標本四分位数の漸近的評価と比較
Asympototic evaluation and comparison for sample quartiles

数学専攻　早崎 久登
HAYASAKI, Hisato

1　はじめに

母集団において確率分布を等確率に四等分する点として四分位数（第 1 四分位数, 第 2 四分位数, 第 3 四分位数）が一意に定義される. これを標本から推定するための推定量として, 様々な標本四分位数が提案されている. 高等学校において「四分位数」として紹介されているものも, そのうちの推定量の一つであるが, これは統計解析ソフト R や Excel, Mathematica において用いられているどの方法とも一致していない. Langford (2006) では 15 種類の標本四分位数の簡易的な比較がされ, Hyndman and Fan (1996) では p 分位数についての比較考察がされているが, それぞれの推定量の数理統計学的な比較はほとんどなされてきていない.

そこで, 本研究では, さまざまな標本四分位数の中からいくつかの推定量に対し, 漸近バイアスを導出し, その数値的比較をする.

2　準備

2.1　母集団の分位数と四分位数

四分位数とは集団を 4 等分する点を意味する. 値が小さい方から順に第 1 四分位数 (25% 点), 第 2 四分位数 (50% 点, 中央値と同義), 第 3 四分位数 (75% 点) となる.

確率変数 X を離散型あるいは連続型とし, その累積分布関数を $F(x)$ とする. 任意の $0 \leq p \leq 1$ に対し, $F(x)$ の p 分位数, あるいは下側 100p パーセント点は

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\},\ 0 \leq p \leq 1$$

と定義される. この $Q(p)$ を分位関数という. とくに, $p = 1/4$ とおいたものを第 1 四分位数といい Q_1 と表し, $p = 3/4$ とおいた分位数を第 3 四分位数といい Q_3 と表す.

2.2　標本の分位数と四分位数

母集団分布 $F(x)$ の母集団からの無作為標本を $X_1, X_2, \ldots, X_n$ とする. このとき, $X_1, X_2, \ldots, X_n$ を昇順に並べ替えた $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ を $X_1,\ X_2,\ \ldots,\ X_n$ の順序統計量という.

この順序統計量を用いて母集団の分位数 $Q(p)$ を推定するための推定量としてさまざまなものが提案されている. その多くは

$$\hat{Q}(p) = (1 - \varepsilon)X_{(j)} + \varepsilon X_{(j+1)}$$

という形により表される. ここで $\varepsilon \in (0, 1)$ である. すなわち, 隣接する 2 つ以下の順序統計量の線形和によって 1 つの分位数を推定している.

2.2.1 Exclusive な推定量

標本四分位数の一つとして, Tukey (1977) のヒンジがある. これは, データを中央値で 2 分割し中央値を含む下半分のデータの中央値, 中央値を含む上半分のデータの中央値をそれぞれ標本第 1 四分位数, 標本第 3 四分位数としている. この推定量は Langford (2006) では中央値を含むという意味で Inclusive な推定量とよばれている. これに対し, 高等学校では, 中央値を含まない前半の中央値, 中央値を含まない後半の中央値をそれぞれ標本第 1 四分位数, 標本第 3 四分位数としている. Langford (2006) ではこの推定量を Inclusive な推定量に対して Exclusive な推定量とよんでいる. Exclusive な推定量は, $n = 4j$ のとき

$$\hat{Q}_1 = \frac{1}{2}X_{(j)} + \frac{1}{2}X_{(j+1)},\ \hat{Q}_3 = \frac{1}{2}X_{(3j)} + \frac{1}{2}X_{(3j+1)},$$

$n = 4j + 1$ のとき

$$\hat{Q}_1 = \frac{1}{2}X_{(j)} + \frac{1}{2}X_{(j+1)},\ \hat{Q}_3 = \frac{1}{2}X_{(3j+1)} + \frac{1}{2}X_{(3j+2)},$$

$n = 4j + 2$ のとき

$$\hat{Q}_1 = X_{(j+1)},\ \hat{Q}_3 = X_{(3j+2)},$$

$n = 4j + 3$ のとき

$$\hat{Q}_1 = X_{(j+1)},\ \hat{Q}_3 = X_{(3j+2)}$$

と表される.

2.2.2 QUARTILE.EXC 関数の推定量

Excel は母四分位数の推定量の一つを

$$\hat{Q}_1 = (1-\varepsilon)X_{(\lfloor \frac{N}{4} \rfloor)} + \varepsilon X_{(\lfloor \frac{N}{4} \rfloor + 1)},$$
$$\hat{Q}_3 = (1-\delta)X_{(\lfloor \frac{3N}{4} \rfloor)} + \delta X_{(\lfloor \frac{3N}{4} \rfloor + 1)}$$

としている. ここで, データ数 n に対して, $N = n + 1$ と置き, $\varepsilon = N/4 - \lfloor N/4 \rfloor$, $\delta = 3N/4 - \lfloor 3N/4 \rfloor$ としている. 上記の式は $n = 4j, n = 4j + 1, 4j + 2, 4j + 3$ の各場合に分けると

$$\hat{Q}_1 = \begin{cases} \frac{3}{4}X_{(j)} + \frac{1}{4}X_{(j+1)} & (n = 4j,\ \varepsilon = 1/4) \\ \frac{1}{2}X_{(j)} + \frac{1}{2}X_{(j+1)} & (n = 4j+1,\ \varepsilon = 1/2) \\ \frac{1}{4}X_{(j)} + \frac{3}{4}X_{(j+1)} & (n = 4j+2,\ \varepsilon = 3/4) \\ X_{(j+1)} & (n = 4j+3,\ \varepsilon = 0) \end{cases}$$

$$\hat{Q}_3 = \begin{cases} \frac{1}{4}X_{(3j)} + \frac{3}{4}X_{(3j+1)} & (n = 4j,\ \delta = 3/4) \\ \frac{1}{2}X_{(3j)} + \frac{1}{2}X_{(3j+1)} & (n = 4j+1,\ \delta = 1/2) \\ \frac{3}{4}X_{(j)} + \frac{1}{4}X_{(j+1)} & (n = 4j+2,\ \delta = 1/4) \\ X_{(j+1)} & (n = 4j+3,\ \delta = 0) \end{cases}$$

と表される. この推定量は, Excel において QUARTILE.EXC 関数とよばれており, 統計解析ソフトウェア MINITAB などでも四分位数の推定量として用いられている.

3　標本四分位数の漸近的評価

母四分位数の推定量としてさまざまな標本四分位数が提案されている. しかし, これらの推定量のうちどれが統計的に優れた性質をもつのかは非常に重要な問題であるが, 過去の研究は非常に少ない. Mudholkar and Huston (1997) では, Excel の QUARTILE.EXC 関数で用いられている標本四分位数について漸近展開を行い, バイアスなどの漸近評価を行っているが, 他の標本四分位数に関しては同様の研究はみられない.

本節では, 主に 2 つの標本四分位数の期待値を導出することによって, 母四分位数とのバイアスを導出する.

3.1　QUARTILE.EXC 関数の推定量

Mudholkar and Huston (1997) では Excel の QUARTILE.EXC 関数で用いられている標本四分位数の漸近展開が与えられている. 標本四分位数 $\hat{Q}_1$, $\hat{Q}_3$ のバイアス $\mathrm{Bias}(\hat{Q}_j) = E(\hat{Q}_j) - Q_j$ は

$$\mathrm{Bias}(\hat{Q}_1) = \frac{3Q_1''}{32n} + \frac{B_1(\varepsilon)}{2048n^2} + O(n^{-3}),$$
$$\mathrm{Bias}(\hat{Q}_3) = \frac{3Q_3''}{32n} + \frac{B_3(\delta)}{2048n^2} + O(n^{-3})$$

で与えられる. ここで,

$$B_1(\varepsilon) = 128[8\varepsilon(1-\varepsilon) - 3]Q_1'' + 64Q_1^{(3)} + 9Q_1^{(4)},$$
$$B_3(\delta) = 128[8\delta(1-\delta) - 3]Q_3'' - 64Q_3^{(3)} + 9Q_3^{(4)}$$

である. ここで $Q_1 = Q(1/4)$, $Q_3 = Q(3/4)$ である. 修士論文において本証明を丁寧に行った.

母集団分布 $F(x)$ が正規分布の場合, $Q(p)$ は, p =1/4 において上に凸, p =3/4 において下に凸であり, 2 次微分 $Q''(p)$ はそれぞれ負, 正となる. すなわち, 第 1 四分位数のバイアスは負であり母第 1 四分位数を過小推定し, 第 3 四分位数のバイアスは正であり母第 3 四分位数を過大推定することがわかる. このことから, 四分位範囲が過大推定されることがわかる.

修士論文では, 正規分布 $N(50, 10^2)$ から標本を生成し, 標本四分位数のバイアスを数値的に調べた結果, 漸近展開の結果と同様の傾向がみられた.

3.2　Exclusive な推定量

Exclusive な推定量 $\hat{Q}_1$, $\hat{Q}_3$ のバイアス $\mathrm{Bias}(\hat{Q}_j) = E(\hat{Q}_j) - Q_j$ は, $n = 4j$ のとき

$$\mathrm{Bias}(\hat{Q}_1) = \frac{8Q_1' + 3Q_1''}{32n} + \frac{-512Q_1' + 64Q_1'' + 112Q_1^{(3)} + 9Q_1^{(4)}}{2048n^2} + O(n^{-3}),$$
$$\mathrm{Bias}(\hat{Q}_3) = \frac{-8Q_3' + 3Q_3''}{32n} + \frac{512Q_3' + 64Q_3'' - 112Q_3^{(3)} + 9Q_3^{(4)}}{2048n^2} + O(n^{-3})$$

$n = 4j + 1$ のとき

$$\mathrm{Bias}(\hat{Q}_1) = \frac{3Q_1''}{32n} + \frac{-128Q_1'' + 64Q_1^{(3)} + 9Q_1^{(4)}}{2048n^2} + O(n^{-3}),$$
$$\mathrm{Bias}(\hat{Q}_3) = \frac{3Q_3''}{32n} + \frac{-128Q_3'' - 64Q_3^{(3)} + 9Q_3^{(4)}}{2048n^2} + O(n^{-3})$$

$n = 4j + 2$ のとき

$$\mathrm{Bias}(\hat{Q}_1) = \frac{8Q_1' + 3Q_1''}{32n} + \frac{-512Q_1' - 192Q_1'' + 112Q_1^{(3)} + 9Q_1^{(4)}}{2048n^2} + O(n^{-3}),$$
$$\mathrm{Bias}(\hat{Q}_3) = \frac{-8Q_3' + 3Q_3''}{32n} + \frac{512Q_3' - 192Q_3'' - 112Q_3^{(3)} + 9Q_3^{(4)}}{2048n^2} + O(n^{-3})$$

$n = 4j + 3$ のとき

$$\mathrm{Bias}(\hat{Q}_1) = \frac{3Q_1''}{32n} + \frac{-384Q_1'' + 64Q_1^{(3)} + 9Q_1^{(4)}}{2048n^2} + O(n^{-3}),$$
$$\mathrm{Bias}(\hat{Q}_3) = \frac{3Q_3''}{32n} + \frac{-384Q_3'' - 64Q_3^{(3)} + 9Q_3^{(4)}}{2048n^2} + O(n^{-3})$$

となる. 修士論文において本証明を丁寧に行った.

修士論文では数値的にバイアスを調べた結果, Exclusive な推定量のバイアスは n が偶数, 奇数のときとでバイアスが反転しており, n が偶数のときは第 1 四分位数を過大推定, 第 3 四分位数を過小推定しているのに対し, n が奇数のときは第 1 四分位数を過小推定, 第 3 四分位数を過大推定していることがわかった. つまり, Exclusive な推定量は n が偶数のとき四分位範囲を過小推定, n が奇数のとき四分位範囲を過大推定しているということである. これは漸近展開の結果とも一致している.

4 まとめ

本研究では, 四分位数に対して, 数理統計学的な推定量としてのよさの評価を, 主としてバイアスの観点から実際に行った. 概して言えば, n が比較的小さいときには, バイアスに若干の優劣はあるが, どの標本四分位数を用いてもそれほど問題はないと考えられる. しかし, 高等学校で用いられている Exclusive な推定量では, 正規分布のとき n が奇数なら四分位範囲 (IQR) を過大推定, n が偶数なら四分位範囲 (IQR) を過小推定していることがわかった. このことから, n が偶数の集団と奇数の集団のばらつきの大きさを四分位範囲で比較する場合には少々注意が必要かもしれない.

今後の課題としては, Exclusive な推定量の漸近分散や漸近 MSE の導出, その他の標本四分位数の漸近展開と比較が考えられる.

参考文献

[1] Hyndman, R.J. and Fan, Y. (1996) Sample quantile in statistical packages, *Statistical Computing*, **50**, 4, 361-365.

[2] Langford, E. (2006) Quartiles in elementary statistics, *Journal of statistics education.*

[3] Mudholkar,G.S. and Huston, A.D. (1997) Improvements in the bias and precision of sample quartiles, *Statistics* , 239-257.

[4] Tukey, J.W. (1977) *Exploratory Data Analysis,* Reading, MA: Addison-Wesley.