
　http://www.bio.ic.ac.uk/research/crawley/statistics/data/zipped.zip
Dalgaard, P. (2002): Introductory Statistics with R, Springer, 267p.
Fisher, R. A. (1925): Statistical Methods for Research Workers, Oliver and Boyd Ltd.
　（遠藤健児・鍋谷清治 共訳 (1970): 研究者のための統計的方法，森北出版）
Fisher, R. A. (1935): The Design of Experiments, Oliver and Boyd Ltd.
　（遠藤健児・鍋谷清治 共訳 (1971): 実験計画法，森北出版）