医学情報処理演習第５回「データの分布と検定の概念」[*1]

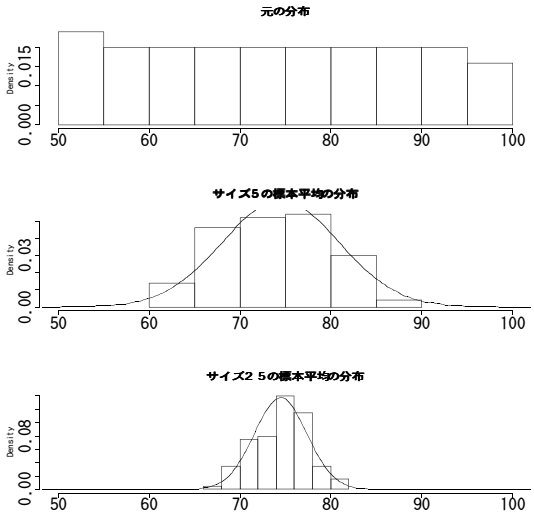
2006 年 11 月 6 日 中澤　港 (nminato@med.gunma-u.ac.jp)

前回の課題の回答例

中心極限定理を確かめてみるための課題であった。解答例を作るためのプログラムは以下の通りである。母集団のデータがすべてあるので，tsd として母集団での標準偏差を計算する関数を定義し，標本平均の分布が近づくであろう理論分布を赤い点線でヒストグラムに重ね描きするものである。

it04-ans-2006.R

```
X <- rep(50:99,1000)
tsd <- function(XX) { sqrt(var(XX)*(length(XX)-1)/length(XX)) }
RNGkind("Mersenne-Twister")
set.seed(1)
layout(1:3)
hist(X,xlim=c(50,100),freq=F,main="元の分布")
Z5 <- rep(0,100)
for (i in 1:100) { Z5[i] <- mean(sample(X,5)) }
hist(Z5,xlim=c(50,100),freq=F,main="サイズ 5 の標本の平均の分布")
curve(dnorm(x,mean(X),tsd(X)/sqrt(5)),add=T,col="red",lty=2)
Z25 <- rep(0,100)
for (i in 1:100) { Z25[i] <- mean(sample(X,25)) }
hist(Z25,xlim=c(50,100),freq=F,main="サイズ 25 の標本の平均の分布")
curve(dnorm(x,mean(X),tsd(X)/sqrt(25)),add=T,col="red",lty=2)
```



標本サイズを大きくすると，標本平均の分布は，平均が母平均で標準偏差が母集団の標準偏差を標本サイズの平方根で割った値の正規分布に近づき，しかもそのばらつきが小さくなることがわかる。

[*1] 本資料は http://phi.med.gunma-u.ac.jp/medstat/it05.pdf としてダウンロード可能である。

はじめに

量的なデータの場合はヒストグラムや正規確率プロットによって分布の様子をみることができるし，カテゴリデータの場合は各カテゴリの度数分布図（あるいは割合）をみることによって，分布の様子をみることができる。今回は，いくつかの理論分布を紹介し，データの分布が理論分布に適合しているかどうかを調べる検定法を紹介し，併せて検定の考え方そのものにも説明を加える。仮説検定という考え方は，伝統的な統計解析の中では，かなり重要な部分を占めてきたので，ここできちんと整理しておく。

ベルヌーイ試行と２項分布

まずはカテゴリデータの分布から説明する。１回の実験で事象 S か事象 F のどちらかが起こり，しかもそれらが起こる可能性が，$Pr(S) = p, Pr(F) = 1 - p = q$ で何回実験しても変わらないとき，これを「ベルヌーイ試行」という。ベルヌーイ試行では，事象 F は事象 S の余事象になっている。

例えば，不透明な袋に黒い玉と白い玉が 500 個ずつ入っていて，そこから中を見ないで１つの玉を取り出して色を記録して（事象 S は「玉の色が黒」，事象 F は「玉の色が白」）袋に戻す実験はベルヌーイ試行である[*2]。

ベルヌーイ試行を n 回行って，S がちょうど k 回起こる確率は，$Pr(X = k) =_n C_k p^k q^{n-k}$ である。${}_nC_k$ は言うまでもなく n 個のものから k 個を取り出す組み合わせの数である。２項係数と呼ばれる。このような確率変数 X は，「２項分布に従う」といい，$X \sim B(n, p)$ と表す。$E(X) = np$, $V(X) = npq$ である。

２項分布のシミュレーション

正二十面体（各面には１から 20 までの数字が割り振られている）サイコロを n 回 $(n = 4, 10, 20, 50)$ 投げたときの，１から４までの目が出る回数を１試行と考えれば，これはベルヌーイ試行である。１回投げたときに１から４までの目が出る確率は理論的には 0.2（＝母比率は 0.2）と考えられるので，試行 1000 セットの度数分布を描く R のプログラムは下記の通り。

it05-1-2006.R

```
times <- function(n) {
    dice <- as.integer(runif(n,1,21))
    hit <- sum(ifelse(dice<5,1,0))
return(hit)}

a <- c(4,10,20,50)
layout(matrix(1:4,nr=2))
for (i in 1:4) {
    y <- 1:1000
    for (k in 1:1000) { y[k] <- times(a[i]) }
    barplot(table(y),main=paste("n=",a[i]))
}
```

[*2] 注：袋に戻さないと１回実験するごとに事象の生起確率が変わっていくのでベルヌーイ試行にならない。なお，サンプリングとみれば，これは復元抽出である。

２項分布の理論分布

この例で，各 n についての理論的な確率分布は，$Pr(X = k) =_n C_k 0.2^k 0.8^{n-k}$ なので，図を描くための R のプログラムは下記の通り。

it05-2-2006.R

```
layout(matrix(1:4,nr=2))
a <- c(4,10,20,50)
for (i in 1:4) {
    n <- a[i]
    k <- 0
    chk <- 1:(n+1)
    names(chk) <- 0:n
    while (k <= n) {
        chk[k+1] <- choose(n,k)*(0.2^k)*(0.8^(n-k))
        k <- k+1
    }
    barplot(chk,main=paste("n=",n))
}
```

ただし，前回も触れた `dnorm()` や `dt()` など，R には様々な確率分布についての関数があり，`choose(n,k)*(0.2^k)*(0.8^(n-k))` は `dbinom(k,n,0.2)` と同値である。このように，確率変数が取りうる各値に対して，その値をとる確率を与える関数を**確率密度関数** (probability density function) という。値が小さいほうからそれを全部足した値を与える関数（つまり，その確率変数の標本空間の下限から各値までの確率密度関数の定積分）を分布関数（あるいは**確率母関数** (probability generating function)，累積確率密度関数）と呼ぶ。

２項分布の確率変数の定義域は整数値なので，飛び飛びの値となる。その意味で，このような分布を**離散分布**という。離散分布には，２項分布の他には，ポアソン分布などがある[*3]。それに対して，正規分布や t 分布など，確率変数の定義域が実数である分布を，**連続分布**という。

正規分布

n が非常に大きい場合は，２項分布 $B(n,p)$ の確率 $Pr(X = np + d)$ という値が，

$$\frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{d^2}{2npq}\right)$$

で近似できる。一般にこの極限（n を無限大に限りなく近づけた場合）である，

$$Pr(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

という形をもつ確率分布を正規分布と呼び，$N(\mu, \sigma^2)$ と書く。

$z = (x - \mu)/\sigma$ と置けば，

$$Pr(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

[*3] ポアソン分布は，独立した事象の生起件数の分布であり，例えば，まったく意図的な出産抑制がない集団での完結出生児数の分布はポアソン分布に従うことが期待される。

となる。これを標準正規分布と呼び，$N(0,1)$ と書く。

既にやったように，R で標準正規分布の確率密度関数を $[-5,5]$ の範囲でプロットするには，`curve(dnorm(x),-5,5)` と打てば良い。`curve()` 関数は，連続分布をプロットするときに，定義域を `x` として，始点と終点をコンマで区切って与えれば曲線を描画してくれるので，とても便利な関数である。重ね描きする場合は，`add=T` を引数リストに加える。例えば，いま書いた標準正規分布の確率密度関数のグラフの上に，同じ範囲で平均 1，標準偏差 2 の正規分布の確率密度関数を赤い破線で重ね描きするには，`curve(dnorm(x,1,2),add=T,col="red",lty=2)` とすればよい。

標準正規分布の 97.5%点（その点より小さい値をとる確率の積分値が 0.975 になるような点。その点を与える関数を分位点関数と呼ぶ）を得るには，`qnorm(0.975)` とすればよいし，-1.96 より小さな値をとる確率を得るには，`pnorm(-1.96)` とすればよい。

R では，一般に，分布名が `hoge` だとすると（注：念のため書いておくが `hoge` などという名前の分布は存在しないが），確率密度関数が `dhoge()`，確率母関数が `phoge()`，分位点関数が `qhoge()` で得られる。また，その分布に従う `n` 個の乱数を得るには，`rhoge(n)` とする。

χ^2 分布

$X_1, X_2, ..., X_v$ が互いに独立に標準正規分布 $N(0,1)$ に従うとき，

$$V = \sum_{i=1}^{v} {X_i}^2$$

の分布を自由度 v の χ^2 分布という。この分布の確率密度関数は，

$$f(x|v) = \frac{1}{2\Gamma(v/2)} \left(\frac{v}{2}\right)^{v/2-1} \exp\left(-\frac{v}{2}\right)$$

である。

なお，言うまでもないが，Γ はガンマ関数で，正の実数 α に対して，

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$$

であり，正の整数 α に対しては $\Gamma(\alpha) = (\alpha-1)!$ である。

$E(x) = v$ であり，$V(x) = 2v$ である。自由度 1 の χ^2 分布を $[0,10]$ の範囲でプロットするには，`curve(dchisq(x,1),0,10)` とすればよい。他の自由度のものを重ね描きするには，例えば自由度 2 の χ^2 分布を赤破線で重ね描きしたければ，`curve(dchisq(x,2),0,10,add=T,col="red",lty=2)` とすればよい。

自由度 1 の χ^2 分布の 95%点を得るには，`qchisq(0.95,1)` とすればよいし，3.84 より小さな値をとる確率を得るには，`pchisq(3.84,1)` とすればよい。

t 分布

標準正規分布に従う確率変数 U と，自由度 v の χ^2 分布 $\chi^2(v)$ に従う確率変数 V があり，それらが独立のとき，

$$T = U/\sqrt{V/v}$$

が従う分布のことをステューデントの t 分布という。この確率密度関数は

$$f(t) = \frac{\Gamma((v+1)/2)}{\sqrt{v}\Gamma(1/2)\Gamma(v/2)}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}$$

である。これは，ステューデントというペンネームで論文を書いていたギネス社の技師ゴセット (Gosset WS) が初めて導いた分布である。

自由度 20 の t 分布の確率密度関数を $[-5,5]$ の範囲でプロットするには，`curve(dt(x,20),-5,5)` とすればよい。これが標準正規分布より裾が長い分布であることを見るために標準正規分布を赤い点線で重ね描きするには，続けて，`curve(dnorm(x),-5,5,add=T,col="red",lty=2)` とすればよい。

また，自由度 20 の t 分布の 97.5%点を得るには，`qt(0.975,20)` とすればよいし，2 より小さな値をとる確率を得るには，`pt(2,20)` とすればよい。

F 分布

V_1 と V_2 が独立で，自由度がそれぞれ ν_1，ν_2 の χ^2 分布に従う統計量であるとする。このとき，

$$F = \frac{V_1/\nu_1}{V_2/\nu_2}$$

が従う分布を自由度 (ν_1,ν_2) の F 分布という。F 分布の確率密度関数は，

$$f(F) = \frac{1}{B(\frac{\nu_1}{2},\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\frac{F^{(\nu_1/2)-1}}{(1+\frac{\nu_1}{\nu_2}F)^{(\nu_1+\nu_2)/2}}$$

で与えられる。$B(\alpha,\beta)$ はベータ関数で，ガンマ関数を用いれば

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

と書ける。自由度 (ν_1,ν_2) の F 分布を $F(\nu_1,\nu_2)$ と書き，その上側 100α%点を $F_\alpha(\nu_1,\nu_2)$ と書く。R で第 1 自由度（$=\nu_1$）9，第 2 自由度（$=\nu_2$）14 の F 分布の確率密度関数を $[0,10]$ の範囲でプロットするには，`curve(df(x,9,19),0,10)` とすればよいし，同じ F 分布の 95%点（上側 5%点）を得るには `qf(0.95,9,19)` とすればよいし，同じ F 分布に従う統計量が 5 より小さな値をとる確率を得るには `pf(5,9,19)` とすればよい。

検定の考え方と第一種，第二種の過誤

検定とは，帰無仮説（一般には，差がない，という仮説）の元で得られた統計量を，既知の確率分布をもつ量と見た場合に，その値よりも外れた値が得られる確率（これを「有意確率」と呼ぶ）がどれほど小さいかを調べ，有意水準[*4]より小さければ，統計的に意味があることと捉え（統計的に有意である，という），帰無仮説

[*4] 分析者が決める一定の確率。当該研究分野の伝統に従うのが普通である。先行研究があればそれに従う。他に基準がなければ 5%か 1%にすることが多い。

がおかしいと判断して棄却する（つまり，「差がないとは言えない」と判断する）という意思決定を行うものである。

この意思決定が間違っていて，本当は帰無仮説が正しいのに，間違って帰無仮説を棄却してしまう確率は，有意水準と等しいので，その意味で，有意水準を第一種の過誤と（αエラーとも）呼ぶ（逆に，本当は帰無仮説が正しくないのに，その差を検出できず，有意でないと判断してしまう確率を，第二種の過誤と（βエラーとも）呼び，1 から第二種の過誤を引いた値が検出力になる)。

注意しなければいけないのは，有意確率の小ささは，あくまで，帰無仮説のありえなさを示すだけであって，**差の大きさを意味するのではない**点である。ここを勘違いしたレポートなどが時折見られるので注意されたい。

両側検定と片側検定

2 つの量的変数 X と Y の平均値の差の検定をする場合（平均値の差の検定についてはまた次回詳しく触れる)，それぞれの母平均を μ_X，μ_Y と書けば，その推定量は $\mu_X = \mathrm{mean}(X) = \sum X/n$ と $\mu_Y = \mathrm{mean}(Y) = \sum Y/n$ となる。

両側検定では，帰無仮説 H0 : $\mu_X = \mu_Y$ に対して対立仮説（帰無仮説が棄却された場合に採択される仮説）H1 : $\mu_X \neq \mu_Y$ である。H1 を書き直すと，「$\mu_X > \mu_Y$ または $\mu_X < \mu_Y$」ということである。つまり，t_0 を「平均値の差を標準誤差で割った値」として求めると，t_0 が負になる場合も正になる場合もあるので，有意水準 5%で検定して有意になる場合というのは，t_0 が負で t 分布の下側 2.5%点より小さい場合と，t_0 が正で t 分布の上側 2.5%点（つまり 97.5%点）より大きい場合の両方を含む。t 分布は原点について対称なので，結局両側検定の場合は，上述のように差の絶対値を分子にして，t_0 の t 分布の上側確率[*5]を 2 倍すれば有意確率が得られることになる。

片側検定は，先験的に X と Y の間に大小関係が仮定できる場合に行い，例えば，X の方が Y より小さくなっているかどうかを検定したい場合なら，帰無仮説 H0 : $\mu_X \geq \mu_Y$ に対して対立仮説 H1 : $\mu_X < \mu_Y$ となる。この場合は，t_0 が正になる場合だけ考えればよい。有意水準 5%で検定して有意になるのは，t_0 が t 分布の上側 5%点（つまり 95%点）より大きい場合である。なお，R で平均値の差の検定を行うための関数は，平均値の信頼区間のところでも出てきた `t.test()` だが，詳しくは次回説明する。

分布の正規性の検定

高度な統計解析をするときには，データが正規分布する母集団からのサンプルであるという仮定を置くことが多いが，それを実際に確認することは難しいので，一般には，分布の正規性の検定を行うことが多い。考案者の名前から Shapiro-Wilk の検定と呼ばれるものが代表的である。

Shapiro-Wilk の検定の原理をざっと説明すると，$Z_i = (X_i - \mu)/\sigma$ とおけば，Z_i が帰無仮説「X が正規分布にしたがう」の下で $N(0,1)$ からの標本の順序統計量となり，$c(i) = E[Z(i)], d_{ij} = Cov(Z(i), Z(j))$ が母数に無関係な定数となるので，「$X(1) < X(2) < ... < X(n)$ の $c(1), c(2), ...c(n)$ への回帰が線形である」を帰無仮説として，そのモデルの下で σ の最良線形不偏推定量 $\hat{(\sigma)} = \sum_{i=1}^{n} a_i X(i)$ と $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ を用いて，$W = (k\hat{\sigma}^2)/S^2$ を検定統計量として検定するものである。k は $\sum_{i=1}^{n}(ka_i)^2 = 1$ より求められる。

[*5] t 分布の確率密度関数を t_0 から無限大まで積分した値，即ち，t 分布の分布関数の t_0 のところの値を 1 から引いた値。R では `1-pt(t0, 自由度)`。

R で数値型変数 X の分布が正規分布にフィットしているかどうかを検定するには,

```
shapiro.test(X)
```

とすればよい。変数 X のデータ数（ベクトルの要素数，R のコードでは `length(X)`）は，3 から 5000 の間でなければならない。2 以下では分布を考える意味がなく，また，検定統計量 W の分布がモンテカルロシミュレーションによって得られたものであるため，あまりに大きなサンプルサイズについては値が与えられていないのである。

例題

`http://phi.med.gunma-u.ac.jp/medstat/p03.txt` にあるパプアニューギニア成人男性の体重データは正規分布に従っているといえるか検定せよ

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p03.txt")
shapiro.test(dat$WT)
```

とすると，`W = 0.9799, p-value = 0.8473` と表示される。`p-value` が 0.05 よりずっと大きいので，この成人男性の体重データが正規分布に従っているという帰無仮説の下でこのようなデータが偶然得られることは十分考えられる。従って，**正規分布に従っているという帰無仮説は棄却されない。**

正規性の検定にはたくさんの方法が提案されているが，ここではもう 1 つだけ紹介しておこう。提案者の名前から，ギアリー (Geary) の検定と呼ばれるものである。現在のところ，R にはデフォルトでは入っていないが，比較的簡便で使いやすい検定である。以下，簡単に原理を説明する。

左右対称な分布について，裾の長さを，平均値のまわりの 1 次の絶対モーメントを 2 次のモーメントの平方根で割ったもので測ることにすると，その一致推定量 G が,

$$G = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{\sqrt{n \sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

となる。この G を用いて帰無仮説 H0:「データ X が正規分布からの標本」を検定することができる。対立仮説の下での分布が正規分布よりも裾が長い対称分布（例えば t 分布のような）であれば $G < g_0$ のとき帰無仮説を棄却する。g のパーセント点については，u_α を標準正規分布の $100\alpha\%$点として，n が大きければ近似的に

$$g(\alpha; n) \simeq \sqrt{\frac{2}{\pi}} + u_\alpha \sqrt{1 - \frac{3}{\pi}} \frac{1}{\sqrt{n}}$$

で得られることがわかっている。R のプログラムは下記のように定義すれば，Geary の正規性の検定を行う関数ができる。

it05-3.R

```
geary.test <- function(X) {
 m.X <- mean(X)
 l.X <- length(X)
 G <- sum(abs(X-m.X))/sqrt(l.X*sum((X-m.X)^2))
 p <- 1-pnorm((G-sqrt(2/pi))/sqrt(1-3/pi)*sqrt(l.X))
 cat("Geary's test for normality:\n G=",G," / p=",p,"\n")
}
```

なお，作図の説明で触れた hist(X) で全体の様子をみたり，qqnorm(X) をしてみるのも，分布の正規性をチェックするにはいい方法である。qqnorm(X) で描かれるグラフは，X が正規分布に従っていれば直線に乗るはずであり，外れているときにどのように外れているかが見える。

課題

MASS ライブラリに含まれている低体重出生についてのデータフレーム birthwt 内に含まれている出生体重を示す変数 (bwt) が正規分布に従っていると言えるかどうか，作図により検討した上で検定せよ。なお，MASS ライブラリ内のデータフレームを使うには，最初に library(MASS) とすればよい。答えだけでなく手順も書くこと。学籍番号と氏名とともにプリンタに出力後，署名して提出するのはいつもと同じである。