Introduction to mathematical statistics ゼミ (第 4 回)

担当: 伊藤真道

未定

4 Chap.4. Some Elementary Statistical Inferences

4.8 The Method of Monte Carlo / モンテカルロ法

特定の分布や標本から観測値を生成する概念はモンテカルロ生成?(Monte Carlo generation) と呼ばれている．これは，複雑な過程の模擬実験や，有限標本における統計的な方法論の性質を調べるのに使われてきた．以降様々な手法を取り扱うが，この節で紹介する方法のほとんどは，一様乱数を生成できることを基にしている．

定理 4.8.1

確率変数 U が $U(0,1)$ に従うとする．また F を連続な分布関数とする．このとき，確率変数 $X = F^{-1}(U)$ は分布関数として F を持つ．

証明 4.1. 一様分布の定義から，U は分布関数として，$u \in (0,1)$ に $F_U(u) = u$ を持つ．これと *distribution-function technique* を利用し，さらに $F(x)$ が狭義単調とすると，X の分布関数は，

$$\begin{aligned} P[X \leq x] &= P[F^{-1}(U) \leq x] \\ &= P[F(F^{-1}(U)) \leq F(x)] \\ &= P[U \leq F(x)] \\ &= F(x) \end{aligned}$$

となる．これは定理の結果である． □

ここで，$F(x)$ が狭義単調であることを仮定したが，これは緩めることができる．テキストの練習問題を参照.

¡note¿

例 4.1. *(Monte Carlo Integration)* $\int_a^b g(x)dx$ を計算したいとする．g の不定積分が存在しない場合，数値積分が用いられる．単純な数値積分の技法のとして，*Monte Carlo* 法がある．問題となっている積分は，

$$\int_a^b g(x)dx = (b-a)\int_a^b g(x)\frac{1}{b-a}dx = (b-a)E[g(x)]$$

とかける．ここで $X \sim U(a,b)$ である．モンテカルロ法は，ここで，$U(a,b)$ からサイズ n の無作為標本を発生させ，$Y_i = (b-a)g(X_i)$ を計算する．以上の手順から得られる $\bar{Y}$ は $\int_a^b g(x)dx$ の不偏推定量である．

以下の例では，データの分布が混合正規分布である際の t 検定の有意水準を，モンテカルロ法で推定するものである．

例 4.2. X を平均 μ の確率変数とし，以下の仮説を考える．

$$H_0 :\ \mu = 0\ versus\ H_1 :\ \mu > 0 \tag{1}$$

この仮説に対して，サイズ $n = 20$ の X の分布からの標本に基づき，以下の棄却ルールのもとで t 検定を行う．

$$Reject\ H_0 :\ \mu = 0\ if\ t > t_{.05,19} = 1.729 \tag{2}$$

ここで，$t = \bar{x}/(s/\sqrt{20})$ である．この例では，X が $\epsilon = 0.25, \sigma_c = 25$ の混合正規分布に従うとする．つまり，標本の 25% は $N(0,25)$ から，75% は $N(0,1)$ から生成されているとする．この仮定から，X の平均は 0 であり，H_0 は正しい．しかし，混合正規分布から標本が生成されているため，検定の正確な有意水準を求めることはかなり複雑である．ここで，X が混合正規分布である時の検定統計量 t の分布を求める代わりに，シミュレーションによって有意水準を推定する．N をシミュレーションの回数とする．

1. $k = 1, I = 1$ とする．
2. 大きさ 20 の無作為標本を X の分布から発生させる．
3. この標本から，検定統計量 t を計算する
4. もし $t > 1.729$ なら，I を 1 つ増やす

5. もし $k = N$ なら *step6* に進み，そうでないなら，k をインクリメントし，*step2* へ戻る．

6. $\hat{\alpha} = I/N$ と，近似的な誤差 $1.96\sqrt{\hat{\alpha}(1-\hat{\alpha})/N}$ を計算する．

こうして求められた $\hat{\alpha}$ がシミュレーションされた α の推定量であり，推定の誤差は，有意水準 α の時の信頼区間の半分の幅に当たる．

4.8.1 Accept-Reject Generation Algorithm

この小節では，累積分布関数の逆関数が，解析的に求まらない確率変数をシミュレートするために頻繁に利用される合否判定法 (accept-reject procedure) について紹介する．X を連続型確率変数とし，その密度関数を $f(x)$ で表すとする．この節では，この密度関数をターゲットとする密度関数，ターゲット密度関数と呼ぶ．また，確率変数 Y を生成することは，X の生成と比較して容易であるとし，Y の密度関数 $g(x)$ と，ある定数 M に対して，

$$f(x) \leq Mg(x),\ -\infty < x < \infty \tag{3}$$

が成り立つとする．このような $g(x)$ を instrumental pdf と呼ぶ．合否判定をアルゴリズムとして，以下にまとめる．

アルゴリズム 4.1. $f(x)$ を *pdf* とする．Y を $g(y)$ を *pdf* として持つ確率変数，U を区間 $(0,1)$ で定義される一様分布を持つ確率変数，Y, U は独立，(3) が成立しているとする．以下のアルゴリズムは，$f(x)$ を *pdf* として持つ確率変数 X を生成する．

1. Y, U を生成する．

2. もし，$U \leq \frac{f(Y)}{Mg(Y)}$ なら，$X = Y$ とする．そうでないなら，1 に戻る．

3. X は $f(x)$ を *pdf* として持つ．

証明 4.0. $-\infty < x < \infty$ とする．この時，

$$\begin{aligned}
P[X \leq x] &= P\left[Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\right] \\
&= \frac{P\left[Y \leq x,\ U \leq \frac{f(Y)}{Mg(Y)}\right]}{P\left[U \leq \frac{f(Y)}{Mg(Y)}\right]} \\
&= \frac{\int_{-\infty}^{x}\int_{0}^{f(y)/Mg(y)} g(y)dudy}{\int_{-\infty}^{\infty}\int_{0}^{f(y)/Mg(y)} g(y)dudy} \\
&= \frac{\int_{-\infty}^{x}\frac{f(y)}{Mg(y)}g(y)dy}{\int_{-\infty}^{\infty}\frac{f(y)}{Mg(y)}g(y)dy} \\
&= \frac{\int_{-\infty}^{x}\frac{f(y)}{M}dy}{\int_{-\infty}^{\infty}\frac{f(y)}{M}dy} \qquad (4.8.6) \\
&= \frac{M}{M}\int_{-\infty}^{x} f(y)dy = F(x) \qquad (4.8.7)
\end{aligned}$$

よって，両辺を y で微分することで，X の pdf が $f(x)$ であることを得る． □

合否判定アルゴリズムの利用例として，Robert&Casella(1999) の方法を紹介する．この方法はガンマ分布 $\Gamma(a,b)$ をシミュレートするものである．もし，$X \sim \Gamma(a,1)$ ならば，$bX \sim \Gamma(a,b)$ であることに注意すると，一般性を失わずに $b = 1$ とすることができる (後で b かければいいから)．もし，a が整数ならば，3 章の定理から，$X = \sum_{i=1}^{a} Y_i,\ Y_i \sim \Gamma(1,1) = Exp(1)$ である．指数分布の cdf の逆関数は閉じた形で得られ，それを使えば良いため，X は容易に生成できる．

a が整数でなく，$X \sim \Gamma(a,1)$ とする．また，$Y \sim \Gamma([a],1/b)$ を持つとする．ただしここで，のちに $b < 1$ なる b を選択し，$[a]$ は a を超えない最大の整数とする．例えば，$a = 36/13 = 2.7692...$ の時，$[a] = [2.7692...] = 2$ である．(3) のような規則を導くために，X, Y の pdf$h(x), t(x)$ の比

$$\frac{h(x)}{t(x)} = b^{-[a]}x^{a-[a]}e^{-(1-b)x} \qquad (4.8.9)$$

を考える．ただし，正規化定数は考慮しない (x, y の関数じゃないから)．次に，定数 b を決定する方法を考える．

(4.8.9) の導関数は，

$$\frac{d}{dx}(4.8.9) \text{ の式} = b^{-[a]}e^{-(1-b)x}\left[(a-[a]) - x(1-b)\right]x^{a-[a]-1}$$

であり，これは $x = (a-[a])/(1-b)$ で最大値をとる．よって，(4.8.9) はその最大値に上から抑えられて，

$$\frac{h(x)}{t(x)} \leq b^{-[a]}\left[\frac{a-[a]}{1-b}e\right]^{a-[a]} \qquad (4.8.11)$$

となる．次にしなければならないのは，b を決定することである．上の不等式の右辺を b について微分すると，

$$\frac{d}{db}b^{-[a]}(1-b)^{-(a-[a])} = -b^{-[a]}(1-b)^{[a]-a}\left[\frac{[a]-ab}{b(1-b)}\right] \tag{4.8.12}$$

を得る．これは，$b = [a]/a < 1$ にて極値を持つ．章末の練習問題から分かるように，この b は (4.8.11) の右辺の最小値を与える．ゆえに，もし，$b = [a]/a$ とするならば，(4.8.11) の等号は成立し，この時不等式は，可能な限り厳しくなっている．最終的な M の値は (4.8.11) 右辺の $b = [a]/a$ における値である．

¡note¿

4.9 Bootstrap Procedures / ブートストラップ法

4.9.1 Percentile Bootstrap Confidence Intervals

X を，pdf として $f(x;\theta)$, $\theta \in \Theta$ を持つ連続型確率変数とする．$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ を X の無作為標本，$\hat{\theta} = \hat{\theta}(\mathbf{X})$ を θ の点推定値とする．ここでは，リサンプリング手法であるパーセンタイルブートストラップ法を紹介する．とりあえず一旦

$$\hat{\theta} \sim N(\theta, \sigma^2_{\hat{\theta}}) \tag{4.9.1}$$

とする．すると，θ の $(1-\alpha)100\%$ の信頼区間 $(\hat{\theta}_L, \hat{\theta}_U)$ は，

$$\hat{\theta}_L = \hat{\theta} - z^{(1-\alpha/2)}\sigma_{\hat{\theta}},\ \hat{\theta}_U = \hat{\theta} - z^{(\alpha/2)}\sigma_{\hat{\theta}} \tag{4.9.2}$$

と表される．ここで，$z^{(\gamma)} = \Phi^{-1}(\gamma)$ は，標準正規分布の 100γ パーセンタイル点を表す．次に $\hat{\theta}^*$ を (4.9.1) と同様の分布に従う確率変数とする．(4.9.2) から，

$$P(\hat{\theta}^* \le \hat{\theta}_L) = P\left(\frac{\hat{\theta}^* - \hat{\theta}}{\sigma_{\hat{\theta}}} \le -z^{(1-\alpha/2)}\right) = \alpha/2 \tag{4.9.3}$$

同様に $P(\hat{\theta}^* \leq \hat{\theta}_U) = 1 - \alpha/2$ となる． 以上のパーセンタイルの考えを，正規分布を仮定せずに行いたい．$\hat{\theta}$ の cdf を $H(t)$ とする．いま，無限個の標本 $\mathbf{X}_1, \mathbf{X}_2, \ldots$ をとることができ，標本 $\mathbf{X}^*$ に対して $\hat{\theta}^* = \hat{\theta}(\mathbf{X}^*)$ を計算でき，θ^* の推定量のヒストグラムを形成できるとする．このヒストグラムのパーセンタイルが，(4.9.3) の信頼区間に対応する．

ブートストラップ法は，**一つの標本によって定められる経験分布から単純にリサンプルする**という考え方に基づく．リサンプリングは，**サイズは元の標本と同一で，無作為かつ，復元抽出**で行う．$\hat{F}_n$ を標本の経験的な分布関数とする．

以下に信頼区間のブートストラップ法のアルゴリズムを紹介する．B はブートストラップ標本をサンプルする回数を表す．一般的には $B \geq 3000$ らしい．

アルゴリズム 4.2. 以下の手順を行う．

1. $j = 1$ とする．
2. $j \leq B$ の間，以下のステップを繰り返す．
3. $x_1, x_2, \ldots, x_n$ から復元抽出してサイズ n の標本 $\mathbf{x}_j^*$ を得る．
4. パラメータ $\hat{\theta}_j^* = \hat{\theta}(\mathbf{x}_j^*)$ を計算する．
5. j をインクリメントする．
6. $\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$ のように，パラメータの順序統計量を求め，

$$(\hat{\theta}^*_{(m)}, \hat{\theta}^*_{(B+1-m)}) \tag{4.9.4}$$

という区間を形成する．これが，$\alpha/2 * 100\%, (1 - \alpha/2) * 100\%$ 点である．ただし，$m = [(\alpha/2)B]$ としており，パーセンタイルは m に依存する．

(4.9.4) は信頼区間のパーセンタイルブートストラップと呼ばれている．

¡note¿

4.9.2 Bootstrap Testing Procedures

ブートストラップ法は，仮説検定にも利用できる．まず 2 標本問題について考える．$\mathbf{X}' = (X_1, X_2, \ldots, X_{n_1})$ を $\mathrm{cdf}F(x)$ を持つ分布からの無作為標本とし，$\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_{n_2})$ を $\mathrm{cdf}F(x - \Delta)$ を持つ分布からの無作為標本とする．Δ は二つの標本の位置のずれを表すパラメータである．特に，それぞれの分布の平均 μ_X, μ_Y が存在するとすると，$\Delta = \mu_Y - \mu_X$ で与えられる．以下の片側検定を考える．

$$H_0 : \ \Delta = 0 \ v.s. \ H_1 : \ \Delta > 0 \tag{4.9.6}$$

この問題の検定統計量は，標本平均の差

$$V = \bar{Y} - \bar{X} \tag{4.9.7}$$

である．この検定の p 値は

$$\hat{p} = P_{H_0}[V \geq \bar{y} - \bar{x}] \tag{4.9.8}$$

であり，これをブートストラップで推定することを考える．以下にアルゴリズムを記す．

アルゴリズム 4.3. 以下の手順を行う．

1. $\mathbf{z}' = (\mathbf{x}', \mathbf{y}')$ のように標本を結合する．
2. $j = 1$ とする．
3. $j \leq B$ の間，以下を繰り返す．
4. $\mathbf{z}$ から復元抽出でサイズ n_1 の無作為標本を得る．その標本を $\mathbf{x}^{*\prime} = (x_1^*, x_2^*, \ldots, x_{n_1}^*)$ とし，$\bar{x}_j^*$ を計算する．
5. $\mathbf{z}$ から復元抽出でサイズ n_2 の無作為標本を得る．その標本を $\mathbf{y}^{*\prime} = (y_1^*, y_2^*, \ldots, y_{n_2}^*)$ とし，$\bar{y}_j^*$ を計算する．
6. $v_j^* = \bar{y}_j^* - \bar{x}_j^*$ を計算する．
7. p 値のブートストラップ推定量は，

$$\hat{p}^* = \frac{\#_{j=1}^{B}\{v_j^* \geq v\}}{B} \tag{4.9.9}$$

で与えられる．

¡note¿

4.10 Tolerance Limits for Distribution / 分布の許容限界

定理 4.10.1

$Y_1, Y_2, \ldots, Y_n$ を無作為標本の順序統計量とする．確率変数 $Z_i = F(Y_i)$ の同時 pdf は，

$$h(z_1, z_2, \ldots, z_n) = \begin{cases} n!\ 0 < z_1 < z_2 < \cdots < z_n < 1 \\ 0\ elsewhere \end{cases} \tag{4.10.1}$$

証明 4.0. $a < x < b$ にて連続で正の値をとる *pdf* を持つ分布からのサイズ n の無作為標本を $X_1, X_2, \ldots, X_n$ で表す．$F(x)$ をそれらの分布関数とする．確率変数 $F(X_1), F(X_2), \ldots, F(X_n)$ を考える．これらの確率変数は独立で，それぞれ $U(0,1)$ に従う．つまり，これらは，サイズ n の一様分布からの標本である．次に $F(X_1), F(X_2), \ldots, F(X_n)$ の順序統計量 Z_i を考える．元の $X_1, X_2, \ldots, X_n$ の順序統計量を $Y_1, Y_2, \ldots, Y_n$ とすると，分布関数 F は，狭義増加関数であることから，$Z_1 = F(Y_1), Z_2 = F(Y_2), \ldots, Z_n = F(Y_n)$ となる．よって，(4.4.1) から $Z_1, Z_2, \ldots, Z_n$ の同時 *pdf* は定理のものとなる． □

もし $F(y_j) - F(y_i) \geq p$ ならば，少なくとも X の分布の確率の $100p\%$ は y_i, y_j の間に存在すると言える．$\gamma = P[F(Y_j) - F(Y_i) \geq p]$ が与えられているとする．その時，確率変数の区間 (Y_i, Y_j) は X の分布の確率の少なくとも $100p\%$ を確率 γ で含む．今，y_i, y_j をそれぞれ，Y_i, Y_j の実現値とする．区間 (y_i, y_j) は X の分布に対する確率の $100p\%$ の $100\gamma\%$ の tolerance interval と呼ばれている．

¡note¿