Chapter 3; Statistical Inference (前半 :pp25-36)

コイン投げの話は、(ベルヌーイ試行 *bernoulli trials* の結果生じる) 様々な 2 値のイベントに応用できる。

例) Shealar & Spendelow (2002)

対象種：roseate tern (*Sterna dougalli*) ベニアジサシ、絶滅危惧種
データ：観察(748h)から習慣性の盗み寄生者(Habitual kleptoparasite)10 個体中 8 個体がメス、2 個体がオス
結果：ランダム抽出で 10 個体中 8 個体以上が同じ性別になる確率は低い (2 項検定, P = 0.055)

このデータを使って、頻度主義とベイズ主義の原理を比較する。

Question:ベニアジサシの習慣性盗み寄生(*K*)はオス(*M*)よりもメス(*F*)で多く観察されるのか。

$$\Pr(K|F) > \Pr(K|M), \qquad (3.1)$$

10 個体の習慣性寄生者の性別が独立のベルヌーイ試行で、"メス"が"成功"であり、未知の成功パラメータを p と仮定。

最初に注意すべきことは p は $\Pr(K|F)$ではなく、$\Pr(F|K)$であるということ。式(3.1)と $\Pr(F|K) > \Pr(F)$ は同じである[*1]。$\Pr(F)$ が既知の場合、Eq.(3.1)の推定は習慣性盗み寄生者間の性別の観察をもとに考えられる。Shealar & Spendelow (2002)は $\Pr(F) = 0.5$ と仮定し、データが $\Pr(F|K) = \Pr(F)$の仮定と一致するかを調べた。

たぶんより根本的な他の問題は、実際に、$p = \Pr(F|K)$が成立するかどうかということである。より正確には、"鳥が観察される"ことを O とすると、$p = \Pr(F|K, O)$と表わせる。オスの習慣性盗み寄生がメスよりも見つけられにくい場合、残念ながら、メスで多く観察されることとなる。よって $p = \Pr(F|K)$が成立するためには、K において O と F は独立であると仮定しなければならない。その時、以下の式が成立。

$$p = \Pr(F|K,O) = \frac{\Pr(F,O|K)}{\Pr(O|K)}$$

$$= \frac{\Pr(F|K)\Pr(O|K)}{\Pr(O|K)} = \Pr(F|K).$$

ここでの必要条件は、習慣性盗み寄生者(K)の性別と観察が独立であること。生態学では、興味のある事柄の観察プロセスが潜在的に複雑であるという問題が共通して生じる。これは、本の後の方で扱うテーマである。

これらの仮定について考えることは、推定プロセスで頻度主義とベイズ主義のどちらを選ぶかを決める上で重要な部分であるが、今はそれを考えずに、確率変数 $X \sim B(10, p)$の観察 $X = 8$ について、p の統計推定プロセスを考えることにする。10 個体の習慣性盗み寄生者アジサシ内のメスの数 X はランダムにおこる出来事で、その結果として $X = x$, $\{x \in 0, 1, .., 10\}$が生じる。この様々な結果が起こる確率は未知の量 p によって決定される。

$$\Pr(X = x) = B(x; 10, p) = \binom{10}{x} p^x (1-p)^{10-x}, \qquad (3.2)$$

我々の目的は観察 $x = 8$ における p を求めることである。

3.1 LIKELIHOOD

確率計算に、2 項分布 $B(x; n, p)$を用い、パラメータの既知の値が与えられた場合の特定の出来事の確率を明らかにする。観察された出来事が与えられたとき、それらを規定する未知のパラメータについて何

がいえるだろうか？盗み寄生のデータを考える。推定プロセスはTable 3.1のようなinspectionから始まる。Table 3.1；$n = 10$ の 2 項分布における、$X = x$ の確率を示したもの。

ここで興味があるのは観察と対応した $x = 8$ の**列**。この列をみると p が小さいと確率は小さく、p が大きくなると確率が大きくなり、$p = 0.8$ の時に最大、その後、確率は減少する。これを基盤として、0.8 を用いて、未知の値を推定するのが妥当なように思える。ここで、真の値 p の推定量 $\hat{p}= 0.80$ と書ける。

$\hat{p}$の値を選ぶプロセスについてみていく。p の既知の値をもとにした結果 x の確率を見つけるためにはTable 3.1 の**行**を使う。よって、Shealar & Spendelow (2002)の 2 項検定 P = 0.055 は X~$B(10, 0.50)$の Pr ($X \geq 8$) から計算できる。さっきは統計推定の目的のために、Table 3.1 の**列**を用いた。結果を知っていたため、p の未知の値を経験に基づいて推定することができたのである。

Table 3.1 の**行**よりも**列**について考えてみると、$B(x; 10, p)$は固定された x における p の関数としてみなせる。確率を計算するためには、$B(x; 10, p)$は固定された p における x だけの関数として使われる（**行**）。統計推定のためには、$B(x; 10, p)$は固定された結果 x をもつ p だけの関数と考えるのである（**列**）。

x を固定して p を変化させると（**行**）、$B(x; 10, p)$は分布関数ではなくなる(1 に到達しないので)。この場合、推定する上で重要なのは、各値の相対値である。例えば、観察 $X = 8$ の確率は $p = 0.5$ で $p = 0.4$ の 4 倍大きい(0.044 vs. 0.011)。データは、$p = 0.4$ の 4 倍 $p = 0.5$ を支持したことになり、ここでは、各確率の差よりも割合が比較の基盤となる。

Table 3.1 の代わりに手計算すると、

$$\frac{B(8;10,0.5)}{B(8;10,0.4)} = \frac{\binom{10}{8}0.5^8(1-0.5)^{10-8}}{\binom{10}{8}0.4^8(1-0.4)^{10-8}}$$

$$= \frac{0.5^8(1-0.5)^{10-8}}{0.4^8(1-0.4)^{10-8}} = 4.14,$$

ここで組み合わせの項は分子と分母でキャンセルされるので計算する必要はなく、x の固定値のみに依存する。このため、統計学者は興味のあるパラメータと独立した確率分布の掛け算部分を省略し、尤度関数(*likelihood function*)を以下のように定義した。

$$L(p) = p^x(1-p)^{n-x}. \tag{3.3}$$

Table 3.1 によれば、尤度の最大値は $p = 0.8$ の時に得られる。尤度そのものよりも、尤度の対数を最大化する方が簡単である。$\ln(L(p))$と $L(p)$の極値は一致するので、$\ln(L(p))$を最大化すればよい。

$$l(p) = \log(L(p)) = x\log(p) + (n-x)\log(1-p)$$

$$l'(p) = \frac{dl(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p} = 0,$$

上記のように微分して、p の解を求めると$\hat{p} = x/n$。この推定量は最尤推定量(*maximum likelihood estimator*; **MLE**)として知られている。

Interval Estimates Needed

MLE はどのくらい信頼できるのか調べるべきである。Table 3.1 によれば、$p = 0.80$ のとき最もあり得る結果は $X = 8$ である。X =7, or 9 は期待されないわけではない。結果には確率的不確実性が存在し、$p = 0.80$ の**行**の様々な結果は単一の固定されたパラメータと一致する。

同様にして、Table 3.1 の特定の**列**、$x = 8$ の列に注目する。p は最も尤もらしい値である（i.e., 尤度を最大化する）が、他の値が高い尤度を持つこともありうる。観察 $X = 8$ は $p = 0.70$ とも 0.60 とも一致しないことはない。パラメータ p には統計的な不確実性が存在する。

このようにして、MLE だけをみると p の推定が限られてしまう。その代わりに、p の集まり、区間推定（観察データと一致するみなされるすべての値）を調べる。

p の集まりを決める方法は頻度主義とベイズ主義のどちらの統計推定を使うかによって異なるが、それを考える前に、尤度関数のみを基盤とした区間推定を示すことにする。

Likelihood Intervals

“scaled likelihood”は、尤度を最大尤度で割ったものであり、以下のように定義される。

$$L^*(p) = \frac{L(p)}{L(\hat{p})}.$$

よって、scaled likelihood は 0 から 1 の間の値をとり、MLE の scaled likelihood は 1 となる（Fig.3.1）。$n = 10$ で $X = 8$ のとき、$L^*(p) \geq 0.25$ となるのは $0.549 \leq p \leq 0.950$ のときで、この区間における p の値は”少なくとも”25%の割合で MLE となりうる。(0.549, 0.950)は p の区間推定で、MLE のような単なる点推定よりも統計的不確実性を定量化するのに使われる。他の方法の区間推定と区別するために、$\{p \in [0, 1] | L^*(p) \geq \alpha\}$ を 100(1−α)% scaled likelihood 区間とする。α は不確実性の範囲（区間）を決める信頼係数であり、α が小さくなると区間は大きくなる。

$n = 100, X = 80$ の場合でも MLE は 0.80 であるが、区間推定量は(0.728, 0.861)と狭まる(Fig. 3.2)。$n = 10$ で支持された適度な p の値の大半はサンプルサイズが大きくなると支持されなくなるのである。サンプルサイズが大きくなると予測される統計的不確実性は減少する。データセットの推測値はおおまかにサンプルサイズの平方根に比例して増加することがよく知られている。

今までの区間推定には二つの欠点があった。一つ目は、計算が決して楽ではないこと、二つ目は 0.25 を選ぶ上で任意の裁量があるということ（主観的）。任意の裁量は仕方がないことで、保守的な人は 0.25 よりも低い値を選ぶかもしれないが、”p の値が少なくとも、最も尤もらしい値を 1/4 の割合でとる”ということはややあいまいである。

では、区間推定をどう行うべきか？

これまでと同様に、2 項確率 p の推定という例で、頻度主義の考え方で考えてゆく。

3.2 CONFIDENCE INTERVALS

頻度主義の区間推定を信頼区間(confidence interval; CI)という。CI はある基準によって定義された観察 X によって決定される確率変数であり、信頼限界(l_X, u_X)間の区間である。まず、CI 自体が、X のようにただの確率変数であることを強調しておきたい。

確率変数としてのデータの概念は統計推定では重要である。データが生成される確率過程の固定されたセットが存在するとする；よって、これらの過程を規定するパラメータの固定されたセットが存在する。さらに、観察データセット、つまり、これらのランダム過程から得られた単一のセットが存在する。例えば、コイン投げをして、結果が“表”だったとする。コイン投げというランダム過程 X とその結果”表”を区別して考える。簡単には、X = “表”と書けるかもしれないが、2 つの概念を区別して理解してほしい。

つまり、”ランダム過程 X”と”その結果の 1 つであるデータ”が存在する。

もう一度、10 のベルヌーイ試行で 8 の成功が観察された例を考えてみる。X~B (10, p), X = 8 と書ける。数字 8 にはランダムな部分はないが、確率的なメカニズムから生成されており（X は 5 でも 9 でもとり得る）、8 が得られたのは偶然なので、X を確率変数として考えることができる。

これらの観察から CI を考えてみる。X = 8 が与えられると、CI は(l_8, u_8)という固定値となる。この特定の区間は確率過程の単一の結果である。同じランダムメカニズムが X = 5 でも起こりうるし、その場合の CI は(l_5, u_5)となる。CI の定義は、特定の結果による区間よりもむしろ、ランダムな区間(l_X, u_X)を基盤としている。

ランダムな区間の概念について簡単な例で考えてみる。単位区間(0, 1)においてランダムに二つの数字を選び、その区間を考える。ランダムな区間について何が言えるだろうか？その特徴はランダムで、その長さは 0（2 変数が一致したとき）から 1（ふたつの値ができる限り遠いとき）まで変化する。平均すると区間の長さは 1/3 となるだろう。区間が 0.6 のような特定の値を含んでいるかを知りたいかもしれない。その答えは、時にはそうだろうし、時にはそうでないかもしれない。区間の両端（信頼限界）が両方とも 0.6 よりも小さい場合(確率 = 0.6^2 = 0.36)または、0.6 よりも大きい場合(確率 = 0.4^2 = 0.16)、区間は 0.6 を含まないし、そうでない場合は 0.6 を含む。このように、区間が 0.6 を含むチャンスは 0.48 (1－0.36－0.16)なのである。

上記のようなランダムな区間を生成するコンピュータプログラムを想像してみる。長い時間計算させれば、区間の 48%は 0.6 を含むだろう。プログラムが正しく書かれ、テストされ、まさにいま最後の区間を生成しようとしているとする。この場合、前もって、0.6 の値を含む区間は 48%の信頼があると言えるだろう。

シナリオをちょっと変えてみる。未知の定数 p と、p を 0.48 の確率で含むランダムな区間を生成するプログラムがある。プログラムの書き手は p を知っており、プログラムは正確に動くとする。長い計算ののち、区間の 48%が p を含んだとする。その区間の両端の値は 0.328 と 0.832 だった。よって、p は 48%のチャンスで 0.328 と 0.832 の間に存在することになる。が、頻度主義の考えでは、状態(statement)が正しくない ; p は固定量で、区間の中に存在するかどうかのどちらかでなければならない ; そこには p 自身に関わる確率は存在しないのである。唯一の確率は区間の生成に関連した確率だけである。区間(0.328, 0.832)は p に対する 48%CI として示される; “信頼”は区間そのものに対してよりは、区間を生成するメカニズムに対して示される。では、ここで、2 項成功率(binomial success rate)の CI の定義について示す。

CI for Binomial Success Rate

X~B(n, p), すべての p について Pr(p∈C_x|p)の場合、p の区間 C_x を 100(1−α)%CI と仮定する。

ここでα=0.05 とする。95%CI はどのような p でも問題がないように定義されたランダムな区間であり、観察データ X は区間 C_x（C_x は”少なくとも”95％の確率で p を含む）から生成されたものであるとする。

そのような区間の生成方法を次の section に記す。これからしばらくは、CI で信頼すべきものは**方法**に関してであり、**特定の結果**についてではないことを強調したい。以下に記す方法では、10 のベルヌーイ試行での 8 の成功における p の 90％CI は(0.541, 0.931)である。これは正確には” 90%の可能性で p が 0.541 と 0.931 の間に存在する”とは言えない。むしろ、この方法は同様の状況の 90%でうまくいくと言えるかもしれない。

詳細に入る前のもう一つ注意すべきことは、CI の定義は条件付き確率 $\Pr(p \in C_x|p)$と関係しているということである。p に関する条件は、p を固定量として扱うということである。確率の状態は、ほとんどデータ X だけに関連したランダム変動である。このポイントについて長々と説明するのは、これが統計の基本のトピックで、頻度主義パラダイムの基盤となる概念であり、非常によく誤解されている部分であるからである。頻度主義パラダイムの他の方法 (e.g. 仮説検定) でも同様の認識論を基盤としているので、これらの一般的なパラダイムもまた共通して同様に誤解されていてもおかしくない。

3.2.1 Approximate CI's for Binomial Success Rate

ここで、もっとも普通な方法で 2 項成功パラメータの CI を求める方法を示す。

2 項分布の正規近似(see B.5)を用いて、MLE $\hat{p} = x / n$ が、平均 p（真の値）と標準偏差$\sigma(\hat{p}) = \sqrt{p(1-p)/n}$をもつ正規確率変数とほとんど同じ分布をもつとする。その結果、$Z_{\alpha/2}$ は標準正規分布の上限(1−α/2)変位値(quantile)を示すとすると、

$$1-\alpha \approx \Pr\left(-z_{\alpha/2} \le \frac{\hat{p}-p}{\sigma(\hat{p})} \le z_{\alpha/2}\right), \qquad (3.4)$$

並べ替えると

$$1-\alpha \approx \Pr\left(\hat{p}-z_{\alpha/2}\,\sigma(\hat{p}) \le p \le \hat{p}+z_{\alpha/2}\,\sigma(\hat{p})\right). \qquad (3.5)$$

$\sigma(\hat{p})$は未知の値 p に依存するため、$\hat{p}$を代入して、

$$\hat{\sigma}(\hat{p}) \doteq \sqrt{\hat{p}(1-\hat{p})/n}.$$

式 3.5 にこの推定量を代入し、観察 X における p の近似の 100(1−α)%CI を求める。

$$I_X = \hat{p} \pm z_{\alpha/2}\,\hat{\sigma}(\hat{p}).$$

区間 I_x が p を含むチャンスは、おおよそ 100(1−α)%である。n が適度に大きく、p が 0 または 1 に非常に近い場合は近似は問題ない。区間の長さは n の平方根に反比例するので、尤度区間で見られたように、区間の長さ（統計的不確実性）はサンプルサイズの平方根に比例して減少することに注意。

区間 I_x は近似した CI である。近似はどのくらい実際に近いのだろうか。CI の定義では、p の各値に対して、p を含む推定量を導く X のセットの確率が少なくとも(1−α)でなければならない。近似した 90%CI の coverage rate (n= 100)が Fig. 3.3 である。全体としてはちょっと低すぎるが、$0.10 \le p \le 0.90$ では coverage rate は適度に 90%に近いことに注目してほしい。

すべての p における十分な performance を求める CI の定義では、多くの p に対してうまくいく近似を使うことが明らかに変である（すべての p に対してではないので）。

式 3.4 からを変形して得られる I_x よりも良い coverage 特性をもつ近似 CI では、$\sigma(\hat{p})$に$\hat{\sigma}(\hat{p})$を代入するのを避ける。その結果、p に対する CI は、

$$I_X^* = \frac{2n\hat{p}+z_{\alpha/2}^2}{2(n+z_{\alpha/2}^2)} \pm \frac{z_{\alpha/2}}{2(n+z_{\alpha/2}^2)}\sqrt{4n\hat{p}(1-\hat{p})+z_{\alpha/2}^2}.$$

Fig. 3.4 が I_x と I^*_x の coverage rate の比較を示す。I^*_x の方が I_x よりも coverage 特性がかなりよいが、めったに使用されない。これはたぶん式が難しそうに見えるから。これら方法は両方とも近似であり、coverage rate が p のすべての値に対して"少なくとも"(1−α)を満たさなければならないという CI の定義を

満たしていないことに注意する必要がある。理想的には、精密な(exact) CI を普通に使えばいいはず。

3.2.2 Exact CI for Binomial Success Parameter

$$F_U(p) = \sum_{k=x}^{n} B(k;n,p) \quad \text{and} \quad F_L(p) = \sum_{k=0}^{x} B(k;n,p);$$

上の式は、$X \sim B(n, p)$における upper tail probability, $\Pr(X \geq x)$, と lower tail probability, $\Pr(X \leq x)$を示している。ベニアジサシデータ($n = 10, x = 8$)についてのこの確率の図が Fig.3.5。

曲線 $F_U(p), F_L(p)$は観察データと一致する p の値を求める際に使える。$F_U(p)$ (赤線)は、すべての $p \geq 0.493$ に対し、$F_U(0.493) = 0.05$, $\Pr(X \geq 8) \leq 0.05$ となるため、p の増加関数となる。$p \leq 0.493$ の時、ランダムに選択された 10 個体の中でメスが 8 個体以上であるチャンスはたった 5%しかない。同様にして、$F_L(p)$(青線)は、すべての $p \geq 0.963$ に対し、$F_L(0.963) = 0.05$, $\Pr(X \leq 8) \leq 0.05$ となるため、p の減少関数となる。これらの大きな値をもつ p では ($p \geq 0.963$)、ランダムに選択された 10 個体中にメスが 8 個体以下であるチャンスはたったの 5%しかないこととなる。

p の 90% exact CI は次の要素から求められる。8 個体以下のメスを観察するという大きすぎる p と 8 以上のメスを観察するという小さすぎる p を除いた区間(0.493, 0.963)。このような区間の形式的な定義は、2 項成功率の exact CI に従う。

Exact CI for Binomial Success Rate

$X \sim B(n, p)$と仮定。$p_L(0, \alpha/2)$とすると x = 1, 2, …., n について以下のように定義される。

$$p_L(x, \alpha/2) = \max_p \{p : F_U(p) \leq \alpha/2\}.$$

$p_U(n, \alpha/2) = 1$ とすると、x = 1, 2, …., n について以下のように定義される。

$$p_U(x, \alpha/2) = \min_p \{p : F_L(p) \leq \alpha/2\}.$$

このとき、区間 $J_X = (p_L(X, \alpha/2), p_U(X, \alpha/2))$ は p の exact 100(1−α)%CI である。
この区間のことを”exact conference interval”と名付けたが、これがすべての p に対して $\Pr(p \in J_X | p) \geq 1-\alpha$ を満たすかはまだ示していない。2 項分布は離散性をもつために、p の多くの値が 1−α よりも大きくなってしまうという嫌な結果が導かれる。例えば、サンプルサイズ $n = 25$ の p の exact 70% CI を考えてみる。X には 26 の起こりうる結果が存在する。26 の起こりうる CI の要約を Table 3.2 に示した。

p のすべての値に対して、exact CI の coverage rate を計算するのは簡単である。例えば、$p = 0.3595$ の時、$7 \leq X \leq 11$ の場合だけの p の区間を見ればよく、その時の $\Pr(7 \leq X \leq 11) = 0.703$ であり、期待した水準(70%)に近い。しかし、少し注目する点を変えて、$p = 0.3591$ について考えると、$6 \leq X \leq 11$ の時の p の区間をみることになり、$\Pr(6 \leq X \leq 11) = 0.792$ となるため、区間の 70% coverage rate よりかなり大きくなる。

Fig 3.6 は $n = 25, \alpha = 0.30$ の時の p の関数の真の coverage rate を示している。

Coverage rate は最低限でも常に、特定の水準と同じくらいの大きさになること期待しているにもかかわらず、多くの p の値が必要以上に大きくなってしまうので、ややイライラしてしまう。この現象はここで取り上げた $n = 25, \alpha = 0.30$ の場合だけに限ったことではない。離散的確率変数を用いて、パラメータのすべての値の”少なくとも”(1−α)100%の coverage をもつ信頼区間を求める際には避けられない結果なので

ある。

3.2.3 Confidence Intervals – summary

Scaled likelihood interval のように、CI によって、統計推定を行う上で内在する不確実性を見ることができた。こちらの方が尤度区間よりも興味深い。

CI には使い手にあまり知られていない様々な複雑な問題が存在する。第一に、$\hat{p} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}$のようなよく知られている単純な形は疑わしい値の近似を基盤としている（たとえかなり大きなサンプルサイズを用いたとしても）。厳密に言えば、これらの近似では、特定の信頼水準の CI を求めることはできない。なぜなら、パラメータのすべての値が”少なくとも”(1−α)の coverage probability を持つ必要があるからである。二つ目は、パラメータのすべての値が少なくとも(1−α)の coverage probability を持つ（定義を満たす）場合、パラメータの大半の値が coverage probability が(1−α)をはるかに超えるかもしれないことである。筆者の印象では、この現象は多くの使い手にはよく知られていない。これは CI の批判ではないが、間違いなくその説明の透明性を低下させるものである。

この節で扱った 3 つの区間のうち、J_x のみが本物の 95%CI であり、p のすべての値で coverage probability ≥ 95%を満たしている。この最小限補償された 95% coverage には、代償がある：区間が長くなるにつれて、coverage はより大きくなる。Fig 3.7 に 3 つすべての coverage probability を示した。適当なサンプルサイズ (n = 100)であるにもかかわらず、不適当な部分は見つからず、これはより小さなサンプルサイズでも顕著にみられる。

Fig 3.7 は 0 から 1 の間を均一に占める様な 1000 個の p の値のグリッドを使って生成されている。この coverage probability を要約したのが Table 3.3 である。多くの使い手が”typical(標準的な？) 90% coverage”のために”minimal 90% coverage”を捨て駒にすることを厭わずに、区間 I^*_x を”90%信頼”の直観的に理解できる概念を表現したものとして好んでしまうのではないかと著者は懸念している。しかし、頻度主義パラダイムでは、p の値の間の”typical”な動きについては何も考えないが、最悪の場合のシナリオにならないようにするということを考える。そこで、今、p の typical な動きを考慮するベイズ主義の区間推定についてみてみようと思う。