成蹊大学博士論文

不完全データに基づく平均への回帰に関する研究

2011年 3月

（初版　2010年12月22日）

成蹊大学大学院　工学研究科　情報処理専攻

博士後期課程　D083401

河　田　祐　一

目次

1. はじめに

1.1 研究の背景

ある処置の効果を定量的に評価する場合，同一個体に対しその処置を施す前後の観測値の比較が行なわれることが多い。たとえば，新しい降圧剤の臨床試験では，同じ患者の薬剤投与前の血圧値と投与後の血圧値を比較する。教育現場では，インターネットの活用といった新しい教育方法もしくは逆に補習授業の効果を知るため，同じクラスにおけるその教育方法の適用前の試験結果と教育終了後の試験結果の間の変化が問題となる。また，処置が2種類以上あり，それらの間の違いを評価する場合には，それぞれの処置を施した群における処置前後のデータから各処置効果を比較する（臨床試験ではこの種の試験が多い)。

処置の前後でデータを取る研究計画は pre-treatment and post-treatment design，pretest-posttest design あるいは test-retest design などと呼ばれる。また処置前値はベースライン値ということもある。本論文ではこの種の計画によって取られたデータを「処置前－処置後」データと呼び，データの観測される対象を個体あるいは被験者という。処置前後で値をとる研究計画を立てる理由は，同じ個体で2回あるいはそれ以上の観測値をとることにより各個体間のばらつきを減じ，処置効果を適切に評価しようとするものである。すなわち同一個体で対応付けてデータを取るのである。

処置前－処置後の比較研究では，処置前値によって個体が選別されることが多い。臨床試験では血圧値やコレステロール値の高い被験者のみが臨床試験の対象とされ，予備校や大学などの補習授業では成績の振るわない学生だけが補習の対象とされる。これを処置前値によるスクリーニング（screening）という。スクリーニングのあるデータの解析では，いわゆる「平均への回帰」(regression to the mean) の現象が問題となる。

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 $E[Y \mid X = x]$ としたとき，

$$| E[Y \mid X = x] - E[Y] | \leq | x - E[X] | \tag{1}$$

もしくは

$$\frac{E[Y \mid X = x] - E[Y]}{x - E[X]} \leq 1 \tag{2}$$

となる現象で，処置前値 x のその期待値 $E[X]$ からの乖離に比べ，条件付き期待値 $E[Y \mid X = x]$ 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

での補習授業の効果を見る研究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－処置後」データの解析と平均への回帰についてはこれまで多くの解説的な論文が書かれている。たとえば，Chuang-Stein (1993)，Davis (1976)，Ederer (1972)，Furby (1973)，Labouvie (1982)，McDonald, and Mazzuca (1983)，Nesselroade, Stigler and Baltes (1980)，Newell and Simpson (1990) などがあり，新薬開発の臨床試験や疫学などの医学関係と心理および教育心理学の分野に多く見られる。特に，医学統計分野の学術雑誌 Statistical Methods in Medical Research の1997年の第6巻，第2号は Regression to the Mean の特集号で，オーガナイザーの S. Senn 自身の論文 Senn (1997) に始まり，Stigler (1997)，Chuang-Stein and Tong (1997)，Lin and Hughes (1997)，Chesher (1997) および Copas (1997) の論文が集められており大変参考になる。さらに，単行本としては Bonate (2000) があるが，本書物には豊富な参考文献が載せられていて文献検索に重宝する。

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おり，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，その後に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) が取り上げた例題に関しては，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。また，岩崎 (2010) ではカウントデータの統計解析について網羅的にまとめられており非常に参考になる。

本論文の目的は以下の2つである。第一に平均への回帰モデルに対して Bayes 流のモデルを当てはめることにより定式化を行い，正規分布だけでなくポアソン分布等のカウントデータにおいてもそのモデルが成り立つことを示すことである。第二には不完全データの問題の中で主要なトピックの1つである打ち切りやトランケーションがあるデータに関する統計的な推測について議論することである。特に処置前値に対するスクリーニングなどのために，データが不完全である場合のパラメータ推定方法，処置後値の分布の形状の研究を行う。

本研究の動機は，新薬の有効性および安全性を評価するための臨床試験を実際に計画，あるいはそこから得られる不完全データに直面した結果から生じたものである。問題設定としては，3章では Bayes 流モデリングによる平均への回帰の定式化を行い，それをカウントデータに対しても適応させる問題を考える。4章では，まず処置前後データが2変量正規分布に従う場合に，処置前値に対しある種のスクリーニングが施された場合の処

置後値の分布の正規分布からの乖離について検討する。さらに，ベータ二項分布に従うと仮定したカウントデータに対し，処置前値に対しスクリーニングが施された場合の分布のパラメータ推定方法ならびに処置後値の平均への回帰の大きさについて考える。また，ベータ二項分布の一般化された分布であるディリクレ多項分布についても同様に検討する。

1.2　論文の構成

本論文では，1章に研究の背景および動機を示し，2章に平均への回帰の歴史的背景の整理，およびスクリーニングの種類とその不完全データ解析としての用語の整理を行う。

3章では平均への回帰モデルに焦点を当てる。3.3節でまず通常検討される正規分布についてのモデルを示し，3.4節以降でカウントデータについて議論する。3.4節ではガンマポアソン分布，3.5節ではベータ二項分布を取り扱う。さらには3.6節ではベータ二項分布の一般化としてディリクレ多項分布を議論する。4章では打ち切りやトランケーションが生じた場合の3つのトピックスについて議論する。4.2節では処置前後値が2変量正規分布に従うとし，処置前値にトランケーションが生じる場合の処置後値の分布の正規性の評価を行う。4.3節では処置前後の値が2変量ベータ二項分布に従うとし，処置前値にある種のスクリーニングが施された場合のパラメータ推定方法について議論する。4.4節では4.3節の議論を一般化し，ディリクレ多項分布に従うカウントデータを同様に議論する。最後に5章にて本論文の総括を行う。

なお，3章は岩崎・河田 (2007)の内容を修正したものである。4.2節，4.3節はそれぞれ，Kawata and Iwasaki (2008) と河田・岩崎 (2009) を修正したものである。

2. 平均への回帰とスクリーニング

2.1 イントロダクション

本章では，平均への回帰の概念の説明と，スクリーニングの種類の定義を行う。特にスクリーニングに関しては，不完全データの解析の立場からの説明を行う。特に欠測メカニズム（missing mechanism）と無視可能性（ignorability）について述べる。

2.2 回帰と回帰効果

「処置前－処置後」データの解析では，いわゆる平均への回帰あるいは回帰効果とよばれる現象が問題となる。2.2.1節で平均への回帰とは何かを述べた後，2.2.2節で平均への回帰にまつわる歴史的に有名な論文に言及する。

2.2.1 平均への回帰

単回帰モデル (simple regression model) を考える。回帰モデルの定式化には正規分布の仮定は必ずしも必要でないが，ここでは見通しの最もよい正規性の仮定の下で結果を導いておく。2変量確率変数 (X, Y) が2変量正規分布 $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$ に従うとする。また，相関係数を $\rho = \sigma_{YX}/\sigma_X\sigma_Y$ とする。このとき，$X = x$ が与えられた下での Y の条件付き分布は，期待値

$$E[Y \mid X = x] = \mu_Y + \beta(x - \mu_X)$$

分散

$$V[Y \mid X = x] = \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2 = \sigma_Y^2(1 - \rho^2)$$

をもつ正規分布となる。ここで $\beta = \sigma_{XY}/\sigma_X^2 = \rho(\sigma_Y/\sigma_X)$ となる回帰係数であり，$\alpha = \mu_Y - \beta\mu_X$ を定数項（切片）とした直線 $y = \alpha + \beta x$ を回帰直線 (regression line) という（詳細は岩崎 (2006) 等を参照）。

$X = x$ とした Y の条件付き期待値 $E[Y \mid X = x]$ と条件付きでない全体の期待値 μ_Y との差は

$$E[Y \mid X = x] - \mu_Y = \beta(x - \mu_X) \tag{3}$$

となる。X と Y が「処置前－処置後」データのように同じ対象の2つの測定値で両変数の分散が等しく $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ とすると，$\beta = \rho$, $V[Y \mid X = x] = \sigma^2(1 - \rho^2)$ となり，(3) は

$$E[Y \mid X = x] - \mu_Y = \rho(x - \mu_X)$$

となる。多くの場合 $0 < \rho < 1$ であるので，$x > \mu_X$ のとき Y の条件付き期待値 $E[Y \mid X = x]$ は全体の平均値 μ_Y よりも大きいものの μ_Y からの差 $E[Y \mid X = x] - \mu_Y$ は x の μ_X からの差 $x - \mu_X$ 程は大きくない。逆に，$x < \mu_X$ のときは $E[Y \mid X = x]$ は μ_Y よりも小さいものの μ_Y からの差 $E[Y \mid X = x] - \mu_Y$ の絶対値は x の μ_X からの差 $x - \mu_X$ の絶対値ほどは小さくない。2.2.2節で述べる歴史的に有名な父親の身長 (x) と息子の身長 (y) の例でいえば，背の高い父親から生まれる子供の身長の平均は子供全体の平均よりも高いものの父親ほどは高くないことを意味する。Y の条件付き期待値 $E[Y \mid X = x]$ は，x の期待値からの差 $x - \mu_X$ に比べ全体の期待値 μ_Y に近いという意味で，この現象を平均への回帰 (regression

toward the mean, regression to the mean) あるいは回帰効果 (regression effect) という。平均への回帰に関する歴史は Stigler (1997) に詳しい。また，Folks (1981) あるいは Freedman, Pisani, Purves and Adhikari (1991) といった教科書でもよく取り上げられている（岩崎 (2000) でも簡単に議論されている)。

平均への回帰は，プロスポーツなどで1年目に活躍した選手が，2年目には1年目ほどの目立った活躍はできないといういわゆる「2年目のジンクス」の説明にもなる（たとえば Ederer (1972) あるいは Schall and Smith (2000) を参照)。

2.2.2 回帰の起源

平均への回帰，というより回帰そのものを最初に論じたのは Galton (1886) である．この論文では，両親の平均身長とその子供（成長した子供 adult children）の身長の関係を議論しているが（表 1)，子供の身長として男子を基準とし女子は身長を1.08倍して男子に合わせている（単位：インチ)。また，データは1家族につき1組ではなく，同じ両親からの複数の子供がカウントされている．表 1の最後の列 (parents) が家族数を表わしている．たとえば，両親の平均身長が72.5インチの行は，6家族分で計19人の子供の分布を表わしている。

表 1の両親の平均身長を x，子供の身長を y としてデータから回帰式を求めるとおおよそ

$$y = 68.1 + 0.74(x - 68.3) = 17.6 + 0.74x$$

となる。すなわち，両親の平均身長が全体の平均値68.3インチよりも1インチ高い69.3インチの場合，その子供の身長の条件付き平均は全体の平均68.1インチよりも0.74インチ高いに過ぎないという結果である。当時の英国では身長が高いことが貴族などの身分の高さのひとつの象徴であり，背の高い両親から生まれた子供が世代を経るごとに平均値に近づき権威の象徴が失われるとして Galton はこの現象を「凡庸への回帰」(regression towards mediocrity) として心配したのであるが，平均への回帰が何かの特別な意味を持つものでなく単なる数学的事実に過ぎないことから，彼の心配は当然ながら杞憂であった。Galton (1886) の論文では，表 1のデータのほかに，両親の身長の組み合わせの度数データ，えんどう豆の種子の大きさとそこから収穫された豆の大きさとの関係など興味深いデータが掲載され分析の対象となっている。

Pearson and Lee (1903) も歴史的に有名な文献である。ここでは Galton (1886) で議論された親子の身長に加え，各個体の両腕を広げた長さおよびひじから指先までの長さを，親子，兄弟，夫婦などについて広範に調査した結果を掲載している。表 2はその中で最も頻繁に引用される父親と息子の身長の度数分布表である（単位：インチ)。表 2の度数は整数ではなく0.25刻みとなっている。これは，（父親，息子）= (63, 66) の場合には ([62.5-63.5], [65.5-66.5]) のセルに度数1が加えられるのに対し，(63.5, 66) の場合には ([62.5, 63.5], [65.5-66.5]) および ([63.5, 64.5], [65.5-66.5]) の2つのセルに0.5ずつを加え，

(63.5, 66.5) の場合には ([62.5, 63.5], [65.5-66.5]), ([63.5, 64.5], [65.5-66.5]), ([62.5, 63.5], [66.5-67.5]) および ([63.5, 64.5], [66.5-67.5]) の4つのセルに0.25ずつを加えるという集計法をとっているためである。これは，現代流にいえば度数分布表のスムージングに相当している。

表 1 両親の平均身長と子供の身長

(Galton (1886) p. 248, Table I)（表側：両親の平均身長，表頭：子供の身長）

Heights	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Total	Parents
Above												1	3		4	5
72.5								1	2	1	2	7	2	4	19	6
71.5					1	3	4	3	5	10	4	9	2	2	43	11
70.5	1		1		1	1	3	12	18	14	7	4	3	3	68	22
69.5			1	16	4	17	27	20	33	25	20	11	4	5	183	41
68.5	1		7	11	16	25	31	34	48	21	18	4	3		219	49
67.5		3	5	14	15	36	38	28	38	19	11	4			211	33
66.5		3	3	5	2	17	17	14	13	4					78	20
65.5	1		9	5	7	11	11	7	7	5	2	1			66	12
64.5	1	1	4	4	1	5	5		2						23	5
Below	1		2	4	1	2	2	1	1						14	1
Total	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205

表 2　父親と息子の身長

(Pearson and Lee (1903) p. 415, Table XXII)（表側：息子の身長，表頭：父親の身長）

	58.5–59.5	59.5–60.5	60.5–61.5	61.5–62.5	62.5–63.5	63.5–64.5	64.5–65.5	65.5–66.5	66.5–67.5	67.5–68.5	68.5–69.5	69.5–70.5	70.5–71.5	71.5–72.5	72.5–73.5	73.5–74.5	74.5–75.5	totals
59.5–60.5					0.5	0.5	1											2.0
60.5–61.5					0.5				1									1.5
61.5–62.5		0.25	0.25		0.5	1	0.25	0.25	0.5	0.5								3.5
62.5–63.5		0.25	0.25	2.25	2.25	2	4	5	2.75	1.25		0.25	0.25					20.5
63.5–64.5	1		1.5	3.75	3	4.25	8	9.25	3	1.25	1.5	0.75	1.25					38.5
64.5–65.5	2	1	0.5	2	3.25	9.5	13.5	10.75	7.5	5.5	3.5	2.5						61.5
65.5–66.5		0.5	1	2.25	5.25	9.5	10	16.75	17.5	16	5.25	2	2.5	1				89.5
66.5–67.5		1.5	2	4.75	3.5	13.75	19.75	26.5	25.75	19.5	12.5	13.75	3.25	0.5	1			148.0
67.5–68.5			1.5	2	7.5	10	10.25	24.25	31.5	23.5	29.5	13.25	8.5	9.5	2.25			173.5
68.5–69.5			1		5.25	5	12.75	18.25	16	24	29	21.5	10	3.5	2.25		1	149.5
69.5–70.5					1	2.5	5.75	18.75	11.75	19.5	22.5	19.5	14.5	6.25	3.5	1.5	1	128.0
70.5–71.5						3.25	5	8.75	10.75	19	14.75	20.75	10.75	8	5	1	1	108.0
71.5–72.5						0.25	3	1.25	7	7.75	10.75	11.25	10	8.5	2.75	0.5		63.0
72.5–73.5							0.75	0.75	2.5	7.5	6.5	6	7.5	6.25	3.25	0.5	0.5	42.0
73.5–74.5						1		1.5	1.5		5.25	2.25	2.5	6.5	3.25	3.25		2	29.0
74.5–75.5											1	2		2.5	0.75	1.75	0.5		8.5
75.5–76.5											1.25	0.25		0.5	1	1			4.0
76.5–77.5											1.25	0.25	1			1.5			4.0
77.5–78.5												1	1		0.25	0.75			3.0
78.5–79.5															0.25	0.25			0.5
Totals	3	3.5	8	17	33.5	61.5	95.5	142	137.5	154	141.5	116	78	49	28.5	4	5.5	1078

2.3　スクリーニング

「処置前－処置後」研究では，処置前値 x の値により個体の研究への組み入れの可否が決まることがある。これを処置前値によるスクリーニング (screening) と呼ぶ。スクリーニングはデータ解析に大きな影響を及ぼすことから，ここではスクリーニングの種類とその影響を扱う。2.3.1節でスクリーニングの種類を述べ，2.3.2節で不完全データ解析の立場からの議論を行なう。

2.3.1　スクリーニングの種類

処置前値 x によるスクリーニングの種類の区別はその後の解析の上できわめて重要である。ここでは Cohen (1955) および Lin and Hughes (1997) に従い，スクリーニングを「完全」,「トランケーション」,「打ち切り」および「選択」の4種類に分類する。

「完全」(complete) はスクリーニングが行なわれず，処置前値 x および処置後値 y に関するすべてのデータが得られるものを表わす。「トランケーション」(truncation) は，処置前値に設定された条件（たとえば c を予め定められた値として $c \le x$ など）を満たすもののみが研究に組み入れられて処置後値 y が測定されるが，設定条件に合わず研究に組み入れられなかったものはその個数も分からないとされる。「打ち切り」(censoring) では，処置前値に関する条件を満たす個体については (x, y) が測定され，条件に合わないものは x も y も測定されないがその個数のみは分かるというものである。「選択」(selection) では，処置前値 x は全部の測定値が得られるが，処置後値 y は x に関する条件に合うものだけが測定される。この中で，トランケーションと打ち切りはよく混同されるが，処置前値の条件に合わないものの個数の情報の有無によって解析法が異なり，得られる推定値の精度も大きく異なる。パラメータの推定精度は「完全」,「選択」,「打ち切り」,「トランケーション」の順に悪くなる（たとえば Senn and Brown (1989) 参照）。

処置前値が複数ある場合にはその組み合わせによりスクリーニングが行なわれることがある。たとえば薬効評価において，投与前値を2回測定し，それらの平均値がある値以上で差が一定値以下であるもののみを試験に組み入れるなどである。本論文ではこの問題を扱わないが，Davis (1976) に若干の議論がある。

処置前値に関するスクリーニングは事前に定められた研究計画によるものであるが，それ以外に，研究者の意図に反して何らかの理由で処置後値 y が得られないことがある。これを欠測 (missing) という。これも実際上重要な問題であるが，本論文では扱わない。

2.3.2　不完全データ解析の視点

本来得られるべきデータが得られないとき，データは不完全 (incomplete) であるという。欠測は不完全データの大きな要因である。処置前値によるスクリーニングも，本来得られるべきデータが得られないという意味で不完全データとみなすことができる。ここでは，不完全データ解析の立場から上記の各スクリーニングによる影響を考察する。不完全データ解析について詳しくは Little and Rubin (1987)，Schafer (1997)，岩崎 (2002a)，渡辺・山口 (2000) などを参照されたい。

処置前値および処置後値を表わす確率変数をそれぞれ X および Y とし，(X, Y) が2変量正規分布 $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$ に従うとする。$X = x$ が与えられたときの Y の条件付き分布は 2.2.1節で見たように $N(\alpha + \beta x, \tau^2)$ となる。ここで $\beta = \sigma_{xy}/\sigma_X^2$，$\alpha = \mu_Y - \beta\mu_X$，$\tau^2 = \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2$ である。本来推定すべきパラメータは $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$ であるが，μ_X および σ_X^2 は X の周辺分布のパラメータであるため X の観測値のみから推定される。また，

$$\mu_Y = \alpha + \beta\mu_X,\ \sigma_Y^2 = \tau^2 + \beta^2\sigma_X^2,\ \sigma_{XY} = \beta\sigma_X^2 \tag{4}$$

の関数関係により，条件付き分布に関するパラメータ α，β および τ^2 の推定値が得られれば (4) の各右辺への当該推定値の代入により μ_Y，σ_Y^2 および σ_{XY} の推定値が求められ

る。

不完全データ解析の文脈では，欠測になったデータをデータ取得の計画段階から無いものとみなした解析が妥当な場合，欠測メカニズム (missingness mechanism) は「無視可能」(ignorable) であるといい，そうでなくデータが欠測となったことを統計的推測において考慮する必要があるとき欠測は「無視可能でない」(nonignorable) という。さらに細かく，欠測メカニズムは「欠測は完全にランダム」(Missing Completely At Random = MCAR) と「欠測はランダム」(Missing At Random = MAR) に分類される。MCAR であれば欠測は常に無視可能であるが，MAR の場合には推定の対象となるパラメータおよび推定方法に依存して無視可能か否かが定まる。

変量が X のみの1変量のときは，欠測が X の値に依存しなければ欠測は無視可能であり，そうでなく欠測が X に依存して生じる場合には無視可能でない。2変量の (X, Y) では，欠測が Y のみに生じる場合（処置前値でのスクリーニングはこれに当たる），欠測が X の値にも Y の値にも無関係ならば MCAR，欠測が X の値には依存するが Y には無関係ならば MAR，欠測が Y および X の値に依存する場合が nonignorable である。なお，これらの議論では最尤法 (method of maximum likelihood) による推測あるいは Bayes 流の推測 (Bayesian inference) が前提にされることが多い。

X のパラメータ μ_X と σ_X^2 の推測では，スクリーニングが「選択」であれば X に関するデータはすべて得られているので μ_X および σ_X^2 の推測には問題は生じない。しかし，「トランケーション」および「打ち切り」は X の値に依存してスクリーニングが生じているので欠測は無視可能でなく，欠測メカニズムを反映した推測が必要となる。

μ_X と σ_X^2 以外のパラメータの推測では，スクリーニングが「選択」でも「トランケーション」でも「打ち切り」でも欠測は X のみに依存して生じることから欠測メカニズムは MAR であり，条件付き分布のパラメータの最尤推定に関する限り欠測は無視可能となる。したがって，まずこれらのパラメータを回帰分析により推定し (4) の関係式を用いて μ_Y，σ_Y^2 および σ_{XY} の推定値を求めればよい。

3. 平均への回帰モデルの定式化

3.1 イントロダクション

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。もう一つは，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)，渡邉他 (2004)，岩崎・吉田 (2005) などを参照）。

本章では，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）。3.2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。3.3節では正規分布に対し3.2節の一般論を適用する。3.4節および3.5節ではそれぞれガンマポアソン分布とベータ二項分布に対し議論する。3.6節ではベータ二項分布の一般化された分布であるディリクレ多項分布について議論する。最後の3.7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。

3.2 モデルの定式化

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。3.2.1節で処置の効果がない場合を扱い，その後3.2.2節で処置効果がある場合の定式化を示す。処置効果がない場合の考察は，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 θ とする。降圧剤の臨床試験では θ 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 θ はある生徒の教育方法適用前の真の学力とみなされる。θ 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（ξ 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 θ 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，θ 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 θ の個体の処置前値 X が確率密度関数 $h(x\mid\theta)$ 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

$$f(x;\xi)=\int_{-\infty}^{\infty}h(x\mid\theta)g(\theta;\xi)d\theta \tag{5}$$

となる。そして，X の期待値と分散をそれぞれ

$$\mu_X(\xi)=E[X;\xi],\quad \sigma_X^2(\xi)=V[X;\xi] \tag{6}$$

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 $f(x;\xi)$ と呼ぶこともある。

3.2.1　処置効果がない場合

処置の効果がないとすると，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 $\sigma_Y^2(\xi)$ も (6) と同じ値となる。θ 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

$$f(x,y;\xi)=\int_{-\infty}^{\infty}h(x\mid\theta)h(y\mid\theta)g(\theta;\xi)d\theta \tag{7}$$

となる。そして，共分散および相関係数を $\sigma_{XY}=Cov[X,Y]$，$\rho=R[X,Y]=\dfrac{Cov[X,Y]}{\sqrt{V[X]}\sqrt{V[Y]}}$ 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 $f(y\mid x;\xi)$ と条件付き期待値 $E[Y\mid X=x]$ を求める。$X=x$ が与えられたときのパラメータ θ の事後分布は

$$g(\theta\mid x;\xi)=\frac{h(x\mid\theta)g(\theta;\xi)}{f(x;\xi)} \tag{8}$$

となる。$g(\theta\mid x;\xi)$ は処置前値が x であった個体のパラメータ θ 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 θ が (8) 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

$$\begin{aligned} f(y\mid x;\xi)&=\int_{-\infty}^{\infty}h(y\mid\theta)g(\theta\mid x;\xi)d\theta \\ &=\frac{1}{f(x;\xi)}\int_{-\infty}^{\infty}h(y\mid\theta)h(x\mid\theta)g(\theta;\xi)d\theta=\frac{f(x,y;\xi)}{f(x;\xi)} \end{aligned} \tag{9}$$

となり，Y の条件付き期待値は

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

となる。(9) は条件付き密度関数に関する定義 $f(y \mid x\,;\,\xi) = f(x, y\,;\,\xi)/f(x\,;\,\xi)$ に他ならないが，一旦 $g(\theta \mid x\,;\,\xi)$ 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 θ 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 θ の事後分布が $g(\theta \mid x\,;\,\xi) = g(\theta\,;\,\xi(x))$ であったとすると，Y の $X = x$ での条件付き分布は

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

となり，条件付き期待値は

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

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

処置後値 Y の条件付き期待値 $E[Y \mid X = x]$ は条件付き事後分布 $g(\theta \mid x\,;\,\xi)$ の期待値に等しいことから，$g(\theta \mid x\,;\,\xi)$ 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 $h(x \mid \theta)$ の分散は，θ 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，(a) は，θ 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，θ は大きいが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（3.3節参照）。また，条件 (b) も θ 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（3.5節参照）。

3.2.2　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）θ が 処置後に θ^* に変化すると想定する。同じ θ 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 (θ, θ^*) 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 θ 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，(θ, θ^*) 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が，本論文では処置効

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

$$f(y;\xi,c) = \int_{-\infty}^{\infty} h(y \mid \eta(\theta;c)))g(\theta;\xi)d\theta \tag{10}$$

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

$$f(x,y;\xi,c) = \int_{-\infty}^{\infty} h(x \mid \theta)h(y \mid \eta(\theta;c)))g(\theta;\xi)d\theta \tag{11}$$

と求められる。$X = x$ が与えられたときの θ の事後分布は (8) であるので，そのときの Y の条件付き分布は

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

となり，Y の条件付き期待値は

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

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。

3.3　正規分布の場合

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3.2節の定式化の下での結果を示す。パラメータ θ の個体間分布を正規分布 $N(\xi, \tau^2)$ とする。まず処置効果がない場合を考察する。パラメータ θ の個体の処置前後の観測値 X および Y は共に分散が θ 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，(7) 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 θ の事後分布は，(8) 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，θ の事後平均は事前平均 ξ 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 ξ に近い。その理由は θ 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 θ 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，θ の事後平均 $E[\theta | x = 1]$ は事前平均 ξ に近づくのである。

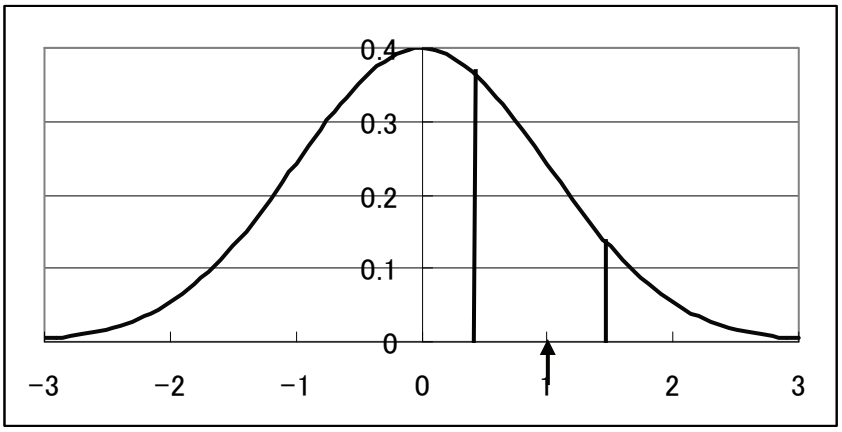


図 1　$x = 1$ を与える θ の事前分布の確率図

処置後値 Y のパラメータ θ が上記の分布に従うとすると，(9) 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，

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

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

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

より，$\tau^2/(\sigma^2 + \tau^2) \le 1$ であるので平均への回帰 (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ても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 θ によらず一定であるので，3.2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 θ 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 θ の個体間分布をパラメータ ξ, τ^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

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

となる.

処置前値が x のときのパラメータ θ 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 $\beta = Cov[X, Y]/V[X] = \tau^2/(\sigma_0^2 + \tau^2)$ であるので,

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

が成り立つ。$x > \xi (= E[X])$ であれば

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

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。

3.4　ガンマポアソン分布の場合

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 λ 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 θ でなく λ とした）。$Po(\lambda)$ の確率関数は

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

であり，期待値および分散は共に λ 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 λ 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 $Gamma(a, b)$ とする（$a > 0$，$b > 0$）。$Gamma(a, b)$ 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。$Gamma(a, b)$ 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 (5) より

$$f(x;a,b)=\Pr(X=x;a,b)=\frac{1}{x!}\frac{\Gamma(a+x)}{\Gamma(a)}\left(\frac{1}{b+1}\right)^a\left(\frac{b}{b+1}\right)^x \tag{12}$$

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,

$$f(x;a,b)={}_{a+x-1}C_x\left(\frac{1}{b+1}\right)^a\left(\frac{b}{b+1}\right)^x \tag{13}$$

となる。(13) 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。(12) および (13) 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，負の二項分布と呼ぶよりガンマポアソン分布としたほうが自然である。(12) 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 (10) 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 (11) 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呼んだほうがよい。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。

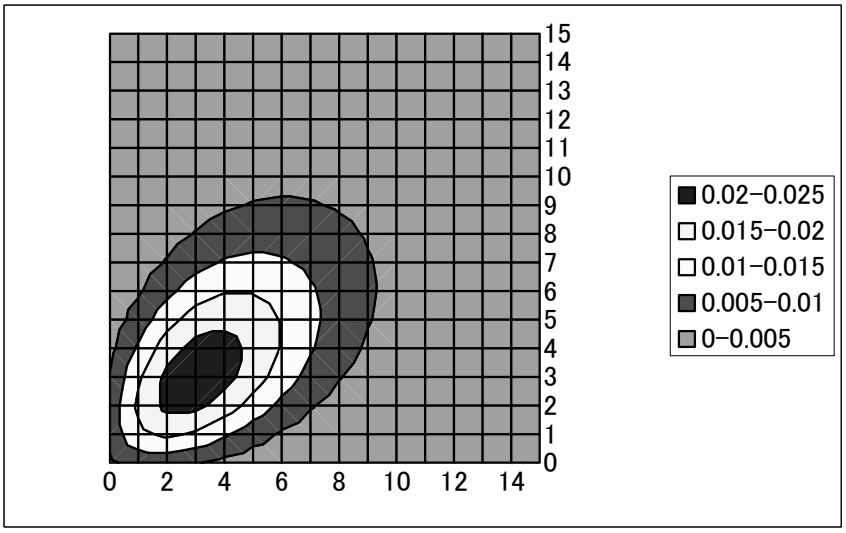


図 2　2変量ガンマポアソン分布の確率の図示（a = 5, b = 1）

処置前値 $X=x$ が与えられたときのパラメータ λ の事後分布は

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

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 λ 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

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

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

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

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

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

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

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

と Y の周辺期待値よりも大きいが,

$$\frac{E[Y \mid X=x]-E[Y]}{x-E[X]}=\frac{\frac{bc}{b+1}(x-ab)}{x-ab}=\frac{bc}{b+1}<c \tag{14}$$

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 (14) は平均への回帰 の(2) に相当する。

ガンマポアソン分布では，3.2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 λ 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。

3.5　ベータ二項分布の場合

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）θ 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 質問票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 θ が小さい稀な事象の場合には，3.4節のポアソン分布による近似が有効である。岩崎・吉田 (2005)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,

$$h(x \mid \theta)=\Pr(X=x \mid \theta)={}_nC_x\theta^x(1-\theta)^{n-x} \quad (x=0,1,\ldots,n)$$

である。そして，二項確率 θ 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

$$g(\theta;a,b)=\begin{cases}\dfrac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1} & (0\le\theta\le1)\\ 0 & (\text{その他})\end{cases}$$

である。ここで $\mathrm{B}(a, b)$ はベータ関数であり，ガンマ関数を用いて $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ と書け，a が自然数のときは $\Gamma(a) = (a - 1)!$ であるので $\mathrm{B}(a, b) = (a + b)/(ab \times {}_{a+b}C_a)$ 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

$$f(x;a,b)=\Pr(X=x;a,b)=\int_0^1 {}_nC_x\theta^x(1-\theta)^{n-x}\frac{1}{\mathrm{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta$$
$$={}_nC_x\frac{1}{\mathrm{B}(a,b)}\int_0^1\theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta={}_nC_x\frac{\mathrm{B}(a+x,b+n-x)}{\mathrm{B}(a,b)}$$
$$={}_nC_x\frac{\Gamma(a+x)\Gamma(b+n-x)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)}$$

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

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

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

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

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

$$f(x,y;a,b)={}_nC_x\times{}_nC_y\frac{1}{\mathrm{B}(a,b)}\int_0^1\theta^x(1-\theta)^{n-x}\theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1}d\theta$$
$$={}_nC_x\times{}_nC_y\frac{1}{\mathrm{B}(a,b)}\int_0^1\theta^{a+x+y-1}(1-\theta)^{b+(n-x)+(n-y)-1}d\theta$$
$$={}_nC_x\times{}_nC_y\frac{\mathrm{B}(a+x+y,b+(n-x)+(n-y))}{\mathrm{B}(a,b)}={}_nC_x\times{}_nC_y\frac{(a)_{x+y}(b)_{(n-x)+(n-y)}}{(a+b)_{2n}}$$

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

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

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。

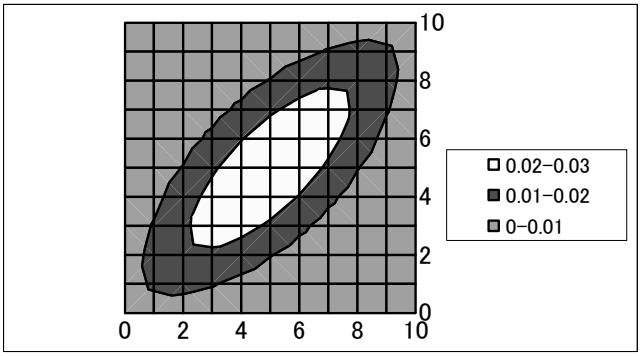


図 3　2変量ベータ二項分布の確率の図示（n = 10, a = b = 3）

処置前値 $X=x$ が与えられたときの二項確率 θ の事後分布は

$$g(\theta \mid x; a,b) = \frac{h(x \mid \theta) g(\theta; a,b)}{f(x; a,b)} = \frac{{}_nC_x \theta^x (1-\theta)^{n-x} \dfrac{1}{\mathrm{B}(a,b)} \theta^{a-1}(1-\theta)^{b-1}}{{}_nC_x \dfrac{\mathrm{B}(a+x, b+n-x)}{\mathrm{B}(a,b)}}$$

$$= \frac{\theta^{a+x-1}(1-\theta)^{b+n-x-1}}{\mathrm{B}(a+x, b+n-x)}$$

と $Beta(a+x, b+n-x)$ になる．処置後値 Y の二項確率 θ 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

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

となる。よって，この分布の期待値は $E[Y \mid X = x] = n(a + x)/(a + b + n)$ 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 $\beta = Cov[X, Y]/V[X] = n/(a + b + n)$ とすると

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

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

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

と Y の周辺期待値よりも大きく，また

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

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（θ の事前分布が左右対称），

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

となる。特に $a = b = 1$ のときは

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

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

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

となる。θ 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，3.2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 θ の事前分布はひと山形であるので，3.2節の条件の (a) および (b) が共に成り立つ状況となる。

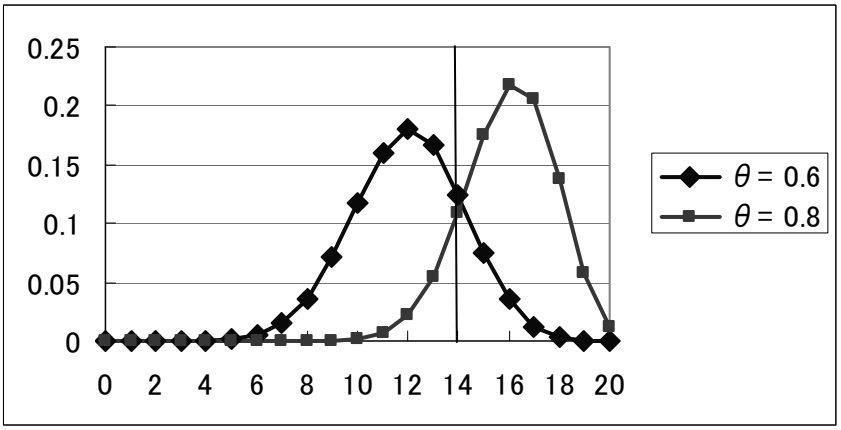


図 4 x = 14 が観測される確率（θ = 0.6 と θ = 0.8 の比較．n = 20）

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。また，処置前後で試行回数が異なる場合に拡張した議論は4.3.2節にて行う。

3.6　ディリクレ多項分布の場合

さらなるカウントデータとして，試行回数 n，取りうる値が k 個の各項の確率が θ_i $(i = 1,\ldots,k)$ である多項分布 $Multinom(n, \theta_1, \ldots, \theta_k)$ を扱う。臨床試験においてある種の QOL 質問票では，各設問の回答を k 個の選択肢で評価する。また，テストにおいて各問題につき，正解，不正解だけでなくその他の取りうる回答（無回答と誤答に対し異なる点数を与える，正答に近い回答に対して部分点を与えるなど）がある場合がある。これらのような場合に用いられる可能性があり，ベータ二項分布の3つ以上の選択肢も可能

であるよう一般化したディリクレ多項分布 (Dirichlet-multinomial distribution)を考慮することが望ましいことが多々ある（Ericson (1969) 等を参照)。また，k 個のそれぞれの値に対し s_i $(i=1,\cdots,k)$ 点が与えられ各個体（N は個体の総数） j $(j=1,\cdots,N)$ の点数は，それぞれの回答の個数を x_{ij} としたとき $z_j=\sum_{i=1}^{k}s_i x_{ij}$ $(j=1,\cdots,N)$ と表され， x_{ij} の線形結合である合計点 z_j に注目することが良くある（s_i は任意の実数で $\min(s_1,\cdots,s_k)<\max(s_1,\cdots,s_k)$ である)。このスコアに対する平均への回帰現象については3.6.1節で報告する。単純に各選択肢に対しスコアを与える方法論で対応も可能であるが，ディリクレ多項分布を考慮することにより，より柔軟なスコアリングが可能となる。

全 n 問の問題の解答の分布がパラメータ $\theta_1,\cdots,\theta_k$ の多項分布に従うとすると，

$$Multinom(n,\theta_1,\cdots,\theta_k)=\frac{n!}{\prod_{i=1}^{k}x_i!}\prod_{i=1}^{k}\theta_i^{x_i}\quad\left(\sum_{i=1}^{k}x_i=n,\sum_{i=1}^{k}\theta_i=1\right)$$

である。ここで， $E(x_i)=n\theta_i$， $Var(x_i)=n\theta_i(1-\theta_i)$， $Cov(x_i,x_j)=-n\theta_i\theta_j$ $(i\neq j)$ である。また，相関係数は

$$R(x_i,x_j)=-\sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}}\ (i\neq j)$$

となる。そして各設問の選択肢の起こりうる確率はすべての設問で同じであると仮定する。この多項確率のパラメータの個体間分布がパラメータ $a_1,\cdots,a_k$ （各 a_i は正値をとる）のディリクレ分布に従うとすると

$$D(a_1,\cdots,a_k)=\frac{1}{B(a_1,\cdots,a_k)}\prod_{i=1}^{k}\theta_i^{a_i-1}\quad\left(\sum_{i=1}^{k}\theta_i=1\right)$$

と表される。ここで $B(a_1,\cdots,a_k)=\dfrac{\prod_{i=1}^{k}\Gamma(a_i)}{\Gamma\left(\sum_{i=1}^{k}a_i\right)}$ は多変量に拡張したベータ関数であり，

$E(\theta_i)=\dfrac{a_i}{\sum_{j=1}^{k}a_j}$， $Var(\theta_i)=\dfrac{a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}}{\left(\sum_{j=1}^{k}a_j\right)^2\left(1+\sum_{j=1}^{k}a_j\right)}$ となる。ディリクレ分布は多項分布の共役事前分布である。

このとき，母集団全体での $X=(x_1,\cdots,x_k)$ の確率分布はディリクレ多項分布となり，(15) のように表される。

$$f(x_1,\ldots,x_k,a_1,\ldots,a_k)=\int_0^1\cdots\int_0^1\left(\frac{n!}{\prod_{i=1}^k x_i!}\prod_{i=1}^k\theta_i^{x_i}\frac{1}{B(a_1,\ldots,a_k)}\prod_{i=1}^k\theta_i^{a_i-1}\right)d\theta_1\cdots d\theta_k$$

$$=\int_0^1\cdots\int_0^1\left(\frac{n!}{\prod_{i=1}^k x_i!}\prod_{i=1}^k\theta_i^{x_i}\frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^k\Gamma(a_i)}\prod_{i=1}^k\theta_i^{a_i-1}\right)d\theta_1\cdots d\theta_k$$

$$=\frac{n!}{\prod_{i=1}^k x_i!}\frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^k\Gamma(a_i)}\int_0^1\cdots\int_0^1\left(\prod_{i=1}^k\theta_i^{a_i+x_i-1}\right)d\theta_1\cdots d\theta_k$$

$$=\frac{n!}{\prod_{i=1}^k x_i!}\frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^k\Gamma(a_i)}B(a_1+x_1,\ldots,a_k+x_k)=\frac{n!}{\prod_{i=1}^k x_i!}\frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^k\Gamma(a_i)}\frac{\prod_{i=1}^k\Gamma(a_i+x_i)}{\Gamma\left\{\sum_{i=1}^k(a_i+x_i)\right\}}$$

$$=\frac{n!}{\prod_{i=1}^k x_i!}\frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\Gamma\left\{\sum_{i=1}^k(a_i+x_i)\right\}}\prod_{i=1}^k\frac{\Gamma(a_i+x_i)}{\Gamma(a_i)}=\frac{n!}{\prod_{i=1}^k x_i!}\frac{\prod_{i=1}^k(a_i)_{x_i}}{\left(\sum_{i=1}^k a_i\right)_n} \tag{15}$$

また期待値と分散は

$$E(x_i)=n\times\frac{a_i}{\sum_{j=1}^k a_j},\quad Var(x_i)=n\times\frac{a_i\left\{\left(\sum_{j=1}^k a_j\right)-a_i\right\}\left(n+\sum_{j=1}^k a_j\right)}{\left(\sum_{j=1}^k a_j\right)^2\left(1+\sum_{j=1}^k a_j\right)}$$

である。 $x_s,x_t\ (s\neq t)$ の共分散は以下のように表される。

$$Cov(x_s,x_t)=-\frac{na_sa_t\left(n+\sum_{j=1}^k a_j\right)}{\left(\sum_{j=1}^k a_j\right)^2\left(1+\sum_{j=1}^k a_j\right)}$$

従って，相関係数は

$$R(x_s,x_t)=-\sqrt{\frac{a_s a_t}{\left\{\left(\sum_{j=1}^{k}a_j\right)-a_s\right\}\left\{\left(\sum_{j=1}^{k}a_j\right)-a_t\right\}}}$$

となる。多変量への拡張の場合において，処置効果をおくことはベータ二項分布の場合と同様に困難であり，処置効果がない場合のみを考えることとする。

X と $Y=(y_1,\cdots,y_k)$ は同じパラメータの分布に従うとする。ただし Y の設問数は m とする。$n=m$ は特に仮定しない。このとき X,Y の同時分布は，以下のとおりとなる。

$$f(x_1,\cdots,x_k,y_1,\cdots,y_k)=\int_0^1\cdots\int_0^1\left[\frac{n!}{\prod_{i=1}^{k}x_i!}\prod_{i=1}^{k}\theta_i^{x_i}\frac{m!}{\prod_{i=1}^{k}y_i!}\prod_{i=1}^{k}\theta_i^{y_i}\frac{\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\prod_{i=1}^{k}\Gamma(a_i)}\prod_{i=1}^{k}\theta_i^{a_i-1}\right]d\theta_1\cdots d\theta_k$$

$$=\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{m!}{\prod_{i=1}^{k}y_i!}\frac{\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\prod_{i=1}^{k}\Gamma(a_i)}\int_0^1\cdots\int_0^1\left[\prod_{i=1}^{k}\theta_i^{a_i+x_i+y_i-1}\right]d\theta_1\cdots d\theta_k$$

$$=\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{m!}{\prod_{i=1}^{k}y_i!}\frac{\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\prod_{i=1}^{k}\Gamma(a_i)}\frac{\prod_{i=1}^{k}\Gamma(a_i+x_i+y_i)}{\Gamma\left(n+m+\sum_{i=1}^{k}a_i\right)}=\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{m!}{\prod_{i=1}^{k}y_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i+y_i}}{\left(\sum_{i=1}^{k}a_i\right)_{n+m}}$$

この分布を2変量ディリクレ多項分布と呼ぶこととする。$n=m$ の場合

$$f(x_1,\cdots,x_k,y_1,\cdots,y_k)=\frac{2(n!)}{\prod_{i=1}^{k}x_i!\prod_{i=1}^{k}y_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i+y_i}}{\left(\sum_{i=1}^{k}a_i\right)_{2n}}$$

となる。同時分布の共分散は以下のとおりとなる。

$$Cov(x_s,y_t)=\begin{cases}mn\times\dfrac{a_s\left\{\left(\sum_{j=1}^{k}a_j\right)-a_s\right\}}{\left(\sum_{j=1}^{k}a_j\right)^2\left(1+\sum_{j=1}^{k}a_j\right)} & (s=t)\\ -mn\times\dfrac{a_s a_t}{\left(\sum_{j=1}^{k}a_j\right)^2\left(1+\sum_{j=1}^{k}a_j\right)} & (s\neq t)\end{cases}$$

従って，相関係数は，

$$R(x_s, y_t) = \begin{cases} \sqrt{\dfrac{mn}{\left(n+\sum_{j=1}^{k} a_j\right)\left(m+\sum_{j=1}^{k} a_j\right)}} & (s=t) \\ -\sqrt{\dfrac{mn}{\left(n+\sum_{j=1}^{k} a_j\right)\left(m+\sum_{j=1}^{k} a_j\right)} \times \dfrac{a_s a_t}{\left\{\left(\sum_{j=1}^{k} a_j\right)-a_s\right\}\left\{\left(\sum_{j=1}^{k} a_j\right)-a_t\right\}}} & (s \neq t) \end{cases}$$

となる。$n = m$ の場合，それぞれ

$$Cov(x_s, y_t) = \begin{cases} n^2 \times \dfrac{a_s\left\{\left(\sum_{j=1}^{k} a_j\right)-a_s\right\}}{\left(\sum_{j=1}^{k} a_j\right)^2\left(1+\sum_{j=1}^{k} a_j\right)} & (s=t) \\ -n^2 \times \dfrac{a_s a_t}{\left(\sum_{j=1}^{k} a_j\right)^2\left(1+\sum_{j=1}^{k} a_j\right)} & (s \neq t) \end{cases}$$

$$R(x_s, y_t) = \begin{cases} \dfrac{n}{n+\sum_{j=1}^{k} a_j} & (s=t) \\ -\dfrac{n}{n+\sum_{j=1}^{k} a_j} \sqrt{\dfrac{a_s a_t}{\left\{\left(\sum_{j=1}^{k} a_j\right)-a_s\right\}\left\{\left(\sum_{j=1}^{k} a_j\right)-a_t\right\}}} & (s \neq t) \end{cases}$$

となる。$x_1, \cdots, x_k$ が与えられたときの，多項分布のパラメータの事後分布は以下のとおりとなる。

$$g(\theta_1, \cdots, \theta_k \mid x_1, \cdots, x_k) = \frac{\dfrac{n!}{\prod_{i=1}^{k} x_i!}\prod_{i=1}^{k}\theta_i^{x_i}\dfrac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\prod_{i=1}^{k}\Gamma(a_i)}\prod_{i=1}^{k}\theta_i^{a_i-1}}{\dfrac{n!}{\prod_{i=1}^{k} x_i!}\dfrac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\Gamma\left\{\sum_{i=1}^{k}(a_i+x_i)\right\}}\prod_{i=1}^{k}\dfrac{\Gamma(a_i+x_i)}{\Gamma(a_i)}}$$

$$= \frac{\Gamma\left\{\sum_{i=1}^{k}(a_i + x_i)\right\}\prod_{i=1}^{k}\theta_i^{a_i+x_i-1}}{\prod_{i=1}^{k}\Gamma(a_i + x_i)} = \frac{1}{B(a_1 + x_1, \cdots, a_k + x_k)}\prod_{i=1}^{k}\theta_i^{a_i+x_i-1}$$

これは，パラメータ $(a_1 + x_1, \cdots, a_k + x_k)$ のディリクレ分布である。$x_1, \cdots, x_k$ が与えられたときの Y の条件付き分布は

$$f(y_1, \cdots, y_k \mid x_1, \cdots, x_k) = \int_0^1 \cdots \int_0^1 \left[\frac{m!}{\prod_{i=1}^{k} y_i!}\prod_{i=1}^{k}\theta_i^{y_i} \frac{\Gamma\left\{\sum_{i=1}^{k}(a_i + x_i)\right\}\prod_{i=1}^{k}\theta_i^{a_i+x_i-1}}{\prod_{i=1}^{k}\Gamma(a_i + x_i)} \right] d\theta_1 \cdots d\theta_k$$

$$= \frac{m!}{\prod_{i=1}^{k} y_i!}\frac{\Gamma\left\{n + \sum_{i=1}^{k} a_i\right\}}{\prod_{i=1}^{k}\Gamma(a_i + x_i)} \int_0^1 \cdots \int_0^1 \left[\prod_{i=1}^{k}\theta_i^{a_i+x_i+y_i-1} \right] d\theta_1 \cdots d\theta_k$$

$$= \frac{m!}{\prod_{i=1}^{k} y_i!}\frac{\Gamma\left\{n + \sum_{i=1}^{k} a_i\right\}}{\prod_{i=1}^{k}\Gamma(a_i + x_i)} \frac{\prod_{i=1}^{k}\Gamma(a_i + x_i + y_i)}{\Gamma\left(n + m + \sum_{i=1}^{k} a_i\right)} = \frac{m!}{\prod_{i=1}^{k} y_i!}\frac{\prod_{i=1}^{k}(a_i + x_i)_{y_i}}{\left(n + \sum_{i=1}^{k} a_i\right)_m}$$

となる。これは，パラメータ $\left(m, a_1 + x_1, \cdots, a_k + x_k\right)$ のディリクレ多項分布である。$x_1, \cdots, x_k$ が与えられたときの y_i の条件付期待値は

$$E(y_i \mid x_1, \cdots, x_k) = m \times \frac{a_i + x_i}{n + \sum_{j=1}^{k} a_j}$$

$$= m \times \frac{a_i}{\sum_{j=1}^{k} a_j} + \frac{m}{n + \sum_{i=1}^{k} a_j}\left(x_i - n \times \frac{a_i}{\sum_{j=1}^{k} a_j} \right) = E(y_i) + \frac{m}{n + \sum_{j=1}^{k} a_j}\left(x_i - E(x_i)\right)$$

となる。$x_i > E(x_i)$ の場合は，

$$E(y_i \mid x_1, \cdots, x_k) - E(y_i) = \frac{m}{n + \sum_{j=1}^{k} a_j}\left(x_i - E(x_i)\right) > 0$$

$$E(y_i \mid x_1, \cdots, x_k) - \frac{m}{n} x_i = m \times \frac{a_i + x_i}{n + \sum_{j=1}^{k} a_j} - \frac{m}{n} x_i = -\frac{m}{n}\frac{\sum_{i=1}^{k} a_i}{n + \sum_{j=1}^{k} a_j}\left(x_i - n \times \frac{a_i}{\sum_{j=1}^{k} a_j} \right) < 0$$

と，各 x_i についての平均への回帰が確認される。

3.6.1　ディリクレ多項分布の線形結合スコア分布の場合

本節では QOL 質問票や学力試験のスコアの点数が $z_x = \sum_{i=1}^{k} s_i x_i$ と線形結合スコアで表された場合の各要約統計量を以下に示す。

$x_1, \cdots, x_k$ から構成される点数を $z_x = \sum_{i=1}^{k} s_i x_i$ とすると，その分布は

$$f(z_x) = \sum_{z_x = \sum_{i=1}^{k} s_i x_i} \left\{ \frac{n!}{\prod_{i=1}^{k} x_i!} \frac{\prod_{i=1}^{k} (a_i)_{x_i}}{\left(\sum_{i=1}^{k} a_i \right)_n} \right\}$$

と表される。これ以上の簡略化は困難である。期待値と分散は

$$E(z_x) = n \times \frac{\sum_{i=1}^{k} (s_i a_i)}{\sum_{j=1}^{k} a_j}$$

$$Var(z_x) = n \times \frac{\sum_{i=1}^{k} \left[{s_i}^2 a_i \left\{ \left(\sum_{j=1}^{k} a_j \right) - a_i \right\} \right] \left(n + \sum_{j=1}^{k} a_j \right)}{\left(\sum_{j=1}^{k} a_j \right)^2 \left(1 + \sum_{j=1}^{k} a_j \right)} + \frac{2n(n-1) \sum_{s<t} \left(s_s a_s s_t a_t \right)}{\left(\sum_{j=1}^{k} a_j \right) \left(1 + \sum_{j=1}^{k} a_j \right)}$$

$$= n \times \frac{\left(n + \sum_{j=1}^{k} a_j \right) \sum_{i=1}^{k} \left[{s_i}^2 a_i \left\{ \left(\sum_{j=1}^{k} a_j \right) - a_i \right\} \right] + 2(n-1) \left(\sum_{j=1}^{k} a_j \right) \sum_{s<t} \left(s_s a_s s_t a_t \right)}{\left(\sum_{j=1}^{k} a_j \right)^2 \left(1 + \sum_{j=1}^{k} a_j \right)}$$

となる。 $z_y = \sum_{i=1}^{k} s_i y_i$ とした時の， z_x, z_y の同時分布は

$$f(z_x, z_y) = \sum_{\substack{z_x = \sum_{i=1}^{k} s_i x_i \\ z_y = \sum_{i=1}^{k} s_i y_i}} \left\{ \frac{n!}{\prod_{i=1}^{k} x_i!} \frac{m!}{\prod_{i=1}^{k} y_i!} \frac{\prod_{i=1}^{k} (a_i)_{x_i + y_i}}{\left(\sum_{i=1}^{k} a_i \right)_{n+m}} \right\}$$

と表される。 $n = m$ の場合は，以下のとおりとなる。

$$f(z_x,z_y)=\sum_{\substack{z_x=\sum_{i=1}^{k}s_ix_i\\ z_y=\sum_{i=1}^{k}s_iy_i}}\left\{\frac{2(n!)}{\prod_{i=1}^{k}x_i!\prod_{i=1}^{k}y_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i+y_i}}{\left(\sum_{i=1}^{k}a_i\right)_{2n}}\right\}$$

z_x,z_y の共分散は

$$Cov(z_x,z_y)=Cov\left(\sum_{s=1}^{k}s_sx_s,\sum_{t=1}^{k}s_ty_t\right)=\sum_{s=1}^{k}\sum_{t=1}^{k}Cov\left(s_sx_s,s_ty_t\right)=\sum_{s=1}^{k}\sum_{t=1}^{k}\left\{s_ss_tCov\left(x_s,y_t\right)\right\}$$

$$=\frac{mn}{\left(\sum_{j=1}^{k}a_j\right)^2\left(1+\sum_{j=1}^{k}a_j\right)}\left[\sum_{i=1}^{k}\left\{s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right\}-2\sum_{s<t}\left(s_sa_ss_ta_t\right)\right]$$

となり，相関係数は

$$R(z_x,z_y)=\frac{\sqrt{mn}\left[\sum_{i=1}^{k}\left\{s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right\}-2\sum_{s<t}\left(s_sa_ss_ta_t\right)\right]}{\left(\sqrt{\left(n+\sum_{j=1}^{k}a_j\right)\sum_{i=1}^{k}\left[s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right]+2(n-1)\left(\sum_{j=1}^{k}a_j\right)\sum_{s<t}\left(s_sa_ss_ta_t\right)}\right.}$$
$$\left.\times\sqrt{\left(m+\sum_{j=1}^{k}a_j\right)\sum_{i=1}^{k}\left[s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right]+2(m-1)\left(\sum_{j=1}^{k}a_j\right)\sum_{s<t}\left(s_sa_ss_ta_t\right)}\right)$$

となる。 $n=m$ の場合は以下のとおりとなる。

$$Cov(z_x,z_y)=\frac{n^2}{\left(\sum_{j=1}^{k}a_j\right)^2\left(1+\sum_{j=1}^{k}a_j\right)}\left[\sum_{i=1}^{k}\left\{s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right\}-2\sum_{s<t}\left(s_sa_ss_ta_t\right)\right]$$

$$R(z_x,z_y)=\frac{n\left[\sum_{i=1}^{k}\left\{s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right\}-2\sum_{s<t}\left(s_sa_ss_ta_t\right)\right]}{\left(n+\sum_{j=1}^{k}a_j\right)\sum_{i=1}^{k}\left[s_i^{\,2}a_i\left\{\left(\sum_{j=1}^{k}a_j\right)-a_i\right\}\right]+2(n-1)\left(\sum_{j=1}^{k}a_j\right)\sum_{s<t}\left(s_sa_ss_ta_t\right)}$$

$x_1,\cdots,x_k$ が与えられたときの $z_y=\sum_{i=1}^{k}s_iy_i$ の条件付期待値は

$$E(z_y\mid x_1,\cdots,x_k)=E(\sum_{i=1}^{k}s_iy_i\mid x_1,\cdots,x_k)=\sum_{i=1}^{k}\left\{s_iE\left(y_i\mid x_1,\cdots,x_k\right)\right\}$$

$$= \sum_{i=1}^{k} \left\{ s_i \times m \times \frac{a_i + x_i}{n + \sum_{j=1}^{k} a_j} \right\} = \sum_{i=1}^{k} \left\{ s_i m \times \frac{a_i}{\sum_{j=1}^{k} a_j} + \frac{s_i m}{n + \sum_{j=1}^{k} a_j} \left(x_i - n \times \frac{a_i}{\sum_{j=1}^{k} a_j} \right) \right\}$$

$$= \sum_{i=1}^{k} \{ s_i E(y_i) \} + \frac{m}{n + \sum_{j=1}^{k} a_j} \sum_{i=1}^{k} \{ s_i (x_i - E(x_i)) \} = E(z_y) + \frac{m}{n + \sum_{j=1}^{k} a_j} \{ z_x - E(z_x) \}$$

$z_x > E(z_x)$ の場合は

$$E(z_y \mid x_1, \cdots, x_k) - E(z_y) = \frac{m}{n + \sum_{j=1}^{k} a_j} (z_x - E(z_x)) > 0$$

$$E(z_y \mid x_1, \cdots, x_k) - \frac{m}{n} z_x = \sum_{i=1}^{k} \left\{ s_i \times m \times \frac{a_i + x_i}{n + \sum_{j=1}^{k} a_j} \right\} - \frac{m}{n} z_x$$

$$= \frac{m}{n + \sum_{j=1}^{k} a_j} \left\{ \sum_{i=1}^{k} (s_i a_i) + z_x \right\} - \frac{m}{n} z_x = -\frac{m}{n} \frac{\sum_{j=1}^{k} a_j}{n + \sum_{j=1}^{k} a_j} \left[z_x - \sum_{i=1}^{k} \frac{s_i n a_i}{\sum_{j=1}^{k} a_j} \right]$$

$$= -\frac{m}{n} \frac{\sum_{j=1}^{k} a_j}{n + \sum_{j=1}^{k} a_j} [z_x - E(z_x)] < 0$$

と，各 z_x についての平均への回帰が確認される。

3.7　議論

本論文では，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 θ 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，3.3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 θ 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が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。

本論文で積み残した問題もいくつかある。その第一は，3.2.2節で述べたように処置が同じ θ に対して異なる効果を持つ，すなわち処置前の θ と処置後の θ^* に確率分布 (θ,θ^*) を想定する場合である。また，ベータ二項分布，ディリクレ多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。正規分布，二項分布，ポアソン分布以外にも，臨床試験では，順序カテゴリーデータ（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。この場合には本論文で取り上げたディリクレ多項分布を用いた平均への回帰の大きさの評価が可能である。また，QOL 質問票などでは，各設問間で強い関連性が認められることが多いが，実際にはその関連の強さは各被験者の選択肢の選び方の違い（軽めにつけやすい人，深刻に考えすぎる人など）が反映されていると考えることが可能であり，ディリクレ多項モデルの考え方が当てはまる可能性がある。

本章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。本章での知見を基に，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。

4. 処置前後データにおける様々な不完全性の問題とその対処

4.1 イントロダクション

本章では，3種類の不完全性の問題設定における統計解析について議論する。4.2節では処置前後値が2変量正規分布に従うとし，トランケーションが処置前値に生じた場合のる場合の処置後値の分布の正規性の評価を行う。4.3節では処置前後の値が2変量ベータ二項分布に従うとし，処置前値にある種のスクリーニングが施された場合のパラメータ推定方法，ならびに平均への回帰の程度の定式化について議論する。4.4節では4.3節の議論を一般化し，ディリクレ多項分布に従うカウントデータにおける議論を行う。

4.2 処置前値にスクリーニングがある場合の処置後値の分布の評価

4.2.1 イントロダクション

処置前後研究のデザインは対応のあるデータ解析の分野の一つである。このようなデザインはさまざまな分野で利用されている。特に，臨床試験のように同じ個体に対し2種類の異なる状況，すなわち薬剤の投与前後において検査を実施するといった場合がこれに該当する。本章では測定値のことを処置前値，処置後値と呼び，X と Y で示す。処置前値においてはスクリーニング検査が実施され，個体が臨床試験に組み込まれるかどうかが決定されることがある。臨床試験においてはしばしばスクリーニング検査は実施される。以下に例を提示する。

例題 1

降圧剤の効果を評価するための臨床試験が実施された。200名の被験者の収縮期血圧が薬剤投与の前後で測定された。スクリーニングの基準の1つが投与前の収縮期血圧が160 mmHg 以上であることであったとし，図 5に投与前後の収縮期血圧の散布図とヒストグラムを示す。投与前のスクリーニング検査により，投与前値のヒストグラムは正規分布から非常に乖離した形状となっている。一方，投与後値のヒストグラムは正規分布に似通った形状を保っている。投与前後の歪度と尖度はそれぞれ1.271，1.204と0.113と −0.220 であった。

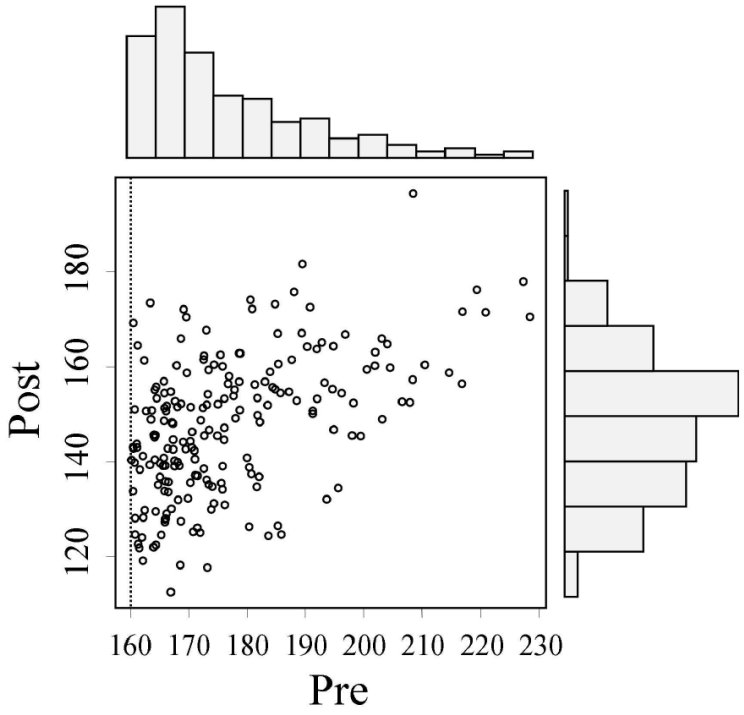


図 5　投与前後の収縮期血圧（㎜Hg）の散布図と各軸の周辺分布ヒストグラム

処置前後研究デザインにおいて，もっとも用いられているパラメトリック検定は処置後値Yに対する Student の t 検定，ならびに処置前後の変化量$Y-X$に対する対応のある t 検定である。本研究デザインに対しては，Fleiss (1986), Stanek (1988), Senn (1997), Bonate (2000), Wei and Zhang (2001), 岩崎 (2002b), 上坂 (2006)等，様々な場で議論がなされている。t 検定を用いる妥当性は測定値の分布が正規分布に従っているかどうかが重要である。しかしながら上記の例の様に処置前値に対するスクリーニング検査が実施された場合には，その分布は正規分布にはもはや従わない。投与前後の同時分布も当然2変量正規分布には従わないが，上記図 5のとおり，処置後値の周辺分布は正規分布に近い形状を保っている。なぜ正規分布に近いのか，t 検定の適用はこの場合適切であるのか，についての疑問について本節で議論する。

本節では，スクリーニングが実施されない状況下で処置前後の測定値の組(X,Y)が2変量正規分布に従うと仮定し，処置前値でスクリーニングが実施される処置前後研究デザインにおける処置後値の分布の非正規性の程度について検討する。より一般化された統計量$Y-\gamma X$を用いて主に議論する。ここでγは定数値であり，$0\le\gamma\le1$である。この統計量には処置後値Y，変化量$Y-X$がその特別な状況として含まれている。共分散分析(ANCOVA)モデルである$Y-\rho X$についても，本統計量の特別な場合と考えることができる。ここで，ρはXとYの相関である。最も興味深い点は，正規分布からの乖離が最も大きい状況とはどのような場合に生じるかである。言い換えると最悪の場合はどのよう

なときかという点である。正規性からの乖離の大きさが t 検定のパフォーマンスへ与える影響の評価は困難であるが，t 検定の有意水準に対する非正規性の影響の検討は 柴田(1981)で行われている。さらに Lehmann (1975)はどのような場合にノンパラメトリック検定がパラメトリック検定と比較し検出力が勝るかについてまとめている。もし正規性からの乖離が大きい場合，それが最悪の場合であったとしても許容可能な場合，正規性の仮定を前提とした手法が安全に利用可能であることがわかる。非正規性を評価する指標として，本節では歪度(skewness)，尖度(kurtosis)，カルバックライブラー (K-L) 情報量（Kullback (1959)参照）を用いる。

4.2.2節においては問題の数学的な定式化を行う。さらにいくつかの必須の数式を提示する。4.2.3節において非正規性の評価方法を示す。$Y-\gamma X$ の議論に加え，実用的な情報として，処置後値 Y と処置前後の変化量 $Y-X$ についても議論する。4.2.4節において結論と簡単な議論を行う。

4.2.2　問題の定式化

処置前値を X，処置後値を Y とし，その同時分布が2変量正規分布 $N(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\sigma_{XY})$ に従っていると仮定する。一般性の損失なしに以下では $(X,Y)\sim N(0,0,1,1,\rho)$ を取り扱う。ここで，統計量 $Y-\gamma X$ において $0\le\rho\le1$，$0\le\gamma\le1$ である。

(X,Y) が $a\le X\le b$ の範囲でのみ値が得られたとする。a と b は定数値である。すなわち処置前値でスクリーニングが行われたとする。言い換えると，(X,Y) は $a\le X\le b$ の範囲外でトランケートされたと表現する。以下ではトランケートされた確率変数を (X^*,Y^*) と示す。(X^*,Y^*) の確率密度関数とそのモーメントは以下の節で与える（以下に示す各式の証明は，Kotz, Balakrishnan and Johnson (2000)，岩崎 (2002b)などを参照）。

4.2.2.1　正規分布 X のトランケートされた分布

$X\sim N(0,1)$ と仮定し X が (a,b) の範囲でのみ観測されたとする。そのとき X^* の確率密度関数は以下のとおり与えられる。

$$f(x\mid a\le x\le b)=\frac{\varphi(x)}{\Phi(b)-\Phi(a)}$$

ここで，$\varphi(x)$ と $\Phi(x)$ はそれぞれ標準正規分布の確率密度関数と累積分布関数である。4次までの X^* のモーメントは以下のとおりである。

$$\mu_X^*=E[X\mid a\le X\le b]=\frac{\varphi(a)-\varphi(b)}{\Phi(b)-\Phi(a)}$$

$$\sigma_X^{*\,2}=V[X\mid a\le X\le b]=1+\frac{a\varphi(a)-b\varphi(b)}{\Phi(b)-\Phi(a)}-\left\{\frac{\varphi(a)-\varphi(b)}{\Phi(b)-\Phi(a)}\right\}^2 \tag{16}$$

$$\mu_{3X}^{*} = E[(X-\mu_X^{*})^3 \mid a \le X \le b] = \frac{(a^2-1)\varphi(a)-(b^2-1)\varphi(b)}{\Phi(b)-\Phi(a)}$$

$$-3\frac{\{\varphi(a)-\varphi(b)\}\{a\varphi(a)-b\varphi(b)\}}{\{\Phi(b)-\Phi(a)\}^2}+2\frac{\{\varphi(a)-\varphi(b)\}^3}{\{\Phi(b)-\Phi(a)\}^3}$$

$$\mu_{4X}^{*} = E[(X-\mu_X^{*})^4 \mid a \le X \le b] = \frac{a(a^2+3)\varphi(a)-b(b^2+3)\varphi(b)}{\Phi(b)-\Phi(a)}$$

$$-2\frac{\{\varphi(a)-\varphi(b)\}\{(2a^2+1)\varphi(a)-(2b^2+1)\varphi(b)\}}{\{\Phi(b)-\Phi(a)\}^2}$$

$$+6\frac{\{\varphi(a)-\varphi(b)\}^2\{a\varphi(a)-b\varphi(b)\}}{\{\Phi(b)-\Phi(a)\}^3}-3\frac{\{\varphi(a)-\varphi(b)\}^4}{\{\Phi(b)-\Phi(a)\}^4}+3 \tag{17}$$

4.2.2.2 $Y^{*}-\gamma X^{*}$の分布

一般化された統計量$T=Y-\gamma X$について検討する。ここでγは定数値で$0\le\gamma\le1$の範囲の値をとる。$\gamma=0$の場合は処置後値Yに対応し，$\gamma=1$の場合は処置前後の変化量$Y-X$に対応する。さらに$\gamma=\rho$の場合は回帰係数が既知の場合の共分散分析 (ANCOVA) に対応する。(X,T)の同時分布は$N(0,0,1,1+\gamma^2-2\rho\gamma,\rho-\gamma)$であるため$T\sim N(0,1+\gamma^2-2\rho\gamma)$である。$X$の$T=t$が与えられた場合の条件付分布は$N(\xi t,\tau^2)$である。ここで，$\xi=(\rho-\gamma)/(1+\gamma^2-2\rho\gamma)$，$\tau^2=1-(\rho-\gamma)^2/(1+\gamma^2-2\rho\gamma)$である。また，$X=x$が与えられた場合の$T$の条件付分布は$N\left\{(\rho-\gamma)x,1-\rho^2\right\}$である。これらの式より，$a\le x\le b$が与えられた場合の$T$の条件付確率密度関数は以下のように表される。

$$f(t)=\frac{1}{\Phi(b)-\Phi(a)}\frac{1}{\sqrt{2\pi(1+\gamma^2-2\rho\gamma)}}$$

$$\times\exp\left[-\frac{t^2}{2(1+\gamma^2-2\rho\gamma)}\right]\left\{\Phi\left(\frac{b-\xi t}{\tau}\right)-\Phi\left(\frac{a-\xi t}{\tau}\right)\right\}. \tag{18}$$

$\gamma=\rho$の場合，$T^{*}\sim N(0,1-\rho^2)$となる。$T^{*}=Y^{*}-\gamma X^{*}$の4次までのモーメントは以下のとおりである。

$$\mu_T^{*}=E[T]=(\rho-\gamma)\mu_X^{*}$$

$$\sigma_T^{*2}=V[T]=(\rho-\gamma)^2\sigma_X^{*2}+1-\rho^2 \tag{19}$$

$$\mu_{3T}^{*}=E[(T-\mu_T^{*})^3]=(\rho-\gamma)^3\mu_{3X}^{*} \tag{20}$$

$$\mu_{4T}^{*}=E[(T-\mu_T^{*})^4]=(\rho-\gamma)^4\mu_{4X}^{*}+6(\rho-\gamma)^2(1-\rho^2)\sigma_X^{*2}+3(1-\rho^2)^2 \tag{21}$$

ここで$\gamma=0$あるいは1とすると，それぞれY^{*}と$Y^{*}-X^{*}$のモーメントが上記式より容易に導出される。具体的に示すと，Y^{*}の確率密度関数は

$$g(y)=\frac{1}{\Phi(b)-\Phi(a)}\frac{1}{\sqrt{2\pi}}\exp\left[-y^2/2\right]\left\{\Phi\left(\frac{b-\rho y}{\sqrt{1-\rho^2}}\right)-\Phi\left(\frac{a-\rho y}{\sqrt{1-\rho^2}}\right)\right\}$$

となり，Y^*の4次までのモーメントは以下のとおりである。

$$\mu_Y^*=E[Y]=\rho\mu_X^*$$

$$\sigma_Y^{*2}=V[Y]=1-\rho^2(1-\sigma_X^{*2})$$

$$\mu_{3Y}^*=E[(Y-\mu_Y^*)^3]=\rho^3\mu_{3X}^*$$

$$\mu_{4Y}^*=E[(Y-\mu_Y^*)^4]=\rho^4\mu_{4X}^*+6\rho^2(1-\rho^2)\sigma_X^{*2}+3(1-\rho^2)^2$$

一方，$Z^*=Y^*-X^*$の確率密度関数は

$$h(z)=\frac{1}{\Phi(d)-\Phi(c)}\frac{1}{\sqrt{2\pi}\sqrt{2(1-\rho)}}$$

$$\times\exp\left[-\frac{z^2}{4(1-\rho)}\right]\left\{\Phi\left(\frac{2d+z}{\sqrt{2(1+\rho)}}\right)-\Phi\left(\frac{2c+z}{\sqrt{2(1+\rho)}}\right)\right\}$$

となり，$Z^*=Y^*-X^*$の4次までのモーメントは以下のとおりとなる。

$$\mu_Z^*=E[Z]=(\rho-1)\mu_X^*$$

$$\sigma_Z^{*2}=V[Z]=(\rho-1)^2\sigma_X^{*2}+1-\rho^2$$

$$\mu_{3Z}^*=E[(Z-\mu_Z^*)^3]=(\rho-1)^3\mu_{3X}^*$$

$$\mu_{4Z}^*=E[(Z-\mu_Z^*)^4]=(\rho-1)^4\mu_{4X}^*+6(\rho-1)^2(1-\rho^2)\sigma_X^{*2}+3(1-\rho^2)^2$$

4.2.2.3　正規性の評価指標

正規分布とそれに対応するトランケートされた分布の乖離の程度は，カルバックライブラー（K-L）情報量，標準化3次モーメント（歪度），標準化4次モーメント（尖度）を用いて評価する。一般的に二つの分布 $p(x)$ と $q(x)$ 間のカルバックライブラー情報量は以下の様に定義される。

$$D_{KL}(p,q)=\int_{-\infty}^{\infty}p(x)\log\frac{p(x)}{q(x)}dx$$

分布 $q(x)$ の歪度と尖度は以下のとおり定義される。

$$\beta_1(q)=\frac{E[(X-\mu)^3]}{\sigma^3}$$

$$\beta_2(q)=\frac{E[(X-\mu)^4]}{\sigma^4}-3$$

ここで，μ と σ^2 はそれぞれ $q(x)$ の平均値と分散を表す。$q(x)$ が正規分布の場合 β_1 と β_2 はともに0となる。

$f(t)$ を4.2.2.2章の(18)で与えられた $T^* = Y^* - \gamma X^2$ の確率密度関数とする。さらに $p(t)$ を同じ平均 μ_T^* と分散 μ_T^* を持つ正規分布とする。このとき分布間のカルバックライブラー情報量は以下のとおりである。

$$D_{KL}(p,f) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{f(x)} dx . \tag{22}$$

ここで，(22)を解析的に求めることは困難であるため，数値計算の手法を(22)の右辺の計算に用いる。本論文では台形法（Monahan (2001), p. 243等を参照）を $g(x)$ の区間 $[a,b]$ の積分において採用する。

$$T_n(g) = \frac{h}{2}\left[g(a) + 2\sum_{i=1}^{n-1} g(a+ih) + g(b) \right]$$

ここで $h = (b-a)/n$ である。計算の際には $[a,b] = [-5 - \mu_T^*, 5 - \mu_T^*]$，$n = 100{,}000$ とした（すなわち $h = 0.0001$ である）。最適化の方法として，黄金比を用いる方法 (Monahan (2001), p. 174)を歪度，尖度，カルバックライブラー情報量の極値（極大値，極小値）の計算に適用した。また，非線形の式を解くために二分法(Monahan (2001), p.175)を用いた。計算と例示をわかりやすくするため，3種類のトランケーションの方法に分けて以下では順次議論する。3種類の方法とは，対称トランケーション $(-a,a)$，片側トランケーション (a,∞)，一般化された両側トランケーション (a,b) である。以下では，それぞれを対称，片側，両側と表記する。

4.2.3　非正規性の評価

4.2.3.1　ガンマが変化したときの考察

非正規性の評価のために一般化された統計量 $T^* = Y^* - \gamma X^*$ を考える。$\gamma = 0$（すなわち Y^*)の場合，$\gamma = 1$（すなわち $Y^* - X^*$）の場合は，様々な実例への応用上特に重要あるので別途4.2.3.2節と4.2.3.3節に記載する。

(20)と(21)から，T^* の分布の正規分布から乖離の評価には対応する2変量正規分布の $|\rho - \gamma|$ の大きさが重要な役割を担っていることが分かる。ρ と γ はともに (0,1) の範囲の値をとるため，正規性からの乖離の大きさは (ρ,γ) が (1,0) あるいは (0,1) の場合にもっとも大きくなる。反対に $\gamma = \rho$ の場合には統計量 T^* は正規分布に正確に従う。これは回帰係数が既知の場合の共分散分析に対応する。

まず ρ と γ の値を固定し，トランケートされた分布の正規性から乖離の程度をカットオフ値の関数として評価する。$\rho = 0.7$，$\gamma = 0$, 0.25, 0.5, 0.75, 1の値をとり，カットオフ値が対称トランケーションの場合は–3から -0.1 の範囲，片側トランケーションの場合は -3 から3の範囲をとるとする。図 6と図 7にそのときの歪度（片側のみ），尖度，カルバックライブラー情報量を示す。$|\rho - \gamma|$ が大きくなるにつれて，評価指標の値は0から

離れていくことがわかる。

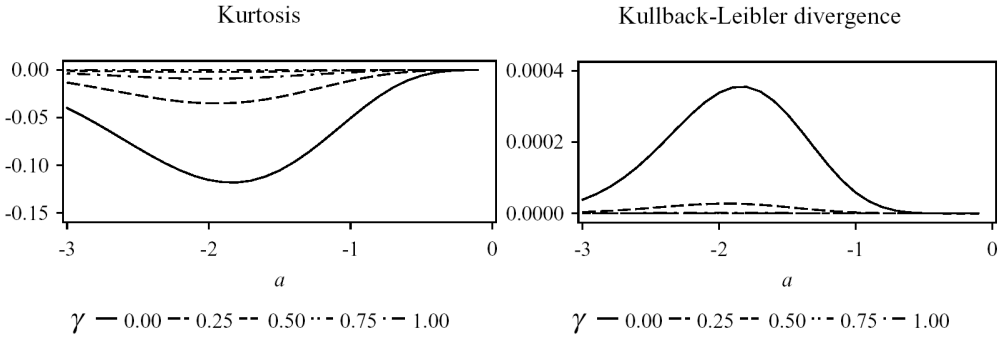
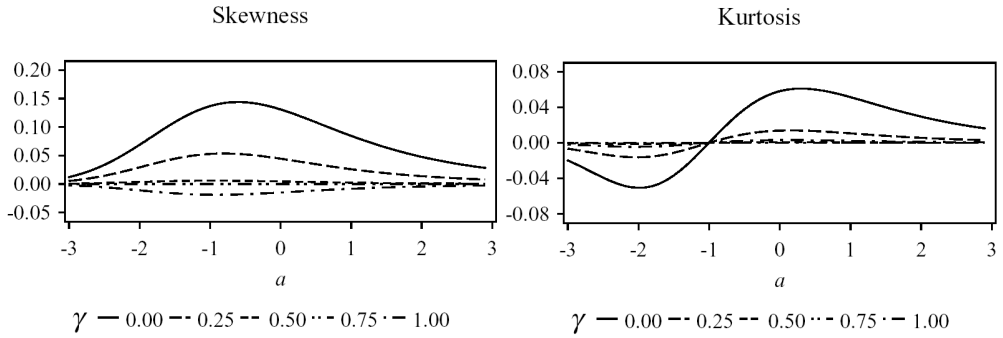


図 6　$Y^* - \gamma X^*$ の尖度，カルバックライブラー情報量（対称，$\rho = 0.7$）



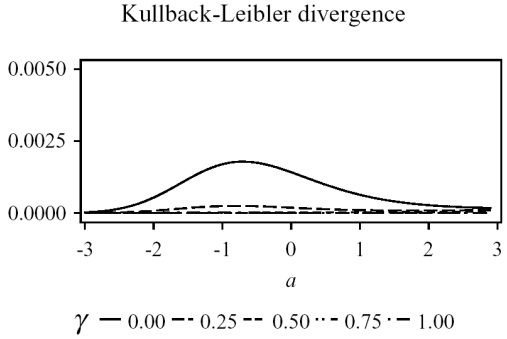


図 7　$Y^* - \gamma X^*$ の歪度，尖度，カルバックライブラー情報量（片側，$\rho = 0.7$）

より一般的である両側トランケーションの場合は，最悪の場合のみを示す。すなわち処置後値Y^*と変化量$Y^* - X^*$の場合についてのみを以下の節で示す。

図 7ではカットオフ値aがおおよそ–1の場合に，任意のγにおいて尖度が0となっている。この事実について以下に注釈を示す。

注釈1　(X,Y)が2変量正規分布$N(0,0,1,1,\rho)$に従うとする。ここでXは(a,∞)の範囲でのみ利用可能であるとする。aが$a^* \approx -1.0024$のとき任意のγとρにおいて，$Y^* - \gamma X^*$の尖度は0となる。

証明

a^* の正確な値を見つけるために，(19)と(21)から構成された以下の式 $\beta_2(T^*)=0$ を解く。

$$\beta_2(T^*)=\frac{\mu_{4T}^*}{\sigma_T^{*4}}-3$$

$$=\frac{(\rho-\gamma)^4\mu_{4X}^*+6(\rho-\gamma)^2(1-\rho^2)\sigma_X^{*2}+3(1-\rho^2)^2}{\{(\rho-\gamma)^2\sigma_X^{*2}+(1-\rho^2)\}^2}-3=0 \tag{23}$$

$\rho=\gamma$ の場合，T^* は上記のとおり正規分布に従う。従って $\beta_2=0$ である。$\rho\neq\gamma$ の場合，(23)は数式の展開後，γ と ρ の含まれない以下の式となる。

$$\mu_{4X}^*-3\sigma_X^{*4}=0 \tag{24}$$

ここで，(23)が γ と ρ に無関係であることを保証している。$b=\infty$ の場合のに(1)と(2)から(9)は以下のようになる。

$$\mu_{4X}^*-3\sigma_X^{*4}=\frac{a(a^2+3)\varphi(a)}{1-\Phi(a)}-2\frac{\varphi(a)(2a^2+1)\varphi(a)}{\{1-\Phi(a)\}^2}$$

$$+6\frac{\varphi(a)^2a\varphi(a)}{\{1-\Phi(a)\}^3}-3\frac{\varphi(a)^4}{\{1-\Phi(a)\}^4}+3-3\left[1+\frac{a\varphi(a)}{1-\Phi(a)}-\left\{\frac{\varphi(a)}{1-\Phi(a)}\right\}^2\right]^2$$

$$=\frac{a(a^2-3)\varphi(a)}{1-\Phi(a)}-\frac{(7a^2-4)\{\varphi(a)\}^2}{\{1-\Phi(a)\}^2}+12\frac{a\{\varphi(a)\}^3}{\{1-\Phi(a)\}^3}-6\frac{\{\varphi(a)\}^4}{\{1-\Phi(a)\}^4} \tag{25}$$

関数 $\varphi(a)/\{1-\Phi(a)\}$ は任意の a に対し0より大きい値をとる。$r(a)$ は(25)に $[\{1-\Phi(a)\}/\varphi(a)]^4$ を掛けて得られる $r(a)$ は以下のとおり表される。

$$r(a)=a(a^2-3)\frac{\{1-\Phi(a)\}^3}{\{\varphi(a)\}^3}-(7a^2-4)\frac{\{1-\Phi(a)\}^2}{\{\varphi(a)\}^2}+12a\frac{1-\Phi(a)}{\varphi(a)}-6 \tag{26}$$

$a\to-\infty$ の場合は $r(a)=-\infty$ となり，$a\to\infty$ の場合には $r(a)=0$ となることを示すことは容易である。図 8にて $r(a)$ の分布を示す。$r(a)$ の特徴を明らかにするために，$r(a)$ の一階微分，二階微分の値を列挙する。$r(a)=-1$ の場合，$r(-1)=0.08>0$，$r'(-1)=33.3>0$，$r''(-1)=-369<0$ である。また，$r(0)=0.283>0$，$r'(0)=-0.893<0$，$r''(0)=0.285>0$ であり，$r(1)=0.0146>0$，$r'(1)=-0.0398<0$，$r''(1)=0.116>0$，$r(2)=0.00124>0$，$r'(2)=-0.0277<0$，$r''(2)=0.00673>0$ となる。すなわち $r(a)$ は a が目的となる値 a_0 $(-1<a_0<0)$ より小さい場合は単調に増加し，a が -1 のときに正値をとる。a が a_0 より大きい場合，$r(a)$ は単調に減少し正値を保ったまま漸近的に0に達する。これは厳密な証明ではないが，$-\infty$ to ∞ の範囲で a はおおよそ -1 のときのみ $r(a)=0$ となる。(26)から直接0となる正確な a を解析的に求めることはほぼ不可能であることから，二分法を用いて数値計算を実施し $a^*\approx-1.0024$ を導出した。

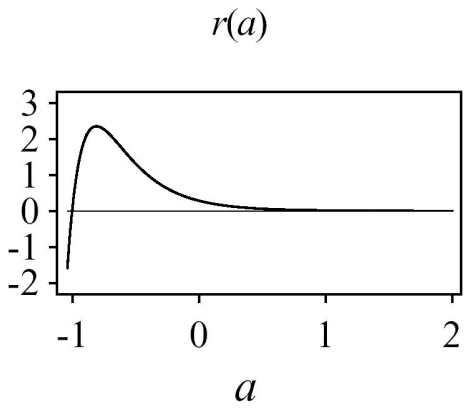


図 8　$r(a)$の分布

4.2.3.2　処置後値の評価

対称トランケートに対して，図 9にて4.2.2.3節で定義した尖度とカルバックライブラー情報量を，$\rho = 0.1, 0.3, 0.5, 0.7, 0.9$，カットオフ値$a$が–3から–0.1の範囲で示す。$\rho$が増加するに従い，尖度とカルバックライブラー情報量は0から離れていく。対称トランケーションであるので歪度は常に0である。表 3では最大の極値を取るカットオフ値，すなわち0からもっとも離れた値をとるカットオフ値を示す。それぞれの指標について，対称トランケーションの下，$\rho = 0.1, 0.3, 0.5, 0.7, 0.9$の値での結果を示す。極値を取る$\rho$と$a$においては，$\rho$が増加するにつれ$a$は増加していることがわかる。

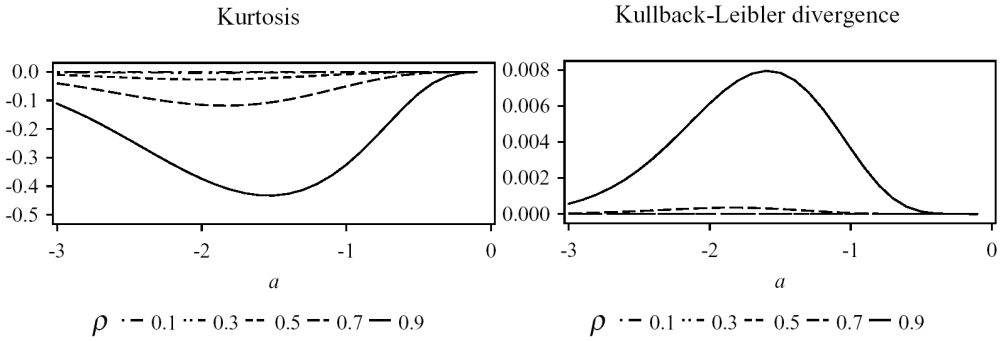


図 9　Y^*の尖度，カルバックライブラー情報量（対称）

表 3　Y^* の分布の正規性の指標の極値（対称）

Indicator	ρ	a	Extreme value*
Kurtosis	0.1	−2.0514	−0.00004
	0.3	−2.0240	−0.00321
	0.5	−1.9615	−0.02671
	0.7	−1.8383	−0.11809
	0.9	−1.5357	−0.43256
Kullback-Leibler divergence	0.1	All values	< 0.00001
	0.3	−2.0068	< 0.00001
	0.5	−1.9567	0.00002
	0.7	−1.8380	0.00036
	0.9	−1.5900	0.00796

*Extreme value は各指標がもっとも0から離れている値を示す。

片側トランケーションに対して，図 10にて $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$，カットオフ値 a が –3から3の範囲での歪度，尖度，カルバックライブラー情報量を示す。ρ が増加するにつれ，これらの指標は0から徐々に離れる。表 4では片側トランケーションにおける $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$のそれぞれの場合の各指標の極値を取るカットオフ値を示す。極値を取る ρ と a においては，ρ が増加するにつれ a は増加している。

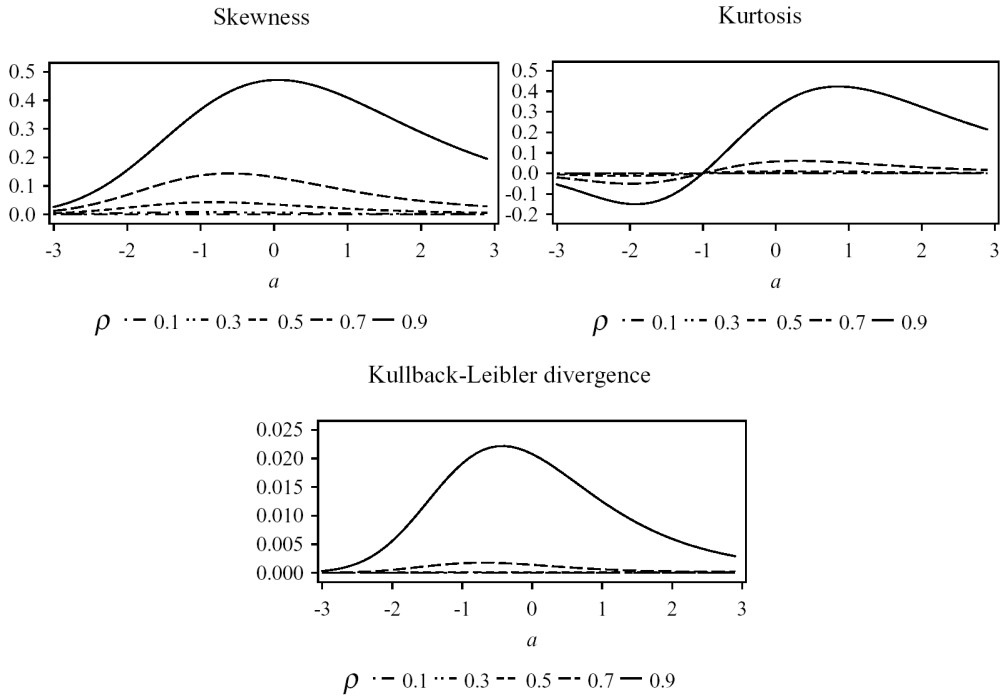


図 10　Y^* の歪度，尖度，カルバックライブラー情報量（片側）

表 4　Y^* の分布の正規性の指標の極値（片側）

Indicator	ρ	a	Extreme value*
Skewness	0.1	− 0.9967	0.00029
	0.3	− 0.9486	0.00841
	0.5	− 0.8361	0.04318
	0.7	− 0.6029	0.14406
	0.9	0.0538	0.47177
Kurtosis (Local maximum)	0.1	0.0333	0.00001
	0.3	0.0627	0.00105
	0.5	0.1353	0.01026
	0.7	0.3011	0.06111
	0.9	0.8538	0.42342
Kurtosis (Local minimum)	0.1	− 2.0698	− 0.00002
	0.3	− 2.0574	− 0.00156
	0.5	− 2.0313	− 0.01245
	0.7	−1.9881	− 0.05058
	0.9	−1.9209	− 0.15047
Kullback-Leibler divergence	0.1	All values $<$ 0.00001	
	0.3	All values $<$ 0.00001	
	0.5	− 0.8080	0.00016
	0.7	− 0.7022	0.00179
	0.9	− 0.4268	0.02219

* Extreme value は各指標がもっとも0から離れている値を示す。

両側トランケーションに対して，図 11にて $\rho = 0.9$，カットオフ値 a，b が –3 から3の範囲での歪度，尖度，カルバックライブラー情報量を示す。$a = –b$ の切断面は対称的なトランケーションを表す。また $b = 3$ の切断面（すなわちこの範囲での最大値）は本質的に片側トランケーションを表す。歪度とカルバックライブラー情報量は $b = 3$ の時に最大値を，尖度は $a = –b$ の時に極小値を，$b = 3$ の時に極大値をとることが容易にわかる。従って Y^* の分布においては，対称トランケーションと片側トランケーションの場合が正規性から乖離の程度が最大になるという意味で，最悪の場合を表しているということになる。

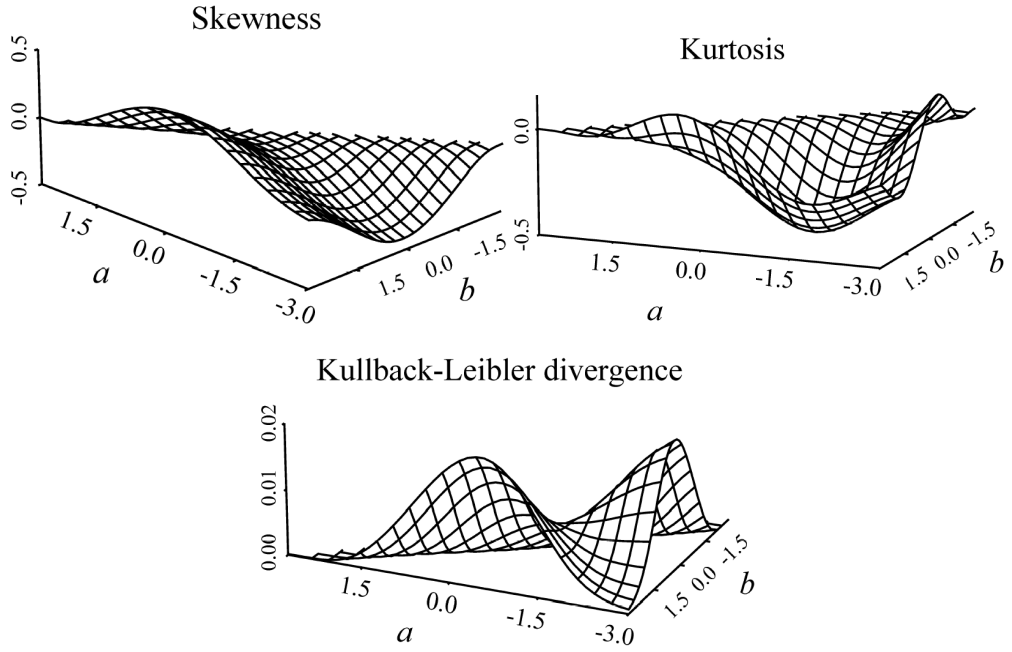


図 11　Y^* の歪度，尖度，カルバックライブラー情報量（両側）

4.2.3.3　変化量の評価

本節では変化量 $Y^* - X^*$ の分布について検討する。対称トランケーションと片側トランケーションにおいて，図 12と図 13にて $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$，カットオフ値 a，b がそれぞれと –3から3の範囲をとる場合の，歪度（片側のみ），尖度，カルバックライブラー情報量を示す。図 12と図 13での各指標の振る舞いは，図 9と図 10での振る舞いと同様である。ただし ρ が減少するにつれ，各指標は0から徐々に離れていく。

表 5は対称トランケーションと片側トランケーションにおける $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$ の場合の各指標の極値を与えるカットオフ値を示す。極大値を取る ρ と a においては，ρ が増加するにつれ a は増加している。表 5における極値は表 3と表 4のほとんどの条件と比較して小さな値をとっていることがわかる。ρ が中程度あるいはそれ以上の場合には特に当てはまる。これは $Y^* - X^*$ の非正規性の程度が，処置後値 Y^* の程度よりも小さくなることを示している。

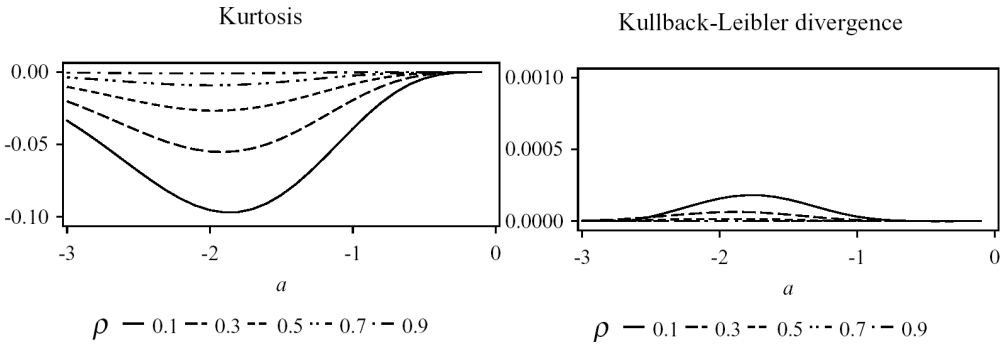


図 12　$Y^* - X^*$ の尖度，カルバックライブラー情報量（対称）

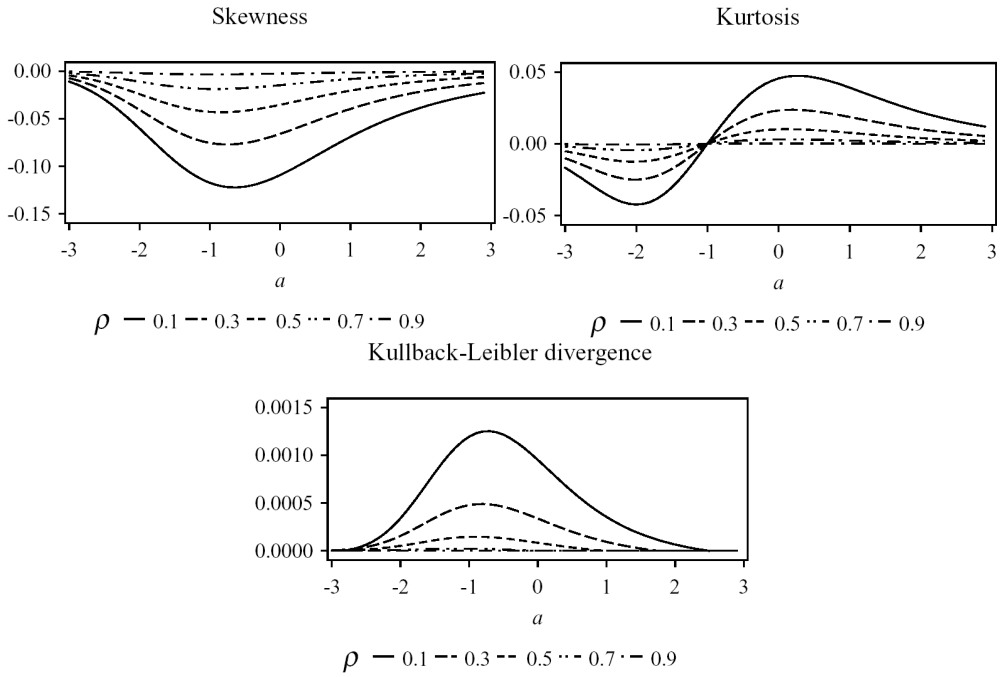


図 13　$Y^* - X^*$ の歪度，尖度，カルバックライブラー情報量（片側）

表 5　$Y^* - X^*$ の分布の正規性の指標の極値

Pattern	Indicator	ρ	a	Extreme value*
Symmetric	Kurtosis	0.1	−1.8623	−0.09701
		0.3	−1.9156	−0.05525
		0.5	−1.9615	−0.02671
		0.7	−2.0019	−0.00916
		0.9	−2.0380	−0.00097
	Kullback-	0.1	−1.7649	0.00018
	Leibler	0.3	−1.8803	0.00006
	divergence	0.5	−1.9567	0.00002
		0.7	−2.0011	< 0.00001
		0.9	−2.0352	< 0.00001
One-sided	Skewness	0.1	−0.6495	−0.12217
		0.3	−0.7510	−0.07704
		0.5	−0.8361	−0.04318
		0.7	−0.9092	−0.01878
		0.9	−0.9732	−0.00340
	Kurtosis	0.1	0.2664	0.04747
	(Local maximum)	0.3	0.1936	0.02378
		0.5	0.1353	0.01026
		0.7	0.0875	0.00317
		0.9	0.0475	0.00031
	Kurtosis	0.1	−1.9957	−0.04225
	(Local minimum)	0.3	−2.0140	−0.02496
		0.5	−2.0313	−0.01245
		0.7	−2.0479	−0.00439
		0.9	−2.0637	−0.00048
	Kullback-	0.1	−0.7299	0.00125
	Leibler	0.3	−0.8136	0.00049
	divergence	0.5	−0.9069	0.00015
		0.7	−1.1118	0.00002
		0.9	All values < 0.00001	

* Extreme value は各指標がもっとも0から離れている値を示す。

両側トランケーションにおいては，図 14で $\rho = 0.1$，カットオフ値 a，b がそれぞれ -3 から3の範囲における歪度，尖度，カルバックライブラー情報量を示す。その指標の振る舞いはほぼ図 11と同様である。

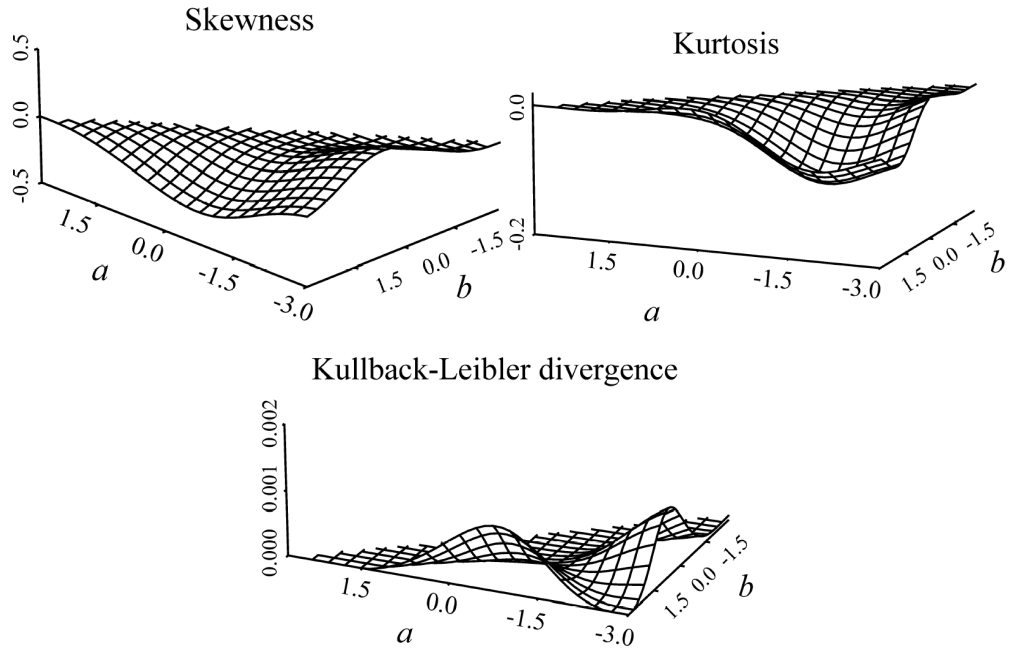


図 14　$Y^* - X^*$ の歪度，尖度，カルバックライブラー情報量（両側）

4.2.4　　議論

本節では処置前後値が2変量正規分布に従っていると仮定された場合に，ベースライン値のトランケーションがエンドポイントの分布の構成に正規分布からの乖離という点でどのように影響するかを明らかにした。歪度，尖度，カルバックライブラー情報量を非正規性の評価指標に用い，より一般化された統計量 $Y^* - \gamma X^*$ の正規分布からの隔たりの程度を示した。本統計量には Y^* と $Y^* - X^*$ が含まれている。その乖離の程度には相関 ρ が非常に1に近い場合でない限り，そこまで大きくはないことが示された。$\gamma = 0$ の場合，すなわち処置後値の場合，ρ が1に近づくにつれ最悪値を徐々に取る。さらに $\gamma = 1$ の場合，すなわち変化量の場合，0に近づくにつれ最悪値を取る。しかしながらその程度は処置後値と比較し緩徐である。ミニマックスの原理に基づくと，$Y^* - X^*$ の使用は Y^* と比べて，エンドポイントの正規性の観点からは適切である。この発見は臨床試験におけるエンドポイントの選択に役に立つであろう。さらに本章で示したとおり，正規分布からの乖離の程度は十分に小さいため，従来の t 検定または分散分析の使用は統計手法の頑健性の立場からも妥当であることがわかった。さらには，尖度は片側トランケーションにおいてカットオフ値 a がおおよそ –1 の場合，γ と ρ の値にかかわらず0となることがわかった。本章においては，確率変数 (X,Y) に対して2変量正規分布を仮定した。ほかの分布における処置前値のトランケーションの効果の評価に関する更なる研究が必要である。

4.3　不完全データに基づく平均への回帰を考慮したテストデータの解析

4.3.1　イントロダクション

教育現場では，学力テストで一定の点数以下（100点満点で30点以下など）の学生に対して補習を行い，学力の底上げを図る試みがしばしば行われる。このように点数が低い学生を集めた場合には，補習後の点数に対し平均への回帰が生じることが岩崎・河田 (2007) 等で指摘されているが，平均への回帰の影響を考慮したデータの評価は実際にはあまり行われていない。そのため，実際には補習の学習効果があまりない場合においても，「その担当教師は効果的な補習を実施した」と誤った評価を与えてしまっている可能性がある。

そこで，本論文では「正答・誤答」データを取り扱い，一定の問題数での正答数のようなカウントデータを問題とする。これは，「中間テスト，期末テスト」や「補習前，補習後」のように，同一学生に対して複数回試験が実施される処置前後研究の問題としてとらえることができる。処置前後研究のデータ解析については，正規分布の枠組みで示されることが多く，岩崎 (2002b) においても，主に正規分布に関する問題が取り扱われている。また，Bonate (2000) では処置前後研究について広範囲な内容がまとめられている。

補習へは一定の正答数以下の学生のみが参加するとしたとき，情報の程度（不完全性）により，以下の3種類の状況が想定される（2.3.1節を参照）。なお，補習後のテストは補習参加者のみが受ける状況を想定している。

選択（Selection）：補習の参加・不参加によらず全員分の補習前のテストの正答数は既知である。

打ち切り（Censoring）：補習に参加しなかった学生の正答数は分からないがその人数は分かる。

トランケーション（Truncation）：補習に参加しなかった学生の正答数も人数も分からない。

本論文では，テストの正答数の分布にベータ二項モデルを当てはめ，まずはパラメータが既知の場合に，補習前（処置前）にある正答数以下であった学生の補習後（処置後）の正答数の分布の期待値，分散を示す。岩崎・河田 (2007) ではベータ二項モデルにおいて，処置前後で問題数が同じ場合が述べられているが，本論文では処置前後で問題数が異なる可能性を考慮した議論に拡張する。また，岩崎・大道寺 (2009)ではゼロ過剰なベータ二項モデルについて取り扱っている。

さらに，上記の不完全性の程度により分類される状況ごとに，モーメント法に基づくモデルのパラメータ推定の方法を提案する。これにより，様々な状況で本モデルの当てはめを可能とする。また，補習参加者の正答率の分布が補習前後で変化するかを検定，

すなわち担当教師の教授方法の効果の有無を検討する場合に対し，平均への回帰を考慮した妥当な方法を適用し，良好な結果が得られたので報告する。

次の4.3.2節では，テストの正答数のモデルの定式化を行い，処置後のテストの正答数の平均値と分散の推定を行う。また，平均への回帰を考慮した検定の方法を提示する。4.3.3節では前述の不完全データの分類ごとにモデルのパラメータ推定方法を定式化する。4.3.4節では適用例を紹介し，最後の4.3.5節で簡単な議論を行う。

4.3.2　モデルの定式化

補習前のテスト（問題数 n）の正答数を処置前値 X，補習後のテスト（問題数 m）の正答数を処置後値 Y とする。正答率 θ を固定したとき，処置前値 X は二項分布 $Binom\,(n\,,\theta)$ に従うとする。これは各問題の正答率が一定とした場合に相当するが，各問題の難易度が多少ばらついても正答数の分布は二項分布でよく近似される（4.3.4節の例参照）。そして，二項確率 θ の個体間分布がベータ分布 $Beta\ (a\ ,\ b)$ に従うとする（$a > 0,\ b > 0$）。このとき，母集団全体での正答数 X はベータ二項分布 (beta-binomial distribution) に従い，その確率分布は

$$BB(x;n,a,b)=\binom{n}{x}\frac{(a)_x(b)_{n-x}}{(a+b)_n}(x=0,1,\cdots,n)$$

である。ここで $(a)_x$ は a の昇べきで，$(a)_x = a\,(a+1)\cdots(a+x-1)$ である。

$BB\,(x\,;n,a,b)$ の期待値と分散は

$$E_{BB}[X\mid n,a,b]=n\times\frac{a}{a+b} \tag{27}$$

$$V_{BB}[X\mid n,a,b]=n\times\frac{ab(a+b+n)}{(a+b)^2(a+b+1)} \tag{28}$$

となる（竹内・藤野 (1981) や Johnson, Kemp and Kotz (2005) を参照）。

以下，ベータ二項分布のパラメータ $a,\ b$ がすでに推定されているとする。データが不完全な場合の推定方法は次節で述べる。

X が正答数 d 以下でのみ観測されたとすると，その確率分布はトランケートされたベータ二項分布 (truncated beta-binomial distribution) となり，確率関数は以下のように表される。

$$TBB(x;n,a,b\mid x\le d)=\frac{BB(x;n,a,b)}{D(a,b)}$$

ここで $D(a,b)=\sum_{x=0}^{d}BB(x;n,a,b)$ である。X の期待値は

$$E_{TBB}\left[X\mid x\le d;n,a,b\right]=\frac{\sum_{x=0}^{\max(d-1,0)}BB(x;n-1,a+1,b)}{D(a,b)}\cdot\frac{na}{a+b} \tag{29}$$

となり，(30)より，X の分散は (31) のように表される。

$$E_{TBB}\left[X(X-1)\,|\,x\le d;n,a,b\right]=\frac{\sum_{x=0}^{\max(d-2,0)}BB(x;n-2,a+2,b)}{D(a,b)}\times\frac{n(n-1)a(a+1)}{(a+b)(a+b+1)} \tag{30}$$

$$\begin{aligned}&V_{TBB}[X\,|\,x\le d;n,a,b]\\&=E_{TBB}[X(X-1)\,|\,x\le d;n,a,b]\quad+E_{TBB}[X\,|\,x\le d;n,a,b]-\left(E_{TBB}[X\,|\,x\le d;n,a,b]\right)^2\end{aligned} \tag{31}$$

次に Y の分布について考える。本論文では検定の問題を主に扱うため，帰無仮説の下での Y の分布として，処置効果はない場合について示す。すなわち，補習前後のテスト間で学生の正答率に違いはないことが以下の議論での前提である。ただし，Y の二項確率は θ のままであるが，試行回数は m とする。3.5 節では $m=n$ の場合のみを扱っており，その拡張となっている。θ を与えた下で X と Y が独立であると仮定すると，X, Y の同時確率は以下のように表される。

$$BB_2(x,y;n,m,a,b)=\binom{n}{x}\binom{m}{y}\frac{(a)_{x+y}(b)_{(n-x)+(m-y)}}{(a+b)_{m+n}}$$

X が d 以下でのみ観測されたとすると，Y の周辺分布は

$$TBB(y;n,m,a,b\,|\,x\le d)=\frac{\sum_{x=0}^{d}\left\{BB(x;n,a,b)\cdot BB(y;m,a+x,b+n-x)\right\}}{D(a,b)}$$

となり，期待値は

$$E_{TBB}[Y\,|\,x\le d;n,m,a,b]=\frac{\sum_{x=0}^{d}\left\{\frac{m(a+x)}{a+b+n}\cdot BB(x;n,a,b)\right\}}{D(a,b)} \tag{32}$$

となる。また，(33) より，Y の分散は (34) のように表される。

$$E_{TBB}[Y(Y-1)\,|\,x\le d;n,m,a,b]=\frac{\sum_{x=0}^{d}\left\{\frac{m(a+x)}{a+b+n}\frac{(m-1)(a+x+1)}{a+b+n+1}BB(x;n,a,b)\right\}}{D(a,b)} \tag{33}$$

$$\begin{aligned}&V_{TBB}[Y\,|\,x\le d;n,m,a,b]\\&=E_{TBB}[Y(Y-1)\,|\,x\le d;n,m,a,b]+E_{TBB}[Y\,|\,x\le d;n,m,a,b]-\left(E_{TBB}[Y\,|\,x\le d;n,m,a,b]\right)^2\end{aligned} \tag{34}$$

なお，(32) より $E\,[Y\,|\,x=d\,;\,n,\,m,\,a,\,b]=m\,(a+d)\,/\,(a+b+n)$ であることから，$d<E\,[X\,|\,n,\,a,\,b]=na\,/\,(a+b)$ ならば，

$$\frac{d}{n}<\frac{(a+d)}{a+b+n}<\frac{a}{a+b}$$

である。$X=d$ の下での Y の期待正答率は，X の観測された正答率 $d\,/\,n$ と X の期待正答率 $a\,/\,a+b$ の中間に位置することから，平均への回帰現象を確認することができる。

続いて，X と Y の変化量について考える。X と Y で試行回数が異なる可能性を想定しているため，X と Y の正答率の変化量 $Z=Y\,/\,m-X\,/\,n$ を以下では取り扱う。正答率の変化量 Z の分布を $f(z)$ とすると，その分布は

$$f(z)=\frac{1}{D(a,b)}\sum_{x=0}^{d}\left[\left\{m\left(z+\frac{x}{n}\right)\right\}^{-1}BB_2\left\{x,m\left(z+\frac{x}{n}\right);n,m-1,a,b\right\}\right]$$

となる。このとき， Z の取り得る範囲は $-d/n \le z \le 1$ である。Z の期待値と分散は

$$E_{TBB}[Z \mid x\le d;n,m,a,b]=\frac{\sum_{x=0}^{d}\left[\frac{a+x}{a+b+n}\cdot BB(x;n,a,b)\right]}{D(a,b)}-\frac{\sum_{x=0}^{\max(d-1,0)}BB(x;n-1,a+1,b)}{D(a,b)}\cdot\frac{a}{a+b}$$

$$V_{TBB}[Z \mid x\le d;n,m,a,b]$$
$$=\frac{1}{m^2}V_{TBB}[Y \mid x\le d;n,m,a,b]+\frac{1}{n^2}V_{TBB}[X \mid x\le d;n,m,a,b]-\frac{2}{mn}Cov_{TBB}[X,Y \mid x\le d;n,m,a,b]$$

となる。XY の期待値は,

$$E_{TBB}(XY \mid x\le d;n,m,a,b)=\frac{\sum_{x=0}^{\max(d-1,0)}BB(x;n-1,a+2,b)}{D(a,b)}\times\frac{mna(a+1)}{(a+b)(a+b+1)} \qquad (35)$$

となることから，X と Y の共分散は (29), (32), (35) より,

$$Cov_{TBB}(X,Y \mid x\le d;n,m,a,b)$$
$$=E_{TBB}[XY \mid x\le d;n,m,a,b]-\left(E_{TBB}[X \mid x\le d;n,m,a,b]\times E_{TBB}[Y \mid x\le d;n,m,a,b]\right)$$

と表される。

実際の観測データと Y のモデル式から算出される期待観測値の当てはまりの度合いをチェックすることが考えられる。その際には，適合度の χ^2 値を求め，当てはまりの良さを比較すればよい。標本数がある程度以上の場合は，1標本 t 検定（母平均の検定）を用いて，処置効果の検定をすることができる。この際，帰無仮説として Y の分布の期待値が母平均であるとして処置後値の平均と比較，あるいは変化量 Z の分布の期待値が母平均であるとして正答率の変化量と比較することにより，平均への回帰を考慮した結果が得られる。期待値が極端に 0 または n に近い場合は当てはまりが悪くなることが予想されるが，一般的なテストであれば問題ない場合がほとんどである。

4.3.3　パラメータ推定

前節で述べたベータ二項モデルの当てはめでは，実際には処置前値 X の観測値からパラメータ a, b を推定しなくてはならない。X が d 以下でのみでしか観測されない状況には，4.3.1節で述べた3種類が考えられる。状況ごとに得られている情報が異なるため，それぞれ別の推定方法を以下に示す。本論文では，打ち切りとトランケーションの各々の場合のパラメータ推定方法を導出し，4.3.3.2節と4.3.3.3節に具体的に提示する。

ベータ二項分布 $BB(n, a, b)$ からの N 個の観測値 $x_1, \ldots, x_N$ がすべて得られた場合のパラメータ a, b の推定法は古くから議論されている。一般の推定問題では最尤法がパラメータの推定法として用いられることが多いが，$BB(n, a, b)$ での最尤法は導出法が難解（4.4.3節参照）であり，代わりにモーメント法が簡便かつ有力な推定法として推奨され

ている。モーメント法の最尤法に対する効率は多くのパラメータ値の範囲で遜色ないことから（たとえば Johnson, Kemp and Kotz (2005), p. 276），以下ではモーメント法による推定を採用する。

4.3.3.1　選択

選択の場合は，すべての処置前の観測値が得られているため，N 個の観測値の標本平均と標本分散を $\bar{x}$ および s^2 とするとき，モーメント法による推定値は (27) と (28) から導かれる以下の式

$$\tilde{a} = \frac{(N - \bar{x} - s^2 / \bar{x})\bar{x}}{(s^2 / \bar{x} + \bar{x} / N - 1)N} \tag{36}$$

$$\tilde{b} = \frac{(N - \bar{x} - s^2 / \bar{x})(N - \bar{x})}{(s^2 / \bar{x} + \bar{x} / N - 1)N} \tag{37}$$

により得られる (Johnson, Kemp and Kotz (2005))。

4.3.3.2　打ち切り

打ち切りの場合は，総観測数 N はわかっているが，データが d 以下でのみ観測され，M 個のみ得られるので，(30) と (31) より，以下の連立方程式が得られる。

$$\frac{\sum_{x=0}^{d}\left\{x \cdot BB(x;n,\tilde{a},\tilde{b})\right\}}{D} - \bar{x} = 0 \tag{38}$$

$$\frac{\sum_{x=0}^{d}\left\{x(x-1) \cdot BB(x;n,\tilde{a},\tilde{b})\right\}}{D} + \frac{\sum_{x=0}^{d}\left\{x \cdot BB(x;n,\tilde{a},\tilde{b})\right\}}{D} - \left[\frac{\sum_{x=0}^{d}\left\{x \cdot BB(x;n,\tilde{a},\tilde{b})\right\}}{D}\right]^2 - s^2 = 0 \tag{39}$$

ここで，$D = M / N$ である。M 個の観測値の標本平均と標本分散を $\bar{x}$ および s^2 とする。(38) を $f(a, b)$，(39) を $g(a, b)$ とする。初期値 a_0, b_0 を適当に置き（$a_0 = b_0 = 1$ など），ニュートン・ラフソン法により得られる以下の漸化式でパラメータを更新し，収束するまで繰り返し計算をする。

$$a_2 = a_1 - \frac{\frac{\partial g(a_1,b_1)}{\partial b_1} \cdot f(a_1,b_1) - \frac{\partial f(a_1,b_1)}{\partial b_1} \cdot g(a_1,b_1)}{\frac{\partial f(a_1,b_1)}{\partial a_1} \cdot \frac{\partial g(a_1,b_1)}{\partial b_1} - \frac{\partial f(a_1,b_1)}{\partial b_1} \cdot \frac{\partial g(a_1,b_1)}{\partial a_1}} \tag{40}$$

$$b_2 = b_1 + \frac{\frac{\partial g(a_1,b_1)}{\partial a_1} \cdot f(a_1,b_1) - \frac{\partial f(a_1,b_1)}{\partial a_1} \cdot g(a_1,b_1)}{\frac{\partial f(a_1,b_1)}{\partial a_1} \cdot \frac{\partial g(a_1,b_1)}{\partial b_1} - \frac{\partial f(a_1,b_1)}{\partial b_1} \cdot \frac{\partial g(a_1,b_1)}{\partial a_1}} \tag{41}$$

各偏微分の各要素は以下のとおりとなる。

$$\frac{\partial f(a,b)}{\partial a} = \frac{\sum_{x=0}^{d}\{x \cdot A(x,n,a,b) \cdot BB(x;n,a,b)\}}{D}$$

$$\frac{\partial f(a,b)}{\partial b} = \frac{\sum_{x=0}^{d}\{x \cdot B(x,n,a,b) \cdot BB(x;n,a,b)\}}{D}$$

$$\frac{\partial g(a,b)}{\partial a} = \frac{\sum_{x=0}^{d}\{x(x-1) \cdot A(x;n,a,b) \cdot BB(x;n,a,b)\}}{D} + \frac{1-2\sum_{x=0}^{d}\{x \cdot BB(x;n,a,b)\}}{D} \cdot \frac{\sum_{x=0}^{d}\{x \cdot A(x;n,a,b) \cdot BB(x;n,a,b)\}}{D}$$

$$\frac{\partial g(a,b)}{\partial b} = \frac{\sum_{x=0}^{d}\{x(x-1) \cdot B(x;n,a,b) \cdot BB(x;n,a,b)\}}{D} + \frac{1-2\sum_{x=0}^{d}\{x \cdot BB(x;n,a,b)\}}{D} \cdot \frac{\sum_{x=0}^{d}\{x \cdot B(x;n,a,b) \cdot BB(x;n,a,b)\}}{D}$$

ここで

$$A(x;n,a,b) = \sum_{i=0}^{x-1}\{1/(a+i)\} - \sum_{i=0}^{n-1}\{1/(a+b+i)\}$$

$$B(x;n,a,b) = \sum_{i=0}^{n-x-1}\{1/(b+i)\} - \sum_{i=0}^{n-1}\{1/(a+b+i)\}$$

である。

4.3.3.3 トランケーション

トランケーションの場合は，総観測数もわかっていないため，打ち切りでは既知であった D も a, b の関数であるので推定する必要がある。(38) と (39) と同様に，以下の連立方程式が得られる。

$$\frac{\sum_{x=0}^{d}\{x \cdot BB(x;n,\tilde{a},\tilde{b})\}}{D(a,b)} - \bar{x} = 0$$

$$\frac{\sum_{x=0}^{d}\{x(x-1) \cdot BB(x;n,\tilde{a},\tilde{b})\}}{D(a,b)} + \frac{\sum_{x=0}^{d}\{x \cdot BB(x;n,\tilde{a},\tilde{b})\}}{D(a,b)} - \left[\frac{\sum_{x=0}^{d}\{x \cdot BB(x;n,\tilde{a},\tilde{b})\}}{D(a,b)}\right]^2 - s^2 = 0$$

ニュートン・ラフソン法による漸化式 (40) と (41) を同様に適用すると，a, b の推定値が得られる。ただし，各偏微分の各要素は打ち切りの場合と比較しさらに複雑になり，以下のとおりとなる。

$$\frac{\partial f(a,b)}{\partial a} = \{1/D(a,b)\}^2 \cdot \left[\sum_{x=0}^{d}\{x \cdot A(x;n,a,b) \cdot BB(x;n,a,b)\} \cdot D(a,b) - \sum_{x=0}^{d}\{x \cdot BB(x;n,a,b)\} \cdot \sum_{x=0}^{d}\{A(x;n,a,b) \cdot BB(x;n,a,b)\}\right]$$

$$\frac{\partial f(a,b)}{\partial b} = \{1/D(a,b)\}^2 \cdot \left[\sum_{x=0}^{d} \{x \cdot B(x;n,a,b) \cdot BB(x;n,a,b)\} \cdot D(a,b) \right.$$
$$\left. - \sum_{x=0}^{d} \{x \cdot BB(x;n,a,b)\} \cdot \sum_{x=0}^{d} \{B(x;n,a,b) \cdot BB(x;n,a,b)\} \right]$$

$$\frac{\partial g(a,b)}{\partial a} = \{1/D(a,b)\}^2 \cdot \left[\sum_{x=0}^{d} \{x(x-1) \cdot A(x;n,a,b) \cdot BB(x;n,a,b)\} \cdot D(a,b) \right.$$
$$\left. - \sum_{x=0}^{d} \{x(x-1) \cdot BB(x;n,a,b)\} \cdot \sum_{x=0}^{d} \{A(x;n,a,b) \cdot BB(x;n,a,b)\} \right]$$
$$+ \left[\{1/D(a,b)\}^2 - 2\{1/D(a,b)\}^3 \cdot \sum_{x=0}^{d} \{x \cdot BB(x;n,a,b)\} \right]$$
$$\times \left[\sum_{x=0}^{d} \{x \cdot A(x;n,a,b)BB(x;n,a,b)\} D(a,b) - \sum_{x=0}^{d} \{x \cdot BB(x;n,a,b)\} \sum_{x=0}^{d} \{A(x;n,a,b)BB(x;n,a,b)\} \right]$$

$$\frac{\partial g(a,b)}{\partial a} = \{1/D(a,b)\}^2 \cdot \left[\sum_{x=0}^{d} \{x(x-1) \cdot B(x;n,a,b) \cdot BB(x;n,a,b)\} \cdot D(a,b) \right.$$
$$\left. - \sum_{x=0}^{d} \{x(x-1) \cdot BB(x;n,a,b)\} \cdot \sum_{x=0}^{d} \{B(x;n,a,b) \cdot BB(x;n,a,b)\} \right]$$
$$+ \left[\{1/D(a,b)\}^2 - 2\{1/D(a,b)\}^3 \cdot \sum_{x=0}^{d} \{x \cdot BB(x;n,a,b)\} \right]$$
$$\times \left[\sum_{x=0}^{d} \{x \cdot B(x;n,a,b)BB(x;n,a,b)\} D(a,b) - \sum_{x=0}^{d} \{x \cdot BB(x;n,a,b)\} \sum_{x=0}^{d} \{B(x;n,a,b)BB(x;n,a,b)\} \right]$$

4.3.4　適用例

157人の学生に対し実施した2回分のテストの正答数のデータ用い，このデータにベータ二項モデルを当てはめた結果を示す。本データは，中間時に問題数15問のテストを行い，全員が同じ授業を2カ月受講後，期末時に問題数18問のテストを行った。どちらの試験も正答数がテストの点数である。

図1に157人全員の補習前後での正答数の分布を示す。ここでは1回目のテストで9点以下であった73人がテスト後に教師の強い指導により発奮し，自発的な学力向上が2回目のテストの成果に表れるかを確認する状況を考える。特別な補習は行っていないが，状況をわかりやすくするため処置前後を補習前後と表す。

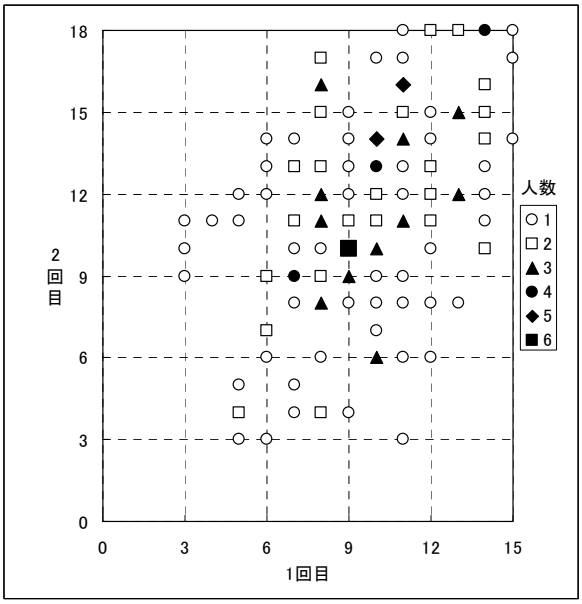


図 15　1回目と2回目のテスト正答数の分布

表 6に157人全員と補習前のテストで9点以下であった73人の補習前後の正答率並びに補習前後の正答率の変化量の標本平均と，標本分散を示す。学生全体での標本平均は補習前後で変化がほとんどないため，テスト全体での難易度は変わらないことが示唆される。個々の問題の正答率には若干のバラツキはあるものの，概ね50%から80%の範囲に含まれていた。全体での正答率は0.646であり，仮に正答数が二項分布 *Binom*(15, 0.646) としたときの分散は 15(0.646)(1 – 0.646) = 3.431 であり，問題ごとに正答率が異なるとしたときの分散は 3.315 となって両者はほぼ一致した。またシミュレーションにより二項分布と正答率が異なる場合の分布の比較をした結果両者はほぼ等しいことが確認された。表 6より点数が低い学生の補習前後の変化量が，全学生での変化量と比較し大きくなっている (– 0.44% → 7.11%) ことがわかり，平均への回帰現象が確認された。

表 6　補習前後の正答率（%）とその変化量の要約

項目	学生数	標本平均	標本分散
補習前	157	64.6	328.8
	73	48.7	106.8
補習後	157	64.2	429.9
	73	55.8	364.8
変化量	157	-0.44	403.6
	73	7.11	379.2

まず157人全員の補習前のデータが既知の場合，すなわち選択の状況を考える。(36) と (37) に対し，表 6に示した157人全員の補習前の標本平均と標本分散を適用し，ベータ二項分布のパラメータを推定する。そのとき各パラメータは $a = 7.17$, $b = 3.93$ と計算され

た。図 16，図 17，図 18に補習前後の正答数並びに正答率の変化量の期待観測値と，実際の観測値を示す。ただし $m \neq n$ であるため，正答率の変化量は取り得る値が非常に多くなる。以下では簡単のため，期待観測値並びに実際の観測値いずれも m 倍のスケールで四捨五入して表示した。

また，補習前テストで9点以下の学生の補習後の正答率並びに補習前後の正答率の変化量の期待値と分散を表 7に，Y の分布に (32) を適用した場合と，X と同じであると仮定した場合の1標本 t 検定の結果を表 8に示す。各モデルの当てはまりを確認するため，適合度の χ^2 値を算出し表 9に示す。その際，補習前の分布に対しベータ二項分布ではなく二項分布を当てはめた場合（$\bar{x} = n\theta$ から θ を推定した）の適合度も併せて表示した。

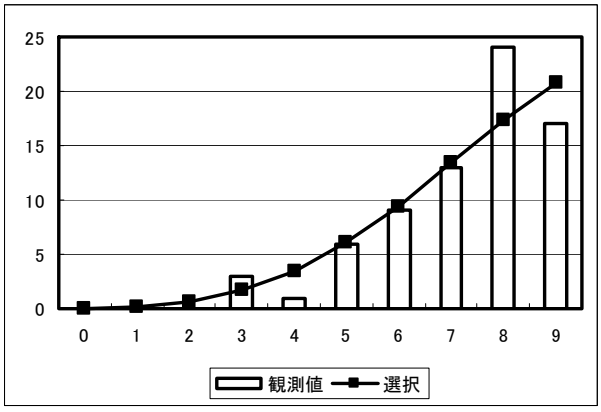


図 16　補習前テストで9点以下の学生の補習前正答数のパラメータ推定による期待観測値（折れ線）と実観測値（棒グラフ）

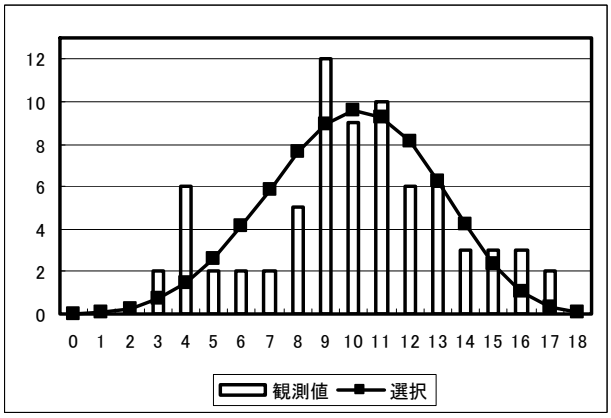


図 17　補習前テストで9点以下の学生の補習後正答数のパラメータ推定による期待観測値（折れ線）と実観測値（棒グラフ）

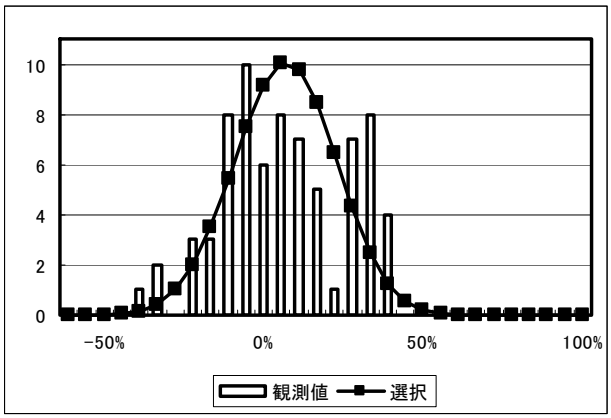


図 18　補習前テストで9点以下の学生の補習前後の正答率の変化量のパラメータ推定による期待観測値（折れ線）と実観測値（棒グラフ）

表 7　補習前テストで9点以下の学生の補習後の正答率（%）並びに補習前後の正答率の変化量の要約

項目	期待値	分散
補習後	55.1	262.7
変化量	7.03	243.2

表 8　補習前テストで9点以下の学生の補習後の正答数並びに補習前後の正答率（%）の変化量に対する1標本 t 検定

項目	母平均	P 値
補習後	補習前の標本平均	0.002
	補習後の期待値	0.771
変化量	0	0.003
	変化量の期待値	0.973

表 9　補習前テストで9点以下の学生の補習前後の正答数並びに正答率（%）の変化量の χ^2 適合度

分布	自由度	χ^2 値
補習前全体 (二項分布)	14	271.88
補習前全体 (ベータ二項分布)	13	16.53
補習前 (9点以下のみ)	7	6.86
補習後	16	35.15
変化量	122	179.83

表 9より157人全員の補習前テスト正答数の分布は，二項分布と比較しベータ二項分布

の当てはまりが非常によかった。また表 8より，補習後の分布に平均への回帰を考慮した検定を適用した場合（帰無仮説：補習後の母平均が各項目の期待値である）には，補習後，変化量のいずれの検定においても有意差が認められなかった。平均への回帰の影響を無視した場合（帰無仮説：補習後の母平均が補習前の標本平均と変わらない）は，本来このデータでは学習効果が得られないことが予測されるにもかかわらず有意差が認められた。すなわち，平均への回帰の影響を考慮しない場合には誤った結果が導かれる可能性が示唆された。

続いて，打ち切り並びにトランケーションの検討結果を示す。データの観測率 D に関しては，打ち切りでは $D = 73/157$ と既知であるが，トランケーションでは総学生数が未知であるため値が得られない。それぞれの場合のニュートン・ラフソン法によるベータ二項分布のパラメータ推定の結果を表 10に示す。また，推定されたパラメータを用いた補習前後の分布とその変化量の分布を図 19，図 20，図 21に示す。

表 10　打ち切りとトランケーションでのベータ二項分布のパラメータ推定の結果

欠測パターン	*a*	*b*
打ち切り	7.95	4.53
トランケーション	17.85	12.86

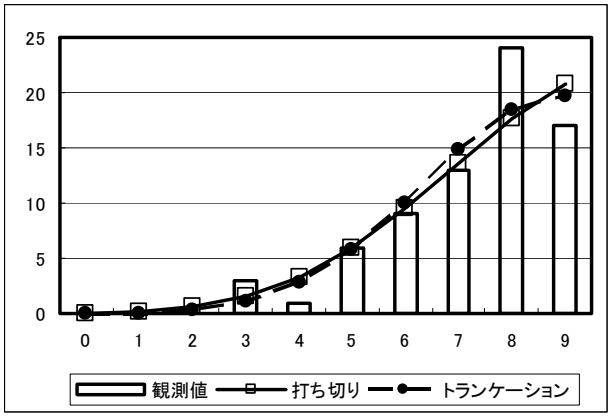


図 19　補習前テストで9点以下の学生の補習前正答数の打ち切りとトランケーションでのパラメータ推定による期待観測値（折れ線）と実観測値（棒グラフ）

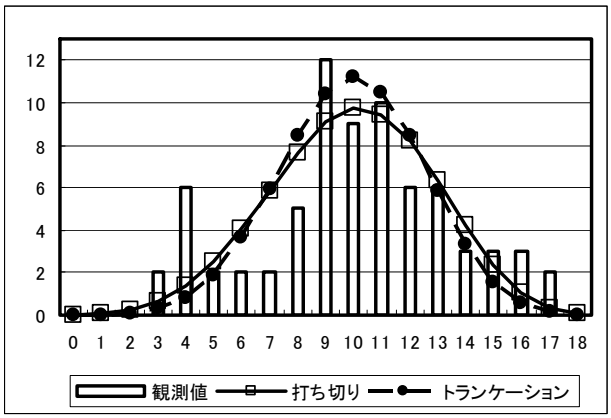


図 20　補習前テストで9点以下の学生の補習後正答数の打ち切りとトランケーションでの
パラメータ推定による期待観測値（折れ線）と実観測値（棒グラフ）

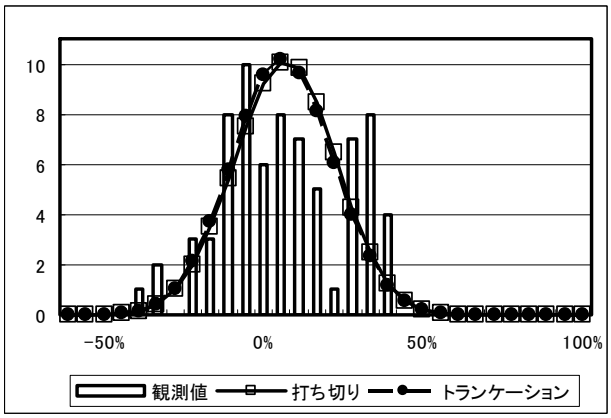


図 21　補習前テストで9点以下の学生の補習前後の正答率（%）変化量の
打ち切りとトランケーションでのパラメータ推定による
期待観測値（折れ線）と実観測値（棒グラフ）

どちらもパラメータの初期値を $a_0 = b_0 = 1$ としたが，10回以内の更新で小数点以下 6 桁の精度で収束した。特にトランケーションは，正規分布等のデータの取り得る範囲が広い分布では D の推定と分布のパラメータ推定は困難であるが，ベータ二項モデルではデータの得られる範囲が0点から15点と限定されることから，パラメータの推定がしやすいと考えられた。

今回のデータでは9点の観測人数が8点に比べて少なかったことが影響し，総学生数が未知であるトランケーションではパラメータ推定の際に D が過大評価されてしまい，選択，打ち切りと比較し補習後の分布並びに補習前後の変化量の分布が少しずれていることが確認できた。

打ち切りでは，すべてのデータが既知である選択と比較し，データの不完全性は増しているにもかかわらず遜色ない推定結果となっている。表 11に打ち切りとトランケーシ

ョンでの，補習後の分布に平均への回帰を考慮した1標本 t 検定の結果を示す。また，各モデルの当てはまりを確認するため，適合度の χ^2 値を算出し表 12に示す。

表 11　打ち切りとトランケーションでの補習前テストで9点以下の学生の補習後正答数並びに補習前後の正答率（%）変化量に対する1標本 t 検定

欠測パターン	項目	P 値
打ち切り	補習後	0.824
	変化量	0.968
トランケーション	補習後	0.736
	変化量	0.741

表 12　打ち切りとトランケーションでの補習前テストで9点以下の学生の補習前後の正答数並びに正答率（%）の変化量の χ^2 適合度

分布	自由度	打ち切り χ^2 値	トランケーション χ^2 値
補習前全体	13	19.13	156.95
補習前 (9点以下のみ)	7	6.77	7.46
補習後	16	38.85	88.38
変化量	122	186.70	262.29

表 8の選択の結果のうち各項目の期待値を母平均とした検定結果と比較し，相違は認められなかった。しかしながら，トランケーションではパラメータ推定が不安定であるため，真値から乖離する可能性が示唆された。表 9の選択での χ^2 適合度の結果と比較し，打ち切りでは同程度の適合度を示した。

トランケーションでは，補習前の分布の 9点以下の当てはまりは良好であるが，補習前全体での当てはまりは選択と比較し乖離が認められた。全体に対する 9点以下の学生の割合の推定が必要なことが，推定精度に影響している。補習後並びに補習前後の変化量の χ^2 適合度の結果も，選択，打ち切りと比較し乖離が認められた。今回は補習前の 9点以下のデータとしたが，得られるデータの割合が増加すればその推定精度は向上する。

4.3.5　議論

処置前後で問題数が異なることを想定したテストでの正答数の分布をベータ二項分布でモデル化した。そして処置前のテストでスクリーニングが実施されることを想定し，ある一定の正答数以下の学生の集団の処置効果がない場合の処置後の期待分布をモデル化し，平均への回帰の影響を考慮した検定方法を適用して良好な結果を得た。また，データの不完全性の状況を分類し，その状況ごとにベータ二項モデルのモーメント法によるパラメータ推定方法を示した。特に，打ち切りとトランケーションの場合には，モー

メント推定量を用いたニュートン・ラフソン法によるパラメータ推定方法を示した。

これらのモデルとパラメータ推定方法を，実際のデータに当てはめたところ，すべてのデータが得られている選択の場合と比べ，打ち切りの場合は，補習に参加しなかった学生の点数は未知であるにもかかわらず，遜色のない推定が可能であることを示した。また，トランケーションの場合にはパラメータ推定は不安定であるものの推定は可能であり，1標本 t 検定の結果は他の状況とさほど違いは認められなかった。

補習後並びに補習前後の変化量の分布をベータ二項分布を用いてモデル化し，その期待値と分散を算出することにより，平均への回帰の影響を考慮した検定方法を示した。平均への回帰の影響を考慮せず検定を行ってしまうと，真実の結果とは異なった検定結果が得られる可能性が示唆された。今回は1標本 t 検定を用いたが，テストデータでは期待値が極端な値を取ることは比較的少ないため，適用可能な場面は多いと考えられる。また，ベータ二項モデルでは，各問題の正答率が一定もしくはほぼ一定であることが前提であるが，それらがある程度異なる場合については今後の研究課題としたい。

現実の教育現場で補習が行われる場合は，選択と打ち切りの場合がほとんどであると考えられる。今回示したように，選択と打ち切りを比較すると，パラメータ推定値はほとんど変わらないという結果が得られた。よって多くの場合，本論文で述べた推定方法は適用可能であり，有益であると考えられる。

本論文では，ある一定の正答数 d 以下のデータが得られる状況を想定したが，ある一定の範囲のデータが得られる状況，すなわち $[c, d]$ $(0 \leq c \leq d \leq n)$ の範囲でデータが得られる場合への拡張は容易である。これまでに行われたデータ分析の中には，平均への回帰の影響が疑われるような場合がある可能性がある。その際には本論文で提案する方法を用いた再検討が必要な場合があるかもしれない。

4.4　QOL 質問票データの解析へのディリクレ多項モデルの適用

4.4.1　イントロダクション

新薬の臨床試験では4.2節で議論したとおり，処置前値にてスクリーニング検査が行われることがある。このような場合には処置後の測定値に対し平均の回帰が生じることがある。4.3節において，「正答・誤答」のようなカウントデータを問題としたが，本節では3.6節で導入したディリクレ多項モデルに従うと仮定したカウントデータが適用される問題を議論する。ディリクレ多項モデルに従うと考えられるデータとして，QOL 調査票データのような1問に付き多種類の回答が可能なデータが考えられる。現在，QOL 調査票についてスクリーニング検査が実施されることは稀であるが，一般的に QOL 調査票の検査結果と主たる目的である検査の変動はある程度の関連性が認められる。したがって QOL 調査票のスコアデータにおいても平均への回帰の問題が潜在的に含まれていると考えられる。

QOL 調査票では，設問ごとに0, 1, 2,…とスコア化し，その合計スコアを用いて臨床症

状を評価することがしばしば行われる。これまで臨床試験においては，このようなデータを単純に正規分布に従っていると仮定したパラメトリックな手法，あるいはノンパラメトリックな手法を用いることが行われてきた。一部，QOL 調査票である SF-36に対しベータ二項分布を適用した例（Arostegui, *et al*. (2007) 参照）も見受けられるが，そのような例はあまり多くない。また，ディリクレ多項モデルにおけるパラメータ推定方法は Chuang and Cox (1985) などで議論されている。

本節では，本来のデータの成り立ちを考慮し QOL データに対しディリクレ多項モデルを当てはめる。すなわち各個体の疾患活動性等のスコアに影響を与える因子が個体ごとに決まっていると考える。正確にはディリクレ多項モデルの線形結合スコアに対する分布に対する当てはまりを検討する。また，2.3.1節で示したようなある種のスクリーニングが生じうると考え，それぞれの問題において議論する。

4.4.2節では，QOL データのディリクレ多項モデルの定式化を行い，処置前後における平均と分散の推定方法を示す。4.4.3節では，スクリーニングの種類ごと，モデルの当てはめ方ごとにパラメータ推定方法を示す。4.4.4節では具体例を示す。4.4.5節では簡単な議論を行う。

4.4.2　モデルの定式化

QOL 調査票（設問数 n）の処置前値を $X_1,\ldots,X_n$，処置後値を $Y_1,\ldots,Y_n$ とする。3.6節では処置前後で問題数が異なる場合を取り扱ったが，本節では処置前後で問題数は同じであるとして以下の議論を進めることとする。全 n 問の問題の解答の分布がパラメータ $\theta_1,\cdots,\theta_k$ の多項分布に従い，この多項確率のパラメータの個体間分布がパラメータ $a_1,\cdots,a_k$（各 a_i は正値をとる）のディリクレ分布に従うとすると，母集団全体での $X=(x_1,\cdots,x_k)$ の確率分布はディリクレ多項分布（Dirichlet-multinomial distribution）に従い，その確率分布は

$$DM(X;n,a_1,\ldots,a_n)=\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i}}{\left(\sum_{i=1}^{k}a_i\right)_n}$$

である。ここで，$(a)_x$ は a の昇べきである。$DM(x_i;n,a_1,\ldots,a_n)$ の期待値と分散は

$$E(x_i \mid n, a_1, \ldots, a_n) = n \times \frac{a_i}{\sum_{j=1}^{k} a_j},$$

$$Var(x_i \mid n, a_1, \ldots, a_n) = n \times \frac{a_i \left\{ \left(\sum_{j=1}^{k} a_j \right) - a_i \right\} \left(n + \sum_{j=1}^{k} a_j \right)}{\left(\sum_{j=1}^{k} a_j \right)^2 \left(1 + \sum_{j=1}^{k} a_j \right)}$$

である。

さらに線形結合スコアの分布は k 個のそれぞれのとり得る値に対し s_i $(i = 1, \cdots, k)$ 点が与えられた場合，N が個体の総数，j $(j = 1, \cdots, N)$ の各個体のそれぞれの設問の回答を x_{ij} とすると，処置前値の線形結合スコアは $z_x = \sum_{i=1}^{k} s_i x_{ij}$ $(j = 1, \cdots, N)$ と表される。その確率分布は

$$f(z_x \mid n, a_1, \ldots, a_n) = \sum_{z_x = \sum_{i=1}^{k} s_i x_i} \left\{ \frac{n!}{\prod_{i=1}^{k} x_i! \left(\sum_{i=1}^{k} a_i \right)_n} \frac{\prod_{i=1}^{k} (a_i)_{x_i}}{} \right\}$$

と表される。期待値と分散は

$$E(z_x) = n \times \frac{\sum_{i=1}^{k} (s_i a_i)}{\sum_{j=1}^{k} a_j}$$

$$Var(z_x) = \frac{\sum_{i=1}^{k} \left[{s_i}^2 a_i \left\{ \left(\sum_{j=1}^{k} a_j \right) - a_i \right\} \right] \left(n + \sum_{j=1}^{k} a_j \right) - 2 \sum_{s<t} s_s s_t a_s a_t \left(n + \sum_{j=1}^{k} a_j \right)}{\left(\sum_{j=1}^{k} a_j \right)^2 \left(1 + \sum_{j=1}^{k} a_j \right)}$$

となる。なお，$k = 2$ でかつ $s_1 = 1, s_2 = 0$ の場合には，ベータ二項分布の議論に一致する。

$z_x \geq d$ でのみ観測されたとすると，その確率分布はトランケートされたディリクレ多項分布となり

$$f(z_x; n, a_1, \ldots, a_n \mid z_x \geq d) = \frac{1}{D} \times \sum_{z_x = \sum_{i=1}^{k} s_i x_i} \frac{n!}{\prod_{i=1}^{k} x_i! \left(\sum_{i=1}^{k} a_i \right)_n} \frac{\prod_{i=1}^{k} (a_i)_{x_i}}{}$$

となる。ここで， $D=\sum_{z_x\geq d}\left\{\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i}}{\left(\sum_{i=1}^{k}a_i\right)_n}\right\}$ である。期待値は

$$E(z_x\,|\,z_x\geq d)=\frac{1}{D}\sum_{z_x\geq d}\left[z_x\sum_{z_x=\sum_{j=1}^{k}s_jx_j}\left\{\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i}}{\left(\sum_{i=1}^{k}a_i\right)_n}\right\}\right]$$

となり，分散は

$$E(z_x(z_x-1)\,|\,z_x\geq d)=\frac{1}{D}\sum_{z_x\geq d}\left[z_x(z_x-1)\sum_{z_x=\sum_{j=1}^{k}s_jx_j}\left\{\frac{n!}{\prod_{i=1}^{k}x_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i}}{\left(\sum_{i=1}^{k}a_i\right)_n}\right\}\right]$$

を用い

$$E(z_x(z_x-1)\,|\,z_x\geq d)+E(z_x\,|\,z_x\geq d)-\left\{E(z_x\,|\,z_x\geq d)\right\}^2$$

と算出可能である。

$\theta_1,\cdots,\theta_k$ が与えられた下で $x_1,\cdots,x_k$ と $y_1,\cdots,y_k$ が独立であると仮定すると， z_x と z_y の同時確率は以下のように表される。

$$f(z_x,z_y\,|\,n,a_1,\ldots,a_n)=\sum_{\substack{z_x=\sum_{i=1}^{k}s_ix_i\\ z_y=\sum_{i=1}^{k}s_iy_i}}\left\{\frac{2(n!)}{\prod_{i=1}^{k}x_i!\prod_{i=1}^{k}y_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i+y_i}}{\left(\sum_{i=1}^{k}a_i\right)_{2n}}\right\}$$

$z_x\geq d$ でのみ観測されたとすると， z_y の周辺分布は

$$f(z_y\,|\,z_x\geq d)=\frac{1}{D}\sum_{\substack{z_y=\sum_{i=1}^{k}s_iy_i\\ z_x=\sum_{i=1}^{k}s_ix_i\geq d}}\frac{2(n!)}{\prod_{i=1}^{k}x_i!\prod_{i=1}^{k}y_i!}\frac{\prod_{i=1}^{k}(a_i)_{x_i+y_i}}{\left(\sum_{i=1}^{k}a_i\right)_{2n}}$$

となり，期待値は

$$E(z_y\,|\,z_x\geq d)=\sum\left\{z_y\times f(z_y\,|\,z_x\geq d)\right\}$$

となり，分散は

$$E(z_y(z_y-1)\,|\,z_x \geq d) = \sum \left\{ z_y(z_y-1) \times f(z_y\,|\,z_x \geq d) \right\}$$

を用い

$$Var(z_y\,|\,z_x \geq d) = E(z_y(z_y-1)\,|\,z_x \geq d) + E(z_y\,|\,z_x \geq d) - \left\{ E(z_y\,|\,z_x \geq d) \right\}^2$$

と算出できる。

実際の観測データと z_y のモデル式から算出される期待観測値の当てはまりの度合いをチェックすることが考えられる。その際には，適合度の χ^2 値を求め，当てはまりの良さを比較すればよい。標本数がある程度以上の場合は，1標本 t 検定（母平均の検定）等を用いて，処置効果の検定に持ち込むことが可能である。

4.4.3　パラメータ推定

4.3.3節で述べたとおりモデルの当てはめでは，実際には処置前の観測値からパラメータ a, b を推定しなくてはならない。z_x が d 以上でのみでしか観測されない状況には，2.3.1節で述べた3種類が考えられる。状況ごとに得られている情報が異なるため，それぞれ別々の推定方法を以下に示す。本節では選択，打ち切り，トランケーションの各々の場合のパラメータ推定方法を導出し，4.4.3.1節，4.4.3.2節，4.4.3.3節に具体的に提示する。4.3節ではモーメント法による推定方法を示したが，ディリクレ多項分布でのモーメント法によるパラメータ推定方法と最尤法による方法のいずれも示すこととする。ベータ二項分布はディリクレ多項分布の特別な場合と位置づけられるので，最尤法による推定方法はそのまま適用可能である。

4.4.3.1　選択

選択の場合はすべての処置前値が得られているため，N 個の観測値の k 個のとり得る値の標本平均と標本分散を $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ ，$Var_1, Var_2, \ldots, Var_k$ とし，標本共分散を $Cov_{12}, Cov_{13}, \ldots, Cov_{(k-1)k}$ とするとき，モーメント法による推定値は

$$a_i = \frac{\bar{x}_i}{N} A \quad (i = 1, \ldots, k)$$

で与えられる。ここで

$$A = -\frac{\dfrac{\sum_{i=1}^{k}\left(s_i^2 Var_i\right) + 2\sum_{s<t} s_s s_t Cov_{st}}{\sum_{i=1}^{k}\left(s_i^2 n \dfrac{\bar{x}_i}{n}\left\{1 - \dfrac{\bar{x}_i}{n}\right\}\right) + 2\sum_{s<t}\left(-s_s s_t n \dfrac{\bar{x}_s}{n}\dfrac{\bar{x}_t}{n}\right)} - n}{\dfrac{\sum_{i=1}^{k}\left(s_i^2 Var_i\right) + 2\sum_{s<t} s_s s_t Cov_{st}}{\sum_{i=1}^{k}\left(s_i^2 n \dfrac{\bar{x}_i}{n}\left\{1 - \dfrac{\bar{x}_i}{n}\right\}\right) + 2\sum_{s<t}\left(-s_s s_t n \dfrac{\bar{x}_s}{n}\dfrac{\bar{x}_t}{n}\right)} - 1}$$

である。ここで $s_i\ (i=1,\cdots,k)$ がすべて同じとなる場合は定義不可能である。なお，4.3.3.1節に示したベータ二項分布でのモーメント法によるパラメータ方法も本式に帰結する。

以下に最尤法によるパラメータ推定方法を示す。最尤法によるパラメータ推定方法には以下の二種類の方法考え方がある。

1) 確率変数として各測定値 x_{ij} を別々に考え，各 x_i を同時に推定しディリクレ多項分布の尤度が最大となるパラメータを求める。
2) 確率変数として各線形結合スコア z_x のみ考え， z_x の分布の尤度が最大となるパラメータを求める。

以下，それぞれ「ディリクレ多項分布を仮定した方法」,「線形結合スコア分布を仮定した方法」と表現する。

ディリクレ多項分布に従うと仮定した場合の対数尤度関数は

$$\log f=\sum_{\mathbf{x}}\left[f_{\mathbf{x}}\log\left\{\prod_{i=1}^{k}\frac{\Gamma(a_i+x_i)}{\Gamma(1+x_i)\Gamma(a_i)}\right\}\right]+N\left[\log\frac{n!\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\Gamma\left(n+\sum_{i=1}^{k}a_i\right)}\right]$$

となり，線形結合スコア分布に従うと仮定した場合の対数尤度関数は

$$\log f=\sum_{z_x}\left[f_{z_x}\log\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left\{\prod_{i=1}^{k}\frac{\Gamma(a_i+x_i)}{\Gamma(1+x_i)\Gamma(a_i)}\right\}\right]+N\left[\log\frac{n!\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\Gamma\left(n+\sum_{i=1}^{k}a_i\right)}\right]$$

と表される。ここで $f_{\mathbf{x}}$ は各 $\mathbf{x}=\begin{pmatrix}x_1 & x_2 & \cdots & x_k\end{pmatrix}^T$ の発現頻度， f_{z_x} は各 z_x の発現頻度を表す。初期値 $a_{1(0)},\ldots,a_{k(0)}$ を適当に置き（モーメント法による推定値を用いるなど），ニュートン･ラフソン法により得られる以下の漸化式

$$\begin{bmatrix}a_{1(2)}\\a_{2(2)}\\\vdots\\a_{k(2)}\end{bmatrix}=\begin{bmatrix}a_{1(1)}\\a_{2(1)}\\\vdots\\a_{k(1)}\end{bmatrix}-\mathbf{H}^{-1}\mathbf{G} \tag{42}$$

によりパラメータを更新し，収束するまで繰り返し計算する。ここで $\mathbf{G}$ は勾配ベクトルであり $\mathbf{H}$ はヘッセ行列である。

前者は

$$\mathbf{G}=\sum_{\mathbf{x}}\{f_{\mathbf{x}}\boldsymbol{\Psi}_{\mathbf{k}}\}-N\Psi_n\times\mathbf{1}_{\mathbf{k}\times\mathbf{1}}$$

$$\mathbf{H}=\sum_{\mathbf{x}}\left[f_{\mathbf{x}}\boldsymbol{\psi}_{\mathbf{k}}\right]\#\mathbf{I}_{\mathbf{k}}-N\psi_n\times\mathbf{1}_{\mathbf{k}\times\mathbf{k}}$$

となり，後者は

$$\mathbf{G}=\sum_{z_x}\left\{\frac{f_{z_x}}{\sum_{z_x=\sum_{i=1}^{k}s_ix_i}A_{x(1)}}\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}\right)\right\}-N\Psi_n\times\mathbf{1}_{\mathbf{k}\times\mathbf{1}}$$

$$\mathbf{H}=\sum_{z_x}\left[f_{z_x}\left\{\frac{\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left\{A_{x(1)}\left(\boldsymbol{\psi}_{\mathbf{k}}\#\mathbf{I}_{\mathbf{k}}+\boldsymbol{\Psi}_{\mathbf{k}}\boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathrm{T}}\right)\right\}}{\sum_{z_x=\sum_{i=1}^{k}s_ix_i}A_{x(1)}}-\frac{\left[\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}\right)\right]\left[\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathrm{T}}\right)\right]}{\left(\sum_{z_x=\sum_{i=1}^{k}s_ix_i}A_{x(1)}\right)^2}\right\}\right]-N\psi_n\times\mathbf{1}_{\mathbf{k}\times\mathbf{k}}$$

となる。ここで，#は行列の要素ごとの乗算を表し，$A_{x(1)}=\prod_{i=1}^{k}\frac{\Gamma(a_i+x_i)}{\Gamma(1+x_i)\Gamma(a_i)}$，

$$\boldsymbol{\psi}_{\mathbf{k}}=\begin{bmatrix}\psi(a_1+x_1)-\psi(a_1)\\ \psi(a_2+x_2)-\psi(a_2)\\ \vdots\\ \psi(a_k+x_k)-\psi(a_k)\end{bmatrix},\boldsymbol{\Psi}_{\mathbf{k}}=\begin{bmatrix}\Psi(a_1+x_1)-\Psi(a_1)\\ \Psi(a_2+x_2)-\Psi(a_2)\\ \vdots\\ \Psi(a_k+x_k)-\Psi(a_k)\end{bmatrix},\quad \Psi_n=\Psi\left(n+\sum_{i=1}^{k}a_i\right)-\Psi\left(\sum_{i=1}^{k}a_i\right),$$

$\psi_n=\psi\left(n+\sum_{i=1}^{k}a_i\right)-\psi\left(\sum_{i=1}^{k}a_i\right)$である。 Ψ と ψ はそれぞれディガンマ関数，トリガンマ関数である。 x が整数の場合は，

$$\Psi(a+x)-\Psi(a)=\sum_{i=0}^{x-1}\frac{1}{a+i},\quad \psi(a+x)-\psi(a)=-\sum_{i=0}^{x-1}\frac{1}{(a+i)^2}$$

と表現可能である。ニュートン・ラフソン法の適用の際には，各個体の値すべてがわかる必要はなく，それぞれ $f_{\mathbf{x}}$ と f_{z_x} がわかるだけで適用可能であることが容易にわかる。

4.4.3.2　打ち切り

打ち切りの場合は，総観測数 N はわかっているがデータは d 以上でのみ N_{obs} 個観測され， N_{mis} 個の値が d 未満であることでのみがわかっている。

ディリクレ多項分布に従うと仮定した場合の対数尤度関数は

$$\log f = \sum_{\mathbf{x}} \left[f_{\mathbf{x}} \log \left\{ \prod_{i=1}^{k} \frac{\Gamma(a_i + x_i)}{\Gamma(1+x_i)\Gamma(a_i)} \right\} \right]$$
$$+ \left(N_{obs} + N_{mis}\right) \left[\log \frac{n!\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\Gamma\left(n + \sum_{i=1}^{k} a_i\right)} \right] + N_{mis} \left[\log \sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left\{ \prod_{i=1}^{k} \frac{\Gamma(a_i + x_i)}{\Gamma(1+x_i)\Gamma(a_i)} \right\} \right]$$

となり，線形結合スコア分布に従うと仮定した場合の対数尤度関数は

$$\log f = \sum_{z_x} \left[f_{z_x} \log \sum_{z_x = \sum_{i=1}^{k} s_i x_i} \left\{ \prod_{i=1}^{k} \frac{\Gamma(a_i + x_i)}{\Gamma(1+x_i)\Gamma(a_i)} \right\} \right]$$
$$+ \left(N_{obs} + N_{mis}\right) \left[\log \frac{n!\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\Gamma\left(n + \sum_{i=1}^{k} a_i\right)} \right] + N_{mis} \left[\log \sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left\{ \prod_{i=1}^{k} \frac{\Gamma(a_i + x_i)}{\Gamma(1+x_i)\Gamma(a_i)} \right\} \right]$$

と表される。選択の場合と同様にニュートン･ラフソン法の漸化式 (42) によりパラメータを更新し，収束するまで繰り返し計算する。前者は

$$\mathbf{G} = \sum_{\mathbf{x}} \{ f_{\mathbf{x}} \boldsymbol{\Psi}_{\mathbf{k}} \} + \frac{N_{mis}}{\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} A_{x(1)}} \sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left(A_{x(1)} \boldsymbol{\Psi}_{\mathbf{k}} \right) - \left(N_{obs} + N_{mis}\right) \Psi_n \times \mathbf{1}_{\mathbf{k} \times \mathbf{1}}$$

$$\mathbf{H} = \sum_{\mathbf{x}} \left[f_{\mathbf{x}} \boldsymbol{\psi}_{\mathbf{k}} \right] \# \mathbf{I}_{\mathbf{k}}$$
$$+ N_{mis} \left\{ \frac{\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left\{ A_{x(1)} \left(\boldsymbol{\psi}_{\mathbf{k}} \# \mathbf{I}_{\mathbf{k}} + \boldsymbol{\Psi}_{\mathbf{k}} \boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathbf{T}} \right) \right\}}{\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} A_{x(1)}} - \frac{\left[\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left(A_{x(1)} \boldsymbol{\Psi}_{\mathbf{k}} \right) \right] \left[\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left(A_{x(1)} \boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathbf{T}} \right) \right]}{\left(\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} A_{x(1)} \right)^2} \right\}$$
$$- \left(N_{obs} + N_{mis}\right) \psi_n \times \mathbf{1}_{\mathbf{k} \times \mathbf{k}}$$

となり，後者は

$$\mathbf{G} = \sum_{z_x} \left\{ \frac{f_{z_x}}{\sum_{z_x = \sum_{i=1}^{k} s_i x_i} A_{x(1)}} \sum_{z_x = \sum_{i=1}^{k} s_i x_i} \left(A_{x(1)} \boldsymbol{\Psi}_{\mathbf{k}} \right) \right\} + \frac{N_{mis}}{\sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} A_{x(1)}} \sum_{z_x = \sum_{i=1}^{k} s_i x_i < d} \left(A_{x(1)} \boldsymbol{\Psi}_{\mathbf{k}} \right) - \left(N_{obs} + N_{mis}\right) \Psi_n \times \mathbf{1}_{\mathbf{k} \times \mathbf{1}}$$

$$\mathbf{H}=\sum_{z_x}\left[f_{z_x}\left\{\frac{\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left\{A_{x(1)}\left(\boldsymbol{\psi}_{\mathbf{k}}\#\mathbf{I}_{\mathbf{k}}+\boldsymbol{\Psi}_{\mathbf{k}}\boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathbf{T}}\right)\right\}}{\sum_{z_x=\sum_{i=1}^{k}s_ix_i}A_{x(1)}}-\frac{\left[\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}\right)\right]\left[\sum_{z_x=\sum_{i=1}^{k}s_ix_i}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathbf{T}}\right)\right]}{\left(\sum_{z_x=\sum_{i=1}^{k}s_ix_i}A_{x(1)}\right)^2}\right\}\right]$$

$$+N_{mis}\left\{\frac{\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}\left\{A_{x(1)}\left(\boldsymbol{\psi}_{\mathbf{k}}\#\mathbf{I}_{\mathbf{k}}+\boldsymbol{\Psi}_{\mathbf{k}}\boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathbf{T}}\right)\right\}}{\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}A_{x(1)}}-\frac{\left[\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}\right)\right]\left[\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}\left(A_{x(1)}\boldsymbol{\Psi}_{\mathbf{k}}{}^{\mathbf{T}}\right)\right]}{\left(\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}A_{x(1)}\right)^2}\right\}$$

$$-\left(N_{obs}+N_{mis}\right)\psi_n\times\mathbf{1}_{\mathbf{k}\times\mathbf{k}}$$

となる。ここで $N_{obs}=N$， $N_{mis}=0$ とすると，選択の場合に対応する。

4.4.3.3　トランケーション

トランケーションの場合は，データは d 以上の観測数が N_{obs} であることのみがわかっており， d 未満のデータの個数 N_{mis} は不明である。この場合， N_{mis} も含めてパラメータ推定することも可能であるが，非常に複雑な式となる。そこでディリクレ多項分布に従うと仮定した場合と線形結合スコア分布に従うと仮定した場合いずれの場合においても，打ち切りの場合のニュートン･ラフソン法の漸化式 (42) にてパラメータを更新の際に，その都度 N_{mis} を以下の式を用いて推定することにより，若干収束は遅くはなるものの対処可能である。

$$N_{mis}=N_{obs}\times\frac{\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}\left\{\frac{n!\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\Gamma\left(n+\sum_{i=1}^{k}a_i\right)}\prod_{i=1}^{k}\frac{\Gamma(a_i+x_i)}{\Gamma(1+x_i)\Gamma(a_i)}\right\}}{1-\sum_{z_x=\sum_{i=1}^{k}s_ix_i<d}\left\{\frac{n!\Gamma\left(\sum_{i=1}^{k}a_i\right)}{\Gamma\left(n+\sum_{i=1}^{k}a_i\right)}\prod_{i=1}^{k}\frac{\Gamma(a_i+x_i)}{\Gamma(1+x_i)\Gamma(a_i)}\right\}}=\frac{N_{obs}\sum_{z_x<d}f(z_x)}{1-\sum_{z_x<d}f(z_x)}$$

4.4.4　適用例

リウマチの臨床試験において，身体障害度の評価指標として HAQ（modified health assessment questionnaire，Fries, *et al*. (1980) を参照）あるいは MHAQ（modified health assessment questionnaire，Pincus, *et al*. (1983) や Matsuda, *et al*. (2003) を参照）といった QOL 調査票がしばしば用いられる。本節では Nishimoto, Ito and Takagi (2010) で用いられているデータのうち，投与前と投与12週後の MHAQ 評価が欠測なしに得られている583名の患者データ（

表 13と表 14を参照）を用い，このデータに対しディリクレ多項モデルを当てはめた結果を示す。本データは表 15の構成となっており，それぞれの項目で「難なくできる (0点)」，「少し難しい (1点)」，「かなり難しい (2点)」，「できない (3点)」を取り得る。すなわち合計スコアは0点から24点まで取りうる。なお

表 13と表 14は，設問に関係なく各得点（0点，1点，2点，3点）の分布ごとに患者の頻度を集計したものである。

表 13　投与前の MHAQ の分布

投与前											
0点	1点	2点	3点	スコア	頻度	0点	1点	2点	3点	スコア	頻度
0	0	3	5	21	3	1	2	4	1	13	1
0	0	4	4	20	3	1	3	4	0	11	3
0	0	5	3	19	3	1	4	3	0	10	7
0	0	6	2	18	1	1	5	1	1	10	1
0	0	7	1	17	1	1	5	2	0	9	7
0	0	8	0	16	9	1	6	1	0	8	15
0	1	2	5	20	1	1	7	0	0	7	30
0	1	3	4	19	2	2	2	4	0	10	2
0	1	4	3	18	1	2	3	2	1	10	1
0	1	5	2	17	1	2	3	3	0	9	3
0	1	6	1	16	2	2	4	2	0	8	6
0	1	7	0	15	3	2	5	1	0	7	11
0	2	1	5	19	1	2	6	0	0	6	26
0	2	2	4	18	3	3	2	1	2	10	1
0	2	3	3	17	3	3	2	2	1	9	1
0	2	4	2	16	5	3	2	3	0	8	4
0	2	5	1	15	3	3	3	2	0	7	2
0	2	6	0	14	5	3	4	1	0	6	9
0	3	3	2	15	3	3	5	0	0	5	29
0	3	4	1	14	2	4	2	2	0	6	2
0	3	5	0	13	14	4	3	1	0	5	8
0	4	2	2	14	2	4	4	0	0	4	37
0	4	3	1	13	1	5	1	0	2	7	1
0	4	4	0	12	12	5	1	1	1	6	1
0	5	0	3	14	1	5	1	2	0	5	2
0	5	1	2	13	1	5	2	1	0	4	5
0	5	2	1	12	4	5	3	0	0	3	34
0	5	3	0	11	20	6	1	1	0	3	1
0	6	1	1	11	3	6	2	0	0	2	46
0	6	2	0	10	31	7	0	1	0	2	2
0	7	1	0	9	23	7	1	0	0	1	38
0	8	0	0	8	38	8	0	0	0	0	41
1	2	2	3	15	1						

表 14　投与後の MHAQ の分布

投与後											
0点	1点	2点	3点	ｽｺｱ	頻度	0点	1点	2点	3点	ｽｺｱ	頻度
0	0	1	7	23	1	1	3	4	0	11	1
0	0	2	6	22	1	1	4	3	0	10	1
0	0	3	5	21	1	1	5	1	1	10	3
0	0	4	4	20	1	1	5	2	0	9	3
0	0	6	2	18	1	1	6	0	1	9	1
0	0	8	0	16	1	1	6	1	0	8	7
0	1	3	4	19	2	1	7	0	0	7	27
0	1	6	1	16	3	2	0	5	1	13	1
0	1	7	0	15	1	2	3	3	0	9	1
0	2	1	5	19	1	2	4	2	0	8	4
0	2	4	2	16	2	2	5	0	1	8	1
0	2	5	1	15	4	2	5	1	0	7	6
0	2	6	0	14	6	2	6	0	0	6	27
0	3	0	5	18	1	3	3	2	0	7	3
0	3	2	3	16	1	3	4	1	0	6	3
0	3	4	1	14	1	3	5	0	0	5	19
0	3	5	0	13	5	4	1	1	2	9	1
0	4	2	2	14	2	4	2	1	1	7	2
0	4	4	0	12	3	4	3	0	1	6	1
0	5	0	3	14	1	4	3	1	0	5	3
0	5	2	1	12	2	4	4	0	0	4	32
0	5	3	0	11	4	5	3	0	0	3	49
0	6	1	1	11	3	6	1	0	1	4	1
0	6	2	0	10	10	6	1	1	0	3	5
0	7	0	1	10	1	6	2	0	0	2	54
0	7	1	0	9	19	7	0	1	0	2	1
0	8	0	0	8	46	7	1	0	0	1	82
1	2	5	0	12	1	8	0	0	0	0	118
1	3	2	2	13	1						

表 15　MHAQ の構成

項目	略語
靴紐を結びボタンかけも含め自分で身支度ができますか？	Dressing
寝床に入ること、寝床からおきることができますか？	Rising
水がいっぱい入っているコップを口元まで運べますか？	Eating
戸外の平坦な地面を歩けますか？	Walking
身体全体を洗いタオルで拭くことができますか？	Hygiene
腰を曲げて床にある衣類を拾えますか？	Reach
蛇口を開けたり閉めたりできますか？	Grip
車の乗り降りができますか？	Activity

ディリクレ多項モデルへの適用の前提として，各設問の選択肢の分布が一定であることが必要である。そこで図 22に投与前後の MHAQ の各選択肢の得点分布を示す。若干のばらつきは認められるものの各選択肢の分布は投与前後ともに均質性を保っていることがわかった。

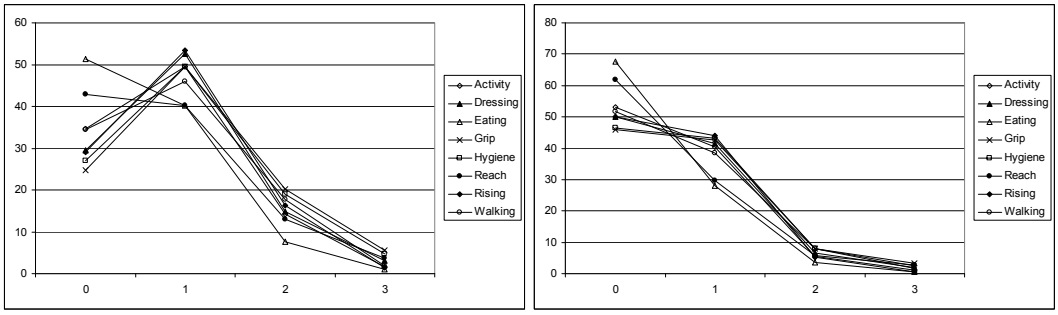


図 22　投与前（左）と投与後（右）の MHAQ の各選択肢の分布

まずは全患者のデータが既知の場合，すなわち選択の場合を考える。表 16と表 17にMHAQ の各選択肢の分布の投与前後の標本平均，標本分散並びに標本共分散を示す。投与前後で比較して，低い点数の方向にシフトしていることがわかった。また，ディリクレ多項モデルにおいては選択肢間の共分散は負値を取るが，本データにおいては「2点」と「3点」において正値を取っていた。このことは「2点」以上の選択肢を選択する患者が非常に限られており，患者間のばらつきがモデルの仮定を超えていることを示している。しかしながら，そのような患者は少ないことから全体に与える影響は少ないと考えられた。

表 16　投与前後の各選択肢の標本平均と標本分散

項目		標本平均	標本分散
投与前	0 点	2.73	7.59
	1 点	3.81	5.71
	2 点	1.23	3.30
	3 点	0.23	0.66
投与後	0 点	4.26	9.38
	1 点	3.08	7.07
	2 点	0.51	1.66
	3 点	0.14	0.46

表 17　投与前後の標本共分散

項目	投与前	投与後
0 点:1 点	-4.26	-6.92
0 点:2 点	-2.75	-1.91
0 点:3 点	-0.58	-0.55
1 点:2 点	-0.96	0.00
1 点:3 点	-0.49	-0.15
2 点:3 点	0.41	0.25

続いて，図 23，図 24，図 25，図 26に MHAQ スコア並びに各選択肢に対し，モーメント法による推定結果，最尤推定法によるニュートン・ラフソン法を用いた方法

（NewtonZ：線形結合分布を仮定した方法，NewtonX：ディリクレ多項分布を仮定した方法）による推定結果を示す。さらには各選択肢の標本平均からパラメータ推定した多項分布を仮定した方法による結果を示す。

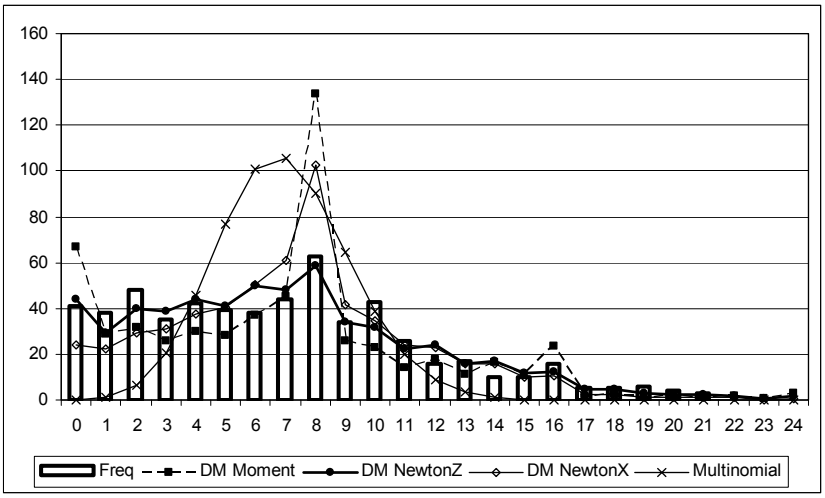


図 23　投与前の MHAQ スコアの期待観測値（折れ線）と実観測値（棒グラフ）

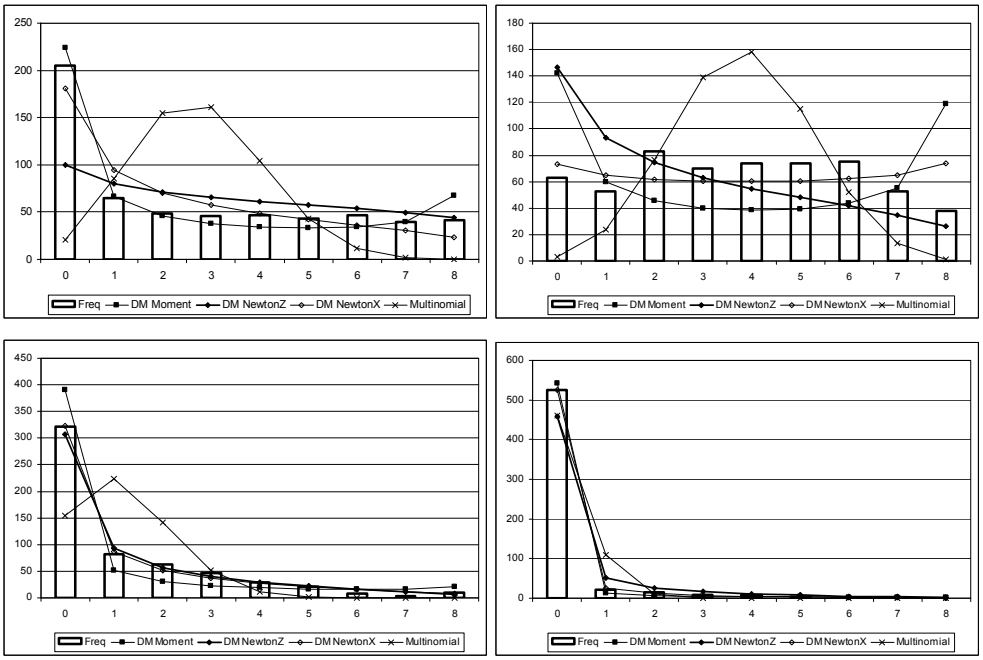


図 24　投与前の MHAQ の各選択肢の期待観測値（折れ線）と実観測値（棒グラフ）

（0点（左上）1点（右上）2点（左下）3点（右下））

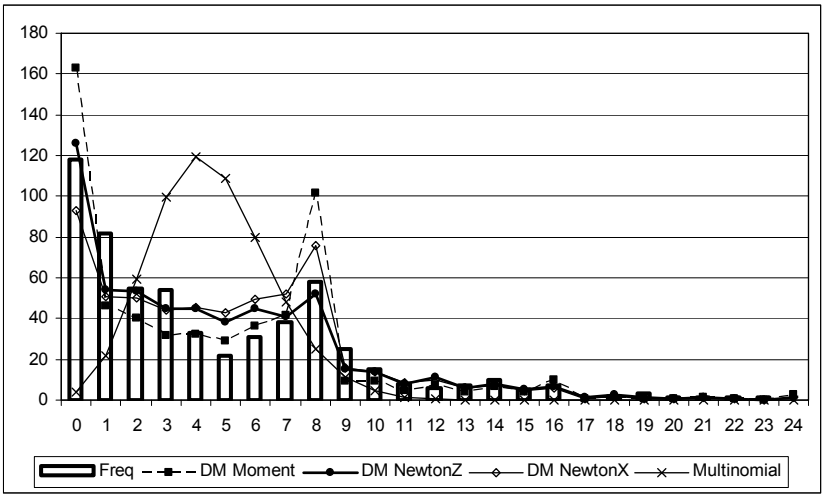


図 25　投与後の MHAQ スコアのパラメータ推定による
期待観測値（折れ線）と実観測値（棒グラフ）

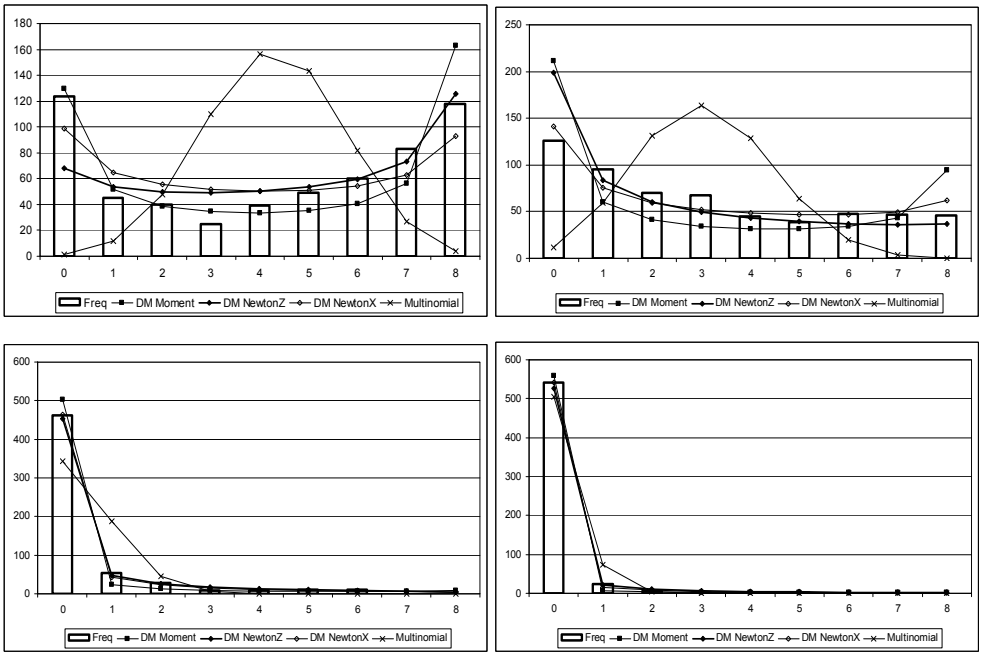


図 26　投与後の MHAQ の各選択肢の期待観測値（折れ線）と実観測値（棒グラフ）
（0点（左上）1点（右上）2点（左下）3点（右下））

また，表 18には推定されたパラメータを，表 19には各パラメータから算出された MHAQ スコアの平均と分散を示す。当然，モーメント法による推定においては観測値から算出された標本平均，標本分散と一致した。ディリクレ多項分布を仮定した方法による最尤推定の結果では，MHAQ データの選択肢間の相関が大きいため，分散が過小推定された。線形結合スコア分布を仮定した推定では，モーメント推定の結果とほぼ同程度

であるが，最尤推定であるため若干の過小推定が見られるが許容できる程度であった。

表 18　パラメータ推定結果

項目		ﾓｰﾒﾝﾄ法	最尤法	
			ﾃﾞｨﾘｸﾚ多項分布	線形結合ｽｺｱ分布
投与前	a_1	0.281	0.813	0.534
	a_2	0.392	0.656	0.872
	a_3	0.126	0.324	0.283
	a_4	0.024	0.126	0.053
投与後	a_1	0.364	0.741	0.627
	a_2	0.263	0.416	0.519
	a_3	0.044	0.106	0.095
	a_4	0.012	0.043	0.030

表 19 投与前後の MHAQ スコアのパラメータから算出される平均と分散

項目		ﾓｰﾒﾝﾄ法	最尤法	
		(観測値)	ﾃﾞｨﾘｸﾚ多項分布	線形結合ｽｺｱ分布
投与前	平均	6.96	7.33	7.01
	分散	22.99	16.56	22.82
投与後	平均	4.54	5.03	4.63
	分散	19.87	17.15	19.48

各モデルの当てはまりを確認するため，表 20と表 21に MHAQ スコア並びに各選択肢に対する適合度の χ^2値を示す。MHAQ スコアでは，線形結合スコア分布を仮定した方法が最も χ^2値が小さかった。ベータ二項分布においては利用可能であったモーメント法による推定では χ^2値は大きく，最尤法による推定方法が優れていることが示唆された。ディリクレ多項分布を仮定した最尤法は，本データの個体間のばらつきの大きさが影響し χ^2値は若干高いがモーメント法の結果と比較し良好であった。MHAQ の各選択肢へ対する当てはまりはでは，ディリクレ多項分布を仮定した最尤推定による χ^2値が最も良好であった。線形結合スコア分布を仮定した最尤推定は若干 χ^2値が大きくなるが，モーメント法による推定比較し良好な結果を得られた。

表 20 MHAQ スコアの χ^2適合度

項目	手法	χ^2値
投与前	DP Moment	151.2
	DP NewtonX	107.0
	DP NewtonZ	29.2
	Multinomial	2880301
投与後	DP Moment	148.1
	DP NewtonX	80.2
	DP NewtonZ	51.8
	Multinomial	5781825290

表 21　MHAQ 各選択肢の χ^2適合度

手法		χ^2値			
		0 点	1 点	2 点	3 点
投与前	DP Moment	26.5	240.6	118.1	32.0
	DP NewtonX	39.5	40.6	14.6	8.9
	DP NewtonZ	137.3	129.2	13.7	48.9
	Multinomial	18248	2194	458424	43830
投与後	DP Moment	43.7	143.1	81.8	47.0
	DP NewtonX	47.1	18.3	16.0	8.4
	DP NewtonZ	65.1	45.2	15.3	10.8
	Multinomial	15281	9239	6335348	357230676

続いて，投与前に MHAQ スコアが5点以上の379名の患者のデータのみが得られたと仮定する。打ち切りの場合，データの得られなかった人数が204名であるとわかっているが，トランケーションの場合は不明である。表 22と表 23に投与前 MHAQ スコアが5点以上の患者の各選択肢の分布の投与前後の標本平均，標本分散並びに標本共分散を示す。表 16と比較し投与前の「1点」以下の標本平均，標本分散が小さくなった。投与後では3.6節で述べた平均への回帰の影響が確認できた。標本共分散は表 17と傾向は同じとなった。

表 22 投与前 MHAQ スコアが5点以上の患者の投与前後の各選択肢の標本平均と標本分散

項目		標本平均	標本分散
投与前	0 点	0.95	1.56
	1 点	4.83	4.79
	2 点	1.87	3.89
	3 点	0.36	0.97
投与後	0 点	2.88	7.77
	1 点	4.14	6.64
	2 点	1.76	2.32
	3 点	0.22	0.68

表 23 投与前 MHAQ スコアが5点以上の患者の投与前後の標本共分散

項目	投与前	投与後
0 点:1 点	-0.34	-5.38
0 点:2 点	-0.96	-1.85
0 点:3 点	-0.26	-0.53
1 点:2 点	-3.33	-0.79
1 点:3 点	-1.11	-0.47
2 点:3 点	0.40	0.32

続いて，図 27と図 28に MHAQ スコア並びに各選択肢に対し，モーメント法による推定結果（比較対象：4点以下の評価はなしと仮定），打ち切りとトランケーションそれぞれでの最尤推定法による方法（Censor_Z, Truncation_Z：線形結合スコア分布を仮定した方法，Censor_X, Truncation_X：ディリクレ多項分布を仮定した方法）による推定結果を示す。

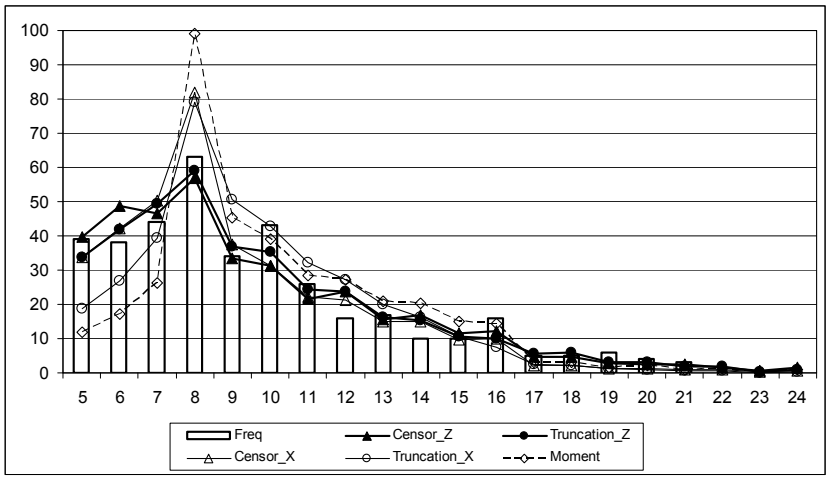


図 27　投与前 MHAQ スコアが5点以上の患者の投与前の MHAQ スコアの期待観測値（折れ線）と実観測値（棒グラフ）

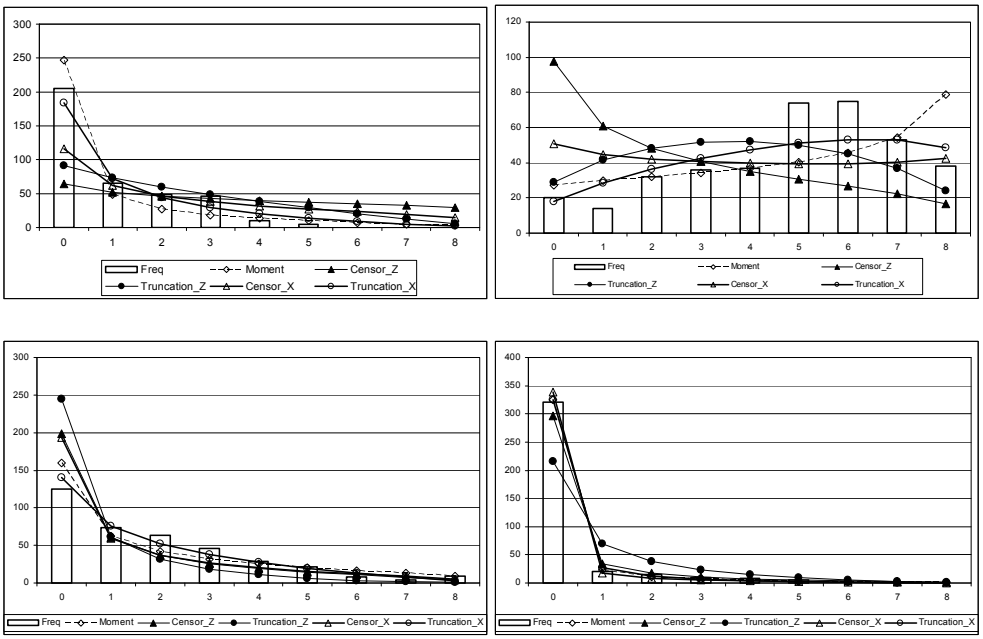


図 28　投与前 MHAQ スコアが5点以上の患者の投与前の MHAQ の各選択肢の期待観測値（折れ線）と実観測値（棒グラフ）
（0点（左上）1点（右上）2点（左下）3点（右下））

また表 24には推定されたパラメータを，表 25には各パラメータから算出された MHAQ スコアの平均と分散を示す。モーメント法による推定においては観測値から算出された標本平均，標本分散と一致した。表 19と比較し平均は大きくなり，分散は小さくなった。ディリクレ多項分布を仮定した方法による最尤推定の結果のほうが，線形結合スコア分布を仮定した方法と比較して平均は小さく，分散は大きくなった。特に打ち切りの線形結合スコア分布を仮定した方法では，真値に非常に近い値となった。

表 24 投与前 MHAQ スコアが5点以上の患者分布のパラメータ推定結果

項目		ﾓｰﾒﾝﾄ法	最尤法　打ち切り		最尤法　ﾄﾗﾝｹｰｼｮﾝ	
			ﾃﾞｨﾘｸﾚ多項分布	線形結合ｽｺｱ分布	ﾃﾞｨﾘｸﾚ多項分布	線形結合ｽｺｱ分布
投与前	a_1	0.207	0.549	0.811	0.459	0.933
	a_2	1.050	0.871	0.643	1.617	1.548
	a_3	0.406	0.327	0.325	0.616	0.307
	a_4	0.078	0.057	0.127	0.103	0.392

表 25投与前 MHAQ スコアが5点以上の患者分布のパラメータから算出される標本平均と標本分散

項目	ﾓｰﾒﾝﾄ法	最尤法　打ち切り		最尤法　ﾄﾗﾝｹｰｼｮﾝ	
	(観測値)	ﾃﾞｨﾘｸﾚ多項分布	線形結合ｽｺｱ分布	ﾃﾞｨﾘｸﾚ多項分布	線形結合ｽｺｱ分布
標本平均	9.63	7.52	7.02	9.04	8.40
標本分散	13.9	17.03	23.13	11.72	18.84

各モデルの当てはまりを確認するため表 26と表 27に MHAQ スコア並びに各選択肢に対する適合度の χ^2値を示した。MHAQ スコアでは，線形結合スコア分布を仮定した方法が最も χ^2値が小さかったが，打ち切りと比較してトランケーションのほうが当てはまりは良い結果となった。これは4点以下の人数の制約がないため，適合度の χ^2値の観点では良い結果が得られたと考えられる。MHAQ の各選択肢へ対する当てはまりはでは，打ち切りでは良い結果が得られたが，トランケーションにおいては MHAQ スコアの適合度が過剰に高いため，各選択肢の分布の当てはまりは悪いという結果が得られた。モーメント法による推定もさほど悪くない結果が得られたが，これは線形結合スコアである MHAQ スコアで5点以上の患者を取り扱ったため，各選択肢の分布への影響が比較的小さくなっていることが示唆された。

表 26 MHAQ スコアの χ^2適合度

項目	手法	χ^2値
投与前	Moment	148.8
	Censor_X	36.9
	Truncation_X	49.9
	Censor_Z	24.4
	Truncation_Z	19.3

表 27　MHAQ 各選択肢の χ^2適合度

	手法	χ^2値			
		0 点	1 点	2 点	3 点
投与前	Moment	93.0	78.6	38.4	9.37
	Censor_X	161.3	109.2	71.8	21.6
	Truncation_X	39.6	33.0	18.7	12.6
	Censor_Z	458.9	322.9	79.5	14.8
	Truncation_Z	226.0	81.7	341.7	121.0

4.4.5　議論

QOL 調査票のように取り得る選択肢が3種類以上ある問題をモデル化した。そして処置前にスクリーニングが実施されることを想定し，その際の処置後の分布を特定した。

さらにデータの不完全性の状況を分類し，その状況ごとにディリクレ多項モデルのパラメータ推定方法を示した。特に，打ち切りとトランケーションの場合には，ニュートン・ラフソン法を用いパラメータの最尤推定の方法を示した。

これらのモデルとパラメータ推定方法を，実際のデータに当てはめたところ，すべてのデータが得られている選択の場合と比べ，打ち切りの場合は未知のデータがあるにもかかわらず遜色のない推定が可能であることがわかった。また，ディリクレ多項分布そのものの最尤推定と線形結合スコア分布の最尤推定のそれぞれの特徴を把握し，前者は各選択肢，後者はMHAQスコアの最尤推定に対してあてはまりが良い。

ディリクレ多項分布の線形結合スコア分布はMHAQの例では0点から3点と各選択肢間のスコアが均等であったが，設定を変更することによって柔軟に対応が可能である。今後，幅広い応用の可能性が示唆された。

また，例としてあげたリウマチの評価指標の多くには臨床的に重要な変動についての議論がWells, *et al*. (2001)等で行われている。MHAQのスコアの評価に関しても同様な議論がなされている（Redelmeier and Lorig (1993)を参照）。臨床試験においてQOL調査票を対象としたスクリーニングが実施されることは非常に稀である。そのような場合に，本節で議論したような明確なカットオフ値は示されないが，評価スコアと主要な評価項目の関連は無視できないことが多いため，評価スコアが良い（MHAQの場合は0点に近い）患者が他の測定におけるスクリーニングで除外されていること（たとえばCRPが2 mg/dL未満の患者など）が容易に想定可能である。そのような場合には，上記論文等で定められた臨床的に重要な変動に対し，本節で提案するような方法での平均への回帰による変動を考慮した再検討が必要かもしれない。

5. 結論

5.1 論文の総括

本論文では，処置前後データをとる研究計画において，研究者の多くが直面する平均への回帰の問題と不完全データに関する統計的問題について議論した。特に臨床試験では薬剤の効果を確認するために処置前後研究は必須であり，またその対象を限定する際には不完全性の問題，特にスクリーニングの問題が生じる。そのような場合に，平均への回帰，スクリーニング検査の影響を慎重に検討することが非常に重要である。それは，もしその臨床成績から医薬品の有効性および安全性が規制当局に認められれば，その薬剤は広く患者に使用されることになるからである。万が一，有効でないあるいは安全でない薬剤が患者に投与されるようなことがあれば，それは社会全体に対する損失となるため，研究者はデータを正しく評価しなくてはならない。

本論文の目的は，第一に平均への回帰モデルに対して Bayes 流のモデルを当てはめることにより定式化を行い，正規分布だけでなくポアソン分布等のカウントデータにおいてもそのモデルが成り立つことを示すことであり，第二には不完全データの問題の中で主要なトピックの一つである打ち切りやトランケーションがあるデータに関する統計的な推測について議論することであった。特に処置前値にスクリーニング等が生じ不完全データである場合のパラメータ推定方法，処置後値の分布の形状の研究を行った。

3章では Bayes 流モデリングによる平均への回帰の定式化を行い，それをカウントデータにも適応させる問題を考えた。4章では，まず4.2節において処置前後データが2変量正規分布に従う場合に，処置前値に対しある種のスクリーニングが施された場合の処置後値の分布の正規分布からの乖離について検討した。さらに，4.3節ではベータ二項分布に従うと仮定したカウントデータに対し，処置前値に対しスクリーニングが施された場合の分布のパラメータ推定方法ならびに処置後値の平均への回帰の大きさについて議論した。また4.4節ではベータ二項分布の一般化された分布であるディリクレ多項分布についても同様に検討した。

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，3章では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，多項分布とその線形結合スコア分布を考察し，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。

処置前後値が2変量正規分布に従っていると仮定された場合に，ベースライン値のトランケーションがエンドポイントの分布の構成に正規分布からの乖離という点でどのように影響するかを4.2節で明らかにした。歪度，尖度，カルバックライブラー情報量を正規性の評価指標に用い，より一般化された統計量 $Y^* - \gamma X^*$ の正規分布からの乖離の程度を

示した。本統計量にはY^*とY^*-X^*が含まれている。その乖離の程度には相関ρが非常に1に近い場合でない限り，そこまで大きくはないことが示された。正規分布からの乖離の程度は十分に小さいため，従来の t 検定または分散分析の使用は統計手法の頑健性の立場からも妥当であることがわかった。さらには，尖度は片側トランケーションにおいてカットオフ値aがおおよそ-1の場合，γとρの値にかかわらず0となることがわかった。

処置前後で問題数が異なることを想定したテストでの正答数の分布を，4.3節ではベータ二項分布でモデル化した。そして処置前のテストでスクリーニングが実施されることを想定し，ある一定の正答数以下の学生の集団の処置効果がない場合の処置後の期待分布をモデル化し，平均への回帰の影響を考慮した検定方法を適用して良好な結果を得た。また，データの不完全性の状況を分類し，その状況ごとにベータ二項モデルのモーメント法によるパラメータ推定方法を示した。特に，打ち切りとトランケーションの場合には，モーメント推定量を用いたニュートン・ラフソン法によるパラメータ推定方法を示した。補習後並びに補習前後の変化量の分布をベータ二項分布を用いてモデル化し，その期待値と分散を算出することにより，平均への回帰の影響を考慮した検定方法を示した。平均への回帰の影響を考慮せず検定を行ってしまうと，真実の結果とは異なった検定結果が得られる可能性が示唆された。

QOL 調査票のようにとり得る選択肢が3種類以上ある問題を4.4節でモデル化した。そして処置前にスクリーニングが実施されることを想定し，その際の処置後の分布を特定した。さらにデータの不完全性の状況を分類し，その状況ごとにディリクレ多項モデルのパラメータ推定方法を示した。特に，打ち切りとトランケーションの場合には，ニュートン･ラフソン法を用いパラメータの最尤推定方法を示した。

本論文では，さまざまな問題設定において平均への回帰とスクリーニング検査の与える影響について議論したが，ベータ二項モデル，ディリクレ多項モデルでは，各問題の回答率が一定もしくはほぼ一定であることが前提であった。それらがある程度異なる場合については今後の研究課題としたい。また，これらのモデルにおいては処置効果の設定方法の検討も同様に課題である。また平均への回帰において，処置が同じ θ に対して異なる効果を持つ，すなわち処置前の θ と処置後の θ^* に確率分布 (θ,θ^*) を想定する場合について，十分な議論はできなかった。本モデルの検討は今後の検討課題である。

謝辞

本論文は筆者が成蹊大学大学院　工学研究科情報処理専攻　博士後期課程に在籍中の研究成果をまとめたものである。

本論文をまとめるにあたり，本研究の実施の機会を与えて頂き，終始暖かい激励とご指導，ご鞭撻を頂いた成蹊大学理工学部情報科学科教授　岩崎学先生には心より感謝申し上げる。

学位論文審査において，貴重なご指導とご助言を頂いた成蹊大学理工学部情報科学科教授　上田徹先生，同　教授　渡邉一衛先生，慶應義塾大学経済学部教授　稲葉由之先生には心より感謝申し上げる。

博士後期課程への進学ならびに研究遂行の機会を与えてくださり，研究と仕事の両立を支援して頂いた中外製薬株式会社　臨床開発本部臨床企画推進部統括マネジャー（生物統計担当）斉藤誠氏には心より感謝申し上げる。

博士後期課程への進学にあたり，温かい激励を頂いた中外製薬株式会社　臨床開発本部臨床開発部統括マネジャー（オンコロジー開発１担当）瀬川耕太郎氏，株式会社中外臨床研究センター　バイオメトリクス部長　菊池かずよ氏には心より感謝申し上げる。

博士後期課程の研究遂行にあたり，温かい激励を頂いた Roche Products Limited, Deputy Global Head of Inflammation Biostatistics の Paul Mahoney 氏，中外製薬株式会社臨床開発本部臨床企画推進部統計解析第1グループ　グループマネジャー　植松弓美子氏には心より感謝申し上げる。

博士後期課程在学中，同期のエーザイ株式会社　信頼性保証本部安全管理部　大道寺香澄氏の存在が，研究を進めていく上で，大きな励みになったことをここに記すとともに，心より感謝申し上げる。

投稿論文作成にあたり，貴重なご助言を頂いた成蹊大学理工学部情報科学科助教　吉田清隆先生，慶應義塾大学医学部クリニカルリサーチセンター講師　阿部貴行先生には心より感謝申し上げる。

4.4.4節の臨床データの利用にあたり，和歌山県立医科大学 免疫制御学講座教授　西本憲弘先生の論文データを利用させて頂いた。心より感謝申し上げる。

また，研究を進めるにあたり，ご支援，ご協力を頂きながら，ここにお名前を記すことができなかった多くの方々に心より感謝申し上げる。

最後に，学位論文をまとめるにあたり，心の支えとなってくれた妻まどか，見守り，時間を与えてくれた長男晃佑，次男晴匡，三男龍吾には心より感謝申し上げる。

本研究の一部は，科学研究費補助金基盤研究(A) No.16200022によった。

参考文献

1) 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.
2) Altham, P. M. E. (1978). Two generalizations of the binomial distribution, *Applied Statistics*, **27**, 162-167.
3) Arostegui, I., Núñez-Antón, V. and Quintana, J. M. (2007). Analysis of the short form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine*, **26**, 1318–1342.
4) 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.
5) Beath, K. J. and Dobson, A. J. (1991) Regression to the mean for nonnormal populations. *Biometrika*, **78**, 431-435.
6) Bonate, P. L. (2000) *Analysis of Pretest-Posttest Designs*. Chapman & Hall, Boca Raton, FL.
7) Chesher, A. (1997) Non-normal variation and regression to the mean. *Statistical Methods in Medical Research*, **6**, 147-166.
8) Chuang, C. and Cox, C. (1985) Pseudo maximum likelihood estimation for the dirichlet-multinomial distribution. *Communications in Statistics – Theory and Methods*, **14**, 2293-2311.
9) Chuang-Stein, C. (1993) The regression fallacy. *Drug Information Journal*, **27**, 1213-1220.
10) 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.
11) Cohen, A. C., Jr. (1955) Restriction and selection in sample from bivariate normal distributions. *Journal of the American Statistical Association*, **50**, 884-893.
12) Copas, J. B. (1997) Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research*, **6**, 167-183.
13) 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.
14) Davis, C. E. (1976) The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology*, **104**, 493-498.
15) 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).
16) Ederer, F. (1972) Serum cholesterol changes: effects of diet and regression toward the mean. *Journal of Chronicle Disease*, **25**, 277-289.
17) Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations (with Discussion). *J.R. Statist. Soc.*, **B**, **31**, 195–233.

18) Fleiss, J. L. (1986). The *Design and Analysis of Clinical Experiments.* New York: John Wiley.
19) Folks, J. L. (1981) *Ideas of Statistics*. John Wiley & Sons, New York.
20) Freedman, D., Pisani, R., Purves, R. and Adhikari, A. (1991) *Statistics, Second Edition*. W. W. Norton & Co., New York.
21) Fries, J. F., Spitz, PW., Kraines, R. G. and Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis Rheum* **23**, 137–45.
22) 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.
23) Furby, L. (1973) Interpreting regression toward the mean in developmental research. *Developmental Psychology*, **8**, 172-179.
24) Galton, F. (1886) Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, **15**, 246-263.
25) Hauer, E. (1980) Selection for treatment as a source of bias in before-and-after studies. *Traffic Engineering and Control*, **21**, 419-421.
26) 岩崎　学 (2000) 統計的データ解析のレシピ．日本評論社.
27) 岩崎　学 (2002a) 不完全データの統計解析．エコノミスト社.
28) 岩崎　学 (2002b). 「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.
29) 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.
30) 岩崎　学 (2006). 統計的データ解析入門　単回帰分析，東京図書.
31) 岩崎　学・河田祐一 (2007). 処置前後研究における平均への回帰とその周辺．日本統計学会誌，**36**，131‐145.
32) 岩崎　学・大道寺香澄 (2009). ゼロ過剰な確率モデルとそのテスト得点の解析への応用. 行動計量学，**36**，25-34.
33) 岩崎　学 (2010). カウントデータの統計解析，朝倉書店.
34) Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Discrete Univariate Distribution*s, Third Edition. NY, John Wiley & Sons
35) Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, John Wiley & Sons, New York.
36) Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Discrete Univariate Distributions*, *Second Edition,* John Wiley & Sons, New York.
37) Kawata, Y. and Iwasaki, M. (2008) Assessment of non-normality in pretest-posttest research under screening of the pretest score. *Japanese Society of Computational Statistics*, 21, 31-44.
38) 河田祐一・岩崎　学 (2009). 不完全データに基づく平均への回帰を考慮したテストデー

タの解析．日本テスト学会誌，5，41-51．

39) Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000). *Continuous Multivariate Distributions, Volume 1*: *Models and Applications,* 2nd Edition. New York: John Wiley.

40) Kullback, S. (1959). *Information Theory and Statistics.* New York: John Wiley.

41) 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.

42) Labouvie, E. W. (1982) The concept of change and regression toward the mean. *Psychological Bulletin*, **92**, 251-257.

43) Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

44) 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.

45) Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

46) Maher, M. J. (1987). Fitting probability distributions to accident frequency data, *Traffic Engineering and Control*, **28**, 356-358.

47) Matsuda, Y., Singh, G., Yamanaka, H., Tanaka, E., Urano, W., Taniguchi, A., et al. (2003). Validation of a Japanese version of the Stanford Health Assessment Questionnaire in 3,763 patients with rheumatoid arthritis. *Arthritis Rheum* **49**, 784–8.

48) McDonald, C. J. and Mazzuca, S. A. (1983) How much of the placebo 'effect' is really statistical regression? *Statistics in Medicine*, **2**, 417-427.

49) McGuigan, D. R. D. (1985). Accident 'migration' – or a flight of fancy? *Traffic Engineering and Control*, **26**, 229-233.

50) Monahan, J. F. (2001). *Numerical Methods of Statistics*. Cambridge: Cambridge University Press.

51) Nesselroade, J. R., Stigler, S. M. and Baltes, P. B. (1980) Regression toward the mean and the study of change. *Psychological Bulletin*, **88**, 622-637.

52) Newell, D. and Simpson, J. (1990) Regression to the mean. *Medical Journal of Australia*, **153**, 166-168.

53) Nishimoto, N., Ito, K. and Takagi, N. (2010). Safety and efficacy profiles of tocilizumab monotherapy in Japanese patients with rheumatoid arthritis – meta-analysis of 6 initial trials and 5 long-term extensions – *Mod Rheumatol*, **20**, 222-232.

54) Pearson, K. and Lee, A. (1903) On the laws of inheritance in man. I. Inheritance of physical characters. *Biometrika*, **2**, 357-462.

55) Pincus, T., Summey, J. A., Soraci, S. A. Jr., Wallston, K. A., Hummon, N. P. (1983). Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment

Questionnaire. *Arthritis Rheum* **26**, 1346–53.

56) Redelmeier, D.A. and Lorig, K. (1993). Assessing the clinical importance of symptomatic improvements — an illustration in rheumatology. *Arch Intern Med* **153**, 1337–42.

57) Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

58) Schall, T. and Smith, G. (2000) Do baseball players regress toward the mean? *American Statistician*, **54**, 231-235.

59) Senn, S. (1997) Editorial: Regression to the mean. *Statistical Methods in Medical Research*, **6**, 99-102.

60) Senn, S. J. and Brown, R. (1989) Maximum likelihood estimation of treatment effects for samples subject to regression to the mean. *Communications in Statistics, Series A*, **18**, 3389-3406.

61) Senn, S. J. and Collie, G. S. (1988). Accident blackspots and the bivariate negative binomial, *Traffic Engineering and Control*, **29**, 168-169.

62) 柴田義貞 (1981). 正規分布，東京大学出版会.

63) Stanek, E. J. (1988). Choosing a pretest-posttest analysis. *The American Statistician* **42**, 178-183.

64) Stigler, S. M. (1997). Regression towards the mean, historically considered, *Statistical Methods in Medical Research*, **6**, 103-114.

65) 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.

66) 上坂浩之 (2006). 医薬開発のための臨床試験の計画と解析，朝倉書店.

67) 渡邉裕之・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ーシグナル検出の最近の手法についてー．計量生物学，**25**，37-60.

68) 渡辺美智子・山口和範（編著）(2000) EM アルゴリズムと不完全データの諸問題．多賀出版.

69) Wei, L. and Zhang, J. (2001). Analysis of data with imbalance in the baseline outcome variable for randomized clinical trials. *Drug Information Journal* **35**, 1201-1214.

70) Wells, G., Beaton, D., Shea, B., Boers, M., Simon, L., Strand, V., Brooks, P. and Tugwell, P. (2001). Minimal clinically important differences: review of methods. *The Journal of Rheumatology* **28**, 406–412

本論文に関係する研究業績一覧

学術論文

1) 岩崎　学・河田祐一 (2007). 処置前後研究における平均への回帰とその周辺. 日本統計学会誌, **36**, 131‐145.
2) Kawata, Y. and Iwasaki, M. (2008) Assessment of non-normality in pretest-posttest research under screening of the pretest score. *Japanese Society of Computational Statistics*, **21**, 31-44.
3) 河田祐一・岩崎　学 (2009). 不完全データに基づく平均への回帰を考慮したテストデータの解析. 日本テスト学会誌, **5**, 41-51.

学会　口頭発表

1) 2008年5月 日本計算機統計学会第22回大会にて「処置前後研究におけるスクリーニングの影響評価」発表
2) 2008年9月 統計関連連合大会にて「処置前後研究におけるカウントデータの解析」発表