

ジャックナイフ統計量に関する高次漸近的性質の検討 ——相関係数の場合——

| | | |
|------------|-------|------------------|
| 千葉大自然科学研究科 | 汪 金芳 | (Jin Fang Wang) |
| 木更津工業高専 | 大内 俊二 | (Shunji Ouchi) |
| 千葉大理学部 | 田栗 正章 | (Masaaki Taguri) |

1 Smooth function model

$\{Z_p = (Z_p^1, \dots, Z_p^r)\}_{p \geq 1}$ を *i.i.d.* r 次元確率ベクトル列, g_1, \dots, g_d を R^r で定義された可測関数とするとき、

$$X_p = (X_p^1, \dots, X_p^d) = (g_1(Z_p), \dots, g_d(Z_p)),$$

$$E(X_p) = \eta$$

で定義される d 次元確率ベクトル X_p を考え、平均ベクトル η の滑らかな関数 $h(\eta) = \theta$ (smooth function model と呼ばれる) を、 $\hat{\theta} = h(\bar{X})$ で推定するものとする。ただし $\bar{X} = (\bar{X}^1, \dots, \bar{X}^d) = \sum_{p=1}^N X_p / N$, N は標本数である。本報告では $h(\bar{X})$ としてとくに標本相関係数を考える。すなわち $r=2$, $d=5$ とし, $X^1 = g_1(Z) = Z^1$, $X^2 = g_2(Z) = Z^2$, $X^3 = g_3(Z) = Z^1 Z^2$, $X^4 = g_4(Z) = (Z^1)^2$, $X^5 = g_5(Z) = (Z^2)^2$ とおけば、標本相関係数は

$$\hat{\theta} = h(\bar{X}) = \frac{\bar{X}^3 - \bar{X}^1 \bar{X}^2}{\sqrt{\bar{X}^4 - (\bar{X}^1)^2} \sqrt{\bar{X}^5 - (\bar{X}^2)^2}}$$

と表せる。

2 ジャックナイフ統計量

N 個の観測値ベクトル X_1, X_2, \dots, X_N から p 番目の観測値ベクトル X_p を除いた $N - 1$ 個に亘る標本平均を $\bar{X}_{(-p)}$ とし、 $\bar{X}_{(-p)}$ の滑らかな関数 $h(\bar{X}_{(-p)})$ を $\hat{\theta}_{(-p)}$ とする。このとき

$$p \hat{\theta}_{(-p)} = N \hat{\theta} - (N - 1) \hat{\theta}_{(-p)} \quad (p = 1, \dots, N)$$

で定義される ${}_{pv}\widehat{\theta}_{(-p)}$ はジャックナイフ pseudo-value と呼ばれ、その平均

$$\widehat{\theta}_J = \frac{1}{N} \sum_{p=1}^N {}_{pv}\widehat{\theta}_{(-p)}$$

によって θ のジャックナイフ推定量は定義される。またここでは $\widehat{\theta}_J$ の分散の推定量としてジャックナイフ推定量

$$\widehat{Var}_J = \frac{1}{N-1} \sum_{p=1}^N ({}_{pv}\widehat{\theta}_{(-p)} - \widehat{\theta}_J)^2$$

を用いる。このとき $\widetilde{\theta}_J = \sqrt{N}(\widehat{\theta}_J - \theta)$, $T = \widetilde{\theta}_J / \sqrt{\widehat{Var}_J}$ と定義する。後者はジャックナイフ-t 統計量と呼ばれている。

3 分布関数の3次のエッジワース展開

3.1 $\widetilde{\theta}_J$, T の分布の場合

Wang(1994) は、一般の smooth function model(多変量) に対して、 X の8次のモーメントの存在と適当な正則条件の下、 $\widetilde{\theta}_J$, T それぞれの4次までのキュミュラントを $1/N$ のオーダーまで求め、分布関数の3次のエッジワース展開を与えた。ここではその応用の一つとして相関係数に対して、 $\widetilde{\theta}_J$, T それぞれの4次までのキュミュラントを具体的に計算し、分布関数の3次のエッジワース展開を求めた。ただし相関係数は、 X の位置および尺度変換に対して不変なので $E(X^1) = E(X^2) = 0$, $E(X^3) = \rho$, $E(X^4) = E(X^5) = 1$ とした。このとき $\theta = h(\eta) = \rho$ となる。次式では、 $\widetilde{\theta}_J$, T を一般的に $\widetilde{\theta}$ と書くものとする。

$\widetilde{\theta}$ の r 次のキュミュラント κ_N^r が

$$\kappa_N^r = N^{-(r-2)/2} (C_0^r + C_1^r N^{-1} + C_2^r N^{-2} + \dots), \quad r \geq 1$$

と $N^{-1/2}$ の巾級数に展開されているとき、

$$\begin{aligned} F_{\widetilde{\theta}}(y) &= Pr(\widetilde{\theta} \leq y) \\ &= \Phi(y; \sigma^2) + \left(\frac{Q_1(y)}{\sqrt{N}} + \frac{Q_2(y)}{N} \right) \phi(y; \sigma^2) + o(N^{-1}), \end{aligned} \quad (1)$$

$$Q_1(y) = -(C_1^1 + \frac{1}{6} C_0^3 H_2(y)), \quad (2)$$

$$Q_2(y) = -\left[\frac{1}{2}(C_1^2 + (C_1^1)^2)H_1(y) + \frac{1}{24}(C_0^4 + 4C_1^1C_0^3)H_3(y) + \frac{1}{72}(C_0^3)^2H_5(y)\right]. \quad (3)$$

ここで $\Phi(y; \sigma^2)$, $\phi(y; \sigma^2)$ は、それぞれ正規分布 $N(0, \sigma^2)$ の分布関数と密度関数、また $H_k(y)$ ($k = 1, 2, 3, 5$) は k 次のエルミート多項式である。

(2), (3) 式における係数 C_k^r ($k = 0, 1; r = 1, 2, 3, 4$) は Z^1, Z^2 の 8 次までの同時モーメントと ρ を用いた式で書けるが、 $Z = (Z^1, Z^2)$ に特定の分布を仮定しない場合、それらの式は膨大なものとなり、ここでは紹介出来ない。したがって以下の議論はすべて $Z = (Z^1, Z^2)$ が 2 次元正規分布 $N_2(0, 0, 1, 1, \rho)$ に従うという仮定の下でのものとする。

$\tilde{\theta} = T$ のとき

$$Q_1(y) = -\rho(y^2 + 1), \quad (4)$$

$$Q_2(y) = \frac{y(1 + 5\rho^2 - 5y^2 + 4\rho^2y^2 - 2\rho^2y^4)}{4}, \quad (5)$$

$$\sigma^2 = 1.$$

$\tilde{\theta} = \tilde{\theta}_J$ のとき

$$Q_1(y) = \frac{\rho\{y^2 - (1 - \rho^2)^4\}}{(1 - \rho^2)^3}, \quad (6)$$

$$Q_2(y) = \frac{y}{4(1 - \rho^2)^6} \{-13 + 43\rho^2 - 35\rho^4 - 37\rho^6 + 93\rho^8 - 79\rho^{10} + 35\rho^{12} - 7\rho^{14} + (3 + 10\rho^2 - 33\rho^4 + 26\rho^6 - 8\rho^8 + 2\rho^{10})y^2 - 2\rho^2y^4\}, \quad (7)$$

$$\sigma^2 = (1 - \rho^2)^2 \quad (\tilde{\theta}_J \text{ の漸近分散}).$$

3.2 T の分布に対する 2 次と 3 次のエッジワース近似の比較

われわれの与えた統計量 T の分布関数のエッジワース展開の妥当性および近似の様子を見るために、2 次の近似 (破線)、3 次の近似 (実線)、および真の分布を同一平面上に描いたものが下図である。ただし真の分布は、2 次元の正規乱数を 1000 回発生させ、1000 個の T の値を計算することから作られた経験分布関数である。図 1 から図 4 は $\rho = 0.8$ の場合について、標本の大きさ N を 10, 25, 50, 100 と変化させたときのものである。

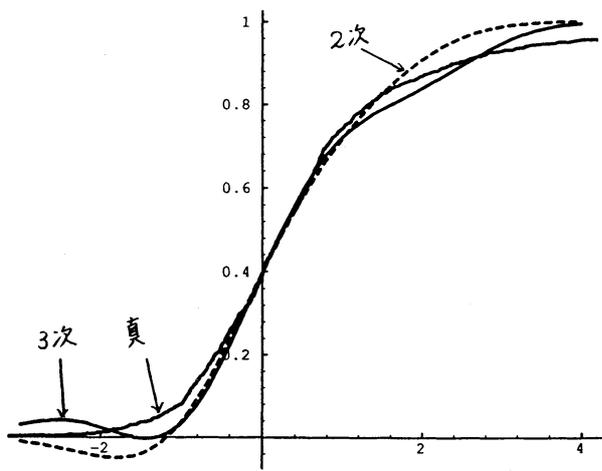


図 1: $\rho = 0.8, N = 10$ の場合

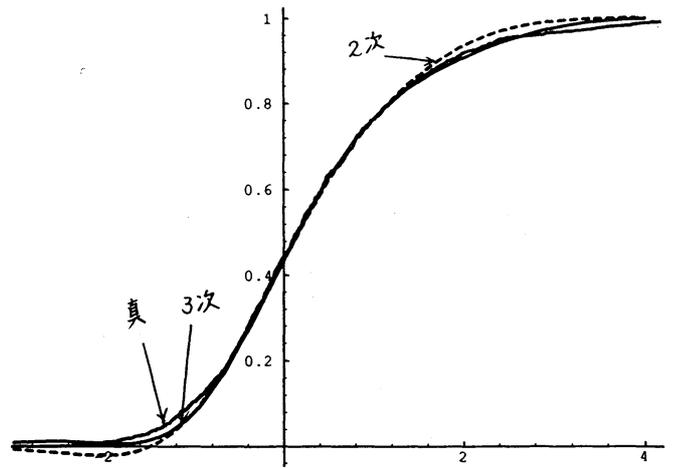


図 2: $\rho = 0.8, N = 25$ の場合

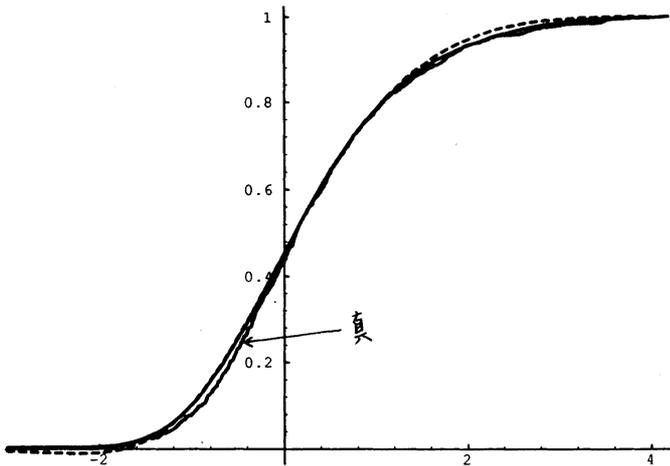


図 3: $\rho = 0.8, N = 50$ の場合

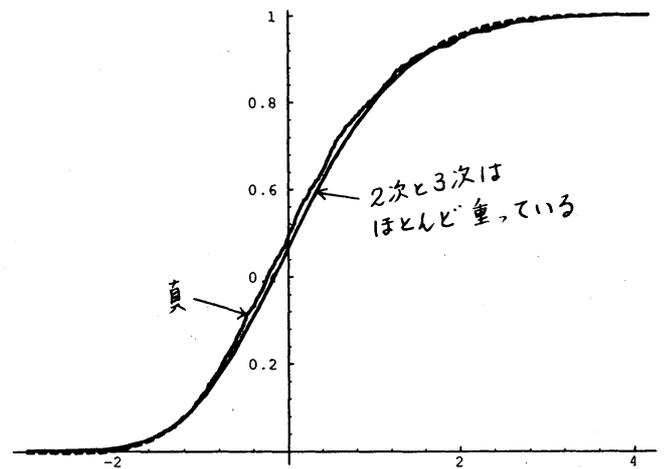


図 4: $\rho = 0.8, N = 100$ の場合

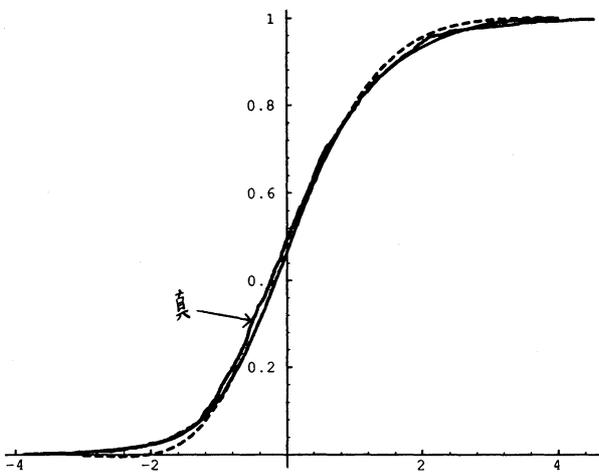


図 6: $\rho = 0.4, N = 10$ の場合

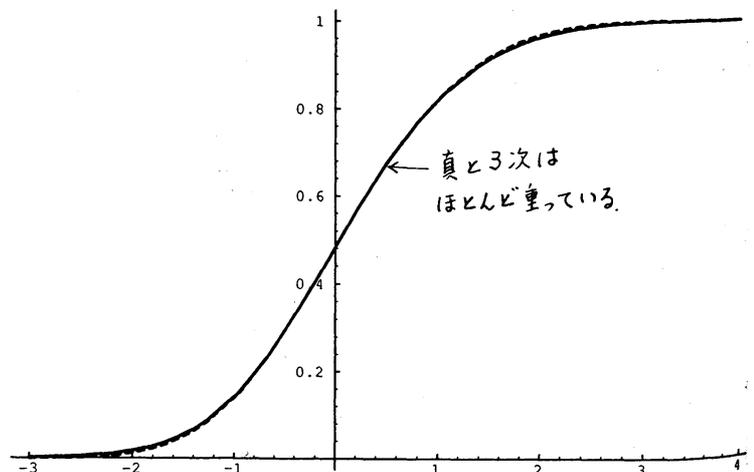


図 5: $\rho = 0.4, N = 10$ の場合

これらの図から、真の分布に対する3次の近似は2次の近似に比べほぼ一様に良いことが読み取れる。特に分布の両裾に近づくほどそれが顕著である。また標本の大きさ N が大きくなるにつれ、2次と3次の近似の差は小さくなり、共に真の分布に近づいていくこともわかる。原点の近傍において、2次と3次の近似がほぼ等しいことが分かるが、このことは奇数次のエルミート多項式がそこで0に近い値を取ることと(3)式から納得できることである。図5と図6は $\rho = 0.4$ ($N = 25, 100$) の場合の図である。 $\rho = 0.8$ の場合と同じ傾向が見られるが、エッジワース展開による近似が $\rho = 0.8$ の場合よりも良くなっている。

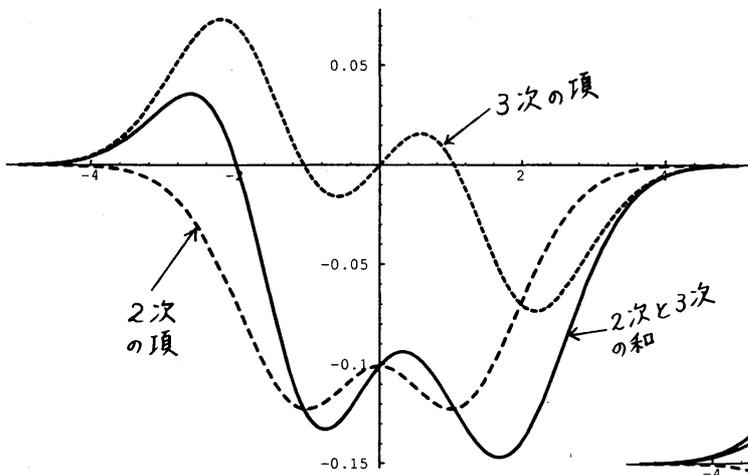


図7: $\rho = 0.8, N = 10$ の場合

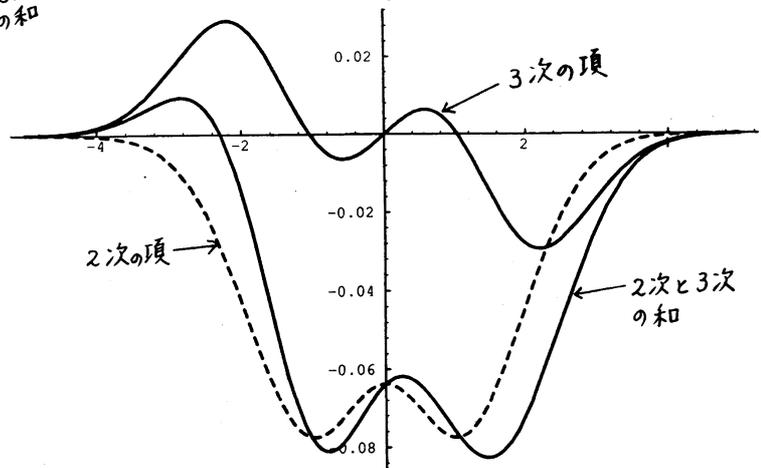


図8: $\rho = 0.8, N = 25$ の場合

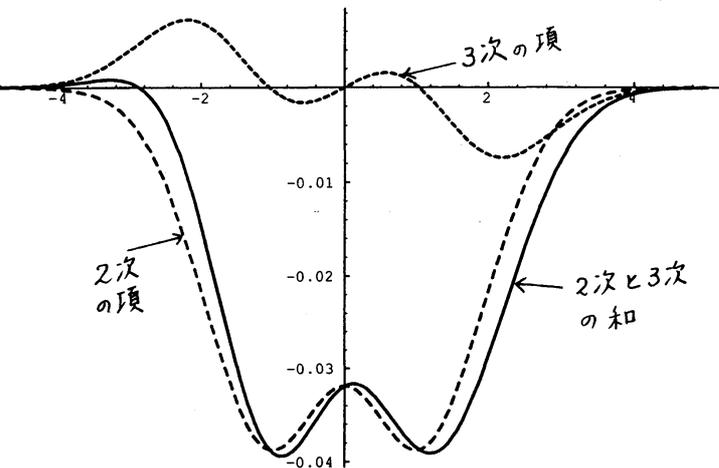


図9: $\rho = 0.8, N = 100$ の場合

図7から図9は、 $\rho = 0.8$ の場合に、展開式(1)における3次の項が2次の項による近似をどのように補正しているかをみるために描いた図であり、各々標本数 N は、10, 25, 100 である。図1と図7, 図2と図8, 図4と図9などを見比べると、3次の項を加えることは、左裾においては2次の項によって負の方向に過大修正されたものを正の方向へ引き戻し、右裾では2次の項による正の方向への過大修正を負の方向へ引き戻していることになっていることがわかる。

4 相関係数に対する信頼区間の構成

4.1 T に対するコーニッシュ・フィッシャー展開

統計量 T の分布関数が(1)式のようにエッジワース展開されているとき、 w_α を T の $100\alpha\%$ 点、すなわち

$$w_\alpha = \inf\{y : Pr(T \leq y) \geq \alpha\},$$

z_α を標準正規分布の $100\alpha\%$ 点とすると、

$$w_\alpha = z_\alpha + \frac{p_1(y)}{\sqrt{N}} + \frac{p_2(y)}{N} + o(N^{-1}), \quad (8)$$

$$p_1(y) = -Q_1(y), \quad (9)$$

$$p_2(y) = Q_1(y)Q_1'(y) - \frac{1}{2}y\{Q_1(y)\}^2 - Q_2(y) \quad (10)$$

が、 $\varepsilon < \alpha < 1 - \varepsilon$ ($0 < \varepsilon < 1/2$) で一様に成り立つ。ここで $Q_1(y)$, $Q_2(y)$ は(4),(5)式で与えられる。ところで(8)式をそのまま使って、現実に統計的推測を行うことは出来ない。なぜなら(9),(10)式より、 $p_1(y)$, $p_2(y)$ には、未知のパラメータ ρ が含まれているからである。したがって現実には $p_1(y)$, $p_2(y)$ をそれらの推定値 $\hat{p}_1(y)$, $\hat{p}_2(y)$ で置き換えた式

$$\widehat{w}_\alpha = z_\alpha + \frac{\hat{p}_1(y)}{\sqrt{N}} + \frac{\hat{p}_2(y)}{N} + o(N^{-1}) \quad (11)$$

を使わなければならない。理論的には $p_1(y)$, $p_2(y)$ はそれぞれ、それらの推定値 $\hat{p}_1(y)$, $\hat{p}_2(y)$ と $O_p(N^{-1/2})$ のズレがあるから、 \widehat{w}_α は真の w_α から $O_p(N^{-1})$ のズレをもつことになる。

コーニッシュ・フィッシャー展開(8)から作られる $\theta = \rho$ に対する信頼係数 $1 - 2\alpha$ の近似(両側)信頼区間は、不等式

$$w_\alpha < T < w_{1-\alpha}$$

の T に $\hat{\theta}_J / \sqrt{\widehat{Var}_J}$ を代入して解き直すことから

$$\hat{\theta}_J - \frac{\sqrt{\widehat{Var}_J}}{\sqrt{N}} w_{1-\alpha} < \theta < \hat{\theta}_J - \frac{\sqrt{\widehat{Var}_J}}{\sqrt{N}} w_\alpha \quad (12)$$

しかし前述したように、 $w_\alpha, w_{1-\alpha}$ には真のパラメータが含まれるので、現実構成可能な信頼区間は、(11)式で与えられる \hat{w}_α を用いた

$$\hat{\theta}_J - \frac{\sqrt{\widehat{Var}_J}}{\sqrt{N}} \hat{w}_{1-\alpha} < \theta < \hat{\theta}_J - \frac{\sqrt{\widehat{Var}_J}}{\sqrt{N}} \hat{w}_\alpha \quad (13)$$

である。次節におけるシミュレーションでは、(11)式における $\hat{p}_1(y), \hat{p}_2(y)$ として、(8)式における $p_1(y), p_2(y)$ に含まれる ρ を、その推定値で置き換えたものを用いた。

4.2 近似信頼区間の数値的検討

表1から表3には、前節で定義した近似信頼区間の被覆確率および正確度が示されている。ここに正確度は、正確度 = 真の信頼限界 - 近似信頼限界で定義されるものである。ただし、ここでは真の信頼限界は、3.2節同様2次元正規乱数を発生させ作られた1000個の T の値を大きさの順に並べ、上・下位 1000α 番目のものを拾い出すことにより求めたものである。また被覆確率は(12)または(13)式で与えられる近似信頼区間を1000個作り、そのうち真の ρ の値がそれらの中に入るものの割合を計算したものである。表中 $\rho, \hat{\rho}, \hat{\rho}_J$ とあるのは、 $p_1(y), p_2(y)$ に含まれる ρ として、真の値、通常の推定値、ジャックナイフ推定値をそれぞれ用いたことを意味するものである。

¹ここでは簡単のため両裾の確率が同じ($=\alpha$)場合を考えた。

表 1: 90%近似信頼区間 ($\rho = 0.8$ の場合)

| N | 近似 | 被覆確率 | | | 正確度 (下側信頼限界) | | | 正確度 (上側信頼限界) | | |
|----|----|--------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
| | | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ |
| 10 | 3次 | 0.862 | 0.899 | 0.896 | 0.6491 | 0.6319 | 0.6274 | -0.7646 | -0.7804 | -0.7859 |
| | 2次 | 0.754 | 0.803 | 0.786 | 0.5695 | 0.5531 | 0.5481 | -0.6851 | -0.7015 | -0.7065 |
| | 正規 | 0.779 | | | 0.4318 | | | -0.8228 | | |
| 25 | 3次 | 0.909 | 0.925 | 0.925 | 0.2966 | 0.2944 | 0.2949 | -0.2877 | -0.2897 | -0.2892 |
| | 2次 | 0.861 | 0.880 | 0.880 | 0.2793 | 0.2772 | 0.2777 | -0.2704 | -0.2725 | -0.2720 |
| | 正規 | 0.867 | | | 0.2319 | | | -0.3177 | | |
| 50 | 3次 | 0.898 | 0.911 | 0.911 | 0.1818 | 0.1815 | 0.1816 | -0.1871 | -0.1874 | -0.1873 |
| | 2次 | 0.878 | 0.885 | 0.885 | 0.1761 | 0.1758 | 0.1759 | -0.1813 | -0.1817 | -0.1816 |
| | 正規 | 0.866 | | | 0.1540 | | | -0.2034 | | |

表 2: 90%近似信頼区間 ($\rho = 0.4$ の場合)

| N | 近似 | 被覆確率 | | | 正確度 (下側信頼限界) | | | 正確度 (上側信頼限界) | | |
|----|----|--------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
| | | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ |
| 10 | 3次 | 0.887 | 0.936 | 0.936 | 1.2912 | 1.2656 | 1.2734 | -1.4130 | -1.4393 | -1.4324 |
| | 2次 | 0.805 | 0.888 | 0.891 | 1.1341 | 1.1082 | 1.1155 | -1.2559 | -1.2818 | -1.2745 |
| | 正規 | 0.800 | | | 0.9930 | | | -1.3970 | | |
| 25 | 3次 | 0.896 | 0.925 | 0.926 | 0.6436 | 0.6390 | 0.6400 | -0.6711 | -0.6757 | -0.6747 |
| | 2次 | 0.859 | 0.898 | 0.899 | 0.6068 | 0.6022 | 0.6032 | -0.6343 | -0.6389 | -0.6379 |
| | 正規 | 0.858 | | | 0.5545 | | | -0.6866 | | |
| 50 | 3次 | 0.911 | 0.930 | 0.931 | 0.4308 | 0.4299 | 0.4301 | -0.4194 | -0.4204 | -0.4201 |
| | 2次 | 0.887 | 0.910 | 0.910 | 0.4180 | 0.4170 | 0.4173 | -0.4066 | -0.4075 | -0.4073 |
| | 正規 | 0.899 | | | 0.3922 | | | -0.4324 | | |

表 3: 90%近似信頼区間 ($\rho = -0.8$ の場合)

| N | 近似 | 被覆確率 | | | 正確度 (下側信頼限界) | | | 正確度 (上側信頼限界) | | |
|----|----|--------|--------------|----------------|--------------|--------------|----------------|--------------|--------------|----------------|
| | | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ | ρ | $\hat{\rho}$ | $\hat{\rho}_J$ |
| 10 | 3次 | 0.880 | 0.907 | 0.898 | 0.7341 | 0.7476 | 0.7562 | -0.6453 | -0.6306 | -0.6229 |
| | 2次 | 0.771 | 0.818 | 0.811 | 0.6545 | 0.6686 | 0.6767 | -0.5657 | -0.5516 | -0.5434 |
| | 正規 | 0.807 | | | 0.7923 | | | -0.4279 | | |
| 25 | 3次 | 0.911 | 0.925 | 0.924 | 0.2944 | 0.2960 | 0.2955 | -0.2871 | -0.2854 | -0.2860 |
| | 2次 | 0.866 | 0.887 | 0.885 | 0.2774 | 0.2790 | 0.2785 | -0.2701 | -0.2684 | -0.2690 |
| | 正規 | 0.876 | | | 0.3240 | | | -0.2234 | | |
| 50 | 3次 | 0.913 | 0.920 | 0.919 | 0.1818 | 0.1822 | 0.1821 | -0.1829 | -0.1826 | -0.1827 |
| | 2次 | 0.888 | 0.899 | 0.899 | 0.1761 | 0.1765 | 0.1764 | -0.1772 | -0.1768 | -0.1769 |
| | 正規 | 0.894 | | | 0.1982 | | | -0.1551 | | |

上記の表から以下のようなことが分かる。

- 標本数が大きいとき ($N = 50$) には、3つの近似にそれほど差がないことは、すでに3.2節で見たが、上記の表からもそのことは読み取れる。
- 小標本 ($N = 10$), 中程度の標本 ($N = 25$) の場合、 ρ が真の値のもとでは、3次の近似は、被覆確率で見る限りすべての場合で他の2つの近似より優れている。2次と正規の近似では、必ずしも2次の方が良いとは限らない。
- ρ を推定値で置き換えた場合、 $\rho = 0.8, -0.8$ で小標本のときには、3次の近似は他の2つに比べかなり良い。
- ρ を推定値で置き換えた場合、 $\rho = 0.4$ のとき小・中程度の標本で2次の近似がもっとも良い。
- ρ を推定値で置き換えた場合、3次の近似は、信頼度より大きめの被覆確率をとる傾向がある。
- 被覆確率は、 $\hat{\rho}, \hat{\rho}_J$ いずれの推定値を用いてもあまり変わらない。
- 被覆確率は、真の ρ をその推定値 $\hat{\rho}, \hat{\rho}_J$ で置き換えることに対して、標本数が小さい場合でも頑健である。
- 正確度はすべての場合において、標本数が大きくなるにつれ良くなっている。

参考文献

- Bhattacharya, R.N. and Gosh, J.K. (1978), On the validity of the formal Edgeworth expansion, *Annals of Statistics* **6**, 434-451.
- Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.
- Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.
- 長尾 壽夫 (1993), ジャックナイフ統計量について, 日本数学会秋季総合分科会統計数学分科会講演アブストラクト, 131-132.
- Wang, J. (1994), Jackknife methods and higher order asymptotic expansions, 本講究録.
- 汪金芳他 (1992), ブートストラップ法—最近まで発展と今後の展望—, 行動計量学, 19 卷, 2 号, 50-81.