2011 年度前期　統計学で考える　第９回

環境ホルモンと出生性比

前回で，統計的推測の基礎を説明する，前半の講義が終わりました。「統計学で考える」の後半は，環境問題など，社会にある問題について考えるための統計学の手法について説明してゆきます。後半１回目の今回は，「最近，環境ホルモンのせいで，女の子が生まれやすくなっているのでは？」という不安に，仮説検定や区間推定を使って答える方法を説明します。

２項分布を使う問題の例－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。

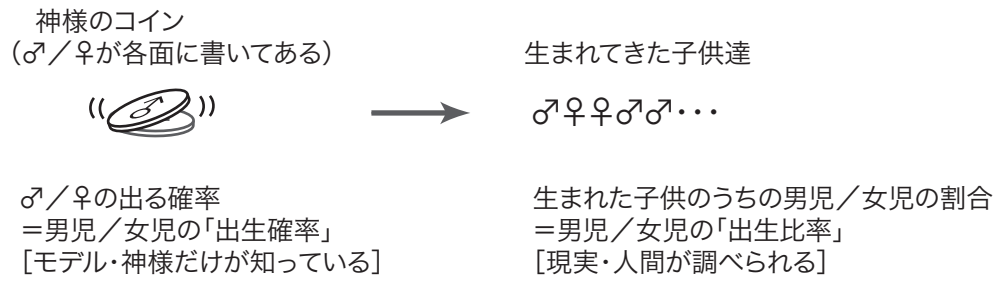

図 1: 「出生確率」と「出生比率」

「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)。

このとき，男児の「出生比率」は S/n で表されます。ここでは，これを $\hat{p}$ という記号で表すことにします[1]。$\hat{p}$ はどういう確率分布にしたがうでしょうか？

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 のとりうる値がすべて $1/n$ になると，期待値も $1/n$ になります。第 3 回の講義で述べたように，S の期待値は np ですから，$\hat{p}$ の期待値は p となります。

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 のとりうる値がすべて $1/n$ になると，分散の計算には S やその期待値を 2 乗する計算が入っているので，分散は $1/n^2$ になります。S の分散は $np(1-p)$ ですから，$\hat{p}$ の期待値は $np(1-p)/n^2$ すなわち $p(1-p)/n$ となります。

さらに，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，全出生数 n がある程度大きければ「出生比率」は概ね期待値 p，分散は $p(1-p)/n$ の正規分布にしたがいます。以上の性質を使っ

[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。

て，この講義の前半で説明した検定の手法を，「出生比率」と「出生確率」について考えてみることにしましょう。次の例題を見てみます。

> 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。

男児の「出生確率」を p とします。上で述べたとおり，「出生比率」$\hat{p}$ は，概ね，期待値 p，分散 $p(1-p)/n$ の正規分布にしたがいます。したがって，

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

のように変換した確率変数 Z は，標準正規分布 $N(0,1)$ にしたがいます[2]。

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 (1) 式に代入すると

$$Z = \frac{0.308 - 0.517}{\sqrt{\frac{0.517(1-0.517)}{13}}} = -1.51 \tag{2}$$

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，(2) 式から，「Z は -1.51 以下」であることがわかります。

第 6 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。

出生確率の区間推定

「出生比率」と「出生確率」の関係について，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。

---

[2]前半の講義で，くじ引きの問題について，あたり本数を S とするとき $Z = \dfrac{S - np}{\sqrt{np(1-p)}}$ という関係を説明しました。$\hat{p} = S/n$ ですから，この式の分母分子を n で割ると，本文中の (1) 式と同じになります。

100 回の出生で，男児が 52 人，女児が 48 人生まれました。男児の「出生確率」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 回の出生で男児が 520 人生まれた場合はどうですか。

前節と同様に考えると，男児の「出生確率」を p,「出生比率」を $\hat{p}$，出生する子供の数を n とすると，

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

は標準正規分布 $N(0,1)$ にしたがいます。

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 $P(-1.96 \leq Z \leq 1.96) = 0.95$ 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，男児の「出生比率」$\hat{p}$ は，出生数 n が大きいときは男児の「出生確率」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，

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

となって，p の 95%信頼区間が求められます。この式に，この問題での数値を入れてやると，$\hat{p} = 52/100 = 0.52, n = 100$ ですから，男児の「出生確率」の 95%信頼区間は $\left[0.52 - 1.96\sqrt{\frac{0.52(1-0.52)}{100}}, 0.52 + 1.96\sqrt{\frac{0.52(1-0.52)}{100}}\right]$ で，すなわち $[0.422, 0.618]$ ということになります。

また，1000 回の出生で男児が 520 人生まれた場合は，$\hat{p} = 520/1000 = 0.52, n = 1000$ ですから，男児の出生確率の 95%信頼区間は $[0.489, 0.551]$ となります。以前説明したとおり，標本サイズが大きくなると，推定が精密になり，信頼区間は狭くなります。