## 2011 年度前期　統計データ解析Ａ　第１２回

# 視聴率調査・出生性比 ー 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，視聴率調査の問題と，出生性比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．

[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．

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．

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 $\times$ 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 $\times$ 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，$\sqrt{\dfrac{p(1-p)}{n}} \leq 0.01$ すなわち $n \geq 10000p(1-p)$ 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５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１０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 $P(-1.96 \leq Z \leq 1.96) = 0.95$ です．よって (3) 式から

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

ということになります．この式から $p$ の範囲を求めると

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

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，

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

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．

---

### 男児／女児の「出生確率」と「出生比率」

数年前，生まれてくる子供の男児／女児の比率が変わってきている，さらに言えば女児の割合が増えているのではないか？ という説が出され，内分泌撹乱物質，いわゆる環境ホルモンが原因ではないかと疑われたことがあります．そのころには，「女の子ばかりが生まれている村がある」というショッキングな記事が週刊誌にでたことがあります．

もし本当に，環境ホルモンのせいで比率が変化しているのなら心配ですが，ここはまず，「何が問題なのか？」をはっきりさせるために，男児／女児の**「出生確率」**と**「出生比率」**の違いを考えてみます．

「男児／女児の出生比率」は，これまでに実際にあった出生のうちの，男児／女児の割合を意味します．男児／女児のどちらが生まれるかは，偶然によって決まる「ランダム現象」ですから，偶然男児ばか

[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．

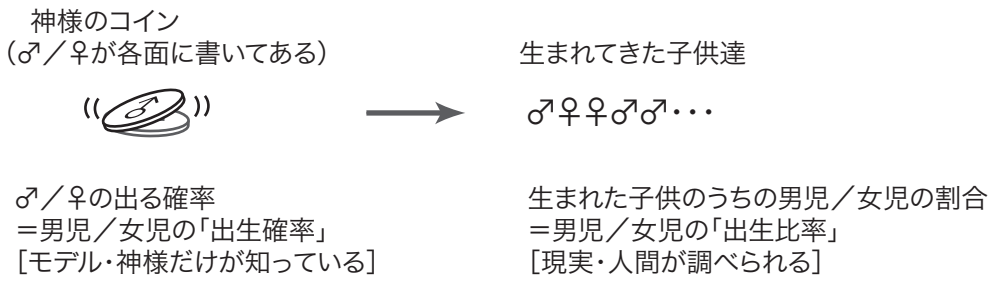


図 1:「出生確率」と「出生比率」

りが生まれたり女児ばかりが生まれたりすることは，兄弟姉妹ぐらいの人数ならよくあることです．私の知っている人には男５人兄弟の人も女５人姉妹の人もいます．このように，出生比率は，偶然によっていろいろな値になる確率変数としてとらえる必要があります．

一方，男女どちらが生まれるかはランダム現象ではありますが，毎回の出産で男児／女児が生まれる確率は概ね一定であると考えられています．また，ある出産の結果（男児か女児か）が，他の出産には影響しない，すなわち各々の出産は独立であるのも，ほぼ認められています．すなわち，１回の出生は第３回の講義で述べたベルヌーイ試行と考えられ，これが出生に関して想定される「モデル」です．このときの男児／女児が生まれる確率が「男児／女児の出生確率」です．つまり，各々の出産について神様がコインを投げて性別を決めていると考えたときの，「男児」／「女児」のそれぞれの面の出る確率に相当します（図 1）．

さて，「男児／女児の出生確率」は一定であるといまのところ考えられていますが，この確率が場所によって異なっていたり，あるいは過去と現在で異なっていれば，何らかの原因－例えば化学物質の影響など－が考えられます．そこで，「男児／女児の出生確率」を知る必要が出てきます．ところが，われわれには「神様のコイン」をのぞき見ることはできません．われわれにできるのは，これまでの「男児／女児の出生比率」を調べることだけです．そこで，現実に調べられる「出生比率」から，区間推定などの方法で「出生確率」を推測する必要があります．

**「出生確率」についての検定**

日本全国くらいの大規模な調査では，「出生比率」は男児のほうがわずかに多い（だいたい男児が 51.7%）という結果が得られています．したがって，日本全体でみれば，男児の「出生確率」も 0.517 くらいであろう，と考えられます．そこで，もしある特定の集団で「出生確率」が 0.517 から大きく隔たっていれば，この集団では何か異変が起きている可能性があります．しかし，上で述べたように，「出生確率」を人間が直接知ることはできません．そこで，その集団での「出生比率」を使って，「『出生確率は 0.517 である』という帰無仮説が棄却できるか」を調べる検定を行います．

前節で，「出生比率」すなわち生まれた子供のうち男児（あるいは女児）の比率は確率変数であると述べました．「出生比率」という確率変数がしたがう確率分布を，何かの確率分布モデルで表すことを考えます．前節で述べたように，出生においては

- 毎回の出生で男児（あるいは女児）が生まれる確率は一定である
- ある出生の結果（男児か女児か）が，他の出生には影響しない，すなわち各々の出生は独立である

という仮定が成り立っていると考えられ，これはベルヌーイ試行です．したがって，$n$ 人の子供が出生し，そのうちでの男児の「出生確率」を $p$ とすると，$n$ 人中の男児の数 $S$ は確率変数で，２項分布 $B(n,p)$ にしたがいます（女児の場合で考えても同じです．以下，男児を例として考えます）．

そこで，下のような問題を，くじびきの問題と同じように解くことができます．

> ある村では，ある期間に出生した子供 13 人のうち 9 人が女児であった．この村では，男児の「出生確率」が 0.517 より小さいといえるかどうか，有意水準 5%で検定せよ．

$n$ 人の子供が出生し，そのうちでの男児の「出生確率」を $p$ とすると，$n$ 人中の男児の数 $S$ は確率変数で，２項分布 $B(n,p)$ にしたがいます．したがって，(1)-(3) 式と同様に，男児の「出生比率」$S/n$ を $\hat{p}$ で表すと，

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{8}$$

は標準正規分布にしたがいます．さて，ここで「男児の『出生確率』$p$ は 0.517 である ($p = 0.517$)」という帰無仮説と，「男児の『出生確率』$p$ は 0.517 より小さい ($p < 0.517$)」という対立仮説を考えます．問題から「出生比率」$\hat{p} = (13-9)/13 = 0.308$ で，帰無仮説が正しいとすると $p = 0.517$ ですから，これらと出生数 $n = 13$ を (8) 式に代入すると

$$Z = \frac{0.308 - 0.517}{\sqrt{\frac{0.517(1-0.517)}{13}}} = -1.51 \tag{9}$$

となります．

いま考えている検定では，「$p = 0.517$」という帰無仮説が棄却されたとき，「$p$ はもっと小さい」という対立仮説が採択されるとしています．つまり，仮に「出生確率」$p$ が 0.517 だとすると，現実の「出生比率」$\hat{p}$ が 0.308 やあるいはそれよりさらに小さくなることはありえないはずだ，本当は「出生確率」はもっと小さいんだろう，と内心考えているわけです．したがって，帰無仮説が棄却されるのは，「$\hat{p}$ が 0.308 以下」という確率が有意水準よりも小さいときです．「$\hat{p}$ が 0.308 以下」とき，(9) 式から，「Z は $-1.51$ 以下」であることがわかります．

第８回の講義の「棄却域」の説明で述べたように，$Z$ が標準正規分布にしたがうとき，$Z$ が $-1.64$ 以下である確率が 5%です．したがって，Z が $-1.51$ 以下である確率は，5%よりも大きいことがわかります．

ですから，「男児の『出生確率』$p$ が 0.517 であるとしたときに，『出生比率』$\hat{p}$ が 0.308 と，異常に小さい」ことは，「確率 5%でしか起きないことが起きている」とはいえません．つまり，帰無仮説は棄却できず，「この村では男児の『出生確率』は 0.517 より小さい」とは言えないという結論になります．つまり，この程度の小さな村では，この程度の男女の偏りは「確率 5%でしか起きないくらい珍しい」ことではない，ということなのです．

---

**今日の演習**

実際の視聴率調査では，広島県の場合，無作為抽出される世帯数は 300 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．