前回のQ＆A

Q1) 先生がパソコンでうつす画面はホームページでみれますか？

A1) 正しい日本語では「見られますか？」と書くべき。とはいえ，尤もなリクエストなので，パワーポイントファイルもダウンロードできるようにしました。

Q2) 説明が速すぎてフォローできなかったのでもう一度説明を希望。

A2) 同様の希望が別々の人から別々の項目についてあったので，それをやっていると今回の講義に入れません。配布資料をじっくり読んで，それでもわからなかったら，E115 まで聞きに来てくれれば個別に対応します。月曜の夕方以降か火曜の午前中が都合がいいです。

Q3) 「Grimm は間違っている」と書いてあるが，それならどうしたらいい？

A3) 同順位の値がある場合の Grimm の考え方は，中央値や四分位範囲，四分位偏差などの順序統計量を正確に計算することを意図していますが，実際には多くの統計ソフトがそこまで考慮しない計算結果を出してきます。ですから，通常は無視して構いません。しかし，同順位の値が多いデータを扱うときに，統計的に意味があるかどうかぎりぎりの結果が出た場合は，Grimm のような考え方をする必要があります。ただし Grimm の教科書の記載は論理的整合性が弱い点があるので，前回配布した資料に書いてあるように修正したアルゴリズムを使って計算すればいいです。

Q4) $\bar{X}$ は X でないという意味ではないのですか？

A4) 集合論では $\bar{X}$ は集合 X の補集合の意味で使われますが，代数では確率変数 X の標本平均が $\bar{X}$ で表されます。同じような記号が別の意味で使われるので混乱するかもしれません。補集合は X^C という表し方がされる場合も多いのですが，標本平均は $\bar{X}$ と表すのが普通です。

統計学第５回　比率に関する推定と検定

（１）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を分析する方法に入る。まずは変数が１つの場合を考える。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[0]。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 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？

（２）推定値の確からしさ

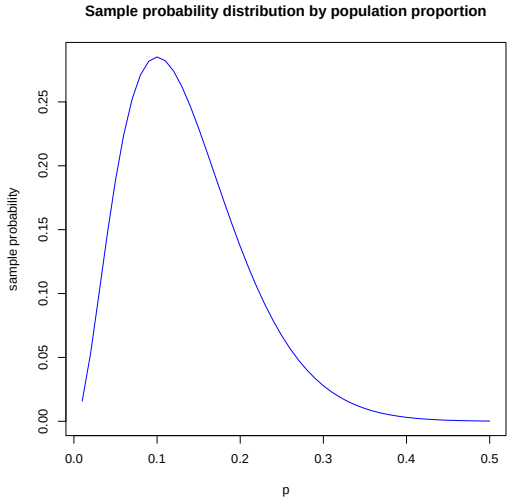
ここで，このようにして求めた推定値がどれほど確からしいか？　を考えよう。例えば，黒石の割合（母比率）が p である容器から 20 個の石を取り出したときに，黒石がちょうど 2 個である確率を考えると，これは二項分布に従う。[1]

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

[0] 配布時は,「白い石が占めていた割合（標本比率）に，加えた黒い石の数をかければ，元の白い碁石の数が推定できる」と書いていたが，もちろんそれは誤りである。

[1] 厳密に考えると，この確率が二項分布に従うためには復元抽出でなければならないので，1 個取り出しては戻す必要がある。CMR ではそうしないが，それは標本に比べて母集団が十分に大きく，抽出操作が母比率に影響を与えないと仮定できるからである。

るので ${}_{20}\mathrm{C}_2$ 回だけそれを足し合わせた確率になる。[2] この値が最大となるのは，下図のように $p = 0.1$ の時である。



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 個とか 324 個に比べて，それほど信頼性は高くない。

（3）信頼区間

では，ある程度の信頼性が見込める範囲を示すことは可能だろうか？　という考え方で示されるのが信頼区間である。例をあげよう。ビデオリサーチ (http://www.videor.co.jp/index.html) によれば，NHK の朝のテレビ小説「ほんまもん」の 10 月 8 日の関東地区の視聴率は 22.9%であった。関東地区の調査対象世帯は 600 だから，137 世帯が見ていたことになる。このとき，関東地区全体の真の視聴率（母比率）は，どのくらいの範囲をとれば，95%の確率でその中に収まるのか？　というのが問題である。

「ほんまもん」を見る／見ないという事象が各世帯独立に起こるとすれば，二項分布で考えることができる。母比率が 137/600 の時にちょうど 137 世帯が見た（裏返して言えば 463 世帯が見なかった）確率は，`choose(600,137)*(137/600)^137*(463/600)^463` で，たかだか 3.9%に過ぎない。

しかし，例えば，母比率が 10%だったのに 137 世帯が見たという確率は，`2.5*10^(-20)` であり，まったくありそうにない。137/600 の前後適当な幅をとれば，かなり高い確率で，ちょうど 137 世帯が見た，という事象が起こることになる。この幅を「信頼区間」という。95%の確率でちょうど 137 世帯が見たという事象が起こるための母比率の推定幅を,「95%信頼区間」という。

95%信頼区間を求めるには，下側 2.5%の点と上側 2.5%の点を求めればよいので，R なら `z<-0; k<-0; while (z<0.025) {k <- k+1; z <- z+zz[k]}` ; `k` として下側 2.5%の点を求め，`z<-0; k<-600; while (z<0.025) {k <- k-1; z <- z+zz[k]}` ; `k` として上側 2.5%の点を求めればよい。

結果として，600 世帯の調査で 22.9%の視聴率だったら，無限母集団の視聴率（真の視聴率）の 95%信頼区間は，19.8%から 26.5%の間と言える。

（4）正規近似による信頼区間の推定

二項分布は，n が大きいときは正規分布で近似できる。このことを利用すれば，母比率 p，標本数（調査世帯数）n で，その標本の中で注目している属性をもつ標本数（「ほんまもん」を見た世帯数）を X，観測比率を $p' = X/n$ とすれば，X が近似的に正規分布 $N(np, np(1-p))$ に従うことになる。正規分布の 95%のサンプルは，平均 $\pm$ 標準偏差$\times$ 1.96 に含まれるので，

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

これから式変形すると，$\mathrm{Prob}[p' - 1.96\sqrt{p'(1-p')/n} \leq p \leq p' + 1.96\sqrt{p'(1-p')/n}] = 0.95$ となるので，母比率 p は 95%の確率で $(p' - 1.96\sqrt{p'(1-p')/n}, p' + 1.96\sqrt{p'(1-p')/n})$ 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。

[2] R では `choose(20,2)*p^2*(1-p)^18` あるいは `dbinom(2,20,`p`)` で得られる。

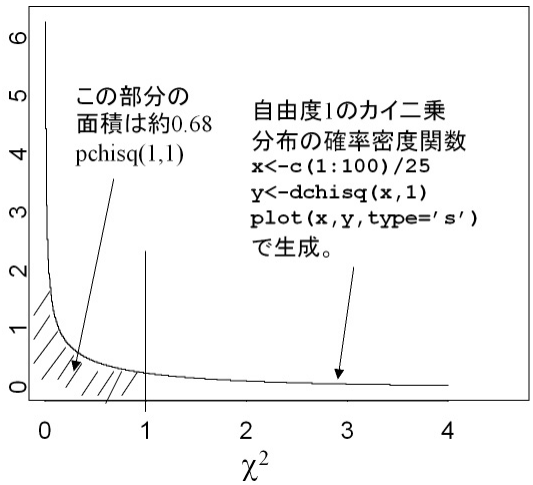
練習： ある大学の正門の前で，ある朝登校して来る学生の男女比を調べてみたところ，300 人中，女子学生が 75 人であった。この大学の女子学生の割合の点推定値と 95%信頼区間を求めよ。[3]

（5）母比率の検定

予め母比率について何らかの期待があるとき（50%であるとか），標本から推定された母比率が，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。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一致するという仮説（帰無仮説）が成り立っている確率を調べて，それが普通では考えられないほど小さい場合（通常は 5%未満）に，滅多にないことだから偶然ではない（これを「有意である」という），と考えて帰無仮説を棄却する。

カテゴリ数が全部で n 個あるとき，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[4]，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，不明な母数があるときは，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）。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（適合しないと判断する）。

ちなみに，自由度 1 のカイ二乗分布は，下図のような形になる。χ^2 値が 1 より大きくなる確率は，約 0.32 ということである。 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，標本の数から，前もって推定する母数の数を引いた値である。[5] 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 1 を引く。



分布関数（確率母関数）は，確率密度関数を積分したものであり，図で見れば面積に当たる。その逆関数，つまり面積に対応する値を与える関数を分位点関数という。自由度 1 のカイ二乗分布の場合，R では，カイ二乗値 x について，確率密度関数が `dchisq(x,1)`，分布関数が `pchisq(x,1)` で与えられ，95%点を与える分位点関数が `qchisq(0.95,1)` で与えられる。これは，χ^2 値が `qchisq(0.95,1)` 以下である確率が 95%であることを意味する。逆にいえば，χ^2 値が `qchisq(0.95,1)` より大きくなることは，確率 5%もない，滅多にないことである。言い換えると，「観測された分布が期待される分布と一致している可能性は 5%もない」ということである。このようなとき，「観測された分布が期待される分布と違いがない」という仮説は有意水準 5%で棄却されたといい，観測された分布は期待される分布と一致するとはいえない，と解釈する。

---

[3] ヒント：95%信頼区間の下限を求める R の式は，`75/300−2*sqrt(75/300*225/300/300)` である。なお，この推定には，朝登校して来る学生に男女の偏りがないという仮定があるので，実は真の値を過大評価することになっている。では，どうすれば正しい推定ができるような標本がとれるだろうか？

[4] χ は「カイ」と発音する。英語では chi-square と書かれるので，間違って「チ 2 乗」とか言わないように。

[5] 合計を母数と考えなければ，標本数から 1 を引いて母数の数を引く，と捉えてよい。前回，不偏分散を求めるときにも標本数から 1 を引いて自由度を求めた。

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）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？[6]

応用： 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？[7]

- 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。
- 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，(0 × 79 + 1 × 61 + 2 × 13 + 3 × 1 + 4 × 1)/155 で得られる。R で書けば，この値を Ehh に保存するとして，cc< −c(0:4); hh< −c(79,61,13,1,1); Ehh< −sum(cc*hh)/sum(hh) となる。
- 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判定すれば良さそうなものだが，そうはいかない。
- epp の値を見ればわかるが，epp[cc==4] が 1 より小さいのである。期待度数が 1 より小さいときはカテゴリを併合しなくてはならないので，epp[cc==4] を epp[cc==3] と併合する。
- 即ち，epp[cc<3]− >ep; epp[cc==3]+epp[cc==4]− >ep[cc==3]; ep< −ep[!is.na(ep)] として期待度数の分布 ep，hh[cc<3]− >h; hh[cc==3]+hh[cc==4]− >h[cc==3]; h< −h[!is.na(h)] 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，与えられたデータの 1 日の交通事故件数がポアソン分布に従っている確率は約 19%あり，ポアソン分布に従っていないとはいえないことになる。

---

[6] （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 人という観察値が得られる確率が計算できる。その確率がきわめて小さければ（通常 5%未満），統計的に意味があるほど有り得なさそうな（「統計的に有意な」という）現象であると考えて，仮説を棄却する。実は，この場合は母比率が 0.5 であるとして 2 項分布で計算してもよい。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 回中ちょうど 480 回起こる確率は，choose(900,480)*0.5^480*0.5^420 で与えられるが，R には二項分布についてもカイ二乗分布と同じように確率密度関数を与える関数があり，この確率は dbinom(480,900,0.5) で与えられる。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(480,900,0.5) で計算できる。420 人以下になる確率は，dbinom(0,900,0.5) + dbinom(1,900,0.5) + ... + dbinom(420,900,0.5) であり，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(480,900,0.5) + pbinom(420,900,0.5) である。

[7] 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。