汚れのあるデータのもとでの統計的検定

M2012MM046 山元一輝
指導教員：木村美善

1　はじめに

データの汚れを記述するための近傍として Kakiuch and Kimura [3] が提唱した (c_1, c_2, γ)-近傍について考察し，汚れのあるデータのもとでの検定に対するこの近傍の応用について研究する．(c_1, c_2, γ)-近傍は 3 つのパラメータを持ち，これらのパラメータを変えることにより様々な近傍が生成されるが，その特徴と有用性について考察する．特に，正規分布，t -分布などの代表的な分布が (c_1, c_2, γ)-近傍にどのように効果的に取り込め得るのかについて研究する．また，1 標本問題と 2 標本問題における正規母集団の平均値の検定と平均値の差の検定が，標本に (c_1, c_2, γ)-近傍で表現されるモデル分布からの「ずれ」が生じるとき，どのような影響を受けるかについて，シミュレーションにより調べる．

2　分布近傍

2.1　様々な分布近傍

$\mathbb{R}$ を実数直線，$\mathscr{B}$ を $\mathbb{R}$ の部分集合からなるボレル集合族，$\mathscr{M}$ を $(\mathbb{R}, \mathscr{B})$ 上の確率分布の全体とする．これまでに提案された $F^{\circ} \in \mathscr{M}$ の分布近傍は多くあるが，その中で代表的な近傍を挙げる．

・ε -近傍:

$$\mathscr{P}_{\epsilon}(F^{\circ}) = \{G \in \mathscr{M} | G\{A\} \leq (1-\epsilon)F^{\circ}\{A\} + \epsilon, \ {}^{\forall}\mathrm{A} \in \mathscr{B}\}, \ 0 \leq \epsilon \leq 1$$

・全変動近傍:

$$\mathscr{P}_{0,\delta}(F^{\circ}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{F^{\circ}\{\mathrm{A}\} + \delta, 1\}, \ {}^{\forall}\mathrm{A} \in \mathscr{B}\}, \ 0 \leq \delta \leq 1$$

・Rider-近傍:

$$\mathscr{P}_{0,1-\epsilon,\epsilon-\delta}(F^{\circ}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{(1-\epsilon)F^{\circ}\{A\} + \epsilon + \delta, 1\}, \ {}^{\forall}\mathrm{A} \in \mathscr{B}\}, \ 0 \leq \epsilon, \ 0 \leq \delta, \ \epsilon + \delta < 1$$

・(c,γ)-近傍 [1]:

$$\mathscr{P}_{c,\gamma}(F^{\circ}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{cF^{\circ}\{A\} + \gamma, 1\}, \ {}^{\forall}\mathrm{A} \in \mathscr{B}\}, \ 0 \leq \gamma < 1, \ 1 - \gamma \leq c < \infty$$

また (c,γ)-近傍は，c と γ に特別な値を入れることで多くの新しい近傍を得ることができる．これらの近傍の定義では上限しか示されていないが下限で定義することもできる．たとえば (c,γ)-近傍の場合は

$$\mathscr{P}_{c,\gamma}(F^{\circ}) = \{G \in \mathscr{M} | G\{A\} \geq \max\{cF^{\circ}\{A\} + 1 - \gamma - c, 0\}, \ {}^{\forall}\mathrm{A} \in \mathscr{B}\}$$

となる．

2.2　特殊容量

Bednaruski [2] は，連続な凹関数 $h : [0, 1] \to [0, 1]$ で $h(1) = 1$ を満たすものに対して，集合関数 $v_h : \mathscr{B} \to [0, 1]$ を

$$v_h\{A\} = \begin{cases} h(F^{\circ}\{A\}) & \phi \neq A \in \mathscr{B} \text{ のとき} \\ 0 & A = \phi \text{ のとき} \end{cases} \quad (1)$$

により定義し，これを特殊容量 (special capacity) と呼んだ ($h(0) = 0$ が成り立つとき 2-alternating Choquet capacity になる)．この確率測度を一般化した特殊容量 v_h を用いて得られる F° の分布近傍

$$\mathscr{P}_h(F^{\circ}) = \left\{G \in \mathscr{M} | G\{A\} \leq v_h\{A\}, {}^{\forall}\mathrm{A} \in \mathscr{B}\right\}$$

は 2 つの近傍間にマクシミン検定が存在するなど優れた性質を持っている．

3　(c_1, c_2, γ)-近傍について

次のような 3 つのパラメータを持つ分布近傍が Kakiuchi and Kimura(2012) により提案された．

$$\mathscr{P}_{c_1,\, c_2,\gamma}(F^{\circ}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{c_2F^{\circ}\{\mathrm{A}\} + \gamma, \, c_1F^{\circ}\{\mathrm{A}\} + 1 - c_1\}, {}^{\forall}\mathrm{A} \in \mathscr{B}\} \quad (2)$$

これは

$$h(x) = \min\{c_2x + \gamma, \, c_1x + 1 - c_1\}, \ 0 \leq x \leq 1$$

としたものに対応している．ただし，$0 \leq c_1 \leq 1 - \gamma \leq c_2 < \infty$, $c_1 \neq c_2$, $0 \leq \gamma < 1$.
この h に対応する F° の分布近傍は (c_1, c_2, γ)-近傍と呼ばれる．また，

1. $h(x) = x$ のとき，$v_h = F^{\circ}$．すなわち，確率測度は特殊容量であり，$\mathscr{P}_h(F^{\circ}) = F^{\circ}$.
2. (c_1, c_2, γ)-近傍は，

$$\mathscr{P}_{c_1,\, c_2,\gamma}(F^{\circ}) = \{G \in \mathscr{M} | c_1F^{\circ}\{A\} \leq G\{A\} \leq c_2F^{\circ}\{A\} + \gamma, {}^{\forall}\mathrm{A} \in \mathscr{B}\}$$

とも表せることに注意する．(c_1, c_2, γ)-近傍は，3 つのパラメータ c_1, c_2, γ を変化させることで多様な近傍を得ることができる．実際に c_1, c_2, γ に値を代入することで次の近傍が得られる．

・$c_1 = 1 - \epsilon, \gamma = \epsilon$ または $c_2 = 1 - \epsilon, \gamma = \epsilon$ のとき，ϵ-近傍に等しい．
・$c_1 = 0$, $c_2 = 1, \gamma = \delta$ のとき，全変動近傍に等しい．
・$c_1 = 0$, $c_2 = c, \gamma = \gamma$ のとき，(c, γ)-近傍に等しい．
・$c_1 = 0$, $c_2 = 1 - \epsilon, \gamma = \epsilon + \delta$ のとき，Rieder 近傍に等

しい.
・$\gamma = 0$ のとき，G 近傍となる．この場合は,

$$h(x) = \min\{c_2 x,\, c_1 x + 1 - c_1\}$$

であり，局外値などを含まない中心部のみが汚れた近傍である.
・$c_1 = 1-\epsilon$，$c_2 = 1$，$\gamma = \delta(\epsilon > \delta)$ のとき ϵ-近傍と全変動近傍を組み合わせることによって $TN\epsilon$-近傍ができ,

$$h(x) = \min\{x+\delta,\, (1-\epsilon)x + \epsilon\},\ 0 \le \delta \le \epsilon \le 1$$

となる，これは，上限の確率が ϵ-汚染近傍と全変動近傍の小さい方で抑えられる分布近傍となる．この分布近傍は ϵ-汚染近傍により決められたと全変動近傍である.

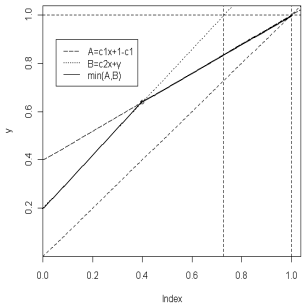

図 1 (c_1, c_2, γ)-近傍グラフ

近傍のまとめ

このように，c_1, c_2, γ の値の取り方により，(c_1, c_2, γ)-近傍は ϵ-近傍，全変動近傍，Rider 近傍，(c, γ)-近傍などを，特別な場合としてすべて含む分布近傍となる.
図 1 における関数 $h(x)$ の切片 ($h(0)$ の値) は空集合を除いた集合に対するその上限の確率を示している．すなわち，極端な値 (局外値，大きな値や小さな値) を取る確率が正であるという分布を示している．したがって，切片の値が大きいほど，外れ値を取る確率が高くなる．また特殊容量の $h(x)$ が 1 となる x の値が小さいほど，正となる確率をもつ集合さえ無視されやすくなることを意味する．その意味で Rider 近傍は，ϵ-近傍や全変動近傍を含む大きな近傍を表していると言える.
図 2 をみると，2 つの直線の最小値を通る直線を選んでいるものが (c_1, c_2, γ)-近傍であることが解かる.

4 (c_1, c_2, γ)-近傍の特徴

4.1 密度関数による特徴

Kakiuchi and Kimura(2012) の Theorem 3．2 は式 (2) により定義された (c_1, c_2, γ)-近傍の密度関数による特徴づけを以下のように与えている.
(c_1, c_2, γ)-近傍は

$$\mathscr{P}_{c_1,\, c_2, \gamma}(F^{\circ}) = \{G = (1-\gamma)F + \gamma K | F \in \mathscr{F}_{c_1,\, c_2, \gamma}(F^{\circ}),\, K \in \mathscr{M}\} \quad (3)$$

として表される．ここで，連続な分布 F° の密度関数を f° とし

$$\mathscr{F}_{c_1,\, c_2, \gamma}(F^{\circ}) = \left\{ F \in \mathscr{M}_c | \frac{c_1}{1-\gamma} f^{\circ} \le f \le \frac{c_2}{1-\gamma} f^{\circ} \right\} \quad (4)$$

とする．ただし，f は F の密度関数とし，$\mathscr{M}_c$ は $(\mathbb{R}, \mathscr{B})$ 上の連続な確率分布からなる集合とする.

4.2 (c_1, c_2, γ)-近傍の限界近傍

F_L, F_R を $\mathscr{F}_{c_1,\, c_2, \gamma}(F^{\circ})$ に含まれる以下の分布として定義する.

$$F_L(x) = \begin{cases} \dfrac{c_2}{1-\gamma} F^{\circ}(x) & (x \le x_L) \\ \dfrac{c_1}{1-\gamma} F^{\circ}(x) + (1 - \dfrac{c_1}{1-\gamma}) & (x > x_L) \end{cases}$$

$$F_R(x) = \begin{cases} \dfrac{c_1}{1-\gamma} F^{\circ}(x) & (x \le x_R) \\ \dfrac{c_2}{1-\gamma} F^{\circ}(x) + (1 - \dfrac{c_2}{1-\gamma}) & (x > x_R) \end{cases}$$

ただし，それぞれ x_L, x_R は

$$x_L = (F^{\circ})^{-1}\left(\frac{1-\gamma-c_1}{c_2-c_1}\right),\ x_R = (F^{\circ})^{-1}\left(\frac{c_2+\gamma-1}{c_2-c_1}\right)$$

である.

5 (c_1, c_2, γ)-近傍の実用化

5.1 (c_1, c_2, γ)-近傍の問題

具体的な分布に対して具体的な近傍を適用することによりその有効性を探る．具体的な分布として標準正規分布と t-分布を用いる．モデル分布は標準正規分布とする．標準正規分布を何倍かすることによって上限と下限を作りその中に t-分布の全ての自由度の密度関数が近傍内に含まれるよう挟み込むことを考える．(図 3) そして，そのとき c_1, c_2, γ の値がどうなるか考察していく.

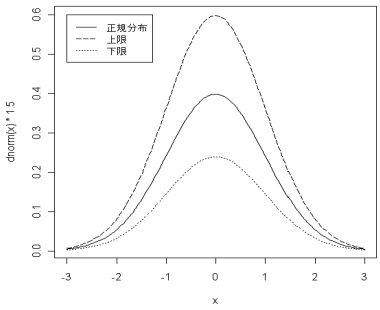

図 2 (c_1, c_2, γ)-近傍の密度関数

5.2 計算方法

標準正規分布と t-分布の分散は異なっている．そこで，分散を合わせるために t-分布の分散を標準正規分布の分散に標準正規分布の分散を t-分布の分散に変換し計算を

行う．
標準正規分布と t-分布の裾の部分を比べると t-分布の方が確率が大きくなる．そのため，ある範囲からは t-分布の密度関数を用いる．この時，標準正規分布の密度関数 f° と自由度 n の t-分布の密度関数 t_n を $\pm x_a$ でつなぎ連続にする為の k を以下で定める．[4]

$$f(x) = \begin{cases} t_n(x), & (x > |x_a|) \\ kf^{\circ}(x), & (-x_a \leq x \leq x_a) \end{cases}$$

ただし k は，$kf^{\circ}(x_a) = t_n(x_a)$ である．

次に，密度関数 $f^*(x) = cf(x)$ と定義する．ただし，$c = 1/\int_{-\infty}^{\infty} f(x)dx$．この $f^*(x)$ を用いて c_1, c_2, γ 値を求める．

5.3 分散を標準正規分布の分散に合わせた場合

表 1 挟み込む最小の自由度による違い ($x_a = 1.96$)

n	$n \geq 3$	$n \geq 5$	$n \geq 10$
k	0.464	0.779	0.949
c	2.065	1.260	1.046
$\frac{c_2}{1-\gamma} f^*$(上限)	1.662	1.249	1.097
$\frac{c_1}{1-\gamma} f^*$(下限)	0.465	0.700	0.853
$c_1(\gamma$=0.01)	0.461	0.693	0.844
$c_2(\gamma$=0.01)	1.645	1.237	1.086
$c_1(\gamma$=0.05)	0.442	0.665	0.810
$c_2(\gamma$=0.05)	1.579	1.187	1.042

表 1 は，挟み込む t-分布の最小の自由度を変化させた時の表である．自由度の最小の値が大きくなるほど上限，下限の幅は狭くなる．

表 2 範囲による違い（$n \geq 3$ 以上）

x_a	1.96	2.3	2.58	3.0
k	0.464	0.568	0.619	1.436
c	2.065	1.711	1.576	0.691
$\frac{c_2}{1-\gamma} f^*$(上限)	1.662	1.641	1.633	1.739
$\frac{c_1}{1-\gamma} f^*$(下限)	0.465	0.460	0.457	0.450
$c_1(\gamma$=0.01)	0.461	0.455	0.453	0.445
$c_2(\gamma$=0.01)	1.645	1.624	1.617	1.722
$c_1(\gamma$=0.05)	0.442	0.437	0.435	0.427
$c_2(\gamma$=0.05)	1.579	1.559	1.551	1.652

表 2 より範囲 x_a を 2．58 まで広げていくと上限と下限は小さくなっていることがわかる．また γ の値に観点を置くと，γ を増やしていくと c_1，c_2 の値は汚染がない値より小さい値になる．x_a の範囲が 3．0 になると上限が大きくなっている．
この理由は，図より明らかとなる．図 4，5 を比べると，t-分布と上限の接点が異なっていることがわかる．このことから確率の範囲が広がることで接点も外に広がり上限が大きくなることがわかる．

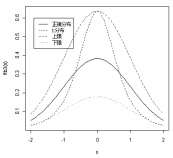

図 3 範囲 x_a=1.96

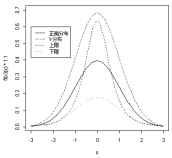

図 4 範囲 x_a=3.0

5.4 分散を t-分布の分散に合わせた場合

自由度が大きくなるにつれて上限と下限の幅が小さくなっていることがわかる．この方法に関しても $\pm x_a$ が t-分布と上限との接点になっている．よって，範囲 x_a が大きくなる事で上限の値も大きくなる．
これらの結果より，標準正規分布の (c_1, c_2,γ)-近傍により t-分布を覆うことができる上限，下限の値は，範囲が大きくなることで幅が広くなる．

表 3 t-分布の分散に変換した場合

	$n \geq 3$	$n \geq 5$	$n \geq 7$	$n \geq 10$
x_a	5.840	4.032	3.499	3.169
k	3.072	2.098	1.585	1.322
c	0.324	0.475	0.628	0.753
$\frac{c_2}{1-\gamma} f^*$(上限)	3.076	2.104	1.589	1.326
$\frac{c_1}{1-\gamma} f^*$(下限)	0.488	0.690	0.783	0.850
$c_1(\gamma = 0.01)$	0.483	0.683	0.775	0.842
$c_2(\gamma = 0.01)$	3.045	2.083	1.573	1.313
$c_1(\gamma = 0.05)$	0.463	0.655	0.744	0.807
$c_2(\gamma = 0.05)$	2.922	1.999	1.509	1.260

6 標準正規分布と t 分布の交点

6.1 ランベルトの W 関数

標準正規分布の密度関数 f° と t-分布の密度関数 t_n を $\pm x_a$ でつなぎ連続にする時の $\pm x_a$ は次の方程式の解である．

$$\frac{1}{\sqrt{2\pi}} e^{\frac{-(x^2)}{2}} = \frac{\Gamma[\frac{n+1}{2}]}{\sqrt{n\pi}\Gamma[\frac{n}{2}]} (1 + \frac{x^2}{n})^{\frac{-(n+1)}{2}}$$

この解はランベルトの W 関数 [5] を使用して求める事が可能であると考えられるが，本研究では求めることは叶わなかった．

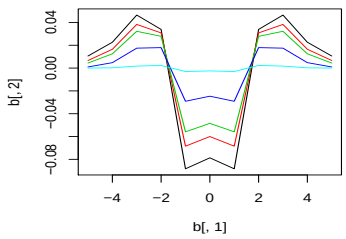

図 5 標準正規分布と t 分布の交点のグラフ

シュミレーションによると x 軸は x_a の値を y 軸は両分布の差を表す図 7 の様になり交点 x_a の値は ± 1.9 付近を取るのではないかと考えられる．

7 検定

7.1 一標本問題

$X_1, X_2, \cdots, X_n$ を $N(\mu, 1)$ からの無作為標本とするとき，平均 μ に関する右側検定問題 $H_0 : \mu = 0$，$H_1 : \mu > 0$ に対する有意水準 $100\alpha\%$ 一様最強力検定は

$$Z = \sqrt{n}\overline{X} > z_\alpha \tag{5}$$

のとき H_0 を棄却するものである．ここで $\overline{X}$ は標本平均であり，z_α は $N(0, 1)$ の上側 $100\alpha\%$ 点である．$X_1, X_2, \cdots, X_n$ の分布 F が $N(\mu, 1)$ からずれたとき，この Z に基づく検定がどのような影響を受けるのかを調べるために次のことを行う．$X_1, X_2, \cdots, X_n$ の真の分布 F は $F_\mu^\circ = N(\mu, 1)$ の $(c_1, c_2, 0)$-近傍 $\mathscr{F}_{c_1,\, c_2,\, 0}(F_\mu^\circ)$ に属すとする．このとき第 1 種の誤りの確率 (有意水準) が最大となるのは F が $\mathscr{F}_{c_1,\, c_2,\, 0}(F_0^\circ)$ の最大分布 $F_{R,\, 0}$ のときであり，検出力が最小となるのは $\mathscr{F}_{c_1,\, c_2,\, 0}(F_\mu^\circ)$ 最小分布 $F_{L,\mu}$ のときであるので (a) $F = F_{R,\, 0}$，(b) $F = F_{L,\mu}$ としてシュミレーションにより，確率 $P(Z > z_\alpha)$ を評価する．今回は，$c_1 = 1 - \eta$, $c_2 = 1 + \eta$,$\gamma = 0$ の場合にシュミレーションを行った．

表 4 $P(Z > z_\alpha)$，$F = F_R$

$\alpha = 0.05$	n=1	n=2	n=5
$\eta = 0.05$	0.0502	0.0547	0.0576
$\eta = 0.1$	0.0537	0.0574	0.0714
$\eta = 0.15$	0.0558	0.0675	0.0791
$\eta = 0.2$	0.0616	0.0708	0.0942
$\eta = 0.25$	0.0625	0.0721	0.1058

η を大きくすると $z_a = 1.645$ をこえる確率が上昇していることが見てわかる．また標本数 n を増やしていくと同じ η の値に対して確率が上昇していることも見てわかる．

7.2 二標本問題

$X_1, X_2, \cdots, X_{n_1}$ を $N(\mu, 1)$ からの無作為標本，$Y_1, Y_2, \cdots, Y_{n_2}$ を $N(\mu, 1)$ からの無作為標本とするとき，平均の差 $\mu = \mu_1 - \mu_2$ に関する検定問題 $H_0 : \mu = 0$，$H_1 : \mu > 0$ に対する有意水準 $100\alpha\%$ 検定は

$$Z = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{X} - \overline{Y}) > z_\alpha \tag{6}$$

のとき H_0 を棄却するものとして与えられる．ここで $\overline{X}$, $\overline{Y}$ はそれぞれ $X_1, X_2, \cdots, X_{n_1}$ と $Y_1, Y_2, \cdots, Y_{n_2}$ の標本平均である，$X_1, X_2, \cdots, X_{n_1}$ の真の分布 F と $Y_1, Y_2, ..., Y_{n_2}$ の真の分布 G がそれぞれ $N(\mu_1, 1)$ と $N(\mu_2, 1)$ からずれたとき，この Z に基づく検定がどのような影響を受けるかを調べる．簡単のために標本数が等しい $n_1 = n_2 = n$ の場合を考え，$F \in \mathscr{F}_{c_1,\, c_2,\, 0}(F_{\mu_1}^\circ)$，$G \in \mathscr{F}_{c_1,\, c_2,\, 0}(F_{\mu_2}^\circ)$ とする．ただし，$F_{\mu_1}^\circ = N(\mu_1, 1)$，$F_{\mu_2}^\circ = N(\mu_2, 1)$．第 1 種の誤りの確率が最大になるのは X が最大分布で Y が最小分布のときであり，検出力が最小になるのは X が最小分布で Y が最大分布のときであるから (a) $F = F_{R,\, 0}$，$G = F_{L,\, 0}$，(b) $F = F_{L,\mu}$，$G = F_{R,\, 0}$ として確率 $P(\sqrt{\frac{n}{2}}(\overline{X} - \overline{Y}) > z_\alpha)$ をシュミレーションにより評価する．$F = F_{L,\, 2}$，$G = F_{R,\, 0}$ で $c_1 = 1 - \eta$, $c_2 = 1 + \eta$, $\gamma = 0$ の場合にシュミレーションを行った．

表 5 $P(Z > z_\alpha)$，$F = F_{L,\mu}$，$G = F_{R,\, 0}$

$\alpha = 0.05$	n=1	n=5
$\eta = 0.05$	0.4066	0.934
$\eta = 0.1$	0.4056	0.9325
$\eta = 0.15$	0.4049	0.932
$\eta = 0.2$	0.4019	0.9281
$\eta = 0.25$	0.3948	0.9276

η を大きくすると 1.645 をこえる確率が減少していることが見てわかる．また標本数 n を増やしていくと同じ η の値の場合でも確率が上昇していることも見てわかる．

8 おわりに

本研究では，(c_1, c_2, γ)-近傍の実用化の一歩として正規分布によって t-分布を格納することを研究した．また，1 標本問題と 2 標本問題の分散が既知の場合の平均値と平均値の差の検定がどのように影響受けるのかについてシュミレーションにより調べた．今後の課題として標準正規分布の密度関数 f° と t-分布の密度関数 t_m を $\pm x_a$ でつなぎ連続にするための $\pm x_a$ を求める問題と分散が未知の場合の平均値と平均値の差の検定が (c_1, c_2, γ) によるずれによってどのよううな影響を受けるかをシュミレーションにより詳しく調べる必要がある．

参考文献

[1] Ando, M.Kakiuchi, I.and Kimura, M.（2009）. Robust nonparametric confidence intervals and tests in the presents of (c,γ)-contamination, J.Statist Plann.inference.139, 1836-1846.

[2] Bednarski, T.(1981). On solutions of minmax tests problems for special capacities. Z. Wahreck. verw Gebiete. 10, 269-278.

[3] Kakiuchi, I.and Kimura, M.（2012）. Robust nonparametric inference for the median under a new neighborhood of distributions, Technical Report of the Nanzan Academic Society Information Sciences and Engineering.

[4] 沖　翔太 (2013).『汚れのあるデータとロバスト推測』，南山大学数理情報研究科修士論文.

[5] R.M.Corless, G.H.Gonnet, D.E.G.Hare, D.J.Jeffrey and D.E.Knuth.(1996). On the Lambert W Function, Computational Mathematics Vol. 5.