医学情報処理演習第９回「計数データと比率の解析」

中澤　港 (nminato@med.gunma-u.ac.jp)

2005 年 1 月 24 日

前回の課題の回答例

まずデータを読んでプロットしてみる。

```
data(airquality)
attach(airquality)
plot(Ozone~Wind)
```

オゾン濃度と風速の相関をみるために `cor.test(Wind,Ozone)` によってピアソンの積率相関係数を求めると，`r=-0.6` であり，有意確率は 10^{-13} のオーダーなので，5%水準で有意な負の相関があるといえる。次に風速からオゾン濃度を予測することを考える。そのためには，オゾン濃度を従属変数，風速を独立変数とした回帰をしなければならないので，

```
res <- lm(Ozone~Wind)
abline(res)
summary(res)
```

とすると，回帰直線が重ね描きされ，回帰の諸パラメータが得られる。回帰係数自体は`-5.5509` となっていて，これがゼロと差がないという帰無仮説の検定結果は有意確率が 10^{-13} のオーダーなので，回帰関係自体に意味はあるけれども，自由度調整済み重相関係数の二乗 (Adjusted R-squared) をみると，0.3563 となっているので，あまり決定係数は大きいとはいえない。つまり説明力は高くない。
Wind が 25 のときの Ozone の値を予測するには，式の通り `25*res$coef[2]+res$coef[1]` とするか，あるいは `predict(res,list(Wind=25))` とすればよいのだが，計算してみると予測値が`-41.9` となってしまう。
オゾン濃度がマイナスになるのは論理的にありえないので，この回帰による予測はおかしい。おかしい理由としては，(1) 独立変数にも測定誤差が含まれている，(2) 回帰関係の決定係数が大きくない，(3) オゾン濃度も風速も正の値しかとらない端の切れた分布である（正の値しかとらない量２つの間に負の相関があれば，無限に外挿できないことは自明である），という３つが考えられる。したがって，この場合，線型回帰分析を使って得られる回帰式を使って風速 25 マイル／時のときのオゾン濃度の予測をすることは不適当である。

前回の課題への付録：非線型回帰と重回帰（演習は省略）

せっかくなので，この関係を使って予測することはできないのだろうか？　ということを考えてみる。[*1]

[*1] 2005 年 1 月 24 日追記。この項は，第 12 回講義資料に，より詳しく系統的に採録したので，そちらを参照されたい。

回帰の結果から残差分析を行うと，回帰がデータから系統的にずれていないかどうかを検討することができる。R では，plot(residuals(res)) として，回帰の結果から残差を取り出してプロットすることができるので，その図をよく見ると，両端でのマイナス方向の残差が大きいことがわかるので，非線型の関係があるのではないかと思われる。R では，nls() という関数を使って非線型の回帰を行うことができる。例えば，二次曲線で回帰するには，

```
res2 <- nls(Ozone ~ a + b*Wind + c*Wind^2, start=list(a=0,b=0,c=1))
lines(x<-seq(min(Wind),max(Wind),length=20),predict(res2,list(Wind=x)),col="red")
summary(res2)
```

とすれば曲線を重ね描きでき，パラメータ推定値 a，b，c を得ることができる。しかし，Wind が 25 マイル／時のときにこの二次曲線回帰から予測される Ozone の値は，

```
predict(res2,list(Wind=25))
```

とすると約 81.9 ppb となり，これもまたありそうにない。二次曲線は常に下に凸なことが原因かと考え，三次曲線で回帰して Wind が 25 マイル／時のときの Ozone の予測値を求めるには，

```
res3 <- nls(Ozone ~ a + b*Wind + c*Wind^2 + d*Wind^3, start=list(a=0,b=1,c=1,d=1))
lines(x<-seq(min(Wind),max(Wind),length=20),predict(res3,list(Wind=x)),col="blue")
summary(res3)
predict(res3,list(Wind=25))
```

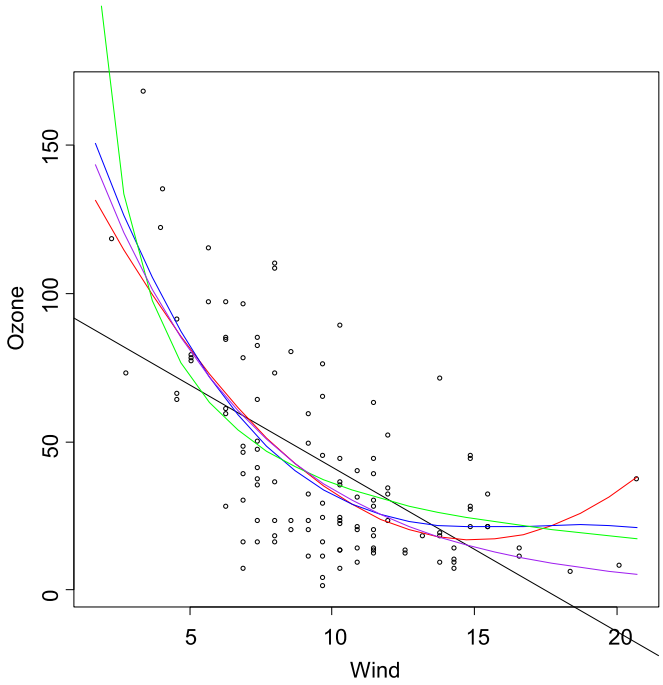
とすれば，約 4.9 ppb となり，何となくもっともらしい。ただし，summary(res3) をすると，d の推定値がゼロと差がないという帰無仮説は棄却されないので，３次回帰もあまりうまくなさそうである。

非線型の関係をよく考えると，風が強いほどオゾンが吹き飛ばされて減るのを積が一定とあらわせば良さそうな気がする。そこで双曲線回帰を試してみる。

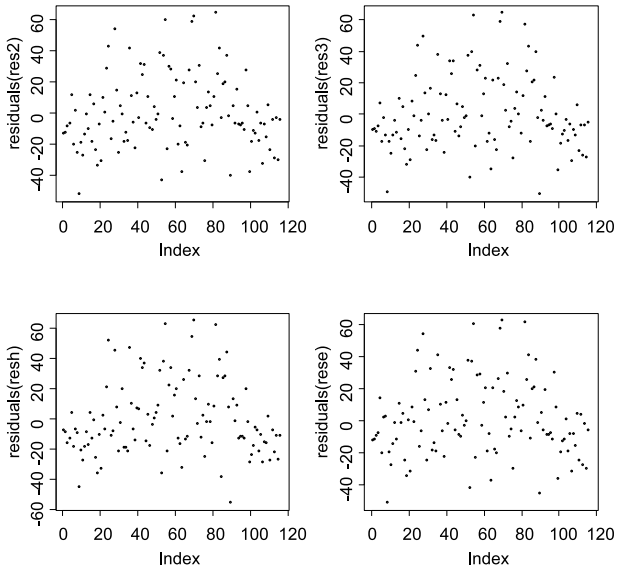
```
resh <- nls(Ozone ~ a/Wind, start=list(a=50))
lines(x<-seq(min(Wind),max(Wind),length=20),predict(resh,list(Wind=x)),col="green")
summary(resh)
predict(resh,list(Wind=25))
```

これだと約 14.4 ppb となってもっともらしいし，パラメータ１つで済む点が優れている。プロットの形からもう一つ考えられるのは係数が負の指数回帰である。これは初期値をうまく決めないと発散してしまって解が得られない点に注意する必要があるが，回帰曲線もかなり当てはまっているように見えるし，風速 25 マイル／時のときのオゾン濃度の推定値も約 2.5 ppb となって，おかしくはない。

```
rese <- nls(Ozone ~ a*exp(-b*Wind), start=list(a=300,b=0.2))
lines(x<-seq(min(Wind),max(Wind),length=20),predict(rese,list(Wind=x)),col="purple")
summary(rese)
predict(rese,list(Wind=25))
```



これらの非線型の回帰モデルのうち，どれを採用すべきかを検討するためには，線型回帰のときとは違って決定係数，すなわち（重）相関係数の二乗は使えない（データとしてある変数に二乗などの非線型の変換を加えた変数を作っておき，無理やりそれを線型回帰に投入するという手法を使えば重相関係数の二乗を計算することはできるが，解釈に注意が必要である）。ではどうやってモデルの採否を決定するかといえば，よく使われるのは残差分析と尤度比検定と AIC である。



残差分析は線型のときと同じようにすればよい。以下のようにしてまとめて表示してみると，係数が負の指数回帰が一番残差が小さく，かつ系統的なずれが少ないように見える。

```
op<-par(mfrow=c(2,2))
plot(residuals(res2))
plot(residuals(res3))
plot(residuals(resh))
plot(residuals(rese))
par(op)
```

３次回帰と２次回帰のように，一方が他方を一般化した形になっている場合（３次回帰で d=0 ならば２次回帰に還元される），一般に，より一般性の低いモデル（２次回帰）の最大尤度を，より一般的なモデル（３次回帰）の最大尤度[*2]で割った値の自然対数をとって-2 を掛けた値が，「より一般性の低いモデル（パラメータの少ないモデル）の方が正しい」という帰無仮説の下で，自由度１（比較するモデル間のパラメータ数の差）のカイ二乗分布に従うことを使って検定できる（これが尤度比検定と呼ばれる）。

```
lambda <- -2*(logLik(res2)-logLik(res3))
1-pchisq(lambda,1)
```

有意水準 5%で帰無仮説は棄却されないので，２次回帰の方が３次回帰よりよいモデルであると判断できる。

さて一方，AIC はパラメータ数と最大尤度からモデルの当てはまりの悪さを表すものとして計算される指標で，数式としては，L を最大尤度，n をパラメータ数として，

$$\mathrm{AIC} = -2\ln L + 2n$$

で表される。AIC が小さなモデルほど当てはまりがいいと考える。この場合は，

```
AIC(res2)
AIC(res3)
AIC(resh)
AIC(rese)
```

とすると，相対的には AIC(rese) の値が一番小さいことがわかる。しかし，AIC の値としては決して小さくなく，あてはまりがそれほどよいとはいえない。メカニズムを考えても，やはり Wind だけから Ozone を予測するところに限界があるので，むしろ他の変数も独立変数として制御する，重回帰分析にする方が筋がよいと思われる。

そこで，オゾン濃度に対して，風速だけが係数が負の指数関数的に作用し，太陽放射と気温が相加的に作用する（それぞれについて散布図を書かせてみると，線型の正の相関がありそうにみえる）という非線型の重回帰モデルを適用してみる。

[*2] 一般に $f(x,\theta)$ で与えられる確率密度関数からの観測値を $\{x_1, x_2, ..., x_n\}$ とするとき，θ の関数として，$L(\theta) = f(x_1,\theta)f(x_2,\theta)\cdots f(x_n,\theta)$ を考えると，確率密度関数の値が大きいところほど観測されやすいため，$L(\theta)$ の値を最大にするような θ を真の θ の推定値とみなすのが最も尤もらしい。この意味で $L(\theta)$ を尤度関数と呼び，この θ のような推定量のことを最尤推定量と呼ぶ。尤度関数を最大にすることはその対数をとったもの（対数尤度）を最大にすることと同値なので，対数尤度を θ で偏微分した式の値をゼロにするような θ の中から $\ln L(\theta)$ を最大にするものが，最尤推定量となる。例えば，正規分布に従うサンプルデータについて得られる尤度関数を母平均 μ で偏微分したものをゼロとおいた「最尤方程式」を解けば，母平均の最尤推定量が標本平均であることがわかる。詳しくは鈴木義一郎 (1995)「情報量基準による統計解析」（講談社サイエンティフィク）を参照のこと。

```
resm <- nls(Ozone ~ a*exp(-b*Wind) + c*Solar.R + d*Temp,
start=list(a=200,b=0.2,c=1,d=1))
summary(resm)
AIC(resm)
```

しかし，d=0 という帰無仮説が棄却されないので，Temp を外してやり直してみる[*3]。

```
resmr <- nls(Ozone ~ a*exp(-b*Wind) + c*Solar.R, start=list(a=200,b=0.2,c=1))
summary(resmr)
AIC(resmr)
```

AIC は若干大きくなるが，独立変数が 1 つだけの場合に比べるとずっと小さい（もっとも，十分に小さいとはいえないので，このデータに含まれていない，他の要因の影響が大きいのであろう）。さきほどと同様に Temp も入れたモデルと尤度比検定をすると有意ではないので，最終的にこのモデルを採用するとよいと考えられる。そこで，predict(resmr,list(Wind=25,Solar.R=mean(Solar.R,na.rm=T))) とすれば，約 16.5 となるので，風速 25 マイル／時のときのオゾン濃度は，太陽放射が平均的な条件なら，約 16.5 ppb になると予測される[*5]。

母比率を推定する方法

今回は，名義尺度や順序尺度をもつカテゴリ変数 1 つを分析する方法を扱う。カテゴリ変数 1 つがもっている情報は，データ数と，個々のカテゴリが占める割合（標本比率）である。したがって，このデータから求める統計的な指標は，母比率，即ち個々のカテゴリが母集団で占めるであろう割合である。通常，標本比率とほぼ一致する。

例えば，手元の容器の中に，数百個の白い碁石があるとする。この概数を手っ取り早く当てるために，数十個の黒い碁石を混ぜる。よくかき混ぜてから 20 個程度の石を取り出してみて（標本），その中で黒い石が占めていた割合（標本比率）を求め，それが母比率と等しいと仮定して加えた黒い碁石の数を割って総数を求め，黒い碁石の数を引けば，元々の白い碁石の数が得られる。生態学で，野原のバッタの数を調べたいときに全数を調べるわけにはいかないので，捕まえてペンキでマークして放して暫く経ってからまた捕まえてマークされているバッタの割合を求めて，マークした数をそれで割って総数を推定する，というリンカーン法（Capture-Mark-Recapture; 略して CMR ともいう）のやり方と同じである。

例題

最初に混入した黒い石の数が 40 個，かき混ぜてから 20 個の石を取り出してみたら黒石 2 個，白石 18 個だった場合，元の白石の数はいくつと推定されるか？

元の白石の数を x とすると，40/(40+x)=2/(2+18) となるので，これを x について解けば，x=360 が得られる。したがって 360 個と推定される。この程度は R を使うまでもないが，40/(2/(2+18))-40 と電卓のよ

[*3] このように，重回帰モデルの独立変数の取捨選択を行うことを変数選択と呼ぶ。非線型重回帰モデルの場合は自動的にはできないが，線型重回帰モデルならば，step(lm(Ozone~Wind+Solar.R+Temp)) のように step() 関数を使って，自動的に変数選択を行わせることができる。この結果はすべての変数が残って AIC が 682 まで小さくなるのだが[*4]，その回帰係数を用いて，Solar.R と Temp がそれぞれの平均値（ただし Solar.R には欠損値が含まれているので，平均を計算するときに，mean(Solar.R,na.rm=T) としなくてはならない）で Wind=25 のときのオゾン濃度を点推定すると約-7.9 となってしまって，やはり採用できない（95%信頼区間はゼロを跨いでいるが）。結局，いくら AIC が小さくなっても，論理的に問題がある線型回帰を適用して予測をしてはいけないということである。

[*5] R-2.0.1 の時点では未実装だが（線型回帰については実装済み），将来的には 95%信頼区間も predict() 関数の中で se.fit=T と interval="confidence"と level=0.95 というオプションをコンマで区切って並べれば求められるようになるらしい。

うに打てば，360 が得られる。

推定値の確からしさ

ここで，このようにして求めた推定値がどれほど確からしいか？　を考えよう。例えば，黒石の割合（母比率）が p である容器から 20 個の石を取り出したときに，黒石がちょうど 2 個である確率を考えると，これは二項分布に従う（個々の抽出を考えると復元抽出でないと二項分布に従わないが，すべての条件付確率を合計すれば問題ない）。

つまり，確率 p の現象が 20 回中 2 回起こり，残りの 18 回は確率 (1-p) の現象が起こったわけだから，その確率をすべて掛け合わせ，20 回中どの 2 回で起こるのかという組み合わせの数だけパタンがありうるので ${}_{20}\mathrm{C}_2$ 回だけそれを足し合わせた確率になる。

R では，この確率は，母比率 p を与えると，`choose(20,2)*p^2*(1-p)^18` あるいは `dbinom(2,20,p)` で得られる。

逆に考えれば，この「母比率 p の現象が 20 回中ちょうど 2 回得られる」確率を最大にするような p が真の母比率として最も尤もらしいと考えられる。0.01 刻みでこの確率を最大にする p を探索するには下枠内のようにする。0.1 が得られる。

```
x<-seq(0,1,by=0.01)
y<-dbinom(2,20,x)
x[which.max(y)]
plot(x,y,type="l")
```

40 個入れて全体の 0.1 を占めるのだから，40/0.1=400 が全体の数で，400-40=360 が元の白石の数だと推定できる。ただし，図を見ればわかるように，$p = 0.09$ だろうが $p = 0.11$ だろうが，黒石がちょうど 2 個である確率には大した差はない。だから，360 個という点推定値は，404 個（$p = 0.09$ の場合）とか 324 個（$p = 0.11$ の場合）に比べて，それほど信頼性は高くない。

母比率の信頼区間

ある程度の信頼性が見込める範囲を示すためには, 平均値の場合と同様, 信頼区間を用いることができる。母比率が $p = 0.1$ のときに，20 個のサンプル中の黒石出現回数がちょうど 2 である確率は，`dbinom(2,20,0.1)` より，約 28.5%に過ぎない（もちろんこれは，母比率が $p = 0.7$ のときに 20 個のサンプル中の黒石出現回数がちょうど 2 である確率である約 3.6×10^{-8} よりもずっと大きい）。ここはやはり，95%くらいの確からしさをもって，母比率はここからここまでの範囲に入るという形で説明したいと考え，95%信頼区間を計算するのがよいだろう。

平均値の場合は正規分布や t 分布を使ったが，比率の場合は二項分布を用いればよい。つまり，サンプルサイズ N のうち，ある事象が観察された個体数が X だったとすると，母比率 p の点推定量は `p <- X/N` で与えられるので，平均値の場合から類推して，95%信頼区間の下限は `qbinom(0.025,N,p)/N` で，上限は `qbinom(0.975,N,p)/N` と考えるのがもっともシンプルである。しかし，二項分布は左右対称ではなく，分位点関数が整数値しかとれないので，N がある程度大きくて，それほど稀でない事象ならばこれでもいいけれども，あらゆる可能性のうち少なくとも 95%を含む最短の区間を 95%信頼区間として求めたいとすると，別の考え方をしなくてはならないだろう。

R では，Clopper CJ, Pearson ES: The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26: 404-413, 1934. に記載されているアルゴリズムでこの信頼区間を計算する関数が

実装済みである。このアルゴリズム[*6]を使っても最短であることは保証されないが，少なくとも 95%を含むことは保証するとされている。`binom.test(X,N,p)` とすれば `N` 個体中 `X` 個体に観察される事象の母比率が `p` と差がないという帰無仮説の検定結果が表示される（`p==X/N` ならば `p-value=1` となる）とともに，Clopper and Pearson の方法による 95%信頼区間が計算される。

正規近似

二項分布 $B(N,p)$ は，`N` が大きいときは正規分布 $N(Np, Np(1-p))$ で近似できる。正規分布は左右対称なので，95%のサンプルは，平均 ± 標準偏差× 1.96（正確には `qnorm(0.975)` であり 1.9599... となる）に含まれると考えてよく，下式が成立する。

$$\mathrm{Prob}[-1.96 \leq (X-Np)/\sqrt{Np(1-p)} \leq 1.96] = 0.95$$

これから $p^* = X/N$ を使って式変形すると，

$$\mathrm{Prob}[p^* - 1.96\sqrt{p^*(1-p^*)/N} \leq p \leq p^* + 1.96\sqrt{p^*(1-p^*)/N}] = 0.95$$

となるので，母比率 p は 95%の確率で下限 $(p^* - 1.96\sqrt{p^*(1-p^*)/N}$，上限 $p^* + 1.96\sqrt{p^*(1-p^*)/N})$ の範囲にあるといえる。即ちこれが，母比率 p の 95%信頼区間となる。

例題

25 匹のマウスに毒物 A を一定量経口投与したところ，5 匹が死亡した。この毒物のその用量によるマウスの致命率の点推定量と 95%信頼区間を求めよ。

点推定量は，5/25 より 20%であることは自明である。シンプルな考え方で 95%信頼区間を求めると，下限が `qbinom(0.025,25,5/25)/25` から 0.04，上限が `qbinom(0.975,25,5/25)/25` から 0.36 となる。`binom.test(5,25,0.2)` によれば [0.068,0.407] となって，シンプルな考え方よりも上側にずれる。正規近似によれば，`0.2-qnorm(0.975)*sqrt(0.2*0.8/25)` が約 0.043，`0.2+qnorm(0.975)*sqrt(0.2*0.8/25)` が約 0.357 となるので，[0.043,0.357] が 95%信頼区間となり，ちょっと幅が狭すぎる。サンプルサイズ 25 程度で正規近似をするのはまずいかもしれない。

カテゴリ２つの場合の母比率の検定

あらかじめ母比率について何らかの期待があるときには（50%であるとか），標本から推定された比率がそれと違っていないかどうかを調べたい，ということが起こる。カテゴリが２つしかない場合は，上で説明した二項分布による推定の裏返しでよい。つまり，「サンプル `N` 個体中 `X` 個体に観察された事象の母比率が `p` と差がない」という帰無仮説を検定するには，`binom.test(X,N,p)` とすればよい。丁寧に考える方法を下の例題で示すが，実際には `binom.test()` を使えば十分である。

例題

ある病院で生まれた子ども 900 人中，男児は 480 人であった。このデータから，（１）男女の生まれる比率は半々であるという仮説，（２）出生性比が 1.06 である（＝男児 1.06 に対して女児 1 という割合で生まれる）という仮説，は支持されるか？（出典：鈴木義一郎「情報量基準による統計解析入門」，講談社サイエンティフィク，1995 年）

（１）母比率が 0.5 であるとして，得られているデータよりも外れたデータが偶然得られる確率（両側に外れることを考えなくてはいけないので，480 人以上になる確率と 420 人以下になる確率の合計）がきわめて小

[*6] R で，括弧もつけずに `binom.test` とすると，どういう計算をしているのかが確認できる。

さければ，「男女の生まれる比率は半々である」という仮説はありそうもないと考えてよいことになる。母比率 0.5 で起こる現象が，900 人中ちょうど 480 人に起こる確率は `dbinom(480,900,0.5)` で与えられ，480 人以上になる確率は，`dbinom(480,900,0.5)` + `dbinom(481,900,0.5)` + ... + `dbinom(900,900,0.5)` となるが，これは分布関数を使えば，`1-pbinom(479,900,0.5)` で計算できる。420 人以下になる確率も同様に分布関数を使って書けば，`pbinom(420,900,0.5)` である。従って，求める確率はこれらの和，即ち，

```
1-pbinom(479,900,0.5)+pbinom(420,900,0.5)
```

である。計算してみると 0.04916... となるので，有意水準 5%で仮説は棄却されることがわかる[*7]。

（2）同じように考えれば，`1-pbinom(479,900,1.06/(1.06+1))+pbinom(446,900,1.06/(1.06+1))` でよいはずであり（446 は帰無仮説の下での母比率の値である `900*1.06/(1+1.06)` が約 463 なので，480 と反対側に同じだけ外れた人数を考えた値である），約 0.271 となる[*8]ので，有意水準 5%で帰無仮説は棄却されない。つまり仮説は支持されるといえる。

カテゴリが複数ある場合の母比率の検定

しかし，注目しているカテゴリ変数のカテゴリは 2 つとは限らず，3 つ以上あるかもしれない。そのうち 1 つの事象に着目して，それが起こるか起こらないかだけを分析することもあるが，それぞれのカテゴリの出現頻度のデータをすべて分析することを考えてみる。こういう場合の基本的な考え方としては，標本データの度数分布が，母集団について期待される分布と差がないという帰無仮説の下で観察データよりも外れたデータが偶然得られる確率を調べて，それが統計的に意味があると考えられるほど小さい場合に帰無仮説を棄却することになる。

具体的には，カテゴリ数が全部で n 個あって，i 番目のカテゴリの観測度数が O_i，期待度数が E_i であるとき，$\chi^2 = \sum (O_i - E_i)^2/E_i$ が[*9]，自由度 $n-1$ のカイ二乗分布に従うことを利用して検定する（但し，期待度数を計算するために不明な母数をデータから推定したときは，その数も自由度から引く。E_i が 1 未満のときはカテゴリ分けをやり直す。また，度数は整数値だけれどもカイ二乗分布は連続分布なので，χ^2 を計算する際に連続性の補正と呼ばれる操作をすることがある）。このような χ^2 が大きな値になることは，観測された度数分布が期待される分布と一致している可能性が極めて低いことを意味する。一般に，χ^2 が自由度 $n-1$ のカイ二乗分布の 95%点よりも大きいときは，統計的に有意であるとみなして，帰無仮説を棄却する。この検定方法を**カイ二乗適合度検定**と呼ぶ。

R で自由度 1 のカイ二乗分布を図示するには，`plot(x<-seq(0,5,by=0.1),dchisq(x,1),type="l")` とすればよい。χ^2 値が 1 より大きくなる確率は `1-pchisq(1,1)` より得られ，約 0.317 である。 参考までに，自由度 n のカイ二乗分布の確率密度関数（R では `dchisq(x,n)` で得られる）は，$x > 0$ について，$f_n(x) = 1/(2^{(n/2)}\Gamma(n/2))x^{(n/2-1)}\exp(-x/2)$ であり，平均 n，分散 $2n$ である。なお，**自由度 (degree of freedom; d.f.)** とは，平均値の検定のところで説明したとおり，標本の数（この場合はカテゴリ数）から，前もって推定する母数の数を引いた値である。この例なら $\sum E_i$ だけを $\sum O_i$ として推定すれば，E_1 から E_{n-1} まで定めて E_n が決まることになるので，自由度は $n-1$ となる。

このやり方は，カテゴリが 2 つのときのデータについても適用できる。上の例題に適用してみると，（1）の場合，χ^2 は，`X <- (480-450)^2/450+(420-450)^2/450` として計算される。この値が自由度 1 のカイ二乗分布に従うので，R で `1-pchisq(X,1)` とすれば，男女の生まれる比率が半々である場合に 900 人中男児 480 人よりも半々から外れた観察値が得られる確率，つまり有意確率が計算できる。実行してみると，0.0455...

[*7] `binom.test(480,900,0.5)` の結果得られる `p-value` と一致する。

[*8] `binom.test(480,900,1.06/(1.06+1))` の結果と一致する。

[*9] χ は「カイ」と発音する。英語では chi-square と書かれるので，英文を読むときに間違って「チ」と読んでしまうと大変恥ずかしい。

となる。したがって，有意水準 5%で「男女の生まれる母比率は半々である」という帰無仮説は棄却される。（2）の場合，

```
EM <- 900*1.06/2.06
EF <- 900*1/2.06
X <- (480-EM)^2/EM+(420-EF)^2/EF
1- pchisq(X,1)
```

より，有意確率は約 0.26 となるので，帰無仮説の下で偶然，男児が 900 人中 480 人以上になる確率は約 26%あると解釈され，この帰無仮説は棄却されない。

ちなみに，出生 900 中男児が 480 人観察されたとき，母集団における出生性比の 95%信頼区間を考えてみると，R Console に下枠内のように入力すれば，[1.0005,1.3059] となることがわかる。

```
res <- binom.test(480,900,480/900)
res$conf.int/(1-res$conf.int)
```

少し複雑な例（演習は省略）

例題

1 日の交通事故件数を 155 日間について調べたところ，0 件の日が 79 日，1 件の日が 61 日，2 件の日が 13 日，3 件の日が 1 日，4 件以上の日が 1 日だったとする。このとき，1 日あたりの交通事故件数はポアソン分布に従うと言えるか？（出典：豊川裕之，柳井晴夫（編著）「医学・保健学の例題による統計学」，現代数学社，1982）[a]

[a] 一般に，稀な事象についてベルヌーイ試行を行うときの事象生起数がポアソン分布に従うことが知られている。交通事故は稀な事象であり，ある日に交通事故が起こる件数と翌日に交通事故が起こる件数は独立と考えられるので，交通事故件数はポアソン分布に従うための条件を満たしている。

R では，ポアソン分布の確率関数（離散分布の場合は，確率密度関数と言わずに確率関数というのが普通）は，`dpois(`**件数，期待値**`)` で与えられる。この例題ではポアソン分布の期待値（これは母数である）がわからないので，データから推定すれば，

$$\frac{(0\times 79+1\times 61+2\times 13+3\times 1+4\times 1)}{155}$$

で得られる。この値を `Ehh` として計算し，観測度数の分布をプロットするためには，下枠内を打てばよい。

```
cc <- 0:4
hh <- c(79,61,13,1,1)
names(hh) <- cc
barplot(hh)
Ehh <- sum(cc*hh)/sum(hh)
```

従って，1 日の事故件数が期待値 `Ehh` のポアソン分布に従うとしたときの，事故件数 0〜4 の期待日数 `epp` は，`epp <- dpois(cc,Ehh)*sum(hh)` で得られる。

こうなれば，`X <- sum((hh-epp)^2/epp)` としてカイ二乗値を求め，これが自由度 3（件数の種類が 5 種類あって，ポアソン分布の期待値が母数として推定されたので，$5-1-1=3$ となる）のカイ二乗分布に従

うとして 1-pchisq(X,3) が 0.05 より小さいかどうかで適合を判定すれば良さそうなものだが，そうはいかない。

epp[5]（この場合，epp[cc==4] と同じものを指すことになるので，以後，この記法を用いる）が 1 より小さいので，カテゴリを併合しなくてはならないのである[*10]。そこで，epp[5] を epp[4] と併合する。

即ち，

```
ep <- epp[cc<4]
ep[4] <- ep[4]+epp[5]
```

として期待度数の分布 ep を得，

```
h <- hh[cc<4]
h[4] <- h[4]+hh[5]
```

として観測度数の分布 h を得る。

後は，XX<-sum((h-ep)^2/ep) としてカイ二乗値を求め，1-pchisq(XX,2) を計算すると（カテゴリが 1 つ減ったので自由度も 1 減って 2 となる），約 0.187 となることがわかる。即ち，1 日の交通事故件数がポアソン分布に従っているという仮定の下でこのデータよりも偏ったデータが得られる確率は約 19%あり，「一日の事故件数がポアソン分布に従っている」という帰無仮説は棄却されない。

R にもカイ二乗適合度検定をやってくれる関数は用意されていて，もし自由度の調整がなければ，

```
chisq.test(as.table(h),p=ep/sum(ep),correct=F)
```

とすればカイ二乗値とその有意確率が計算できるのだが，カイ二乗分布は自由度 2 の場合と自由度 3 の場合では大きく違うので，この場合のように自由度を減らさなくてはいけないときには使えない。なお，2 つの分布が一致しているという帰無仮説を検定する方法としては，コルモゴロフ＝スミルノフ検定という方法もあり，これなら，ks.test(h,ep) で検定できる。なお，このデータについては，ここで示したどのやり方で分析しても，帰無仮説が有意水準 5%で棄却されないという結論は変わらない。

サイコロの正しさの検定

この特別な場合として，どのカテゴリも出現頻度が等しいという帰無仮説を検定することが考えられる。たとえば，サイコロを 900 回振って出た目の回数が下表のようであったとき，このサイコロの各目の出やすさに差はないと考えていいかという問題である。

目	1	2	3	4	5	6
回数	137	163	137	138	168	157

上と同じように考えれば，

```
h <- c(137,163,137,138,168,157)
X <- sum((h-150)^2/150)
1-pchisq(X,4)
```

により，どの目の出やすさにも差がないという帰無仮説（つまり，900 回振ったときの各目の期待頻度は 150 回ずつということ）を検定すると，有意確率は 0.145... となるので，有意水準 5%で帰無仮説は棄却されず，このサイコロの各目の出やすさには差がないといえる。

[*10] もっとも，併合した分布は元の分布と等価ではないので，併合の際にも本当は慎重な検討が必要である。

複数群間の比率の差

この話をもっと一般化して，１つのカテゴリ変数のカテゴリ間の頻度の差ではなく，複数の独立した事象の観察頻度に差があるかどうかを考えてみる。もっとも単純な場合として，患者群 n_1 名と対照群 n_2 名の間で，ある特性をもつ者の人数がそれぞれ r_1 名と r_2 名だったとして，その特性の母比率に差がないという帰無仮説を考える。カイ二乗適合度検定でもいい（ただし特性をもたない者についても期待度数と観測度数の差を考えなくてはいけない）のだが，以下では，二乗しないで正規近似によって検定してみる。

２群の母比率 p_1, p_2 が，各々の標本比率 $\hat{p_1} = r_1/n_1, \hat{p_2} = r_2/n_2$ として推定されるとき，それらの差を考える。差 $(\hat{p_1} - \hat{p_2})$ の平均値と分散は，$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2, V(\hat{p_1} - \hat{p_2}) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ となる。２つの母比率に差が無いならば，$p_1 = p_2 = p$ とおけるはずなので，$V(\hat{p_1} - \hat{p_2}) = p(1-p)(1/n_1 + 1/n_2)$ となる。この p の推定値として，$\hat{p} = (r_1 + r_2)/(n_1 + n_2)$ を使い，$\hat{q} = 1 - \hat{p}$ とおけば，$n_1 p_1$ と $n_2 p_2$ がともに 5 より大きければ，標準化して正規近似を使い，

$$Z = \frac{\hat{p_1} - \hat{p_2} - E(\hat{p_1} - \hat{p_2})}{\sqrt{V(\hat{p_1} - \hat{p_2})}} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} \sim N(0,1)$$

によって[*11]検定できる。

数値計算をしてみるため，仮に，患者群 100 名と対照群 100 名で，喫煙者がそれぞれ 40 名，20 名だったとする。喫煙率に 2 群間で差がないという帰無仮説を検定するには，

```
p <- (40+20)/(100+100)
q <- 1-p
Z <- (abs(40/100-20/100)-(1/100+1/100)/2)/sqrt(p*q*(1/100+1/100))
2*(1-pnorm(Z))
```

より，有意確率が約 0.0034 となるので，有意水準 5%で帰無仮説は棄却される。つまり，喫煙率に 2 群間で差がないとはいえないことになる。

差の 95%信頼区間を求めるには，サンプルサイズが大きければ正規分布を仮定できるので，原則どおりに差から分散の平方根の 1.96 倍を引いた値を下限，足した値を上限とすればよい。この例では，

```
dif <- 40/100-20/100
vardif <- 40/100*(1-40/100)/100+20/100*(1-20/100)/100
difL <- dif - qnorm(0.975)*sqrt(vardif)
difU <- dif + qnorm(0.975)*sqrt(vardif)
cat("喫煙率の差の点推定値=",dif,"  95%信頼区間= [",difL,",",difU,"]\n")
```

より，[0.076,0.324] となる。しかし，通常は連続性の補正を行うので，下限からはさらに $(1/n_1 + 1/n_2)/2 = (1/100 + 1/100)/2 = 0.01$ を引き，上限には同じ値を加えて，95%信頼区間は [0.066,0.334] となる。

[*11] この Z は離散値しかとれないため，連続分布である正規分布による近似の精度を上げるために，連続性の補正と呼ばれる操作を加え，かつ $p_1 > p_2$ の場合（つまり $Z > 0$ の場合）と $p_1 < p_2$ の場合 (つまり $Z < 0$ の場合）と両方考える必要があり，正規分布の対称性から絶対値をとって $Z > 0$ の場合だけ考え，有意確率を 2 倍する。即ち，

$$Z = \frac{|\hat{p_1} - \hat{p_2}| - (1/n_1 + 1/n_2)/2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

として，この Z の値が標準正規分布の 97.5%点（R ならば `qnorm(0.975,0,1)`）より大きければ有意水準 5%で帰無仮説を棄却する。

R には，こうした比率の差を検定するための関数 prop.test() が用意されており，以下のように簡単に実行することができる。

```
smoker <- c(40,20)
pop <- c(100,100)
prop.test(smoker,pop)
```

母比率の推定と，その差があるかどうかの検定[*12]，差の 95%信頼区間を一気に出力してくれる。上で一段階ずつ計算した結果と一致することを確かめてみよう。

prop.test() 関数は，３群以上の間でも，「どの群でも事象の生起確率に差がない」という帰無仮説を検定するのに使える。その帰無仮説が棄却されるときに，どの群間で差があるのかをみるには，検定の多重性が生じるので，平均値の差の場合と同様，第一種の過誤を調整する必要があり，ボンフェローニの方法やホルムの方法を用いることができる。R の関数は pairwise.prop.test() である[*13]。

例題

ある IT 系の企業の健診時に得たアンケート結果の集計によれば，部別の喫煙頻度は，総務部が 214 人中 42 人，営業部が 658 人中 242 人，開発部が 327 人中 122 人だった。この企業の喫煙率は部によって差があるといえるか？

下枠内のように入力すれば，部によって差がないという帰無仮説の検定の結果，得られる有意確率は約 7.5×10^{-6} なので有意水準 5%で帰無仮説は棄却され，「部によって差がないとはいえない」ことと，ホルムの方法で第一種の過誤を調整した多重比較の結果からは，総務部と営業部，総務部と開発部の喫煙率はそれぞれ有意水準 5%で有意な差があるが，営業部と開発部の喫煙率には差がないことがわかる。

```
smoker <- c(42,242,122)
names(smoker) <- c("総務","営業","開発")
pop <- c(214,658,327)
prop.test(smoker,pop)
pairwise.prop.test(smoker,pop)
```

課題

パプアニューギニアのある地方の，内陸，川沿い，海沿いの３つの村で，住民の悉皆調査によってマラリア原虫が血液中に検出される割合を調べた結果，内陸では 180 人中 6 人，川沿いでは 220 人中 10 人，海岸では 80 人中 18 人が原虫陽性だったとする。村によってマラリア原虫陽性割合には差があるか検討せよ。

付加的な情報としては，マラリア原虫を媒介するハマダラカの相対的な密度が，内陸を 1 とすると川沿いでは 2，海沿いでは 4 程度になるということがわかっているものとする。余裕があれば，ハマダラカの密度が高くなるほどマラリア原虫陽性割合が上昇する傾向があるかどうか検討してみよう。

結果は配布する紙に学籍番号，氏名と共に自筆して提出すること。結果の提出をもって出席確認とする。

[*12] 連続性の補正済み，事象が生起しない場合についても考慮してカイ二乗適合度検定をしているのだが，この操作は次回説明する 2 つの変数の独立性のカイ二乗検定と数学的に等価である。

[*13] なお，３群以上の間で事象の生起確率に一定の傾向がみられるかどうかを調べたい場合には，コクラン＝アーミテージの検定という手法があり，R では，
prop.trend.test(**事象生起数，観察総数，傾向を示すためのスコア**) によって実行できる。?prop.trend.test とすれば詳細な説明が表示される。