現代日本学演習 V「実践的統計分析」

第2講　正規分布の利用

田中重人 (東北大学文学部准教授)

[**テーマ**]　二項分布から正規分布へ、正規分布の性質、数表の利用、比率と平均値の区間推定

1　復習と宿題のポイント

- 無作為抽出とは / 区間推定の考えかた / 二項分布とは
- なぜ信頼率 0.95 に対して確率 0.025 が基準になるか

宿題 1 について：

- 95%信頼区間は 0.631 〜 1
- ということは、袋 (= 母集団) のうち 2/3 程度かそれ以上は赤玉だと考えてよい。(この推測はまちがいかもしれないが、その可能性は 5%未満である → 危険率)
- 信頼率は適当に決めている。信頼率 0.95 ＝ 危険率 0.05 にするのが通例だが、根拠は特にない。

宿題 2 について：

- 表裏の組合わせは 16。全部書いて考えてもよいし、「組合せ」(combination) 公式を利用してもよい。
- グラフをえがいてみると?

このような、一定の確率 (課題 2 の場合は確率 0.5) で偶然起こる出来事を n 回繰り返したとき、その出来事が起こる回数を理論的に予測した理論分布が「二項分布」(binomial distribution) である。

→　確率が 0.5 でない場合はどうなるか?

2　棄却域と採択域

理論分布:　一定の仮定の下での確率の分布を理論的に計算したもの

二項分布では、極端なケース (硬貨を 8 回投げて 6 回以上表、など) は起こる確率が低い。非常に確率が低いはずの極端な事象を観測したときは理論分布の仮定を疑う、というのが統計的推測の基本 (教科書 160 頁)。

(1) 「危険率」(α) を決める ($\alpha = 0.05$ にすることが多い →信頼率 0.95 に対応)

(2) 理論分布の上下の端から、確率が $\alpha/2$ を下回る領域を「棄却域」、それ以外の領域を「採択域」とする
(3) 棄却域と採択域との境界を「臨界値」という

区間推定の場合、「一定の仮定」を変化させながら、そのつど臨界値を計算し、実際の観測値と比較することになる。

3 正規分布

二項分布は、試行回数を増やすと、一定の形状に近づいていく (グラフを描くと、左右対称で真ん中にピークを持つなだらかな曲線になる)。試行回数が無限大 (∞) のときの二項分布のことを「正規分布」(normal distribution) という。

真ん中 (= 平均値) が 0 で標準偏差 (SD) が 1 になるように単位を調整して正規分布を描いたものを「標準正規分布」といい、$N(0,1)$ のようにあらわす。これを s 倍して m を足したものもやはり正規分布であり、$N(m,s)$ であらわす。

標準正規分布については、臨界値の表が用意されている (教科書巻末)。

例題: 標準正規分布の α =0.05 に対応する棄却域と採択域を教科書の数表から求めよ。

母比率の推測の場合、それほど比率が偏っていなくて ($0.1 < M < 0.9$)、サンプルサイズが大きければ ($n > 30$)、正規分布で近似できるものと考えて代用することが多い。通常、「比率の区間推定」といえば、この方法を指す。(実際には、平均値の区間推定 (後述) の方法で代用することが多い。)

母集団から無作為に n 人を抽出したところ、標本比率が m であった場合、母比率 M の 95%信頼区間はつぎの式で求められる：

$$m \pm 1.96\sqrt{\frac{m(1-m)}{n}} \tag{1}$$

この式の $\sqrt{\frac{m(1-m)}{n}}$ の部分を「標準誤差」(standard error) という。

臨界値 1.96 は危険率 0.05 に対するものである。この値は、危険率によって変わる (数表で調べる)。

例題: 標本規模 n=400 で標本比率 m=0.6 の場合、母比率 M の 95%信頼区間は?。

4 平均値の区間推定

値がいくつもある (たとえば 1–5) 変数の場合は?

→ すべての組合せについて理論分布を求めることは、事実上不可能

間隔尺度以上の変数の場合には、「母集団においては正規分布している」という仮定を置けば、平均値の区間推定が可能。つまり、標本における平均 m と標準偏差 s から、母集団における平均 M を推測する。この推測プロセスでは、母集団における平均と標準偏差の 2 つを推測しなければならないため、正規分布ではなく、t 分布 (Student's t distribution) を使う。

t 分布の性質 (教科書巻末参照)：

- かたちは標準正規分布に似ているが、正規分布より幅が広い
- 「自由度」(degree of freedom: DF) を持つ。これは標本規模によってきまる (df=$n-1$)。
- 自由度が大きくなると、標準正規分布に近づく (df>200 なら標準正規分布と同じと考えてよい)。

母平均の 95%信頼区間：

$$m \pm \text{臨界値}\frac{\text{SD}}{\sqrt{n}} \tag{2}$$

臨界値は自由度と危険率によって変化する (数表で調べる)。標本規模200以上で信頼率95%なら、1.96と考えてよい。

5 SPSS コマンド

「分析」→「記述統計」→「探索的」

- 「従属変数」を指定
- パネル左下の「統計」だけをチェック

信頼率を変更するには「統計」オプション。「因子」を指定すると、グループ別に分析できる

6 課題 1

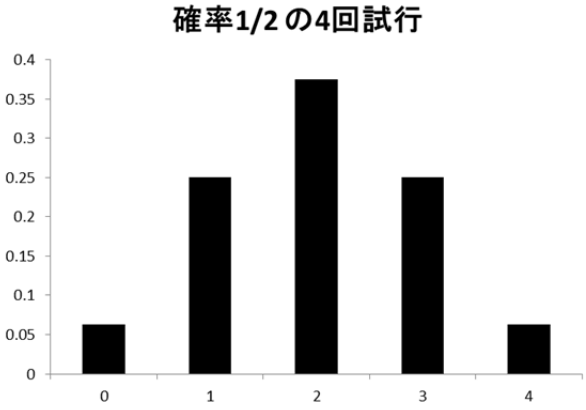
Wikipedia の「二項分布」の項 <http://ja.wikipedia.org/wiki/二項分布 > と「正規分布」の項を読んで、これらの関係を理解する。

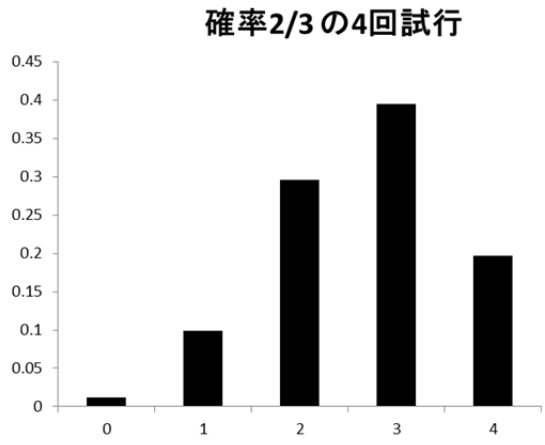
7 課題 2

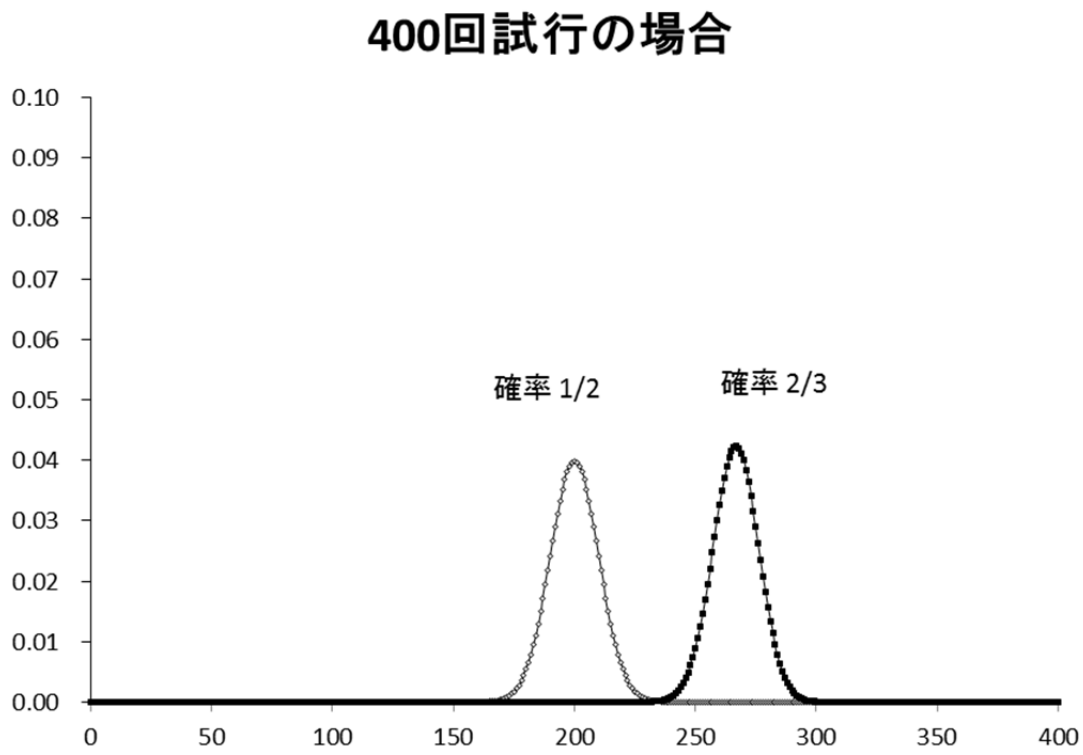
SPSS で、つぎのふたつの分析をおこなう

(1) 適当な変数について平均値の区間推定
(2) 同じ変数について、「因子」を指定して男女別の分析

これらの結果についてコメントをつけて提出 (ISTU で月曜正午まで)。他の人の意見をもらうこと（その人の名前を書く）


確率1/2 の4回試行
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4


確率2/3 の4回試行
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4


400回試行の場合
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
確率 1/2
確率 2/3
0
50
100
150
200
250
300
350
400