

## 処置前後研究における平均への回帰とその周辺

岩崎 学\*, 河田 祐一\*\*

## Regression to the Mean and Related Topics in Pretest-Posttest Designs

Manabu Iwasaki\* and Yuichi Kawata\*\*

新薬開発の臨床試験などのように、処置の効果の有無やその大きさを評価する研究では、同じ被験者に対し処置を施す前と施した後でデータを観測するいわゆる処置前後研究が行なわれることが多い。この種の研究デザインでは「平均への回帰」現象が生じ、ともすると試験結果の解釈に重大な誤りをもたらすこともある。本論では、平均への回帰がなぜ生じるのかに関する考察を行ない、正規分布のみならず、ポアソン分布および二項分布のような離散分布における平均への回帰を議論する。

In studies that aim to assess effectiveness of a certain treatment, such as clinical trials of new drug development, pretest-posttest research designs are often conducted, in which measurements are observed before and after the treatment for the same participant. In such research designs, it is known that the regression to the mean phenomenon might be observed, and it sometimes causes misleading interpretation of the result obtained. This paper first examines the reason why regression to the mean is observed by using a Bayesian model. Then, in addition to the ordinary normal distributions, some discrete distributions such as Poisson and binomial distributions are also considered from the same viewpoint.

*Key Words and Phrases:* Bayesian modeling, before-after design, beta-binomial distribution, gamma-Poisson distribution, screening

## 1. はじめに

ある処置の効果の有無やその大きさを評価するため、同じ個体に対してその処置を施す前と後で観測値を得る、いわゆる処置前後研究 (pretest-posttest study) が多くの分野で標準的に用いられている。新薬開発の臨床試験では、同じ被験者に対し、薬剤の投与前と投与後を比較する研究が主流であり、心理もしくは教育の分野でも新しいカウンセリング法やIT技術を用いた教育方法の効果を見るため、それらの方法を用いる前と用いた後である種のテストを行ない、それらの比較により処置の有効性を見ることが多い。これらは実験研究であるが、ある施策の有効性を判断するためその施策の前後を比較するというタイプの観察研究もよく行なわれる。

処置前後データの解析では、いわゆる平均への回帰 (regression to the mean) の問題が生じ

---

\* 成蹊大学理工学部情報科学科 (〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1).

E-mail: iwasaki@st.seikei.ac.jp)

\*\* 中外臨床研究センターバイオメトリクス部 (〒103-8324 東京都中央区日本橋室町 2-1-1).

E-mail: kawatayui@chugai-pharm.co.jp)

る。平均への回帰とは、処置前後値を表わす確率変数をそれぞれ  $X$  および  $Y$  とし、それらの母集団全体での期待値を  $E[X]$  および  $E[Y]$ 、 $X=x$  のときの  $Y$  の条件付き期待値を  $E[Y|X=x]$  としたとき、

$$|E[Y|X=x] - E[Y]| \leq |x - E[X]| \quad (1.1)$$

もしくは

$$\frac{E[Y|X=x] - E[Y]}{x - E[X]} \leq 1 \quad (1.2)$$

となる現象で、処置前値  $x$  の期待値  $E[X]$  からの乖離に比べ、条件付き期待値  $E[Y|X=x]$  のほうが期待値  $E[Y]$  からの乖離が少ない（平均に回帰する）というもので、Galton (1886) の歴史的な論文に由来する。特に、処置前値によるスクリーニングがある、すなわち処置前値がある基準よりも大きい（もしくは小さい）場合にのみその個体が研究にエントリーされて処置後の値が観測される、という場合には、処置の効果が全くなくても見かけ上効果があったように見えることから、結果の解釈に特に注意を要する。降圧剤の臨床試験では血圧がある値以上の被験者のみが試験にエントリーされる、学校での補習授業では試験の点数がある値以下のものだけが補習授業を受ける、などスクリーニングのある研究は多く、これまで報告された研究結果の効果の大きさの幾分かはこの平均への回帰によるものではないかと推察される。

処置前後研究に関するまとまった書物としては Bonate (2000) がある。そこでは、臨床試験及び心理実験といった実験研究が主として扱われている。岩崎 (2006) では章の1つが平均への回帰に当てられている。また、岩崎 (2002) の解説論文では、処置前値によるスクリーニング下での分布の形状も詳しく議論されている。なお、S. J. Senn の編集による *Statistical Methods in Medical Research* (1997) には平均への回帰に関する様々な観点からの論文が集められていて、きわめて有用である (Senn (1997), Stigler (1997), Chuang-Stein and Tong (1997), Lin and Hughes (1997), Chesher (1997) および Copas (1997) が収められている)。

多くの書物あるいは論文では、平均への回帰の問題は主として2変量正規分布の枠組みで議論されていて、非正規分布に関する研究はあまり多くない (Beath and Dobson (1991), Chesher (1997) はある種の連続型の非正規分布を扱っている)。処置前後研究での観測値は連続型のものだけとは限らない。ある事象の生起回数を観測するカウントデータも重要な評価指標となり得る。上記の Senn (1997) では、その後続く論文が主として医薬関係の実験研究であるにもかかわらず、平均への回帰が生じる例として、交差点などにおける交通事故件数というカウントデータに基づく観察研究が取り上げられている。ちなみに、Senn (1997) が取り上げた話題に関しては、Hauer (1980), Abbess, Jarrett and Wright (1981), McGuigan (1985), Maher (1987), Senn and Collie (1988) などが交通工学の分野の雑誌に掲載され議論されている。

新薬開発の臨床試験でも事象の出現回数といったカウントデータがエンドポイントになる例は多い。カウントデータには2種類ある。第一は、ある事象の生起回数を観測するもので、臨床試験ではある一定期間内での発作の回数などがその例であり、ポアソン分布が仮定されることが多い。今ひとつは、一定の試行数での事象の観測度数を問題とするもので、リウマチの治療における手指の関節の疼痛箇所がその例であり（手指の関節の総数は一定である）、二項分布が仮定される。臨床試験以外でもカウントデータは主要な評価指標であり、上述の交通事故の発生件数はその例である。また、医薬分野では、薬剤の市販後における予期せぬ有害事象の出現回数のデータマイニング・アプローチによる分析が、近年では大きな問題となっている

(藤田他 (2004), 渡邊他 (2004), 岩崎・吉田 (2005) などを参照).

本論では, これまでの正規分布に加え, 上記2種類のカウントデータすなわちポアソン分布および二項分布に関し, ある種のモデル (Bayes 流のモデル) に基づき, 平均への回帰現象の生じる理由を含め議論する (ポアソン分布と二項分布の基礎的な事項に関しては竹内・藤野 (1981) 参照). 第2節で平均への回帰を説明するモデルを導入し, 平均への回帰が起こる条件を示す. 第3節では正規分布に対し第2節の一般論を適用する. 第4節および第5節ではそれぞれガンマポアソン分布とベータ二項分布を議論し, 最後の第6節で簡単なまとめと今後の展望を示す.

なお, 次節以降の平均への回帰での議論では, 観測値が大きいほど状態が悪いとし, 処置前値でのスクリーニングも, 処置前値がある値以上の時のみ処置後の値が観測されるとする. すなわち, 降圧剤の試験では血圧が高いほうが悪い, ある種の発作回数が多いほうが悪い, 事故や有害事象の発生件数が多いほうが悪い, などである. 試験の点数のように値が小さいほうが悪い場合には不等号の向きを逆にするなどにより対処できる.

## 2. 平均への回帰のモデル

ここでは, 平均への回帰現象を説明するひとつのモデルを与える. 2.1節で処置の効果が無い場合を扱い, その後2.2節で処置効果がある場合の定式化を示す. 処置効果が無い場合の考察は, 平均への回帰に起因する量が純粋にどの程度であるのかの情報が得られることから重要である. ここでのモデルは, 母集団における個体間差と, 同一個体における個体内変動とを区別して捉え, 処置前値が与えられたときの個体間差の条件付き分布が, 処置後の観測値の特徴を規定するとの考察に基づくものである.

ある母集団における処置前の特定の個体を特徴付けるパラメータを  $\theta$  とする. 降圧剤の臨床試験では  $\theta$  はある患者の薬剤投与前の血圧の真値であり, 教育方法の有効性の研究では  $\theta$  はある生徒の教育方法適用前の真の学力とみなされる.  $\theta$  は個体ごとに (連続的に) 異なるであろうから, 母集団内におけるその分布を確率密度関数  $g(\theta; \xi)$  で表現する ( $\xi$  は分布を特徴付けるパラメータ).  $g(\theta; \xi)$  によって規定される分布は  $\theta$  の個体間分布 (inter-individual distribution) であり, Bayes 流の定式化では事前分布とみなされる. ただし,  $\theta$  の確率分布は主観的なものではなく, 母集団における個体間差という客観的な意味を持つため純粋な Bayes 流の議論とは異なるが, 事前分布, 事後分布といった Bayes 統計の用語を用いる.

パラメータ  $\theta$  の個体の処置前値  $X$  が確率密度関数  $h(x|\theta)$  を持つ分布に従うとする. これは個体内分布 (intra-individual distribution) である. このとき, 母集団全体での処置前値  $X$  の確率密度関数  $f(x; \xi)$  は

$$f(x; \xi) = \int_{-\infty}^{\infty} h(x|\theta)g(\theta; \xi)d\theta \quad (2.1)$$

となる. そして,  $X$  の期待値と分散をそれぞれ

$$\mu_X(\xi) = E[X; \xi], \quad \sigma_X^2(\xi) = V[X; \xi] \quad (2.2)$$

と書く.  $X$  が離散的な場合には  $f(x; \xi)$  は確率関数となるが, 混乱の恐れがない限り, 以下では離散型の場合でも確率密度関数という. また,  $f(x; \xi)$  によって規定される確率分布を単に分布  $f(x; \xi)$  とよぶこともある.

### 2.1 処置効果がない場合

処置の効果が無いとすると, 処置後の観測値  $Y$  も  $X$  と同じ分布に従い, 期待値  $\mu_Y(\xi)$  と分

散  $\sigma_X^2(\xi)$  も (2.2) と同じ値となる。  $\theta$  を与えた下で  $X$  と  $Y$  が独立と仮定すると、  $(X, Y)$  の同時確率密度関数  $f(x, y; \xi)$  は

$$f(x, y; \xi) = \int_{-\infty}^{\infty} h(x|\theta)h(y|\theta)g(\theta; \xi)d\theta \quad (2.3)$$

となる。そして、共分散および相関係数を  $\sigma_{XY} = \text{Cov}[X, Y]$ ,  $\rho = R[X, Y] = \text{Cov}[X, Y] / (\sqrt{V[X]}\sqrt{V[Y]})$  と置く。処置前の観測値が  $x$  であるとの条件の下で、処置後値  $Y$  の条件付き確率密度関数  $f(y|x; \xi)$  と条件付き期待値  $E[Y|X=x]$  を求める。  $X=x$  が与えられたときのパラメータ  $\theta$  の事後分布は

$$g(\theta|x; \xi) = \frac{h(x|\theta)g(\theta; \xi)}{f(x; \xi)} \quad (2.4)$$

となる。  $g(\theta|x; \xi)$  は処置前値が  $x$  であった個体のパラメータ  $\theta$  の条件付き分布であり、平均への回帰の議論では中心的な役割を果たす。  $X=x$  となった個体のパラメータ  $\theta$  が (2.4) の分布に従うとすると、  $Y$  の  $X=x$  の下での条件付き分布は

$$\begin{aligned} f(y|x; \xi) &= \int_{-\infty}^{\infty} h(y|\theta)g(\theta|x; \xi)d\theta \\ &= \frac{1}{f(x; \xi)} \int_{-\infty}^{\infty} h(y|\theta)h(x|\theta)g(\theta; \xi)d\theta = \frac{f(x, y; \xi)}{f(x; \xi)} \end{aligned} \quad (2.5)$$

となり、  $Y$  の条件付き期待値は

$$E[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f(y|x; \xi)dy = \frac{1}{f(x; \xi)} \int_{-\infty}^{\infty} y \cdot f(x, y; \xi)dy$$

となる。(2.5) は条件付き密度関数に関する定義  $f(y|x; \xi) = f(x, y; \xi) / f(x; \xi)$  に他ならないが、一旦  $g(\theta|x; \xi)$  を経由するところに大きな意味があり、平均への回帰現象を理解する上で重要である。

パラメータ  $\theta$  の事前分布  $g(\theta; \xi)$  が共役事前分布で、  $X=x$  が与えられたときの  $\theta$  の事後分布が  $g(\theta|x; \xi) = g(\theta; \xi(x))$  であったとすると、  $Y$  の  $X=x$  での条件付き分布は

$$f(y|x; \xi) = \int_{-\infty}^{\infty} h(y|\theta)g(\theta; \xi(x))d\theta = f(y; \xi(x))$$

となり、条件付き期待値は

$$E[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f(y; \xi(x))dy = \mu_Y(\xi(x))$$

となる。特に、  $\mu_Y(\xi)$  が  $\xi$  の線形関数であり、かつ  $\xi(x)$  が  $x$  の線形関数であれば  $E[Y|X=x]$  は  $x$  の線形関数となる。

処置後値  $Y$  の条件付き期待値  $E[Y|X=x]$  は条件付き事後分布  $g(\theta|x; \xi)$  の期待値に等しいことから、  $g(\theta|x; \xi)$  の期待値が  $x$  よりも小さくなるための条件が問題となる。それは以下の2つの条件のいずれかもしくは両方が成立するときであることが分かる。

#### 平均への回帰の条件

- (a) 個体間分布  $g(\theta; \xi)$  の単峰性。
- (b)  $X$  の個体内分布  $h(x|\theta)$  の分散は、  $\theta$  が  $g(\theta; \xi)$  の分布の中央のほうが大きい。

条件 (a) は、処置前の観測値  $x$  が  $E[X]$  よりも大きいとき、 $\theta$  は小さいが観測値がたまたま大きくて  $x$  となった個体のほうが、 $\theta$  は大きい観測値がたまたま小さくて  $x$  となった個体よりも多いことを表わしている (第3節参照). また、条件 (b) は  $\theta$  が分布の端であるより中ほどに近いほうが  $x$  の値を取り易いことを示している (第5節参照).

## 2.2 処置効果がある場合

処置の効果がある場合には、処置により処置前の個体のパラメータ (真値)  $\theta$  が処置後に  $\theta^*$  に変化すると想定する. 同じ  $\theta$  を持つ個体でも、個体ごとに処置によって効果の大きさが異なるとする場合には  $(\theta, \theta^*)$  に2次元の確率分布  $g(\theta, \theta^*)$  を想定することになる. それに対し、同じ  $\theta$  をもつ個体に対しては処置の効果は同じであると仮定すると、 $\theta^* = \eta(\theta; c)$  となる. ここで  $c$  は効果の大きさを表わすパラメータである. このとき、 $(\theta, \theta^*)$  は1次元に退化した分布を持つ.  $\eta(\theta; c)$  の具体的な形としては  $\theta + c$  あるいは  $c\theta$  などが考えられるが、もっと複雑な関数が想定されることもある. 本論では処置効果は同じで  $\theta^* = \eta(\theta; c)$  となる場合を考察する.

パラメータ値が  $\theta^*$  のときの処置後値  $Y$  の個体内分布を  $h(y|\theta^*) = h(y|\eta(\theta; c))$  とすると、処置後値  $Y$  の母集団全体での確率密度関数は

$$f(y; \xi, c) = \int_{-\infty}^{\infty} h(y|\eta(\theta; c))g(\theta; \xi) d\theta \quad (2.6)$$

となる.  $Y$  の期待値と分散をそれぞれ

$$\mu_Y(\xi, c) = E[Y; \xi, c], \quad \sigma_Y^2(\xi, c) = V[Y; \xi, c]$$

と書く.  $(X, Y)$  の同時確率密度関数は

$$f(x, y; \xi, c) = \int_{-\infty}^{\infty} h(x|\theta)h(y|\eta(\theta; c))g(\theta; \xi) d\theta \quad (2.7)$$

と求められる.  $X=x$  が与えられたときの  $\theta$  の事後分布は (2.4) であるので、そのときの  $Y$  の条件付き分布は

$$\begin{aligned} f(y|x; \xi, c) &= \int_{-\infty}^{\infty} h(y|\eta(\theta; c))g(\theta|x; \xi) d\theta \\ &= \frac{1}{f(x; \xi)} \int_{-\infty}^{\infty} h(y|\eta(\theta; c))h(x|\theta)g(\theta; \xi) d\theta = \frac{f(x, y; \xi, c)}{f(x; \xi)} \end{aligned}$$

となり、 $Y$  の条件付き期待値は

$$E[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f(y|x; \xi, c) dy = \frac{1}{f(x; \xi)} \int_{-\infty}^{\infty} y \cdot f(x, y; \xi, c) dy$$

で与えられる. これらの具体的な形は次節以降で議論する.

## 3. 正規分布

正規分布における平均への回帰は多くの文献で議論されているが、ここでは第2節の定式化の下での結果を示す. パラメータ  $\theta$  の個体内分布を正規分布  $N(\xi, \tau^2)$  とする. まず処置効果がない場合を考察する. パラメータ  $\theta$  の個体の処置前後の観測値  $X$  および  $Y$  は共に分散が  $\theta$  に依存しない正規分布  $N(\theta, \sigma^2)$  に従うと仮定する (個体内分布). このとき、処置前値  $X$  および処置後値  $Y$  の確率分布は共に  $N(\xi, \sigma^2 + \tau^2)$  となる. よって、 $E[X] = E[Y] = \xi$ ,  $V[X]$

$=V[Y]=\sigma^2+\tau^2$  であり,  $(X, Y)$  の同時分布は, (2.3) の計算より  $N(\xi, \xi, \sigma^2+\tau^2, \sigma^2+\tau^2, \tau^2)$  となって,  $Cov[X, Y]=\tau^2, R[X, Y]=\tau^2/(\sigma^2+\tau^2)$  を得る.

処置前値が  $X=x$  のときの  $\theta$  の事後分布は, (2.4) より

$$N\left(\frac{\tau^2 x + \sigma^2 \xi}{\sigma^2 + \tau^2}, \tau^2 - \frac{\tau^4}{\sigma^2 + \tau^2}\right)$$

となる. すなわち,  $\theta$  の事後平均は事前平均  $\xi$  と実現値  $x$  の加重平均であり, その値は実現値  $x$  より事前平均  $\xi$  に近い. その理由は  $\theta$  の事前分布がひと山形であるためである. 図1は簡単のため  $\theta$  の事前分布を  $N(0, 1)$  とした図であり (どんな正規分布でも同様),  $x=1$  が観測された場合, パラメータ値 (個体の真値) が  $\theta=0.5$  であるが偶然変動 (個体内変動) によりたまたま大きな値となって  $x=1$  となった個体比率 (確率密度関数値で表現される) は,  $\theta=1.5$  であるがたまたま小さな値となって  $x=1$  となった個体比率よりも大きいため,  $\theta$  の事後平均  $E[\theta|x=1]$  は事前平均  $\xi=0$  に近づくのである.

処置後値  $Y$  のパラメータ  $\theta$  が上記の分布に従うとすると, (2.5) より  $Y$  の  $X=x$  の条件下での条件付き分布は

$$N\left(\frac{\tau^2 x + \sigma^2 \xi}{\sigma^2 + \tau^2}, \sigma^2 + \tau^2 - \frac{\tau^4}{\sigma^2 + \tau^2}\right)$$

となることが示される. よって,  $Y$  の条件付き期待値は  $x$  の線形関数 (回帰直線) となり, 回帰係数は  $\beta = Cov[X, Y]/V[X] = \tau^2/(\sigma^2 + \tau^2)$  であるので,

$$E[Y|X=x] = \frac{\tau^2 x + \sigma^2 \xi}{\sigma^2 + \tau^2} = \xi + \frac{\tau^2}{\sigma^2 + \tau^2}(x - \xi) = E[Y] + \beta(x - E[X])$$

というよく知られた式に帰着される.  $x > \xi (=E[X])$  であれば,

$$E[Y|X=x] - E[Y] = \frac{\tau^2}{\sigma^2 + \tau^2}(x - \xi)$$

より,  $\tau^2/(\sigma^2 + \tau^2) \leq 1$  であるので平均への回帰 (1.1) が観察される. 降圧剤や抗コレステロール剤などの臨床試験では, 血圧値やコレステロール値がある基準値よりも大きな被験者のみが

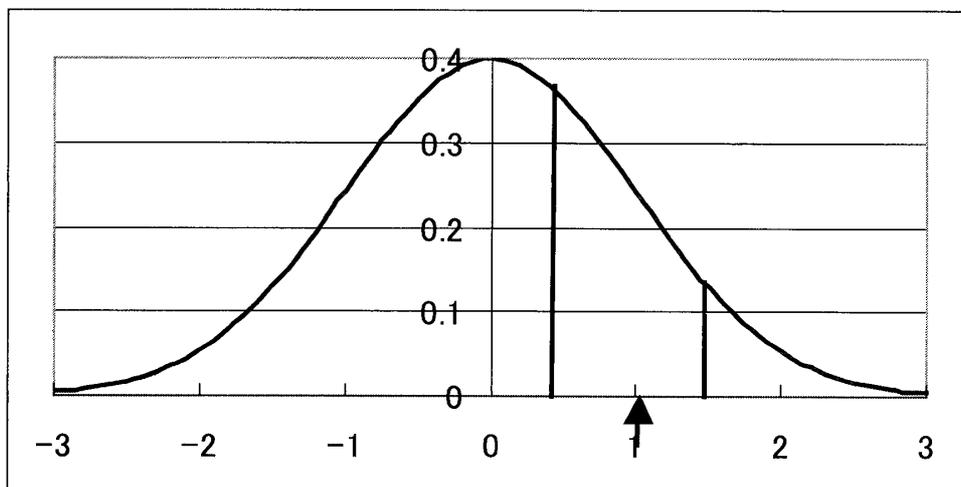


図1  $x=1$  を与える  $\theta$  の事前分布の確率

試験にエントリーされるというスクリーニングがあるため、薬剤の効果が何もなくとも2度目の計測では血圧値は平均的には下がることが多い。

ここで述べた正規分布では、事前分布も正規分布というひと山形の分布で、個体内変動の分布の分散は  $\theta$  によらず一定であるので、第2節で述べた平均への回帰の条件のうち (a) が成り立ち (b) は成り立たない場合に相当する。

次に処置効果がある場合を議論する。パラメータ  $\theta$  を持つ個体の処置前値での観測値の分布を正規分布  $N(\theta, \sigma_0^2)$  とし、処置後値の分布を  $N(\theta+c, \sigma_1^2)$  とする。  $c$  が処置効果で、全ての個体のパラメータを、その値によらず同じ  $c$  だけ変化させるというモデルである。この想定では、処置前後の個体内分布は平均のみが異なり分散が同じ正規分布となるが、ここではそれをやや一般化し、処置前後で個体内分布の分散が異なるとして結果を導く。パラメータ  $\theta$  の個体内分布をパラメータ  $\xi, \tau^2$  の正規分布  $N(\xi, \tau^2)$  とすると、処置前値  $X$  および処置後値  $Y$  の確率分布はそれぞれ  $N(\xi, \sigma_0^2 + \tau^2)$ ,  $N(\xi+c, \sigma_1^2 + \tau^2)$  となり、 $(X, Y)$  の同時分布は  $N(\xi, \xi+c, \sigma_0^2 + \tau^2, \sigma_1^2 + \tau^2, \tau^2)$  となることが示される。よって

$$\text{Cov}\{X, Y\} = \tau^2, \quad R[X, Y] = \frac{\tau^2}{\sqrt{\sigma_0^2 + \tau^2} \sqrt{\sigma_1^2 + \tau^2}}$$

となる。

処置前値が  $x$  のときのパラメータ  $\theta$  の事後分布は

$$N\left(\frac{\tau^2 x + \sigma_0^2 \xi}{\sigma_0^2 + \tau^2}, \tau^2 - \frac{\tau^4}{\sigma_0^2 + \tau^2}\right)$$

となる。このとき、 $Y$  の  $X=x$  の条件付きでの分布は

$$N\left(\frac{\tau^2 x + \sigma_0^2 \xi}{\sigma_0^2 + \tau^2} + c, \sigma_1^2 + \tau^2 - \frac{\tau^4}{\sigma_0^2 + \tau^2}\right)$$

となり、条件付き期待値は  $x$  の線形関数（回帰直線）となる。回帰係数は  $\beta = \text{Cov}[X, Y] / V[X] = \tau^2 / (\sigma_0^2 + \tau^2)$  であるので、

$$E[Y|X=x] = \frac{\tau^2 x + \sigma_0^2 \xi}{\sigma_0^2 + \tau^2} + c = \xi + c + \frac{\tau^2}{\sigma_0^2 + \tau^2} (x - \xi) = E[Y] + \beta(x - E[X])$$

が成り立つ。  $x > \xi (= E[X])$  であれば

$$E[Y|X=x] - E[Y] = \frac{\tau^2}{\sigma_0^2 + \tau^2} (x - \xi)$$

となり、処置効果がない場合と同様、平均への回帰が観察される。

#### 4. ガンマポアソン分布

観測値は稀な事象の生起回数とし、ある個体での処置前の観測値  $X$  はパラメータ  $\lambda$  のポアソン分布  $Po(\lambda)$  に従うとする（伝統に従いパラメータを  $\theta$  でなく  $\lambda$  とした）。  $Po(\lambda)$  の確率関数は

$$h(x|\lambda) = \Pr(X=x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x=0, 1, \dots)$$

であり，期待値および分散は共に $\lambda$ である．同じ個体の処置後の観測値 $Y$ はパラメータ $c\lambda$ のポアソン分布 $Po(c\lambda)$ に従うとする．定数 $c$ が処置の効果を表わし，全ての個体の期待値を $c$ だけ変化させるというモデルである． $c=1$ が無効果を意味する．ポアソン分布の場合には処置効果 $c$ を考慮したほうが議論の見通しがよくなるので，最初から $c$ を導入した．そして，パラメータ $\lambda$ の個体間分布を形状パラメータ $a$ ，尺度パラメータ $b$ のガンマ分布 $\text{Gamma}(a, b)$ とする( $a>0, b>0$ )． $\text{Gamma}(a, b)$ の確率密度関数は $\lambda \geq 0$ の範囲で

$$g(\lambda; a, b) = \frac{1}{\Gamma(a)b^a} \lambda^{a-1} e^{-\lambda/b}$$

である．ここで $\Gamma(a)$ はガンマ関数であり， $a$ が自然数のときは $\Gamma(a) = (a-1)!$ となる． $a=1$ のガンマ分布は平均値 $b$ の指数分布である． $\text{Gamma}(a, b)$ の期待値および分散はそれぞれ $E[\lambda] = ab$ ， $V[\lambda] = ab^2$ で与えられる．また， $a \geq 1$ のときモード（最頻値）は $\lambda = a-1$ となる．このとき，処置前値 $X$ の周辺確率分布は(2.1)より

$$f(x; a, b) = \Pr(X=x; a, b) = \frac{1}{x!} \frac{\Gamma(a+x)}{\Gamma(a)} \left(\frac{1}{b+1}\right)^a \left(\frac{b}{b+1}\right)^x \quad (4.1)$$

となる．これをパラメータ $(a, b/(b+1))$ のガンマポアソン分布 (gamma-Poisson distribution) といい $GP(a, b/(b+1))$ と書く (負の二項分布 (negative binomial distribution)  $NB(a, 1/(b+1))$ ともいう)． $a$ が自然数のときは $\Gamma(a+x)/\Gamma(a) = (a+x-1)!/(a-1)!$ であるので，

$$f(x; a, b) = {}_{a+x-1}C_x \left(\frac{1}{b+1}\right)^a \left(\frac{b}{b+1}\right)^x \quad (4.2)$$

となる．(4.2)は成功の確率 $p$ のベルヌーイ試行で $a$ 回成功するまでに要した失敗の回数の分布でもあり，パスカル分布 (Pascal distribution) ともいう．(4.1)および(4.2)は負の二項分布もしくはパスカル分布と確率関数がたまたま一致するが (Johnson, Kotz and Kemp (1992) を参照)，その成り立ちを考えると，負の二項分布とよぶよりガンマポアソン分布としたほうが自然である．(4.1)の期待値と分散は $E[X] = ab$ ， $V[X] = ab(b+1)$ である．処置後値 $Y$ の確率関数は(2.6)より

$$f(y; a, b, c) = \frac{1}{y!} \frac{\Gamma(a+y)}{\Gamma(a)} \left(\frac{1}{bc+1}\right)^a \left(\frac{bc}{bc+1}\right)^y$$

となり，これは $GP(a, bc/(bc+1))$ である．よって，期待値と分散は $E[Y] = abc$ ， $V[Y] = abc(bc+1)$ となる．ガンマポアソン分布は，薬剤の市販後の安全性情報へのデータマイニング的アプローチ (DuMouchel, 1999) や稀なイベントに関する“Rule of Three” (岩崎・吉田, 2005) など，特に医薬品の安全性の分析で効果的に用いられている．

処置前後の値 $(X, Y)$ の同時確率は(2.7)より

$$f(x, y; a, b, c) = \frac{1}{x!y!} \frac{\Gamma(a+x+y)}{\Gamma(a)} \left(\frac{1}{b+bc+1}\right)^a \left(\frac{b}{b+bc+1}\right)^x \left(\frac{bc}{b+bc+1}\right)^y$$

となる．これは2変量ガンマポアソン分布 (bivariate gamma-Poisson distribution) であり， $GP_2(a, b/(b+bc+1), bc/(b+bc+1))$ と書く．この確率分布は，多変量負の二項分布もしくは負の多項分布ともよばれるが (Bates and Neyman (1952), Johnson, Kotz and Balakrishnan

(1997)などを参照), 前述の理由により, 2変量ガンマポアソン分布とよんだほうがよい. 共分散は  $Cov[X, Y] = ab^2c$  であり, 相関係数は

$$R[X, Y] = \frac{bc}{\sqrt{c(b+1)(bc+1)}}$$

と  $a$  に無関係となる. 特に  $c=1$  とすると  $R[X, Y] = b/(b+1)$  となる. 図2は  $a=5$ ,  $b=1$  の場合のガンマポアソン分布の確率のグラフである.

処置前値  $X=x$  が与えられたときのパラメータ  $\lambda$  の事後分布は

$$g(\lambda|x; a, b) = \frac{\lambda^{a+x-1} \exp[-\lambda/\{b/(b+1)\}]}{\Gamma(a+x) \{b/(b+1)\}^{a+x}}$$

と形状パラメータ  $a+x$ , 尺度パラメータ  $b/(b+1)$  のガンマ分布となる. 処置後値  $Y$  のパラメータ  $\lambda$  がこの分布に従うとすると,  $Y$  の  $X=x$  の下での条件付き分布は

$$f(y|x; a, b, c) = \frac{1}{y!} \frac{\Gamma(a+x+y)}{\Gamma(a+x)} \left(\frac{b+1}{b+bc+1}\right)^{a+x} \left(\frac{bc}{b+bc+1}\right)^y$$

となる. これは  $GP(a+x, bc/(b+bc+1))$  であり, 条件付き期待値は

$$E[Y|X=x] = \frac{bc}{b+1}(a+x)$$

と処置前の観測値  $x$  の線形関数になる. このとき, 回帰係数は  $\beta = Cov[X, Y]/V[X] = bc/(b+1)$  であるので

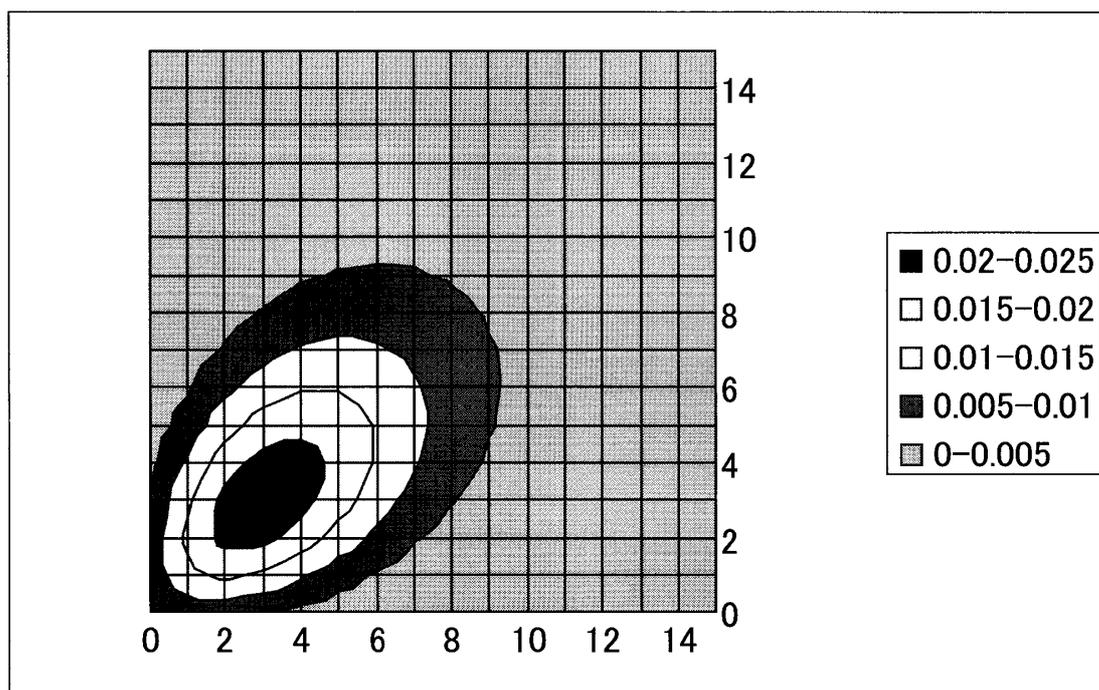


図2 2変量ガンマポアソン分布の確率の図示 ( $a=5$ ,  $b=1$ )

$$E[Y|X=x] = abc + \frac{bc}{b+1}(x-ab) = E[Y] + \beta(x - E[X])$$

が成り立つ。  $x > ab (= E[X])$  であれば、  $Y$  の条件付き期待値も

$$E[Y|X=x] - E[Y] = \frac{bc}{b+1}(x-ab) > 0$$

と  $Y$  の周辺期待値よりも大きいが、

$$\frac{E[Y|X=x] - E[Y]}{x - E[X]} = \frac{\frac{bc}{b+1}(x-ab)}{x-ab} = \frac{bc}{b+1} < c \quad (4.3)$$

とパラメータの比  $c$  よりも小さくなる。 処置効果がない、すなわち  $c=1$  の場合には (4.3) は平均への回帰 (1.2) に相当する。

ガンマポアソン分布では、第2節の平均への回帰の条件のうち (a) は成り立つが、(b) の逆、すなわち  $\lambda$  が大きいほど個体内分散は大きくなる。したがって、正規分布に比べ、平均への回帰現象はやや小さくなる。

## 5. ベータ二項分布

ここでは今ひとつのカウントデータとして、試行回数  $n$ 、成功の確率 (二項確率)  $\theta$  の二項分布  $Binom(n, \theta)$  を扱う。臨床試験での主要なエンドポイントに二項分布が想定される場合はあまり多くないが、決められた個数中でのある種の反応の個数や QOL 質問表でのチェックの個数などに用いられる可能性がある。一方、学力試験では、テストの「正答・誤答」データとして普通に見られる。二項分布で  $n$  が大きく  $\theta$  が小さい稀な事象の場合には、第4節のポアソン分布による近似が有効である。岩崎・吉田 (2005)、Yoshida and Iwasaki (2007) では、市販後の薬剤の稀で重篤な有害事象の検出に関する研究で、二項分布のポアソン分布による近似の精度に言及している。また、二項分布の拡張としては Altham (1978)、Kupper and Haseman (1978)、Danaher and Hardie (2005) などがある。

処置前値  $X$  は二項分布  $Binom(n, \theta)$  に従うとする。すなわち、

$$h(x|\theta) = \Pr(X=x|\theta) = {}_n C_x \theta^x (1-\theta)^{n-x} \quad (x=0, 1, \dots, n)$$

である。そして、二項確率  $\theta$  の個体間分布 (事前分布) をパラメータ  $a$  および  $b$  のベータ分布  $Beta(a, b)$  とする ( $a > 0$  および  $b > 0$ )。  $Beta(a, b)$  の確率密度関数は

$$g(\theta; a, b) = \begin{cases} \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1} & (0 \leq \theta \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

である。ここで  $B(a, b)$  はベータ関数であり、ガンマ関数を用いて  $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$  と書け、 $a$  が自然数のときは  $\Gamma(a) = (a-1)!$  であるので  $B(a, b) = (a+b)/(ab \times {}_{a+b} C_a)$  となる。特に  $a=b=1$  のときの  $Beta(1, 1)$  は区間  $(0, 1)$  上の一様分布となり、 $a > 1$  および  $b > 1$  のときはひと山形の分布となる。  $Beta(a, b)$  の期待値と分散は  $E[\theta] = a/(a+b)$ 、 $V[\theta] = ab/\{(a+b)^2(a+b+1)\}$  であり、 $a > 1$ 、 $b > 1$  のとき、モード (最頻値) は  $\theta = (a-1)/(a+b-2)$  で与えられる。ベータ分布は二項確率の共役事前分布である。

このとき、母集団全体での  $X$  の確率分布は

$$\begin{aligned} f(x; a, b) &= \Pr(X=x; a, b) = \int_0^1 {}_n C_x \theta^x (1-\theta)^{n-x} \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \\ &= {}_n C_x \frac{1}{B(a, b)} \int_0^1 \theta^{a+x-1} (1-\theta)^{b+n-x-1} d\theta = {}_n C_x \frac{B(a+x, b+n-x)}{B(a, b)} \\ &= {}_n C_x \frac{\Gamma(a+x)\Gamma(b+n-x)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)} \end{aligned}$$

となる。  $\Gamma(a+x)/\Gamma(a) = (a+x-1)(a+x-2)\cdots(a+1)a$  であるので、これを  $a$  の昇べきとして記号  $(a)_x$  で表わすと、結局

$$f(x; a, b) = {}_n C_x \frac{(a)_x (b)_{n-x}}{(a+b)^n} \quad (x=0, 1, \dots, n)$$

となり、これはパラメータ  $(n, a, b)$  のベータ二項分布 (beta-binomial distribution) である (Johnson, et al. (1992) 参照)。  $a$  および  $b$  が共に自然数のときは

$$f(x; a, b) = \frac{{}_{a+x-1} C_{a-1} \times {}_{b+n-x-1} C_{b-1}}{{}_{a+b+x-1} C_{a+b-1}} = \frac{{}_{a+x-1} C_x \times {}_{b+n-x-1} C_{n-x}}{{}_{a+b+x-1} C_n}$$

とも表わされる。  $a=b=1$  のときは  $f(x; 1, 1) = 1/(n+1)$  と離散一様分布になる。パラメータ  $(n, a, b)$  のベータ二項分布の期待値と分散は

$$E[X] = n \times \frac{a}{a+b}, \quad V[X] = n \times \frac{ab(a+b+n)}{(a+b)^2(a+b+1)}$$

である。処置の効果がないときは処置後値  $Y$  の確率分布も同じとなる。

次に、処置効果がないとして  $(X, Y)$  の同時分布を求める。同時確率  $f(x, y; a, b) = \Pr(X=x, Y=y; a, b)$  は

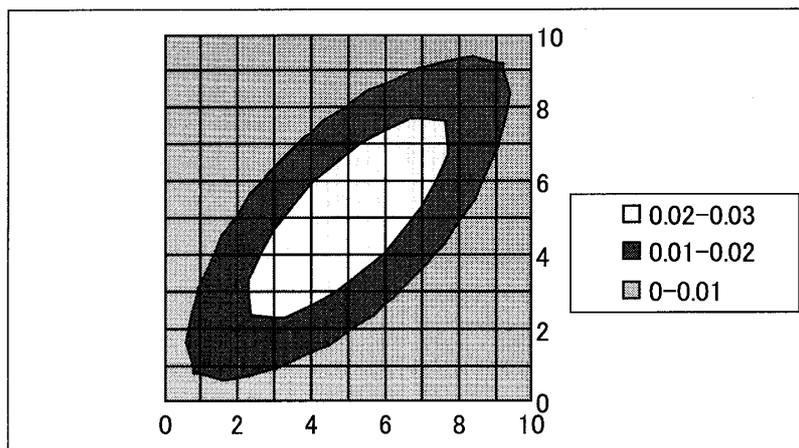
$$\begin{aligned} f(x, y; a, b) &= {}_n C_x \times {}_n C_y \frac{1}{B(a, b)} \int_0^1 \theta^x (1-\theta)^{n-x} \theta^y (1-\theta)^{n-y} \theta^{a-1} (1-\theta)^{b-1} d\theta \\ &= {}_n C_x \times {}_n C_y \frac{1}{B(a, b)} \int_0^1 \theta^{a+x+y-1} (1-\theta)^{b+(n-x)+(n-y)-1} d\theta \\ &= {}_n C_x \times {}_n C_y \frac{B(a+x+y, b+(n-x)+(n-y))}{B(a, b)} = {}_n C_x \times {}_n C_y \frac{(a)_{x+y} (b)_{(n-x)+(n-y)}}{(a+b)_{2n}} \end{aligned}$$

で与えられる。共分散は  $Cov[X, Y] = n^2 ab / \{(a+b)^2(a+b+1)\}$  であり、相関係数は

$$R[X, Y] = \frac{Cov[X, Y]}{\sqrt{V[X]V[Y]}} = \frac{n^2 ab}{(a+b)^2(a+b+1)} \frac{(a+b)^2(a+b+1)}{nab(a+b+n)} = \frac{n}{a+b+n}$$

と簡潔な表現になる。図3は  $n=10, a=b=3$  の場合の2変量ベータ二項分布の確率の図示である。

処置前値  $X=x$  が与えられたときの二項確率  $\theta$  の事後分布は

図3 2変量ベータ二項分布の確率の図示 ( $n=10, a=b=3$ )

$$g(\theta|x; a, b) = \frac{h(x|\theta)g(\theta; a, b)}{f(x; a, b)} = \frac{{}_n C_x \theta^x (1-\theta)^{n-x} \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-x}}{{}_n C_x \frac{B(a+x, b+n-x)}{B(a, b)}} \\ = \frac{\theta^{a+x-1} (1-\theta)^{b+n-x-1}}{B(a+x, b+n-x)}$$

と  $Beta(a+x, b+n-x)$  になる。処置後値  $Y$  の二項確率  $\theta$  がこの分布に従うとすると、 $Y$  の  $X=x$  の下での条件付き分布はパラメータ  $(n, a+x, b+n-x)$  のベータ二項分布

$$f(y|x; a, b) = \Pr(Y=y|X=x) = {}_n C_y \frac{(a+x)_y (b+n-x)_{n-y}}{(a+b+n)}$$

となる。よって、この分布の期待値は  $E[Y|X=x] = n(a+x)/(a+b+n)$  となり、処置前の観測値  $x$  の線形関数になることが分かる。回帰係数を  $\beta = \text{Cov}[X, Y]/V[X] = n/(a+b+n)$  とすると

$$E[Y|X=x] = E[Y] + \beta(x - E[X])$$

と正規分布と同様の表現が得られる。 $x > E[X] (= na/(a+b))$  であれば、 $Y$  の条件付き期待値は

$$E[Y|X=x] - E[Y] = \frac{n(a+x)}{a+b+n} - \frac{na}{a+b} = \frac{n\{(a+b)x - a\}}{(a+b)(a+b+n)} > 0$$

と  $Y$  の周辺期待値よりも大きく、また条件付き期待値は

$$E[Y|X=x] - x = \frac{n(a+x)}{a+b+n} - x = \frac{bx}{-a+b+n} < 0$$

と1度目の観測値  $x$  よりも小さく、平均への回帰が観察される。 $x < na/(a+b)$  のときは逆向きの不等号となる。 $a=b$  のときは ( $\theta$  の事前分布が左右対称)、

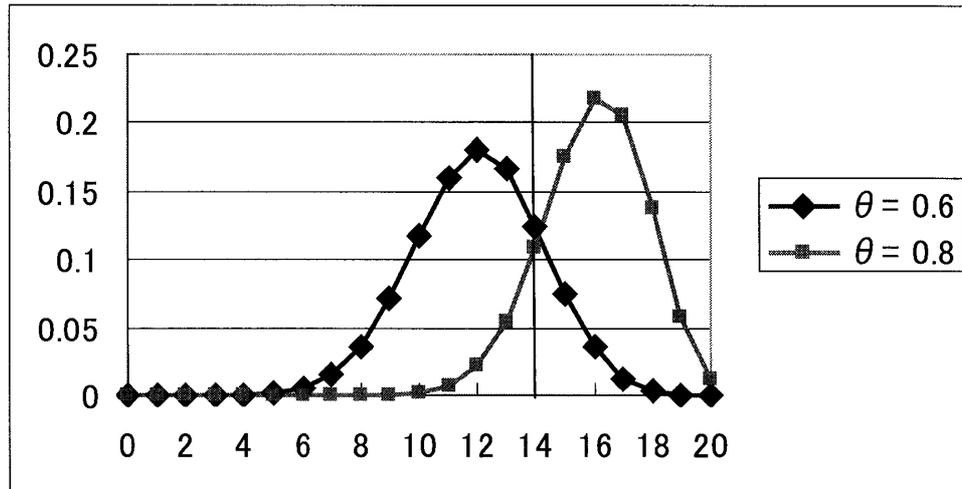


図4  $x=14$  が観測される確率( $\theta=0.6$  と  $\theta=0.8$  の比較,  $n=20$ )

$$E[Y|X=x] - E[Y] = \frac{n(x-1/2)}{2a+n}, \quad E[Y|X=x] - x = -\frac{ax}{2a+n}$$

となる. 特に  $a=b=1$  のときは

$$f(y|x; 1, 1) = \Pr(Y=y|X=x) = {}_n C_y \frac{(1+x)(1+n-x)_{n-y}}{(2+n)_n}$$

であり, 条件付き期待値は  $E[Y|X=x] = n(1+x)/(2+n)$  となるので,

$$E[Y|X=x] - E[Y] = \frac{n(x-1/2)}{2+n}, \quad E[Y|X=x] - x = -\frac{x}{2+n}$$

となる.  $\theta$  の事前分布が一様分布であっても平均への回帰が観察され, 第2節の平均への回帰の条件のうち, (a) は成り立たないが, 二項分布では二項確率が 0.5 に近いほど分散が大きいので (b) が成り立つ例となっている. 図4 は  $n=20$  のとき  $x=14$  が観測される確率を  $\theta=0.6$  と  $\theta=0.8$  で比較したものである.  $\theta=0.6$  のほうが  $\theta=0.8$  に比べ分散が大きいので  $x=14$  となる確率が大きい.  $a>1$  および  $b>1$  のときは  $\theta$  の事前分布はひと山形であるので, 第2節の条件の (a) および (b) が共に成り立つ状況となる.

処置効果がある場合の処置効果  $c$  を, 正規分布では  $\theta+c$ , ポアソン分布では  $c\lambda$  と, いずれも簡単な関数として導入した. ところが二項分布では, 二項確率の存在範囲が  $(0, 1)$  であるとの制約のため, 処置効果の定義は自明ではない. 処置効果を加味したベータ二項分布については, 数学的な議論も必要であることから別の場所で論じたい (Iwasaki, 2007).

## 6. おわりに

本論では, 処置前後研究に不可避でかつ結果の解釈に注意を要する平均への回帰現象について, その発生のメカニズムを Bayes 流のモデル化により考察した. モデルでは, 母集団内の個体間分布とそれぞれの個体の繰り返し測定における個体内分布を区別し, 平均への回帰は (a) 個体間分布のひと山性, (b) 個体内分布の不均一分散性, によるものであるとした. 分布の具体例として, 臨床試験を始め多くの分野で観察される正規分布, ポアソン分布, 二項分布

を考察し、いずれの分布でも処置後値と処置前値との関係は形式的に同じ形であることを指摘した。

ここで議論したような個体間分布と個体内分布を区別するアプローチにより、より現実に即したモデル化も可能になる。たとえば、降圧剤の臨床試験では、高血圧の患者の血圧値  $\theta$  が高いほど血圧の変動が大きいすなわち個体内分散が大きいことが経験上知られている。ところが、処置前後の血圧値に2変量正規分布を想定する従来のアプローチは、第3節で見たように個体内分布の分散を全て等しいとしたものである。したがって、より現実的なモデル化は、分散が  $\theta$  の単調増加関数であるとするものであり、その際は、第2節の平均への回帰の条件のうち (b) の逆、すなわち分布の端のほう分散が大きいという状況になり、平均への回帰の大きさはやや緩和されることになる。実際にどの程度緩和されるのかは今後の研究課題である。

本論で積み残した問題もいくつかある。その第一は、2.2節で述べたように処置が同じ  $\theta$  に対して異なる効果を持つ、すなわち処置前の  $\theta$  と処置後の  $\theta^*$  に確率分布  $(\theta, \theta^*)$  を想定する場合である。また、ベータ二項分布での処置効果の影響評価も課題として残っている。さらには、本論で取り上げた正規分布、二項分布、ポアソン分布以外にも、臨床試験では、順序カテゴリデータ（病気の症状の重さが「かなり重症」、「重症」、「やや重症」、「軽症」、「正常」など）が多く用いられる。この場合にも平均への回帰の大きさの評価は必要であり、これも今後の課題である。

本論では平均への回帰の基礎的な事柄について議論した。平均への回帰は、実際のデータ解析でともすると見過ごされ、結果の解釈を誤らせるものである。本論での知見を基に、これまでの研究結果の再検討も必要であるかもしれない。

## 謝辞

本論は Professor R. J. A. Little (University of Michigan) を招聘して2006年1月に京都で開催した科研費セミナーでの講演資料に加筆したものである。その際ご議論いただいた Professor Little および大阪大学の狩野裕教授に感謝します。また、当初の原稿の不備を指摘いただいた査読者にもお礼申し上げます。本研究は科学研究費補助金基盤研究 (A) No. 16200022 の援助を受けました。

## 参 考 文 献

- Abbess, C., Jarrett, D. and Wright, C. C. (1981). Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the 'regression-to-mean' effect, *Traffic Engineering and Control*, **22**, 535-542.
- Altham, P. M. E. (1978). Two generalizations of the binomial distribution, *Applied Statistics*, **27**, 162-167.
- Bates, G. E. and Neyman, J. (1952). Contribution to the theory of accident proneness I. An optimistic model of the correlation between light and severe accidents, *University of California Publications of Statistics*, **1**, 215-254.
- Beath, K. J. and Dobson, A. J. (1991). Regression to the mean for nonnormal populations, *Biometrika*, **78**, 431-435.
- Bonate, P. L. (2000) *Analysis of Pretest-Posttest Designs*, Chapman & Hall, Boca Raton, FL.
- Chesher, A. (1997) Non-normal variation and regression to the mean, *Statistical Methods in Medical Research*, **6**, 147-166.
- Chuang-Stein, C. and Tong, D. M. (1997). The impact and implication of regression to the mean on the design and analysis of medical investigations, *Statistical Methods in Medical Research*, **6**, 115-128.
- Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean, *Statistical Methods in Medical Research*, **6**, 167-183.
- Danaher, P. J. and Hardie, B. G. S. (2005). Bacon with your eggs? Applications of a new bivariate beta-binomial distribution, *The American Statistician*, **59**, 282-286.
- DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with application to the FDA spontaneous

- reporting system, *The American Statistician*, **53**, 177-202 (with discussion).
- 藤田利治・岩崎 学・林 邦彦・佐藤俊哉・大森 崇 (2004). 医薬品の副作用自発報告によるシグナル検出の実用化に向けての検討, 厚生労働科学研究費補助金 2003 年度分担研究報告書.
- Galton, F. (1886) Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute*, **15**, 246-263.
- Hauer, E. (1980) Selection for treatment as a source of bias in before-and-after studies, *Traffic Engineering and Control*, **21**, 419-421.
- 岩崎 学 (2002). 「処置前—処置後」データの解析と平均への回帰, 行動計量学, **29**, 247-273.
- 岩崎 学 (2006). 統計的データ解析入門 単回帰分析, 東京図書.
- Iwasaki, M. (2007). Some generalizations of the beta-binomial distribution (in preparation).
- 岩崎 学・吉田清隆 (2005). 稀な事象の生起確率に関する統計的推測—Rule of Three とその周辺—, 計量生物学, **26**, 53-63.
- Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, John Wiley & Sons, New York.
- Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Discrete Univariate Distributions, Second Edition*, John Wiley & Sons, New York.
- Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics*, **34**, 69-76.
- Lin, H. M. and Hughes, M. D. (1997). Adjusting for regression toward the mean when variables are normally distributed, *Statistical Methods in Medical Research*, **6**, 129-146.
- Maher, M. J. (1987). Fitting probability distributions to accident frequency data, *Traffic Engineering and Control*, **28**, 356-358.
- McGuigan, D. R. D. (1985). Accident 'migration' — or a flight of fancy? *Traffic Engineering and Control*, **26**, 229-233.
- Senn, S. J. (1997). Editorial: Regression to the mean, *Statistical Methods in Medical Research*, **6**, 99-102.
- Senn, S. J. and Collie, G. S. (1988). Accident blackspots and the bivariate negative binomial, *Traffic Engineering and Control*, **29**, 168-169.
- Stigler, S. M. (1997). Regression towards the mean, historically considered, *Statistical Methods in Medical Research*, **6**, 103-114.
- 竹内 啓・藤野和健 (1981). 2 項分布とポアソン分布, 東京大学出版会.
- 渡邊裕之・松下泰之・渡辺 篤・前田敏郎・温井一彦・小川嘉正・澤 淳悟・前田 博 (2004). 重要な安全性情報を早期に検出する仕組み—シグナル検出の最近の手法について—, 計量生物学, **25**, 37-60.
- Yoshida, K. and Iwasaki, M. (2007). A new testing procedure for the probability of rare events, (submitted for publication).