2008 年度前期　統計データ解析Ａ　第１３回

視聴率調査 − ２項分布の応用

今回は，前半の講義で説明した２項分布を応用する例として，視聴率調査の問題をとりあげます．

視聴率調査

ある番組の視聴率とは，本来は，対象の地域の世帯のうち，その番組を見ている世帯の割合です．しかし，対象の地域のすべての世帯を調査することは，現実には費用と時間がかかりすぎてできません．そこで，標本調査を行ないます．すなわち，対象の地域から無作為抽出されたいくつかの世帯を調査し，そのうちでその番組を見ている世帯の割合を求めて，これを視聴率ということにしています．

このような標本調査で求められた視聴率は，信用できるのでしょうか？ ここでは，この問題を２項分布を使って考えます．

視聴率調査では，標本として選ばれた世帯が，ある番組を見ているかいないかを問題にしています．この標本が無作為抽出されているとすると，

- １回の標本抽出で，「その番組を見ている世帯」または「見ていない世帯」のどちらかが選ばれる．
- 各回の抽出で，どの世帯も選ばれる確率は同じであり，つねに一定である．
- ある回の抽出でどの世帯が選ばれても，他の回の抽出結果には影響を及ぼさない．

のであれば，この無作為抽出をベルヌーイ試行と考えることができます．

このとき，対象の地域の世帯全体のうち，その番組を見ている世帯の割合が p だとしましょう．すると，どの世帯も選ばれる確率は同じですから，１回のくじびきで取り出される世帯が「その番組を見ている世帯」である確率は p です．したがって，n 世帯を抽出するとき，そのうち「その番組を見ている世帯」の数を S とすると，S は確率変数で，２項分布 $B(n, p)$ にしたがいます．

「対象の地域の世帯全体のうち，その番組を見ている世帯の割合」p は，どういう調査をしても変わらずひとつの数に決まっていますが，その値は対象の地域の世帯全体を調べなければわかりません．一方，一般に「視聴率」とよばれている「取り出された n 世帯のうちで，その番組を見ている世帯の割合」は，上の記号を使うと S/n となり，これを $\hat{p}$ という記号で表すことにします[1]．「視聴率」$\hat{p}$ は p とは違って確率変数で，偶然に左右されます．つまり，取り出した n 世帯の中に，その番組を見ている世帯が偶然多く含まれれば視聴率 $\hat{p}$ は p よりも大きくなるし，偶然少なければ視聴率は小さくなります．

では，このように調べた視聴率は，どの程度信用できるのでしょうか？ 次の問題で考えてみましょう．

> ある視聴率調査で，ある地方から n 世帯を無作為抽出して，ある番組を見ていたかどうかを調査しました．このとき，「この n 世帯のうち，その番組を見ていた世帯の割合」つまり視聴率の標準偏差を 0.01(1%) 以下にするには，n は少なくともいくらでなければならないでしょうか．

上で述べたように，「n 世帯からなる標本のうち，その番組を見ていた世帯の数」を S とすると，S は２項分布 $B(n, p)$ にしたがいます．したがって，S の分散は $np(1-p)$ となります．

[1] $\hat{p}$ は「p ハット」と読みます．「ハット」は「推定値」を表します．

このとき，$\hat{p}=S/n$ の分散は，いくらになるでしょうか．確率変数 S を，ある数 n で割るということは，S がとりうるさまざまな値が「いっせいに」$1/n$ になる，ということになります．S の期待値は「『S のとりうる値×その値をとる確率』の合計」です．ですから，S のとりうる値がいっせいに $1/n$ になると，期待値も $1/n$ になります．また，S の分散は「『(S のとりうる値 − 期待値）の２乗×その値をとる確率』の合計」です．ですから，S のとりうる値がいっせいに $1/n$ になると，２乗の計算が入っているため，分散は $1/n^2$ になります..

S の分散は $np(1-p)$ ですから，$\hat{p} = S/n$ の分散は $\dfrac{np(1-p)}{n^2}$ すなわち $\dfrac{p(1-p)}{n}$ となり，標準偏差は $\sqrt{\dfrac{p(1-p)}{n}}$ となります．この値が 0.01 以下でなければならないので，$\sqrt{\dfrac{p(1-p)}{n}} \le 0.01$ すなわち $n \ge 10000p(1-p)$ となります．p は 0 から 1 の範囲ですから，$p(1-p)$ の最大値は 1/4 です（$p = 1/2$ のとき）．よって，n は 2500 以上でなければなりません．

逆に言うと，少なくとも 2500 世帯程度を調べれば，調査によって測られる視聴率は，偶然によって左右されはするものの，たいてい 1%程度の違いしかない，ということになり，かなり信用できる数値になるといえます．

２項分布の区間推定

視聴率調査を題材にして，２項分布の場合の，区間推定の手法を使って考えてみましょう．次の問題を考えてみます．

> ある地域から 100 世帯を無作為抽出して調査すると，ある番組を見ていたのは 20 世帯でした．地域全体でその番組を見ていた世帯の割合を p とするとき，p の 95%信頼区間を求めてください．また，1000 世帯を調査して，その番組を見ていたのが 200 世帯だった場合はどうですか．

前節と同様に，n 世帯を抽出するとき，そのうち「その番組を見ている世帯」の数を S とすると，S は確率変数で，２項分布 $B(n,p)$ にしたがいます．

第４回の講義で説明したド・モアブル＝ラプラスの定理によって，抽出された世帯数 n がある程度大きければ，S は概ね期待値 np，分散は $np(1-p)$ の正規分布にしたがいます．よって，

$$Z = \frac{S - np}{\sqrt{np(1-p)}} \tag{1}$$

とすると，Z は標準正規分布にしたがいます．さらに，この分母分子を n で割ると，

$$Z = \frac{\frac{S}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{2}$$

となります．S/n は前節で述べた「視聴率」で，同様にこれを $\hat{p}$ で表すことにすると，

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{3}$$

となります[2].

さて，第 7 回の講義で説明したように，Z が標準正規分布にしたがうとき，Z が -1.96 から 1.96 に入る確率は 95%，すなわち $P(-1.96 \le Z \le 1.96) = 0.95$ です．よって (3) 式から

$$P(-1.96 \le \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \le 1.96) = 0.95 \tag{4}$$

ということになります．この式から p の範囲を求めると

$$P(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}) = 0.95 \tag{5}$$

となります．

この式で p の 95%信頼区間が求められたように見えますが，「p の範囲」を求めているはずなのに，その両端を表す式の中に p が入っています．これでは「p の範囲」になりません．そこで，調査した世帯数 n が多ければ，調査した世帯中での視聴率 $\hat{p}$ は，地域全体での視聴世帯の割合 p に近いはずですから，両端の式の p を $\hat{p}$ でおきかえます．そうすると，

$$P(\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n} \le p \le \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n}) = 0.95 \tag{6}$$

となるので，この式のカッコ内が p の 95% 信頼区間となります．

この式に，この問題での数値を入れてやると，$\hat{p} = 20/100 = 0.2, n = 100$ ですから，95%信頼区間は

$$[0.20 - 1.96\sqrt{0.20(1-0.20)/100}, 0.20 + 1.96\sqrt{0.20(1-0.20)/100}] \tag{7}$$

となり，計算すると，地域全体での視聴世帯の割合の 95%信頼区間は「0.121 以上 0.278 以下」となります．

また，1000 世帯の調査で，そのうちその番組を見ていたのが 200 世帯だった場合は，$\hat{p} = 200/1000 = 0.20, n = 1000$ です．このとき，地域全体での視聴世帯の割合の 95%信頼区間は「0.175 以上 0.225 以下」となります．以前説明したとおり，標本サイズが大きいと，推定が精密になり，信頼区間は狭くなります．

今日の演習

実際の視聴率調査では，新聞等でよく報道されている関東地方の調査の場合，無作為抽出される世帯数は 600 だそうです．このとき，視聴率の標準偏差は最大いくらですか．また，$p = 0.15$ のときの視聴率の標準偏差はいくらですか．

[2] この式は，前節で示した通り，$\hat{p}$ の期待値が p，標準偏差が $\sqrt{\frac{p(1-p)}{n}}$ であることに対応しています．