統計学第５回
「比率に関する検定と推定」

- 中澤　港 <minato@ypu.jp>

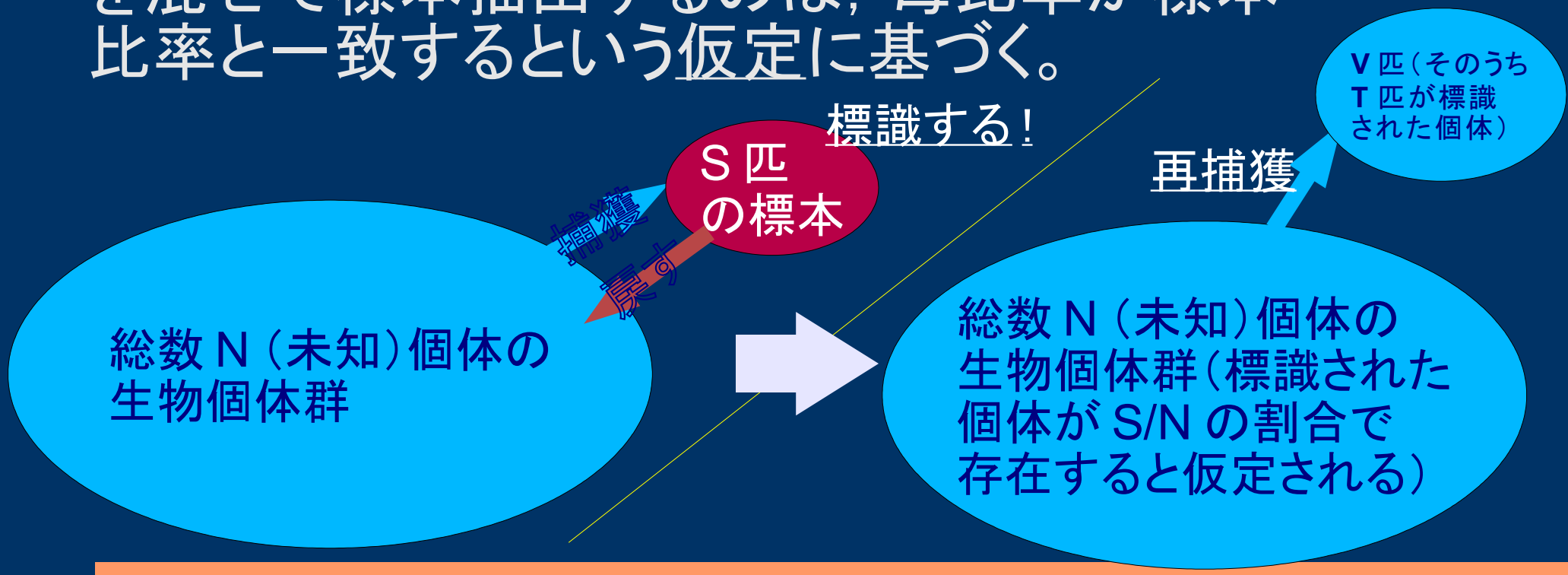
- http://phi.ypu.jp/statlib/l5-2003.pdf
- 今回のテーマは，標本から得られる比率（proportion）のデータを扱う方法である。
- 母集団における期待値と信頼区間を推定したり，与えられた母比率と有意に違うかどうかを検定したりする。

比率とは？

- カテゴリ変数（データは名義尺度をもつ）1つについて得られる情報は，データの総数と，個々のカテゴリが占める割合である。
- 総数に対して個々のカテゴリが占める割合を，「比率」(proportion) と呼ぶ。
- 通常，得られているデータは標本のデータなので，直接得られる比率は標本比率である。
- 知りたいのは，個々のカテゴリが母集団で占めるであろう割合，つまり母比率である。

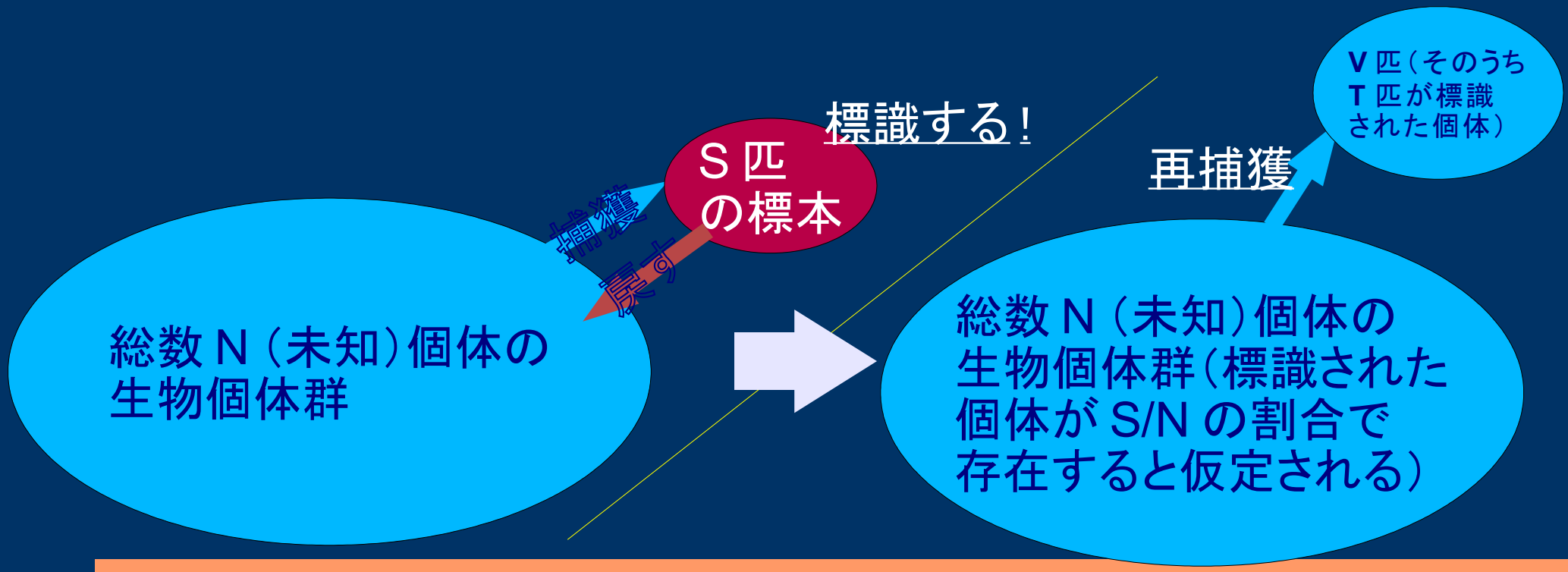
母比率を推定する方法

- 通常，母比率は標本比率と一致する。
- 生態学のリンカーン法（Capture-Mark-Recapture Method）で，母集団の個体数を推定するのに，既知の数のマークした個体を混ぜて標本抽出するのは，母比率が標本比率と一致するという仮定に基づく。



CMR での母集団個体数推定式

- N は未知。S，V，T は既知。
- 仮定: S/N = T/V
- 簡単な式変形で N = SV/T と計算されることがわかる。



例: 白い碁石の数は？

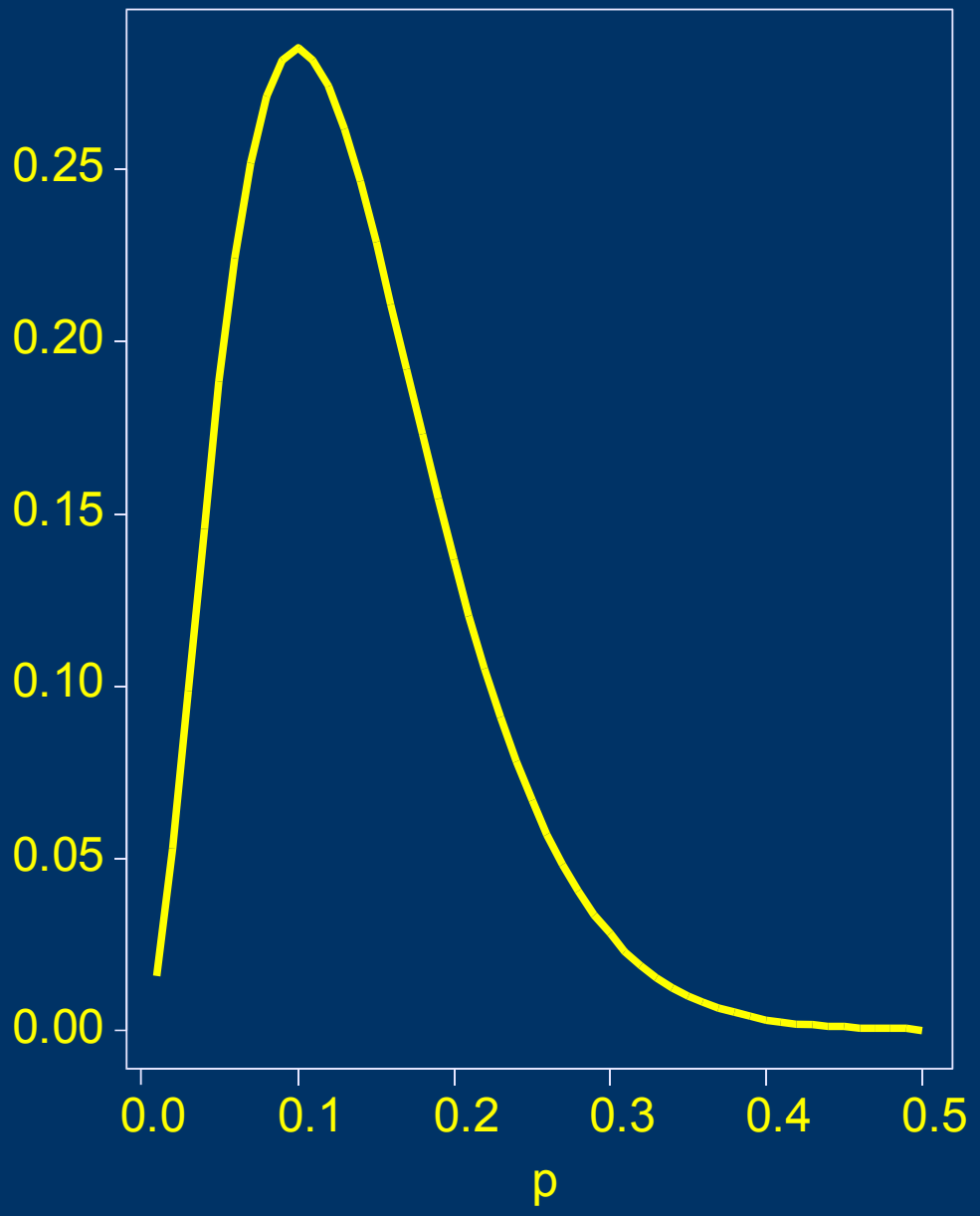
- 多数の白い碁石に 40 個の黒石を混ぜ，20 個取り出したときに黒石 2 個，白石 18 個だったときの，元の白石の数は？
  - 黒石の標本比率は 2/20=0.1
  - 母集団での黒石の個数が 40 個だから，40/0.1=400 個が母集団の総数。
  - 白石の数は，400-40=360 個
  - CMR と似ているが，最初の捕獲がないので，母集団総数から，加えた黒石の数を引かないと白石の数が出ないことに注意。

推定値の確からしさ

- 母比率が標本比率と一致するという仮定は尤もらしいので，推定値として標本比率を使うのは，まあいいことにしよう。
- 次に問題になるのは，その値がどれほど確からしいか？　である。
- サンプルサイズが 10 しかなければ，標本比率は 0%, 10%, ..., 100% の 11 通りの値しかとれない。真の母比率が 0.15 だったら，標本比率が一致することはありえなくなる。しかし，サンプルサイズが 100 ならば，真の母比率との差がより少ない標本比率が得られると期待される。

二項分布のプロット

- 母集団での黒石の割合が p のとき，20 個の標本を取り出して，ちょうど 2 個の黒石が得られる確率（二項分布に従う）を図示すると右図のようになる（p=0.1 は 0.11 とか 0.09 に比べて際立って高い可能性をもつ値とは言えない。0.01 の差は小さいと思うかもしれないが，推定される白石の数に直すと 36 個の差に相当する）。
- R では，
  ```
  p<-c(1:50)/100
  prob<-dbinom(2,20,p)
  plot(p,prob,type="l")
  ```
  で描ける。
- なお，dbinom(2,20,p) は ${}_{20}C_2 * p^2 * (1-p)^{(20-2)}$ と同値。



母比率の推定値の信頼区間

- では，どうやって推定値の確からしさを示す？
- 適当な幅をもって母比率 p を推定すれば，かなり高い可能性をもって真の母比率がその幅に入るということができる。
- この「適当な幅」が信頼区間である（「信頼限界」ということもあるが，「信頼区間」の方が普通）。
- 通常は，「かなり高い可能性」を 95% とした「 95% 信頼区間」を示す。

信頼区間の計算手順

1)分布を求める
2)下側 2.5% 点として 95% 信頼区間の下限を求め，上側 2.5% 点として 95% 信頼区間の上限を求める。
3)二項分布する変数など，計算が面倒だが，標本サイズが大きいとき（目安としては 50 以上）は正規近似すると楽。

例）視聴率の信頼区間

- ビデオリサーチ調べによると，2003 年 11 月 9 日 22:00 からの NHK 総合「衆院選開票速報」の関東地区の視聴率は 17.2% であった。
- 関東地区の調査対象世帯数（サンプルサイズ）は 600 だから，103 世帯が見ていたことになる。
- 母比率が 103/600 のとき，ちょうど 600 世帯中 103 世帯が見た確率は，`dbinom(103,600,103/600)` となるので，約 4.3% に過ぎない。その前後のケースを加えて，その範囲に入る確率が 95% になるような世帯数の範囲を求めるには，両端の 2.5% 点を求めればいい。

上側 2.5% 点と下側 2.5% 点の求め方

- テキストにあるように個々の確率を足せばいい。Rでは,

```
z <- 0; k <- 0; while (z<0.025) {
z <- z+dbinom(k,600,103/600); k <- k+1 }
ll <- (k-1)/600
z <- 0; k <- 600; while (z<0.025) {
z <- z+dbinom(k,600,103/600); k <- k-1 }
ul <- (k+1)/600
cat('95% CI=[',ll,',',ul,']\n')
```

- とすれば, この視聴率の 95% 信頼区間が 14.2% から 20.2% であることがわかる。
- `binom.test(103,600,103/600)` でも計算できる
- （但し若干計算方式が違うので微妙に違う結果）。

テキスト練習問題の回答例

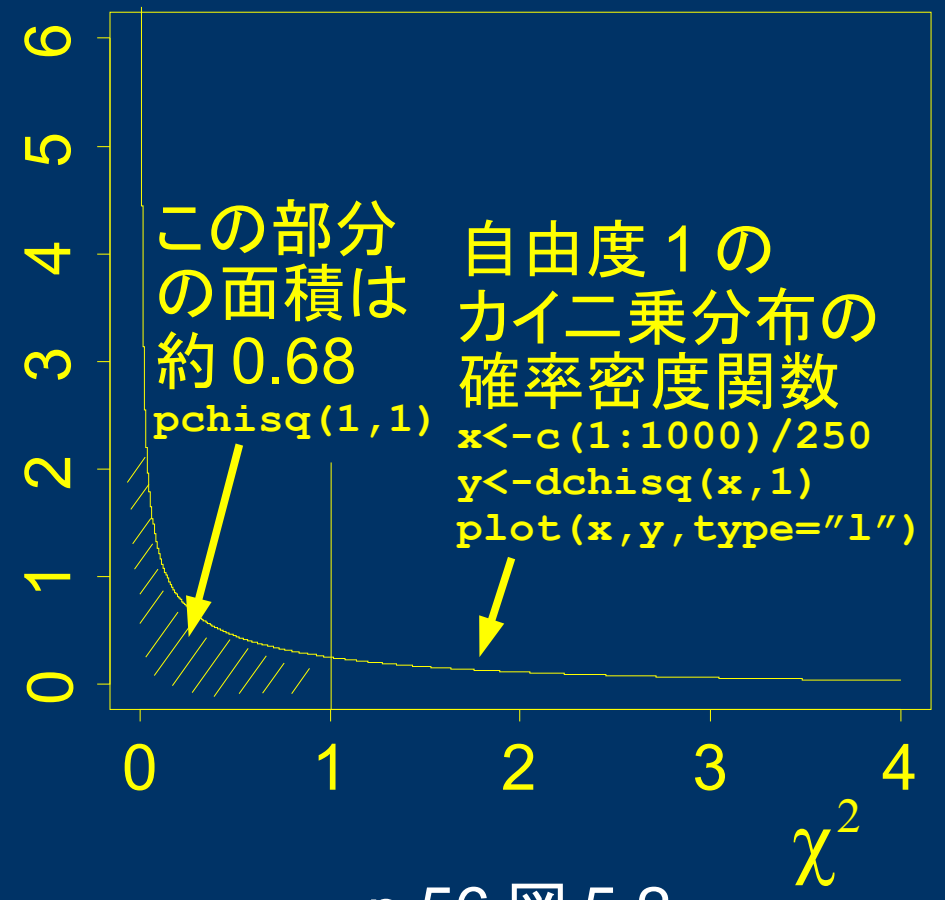
- 問題「ある大学正門前で学生の男女比を調べたら 300 人中 75 人が女性だった場合，大学の女子学生割合の点推定値と 95% 信頼区間を求めよ」
- この大学の女子学生の割合の点推定値は，標本比率と一致するはずなので，75/300=0.25 ，つまり 25% である。
- 95% 信頼区間の下限は，
  `75/300-2*sqrt(75/300*225/300/300)`
  と計算すれば（式中の 2 は正規分布の 97.5% 点である 1.96 の近似値），0.25-0.05=0.2 より 20% である。
- 上限は当然 30% となる。
- したがって，95% 信頼区間は，(20%, 30%) となる。

検定の考え方

- 検定とは，仮説が正しいかどうか確かめること。
- 「差がある」仮説を直接証明することは困難なので（どの程度の差があったら差があるとみなすのか？），「差がない」仮説（これを帰無仮説という）を検証する
- 母比率についての検定なら，標本比率が期待される母比率と差がない，という帰無仮説を調べる。
- 帰無仮説が成り立っている確率が統計的に意味があるほど小さい（そのレベルを有意水準といい，通常は5%未満とする）なら，帰無仮説を棄却する（＝標本比率は期待される母比率と差がないとは言えないことになり，その標本データから考えると期待される母比率が違っていると解釈する）。

母比率の検定

- n 個のカテゴリがあって，i 番目のカテゴリの観測度数（実際の標本数）が Oi，期待度数（期待される母比率と標本比率が一致した場合に標本が示すであろう度数）が Ei ならば，(Oi-Ei)^2/Ei をすべてのカテゴリについて足し合わせて得られる値 X は，自由度 n-1 のカイ二乗分布に従う（自由度は，n から前もって推定する母数の数を引いた値）。
- ちなみに，自由度 n のカイ二乗分布とは，独立に標準正規分布（平均 0，分散 1 の正規分布）に従う n 個の確率変数があったとき，それらの二乗の和が従う分布である。



p.56 図 5.2

例題2）出生 900 人中男児 480 人のとき，（1）男女半々仮説，（2）男児 1.06 倍仮説，は支持されるか？

- (1) 男女半々ならば期待される男児出産数は 900*0.5=450 。公式からカイ二乗値を計算すると，`(480-450)^2/450 + (420-450)^2/450 = 4` この値はカイ二乗分布の 95% 点 (3.84) より大きいので 5% もない現象と考え，仮説を棄却する（ 1-pchisq(4,1) より p=0.0455 ）。
- (2) 男児が女児の 1.06 倍なら，期待される出産数は男女それぞれ， EM<-900*1.06/2.06 と EF<-900*1/2.06 で得られる。カイ二乗値は (480-EM)^2/EM+(420-EF)^2/EF で得られる。

交通事故への応用例について

- １日あたりの交通事故件数の分布 f{0,1,2,3,>=4}={79,61,13,1,1} はポアソン分布に従うか？
- ポアソン分布は稀な事象についてのベルヌーイ試行を行うときの事象生起数が従う分布。意図的な出産抑制をしない集団では生涯出産数の分布もポアソン分布に従うことが知られている。
- 基本は，総数をポアソン分布 * に掛けて期待度数を計算し，それと実測値の差の二乗を期待度数で割ったものの総和が自由度 n-2 のカイ二乗分布に従うと考えて検定すればいい。
  * ポアソン分布を確定するにはその期待値（母数である）を推定しなくてはいけないので，平均事故件数を期待値とする。
- やってみよう！