Journal of Culture and Information Science, 11(2), 103-109,（March 2016）

研究ノート

オッズ比の高次元への拡張とその性質

矢野　環

オッズ比（OR）は、近年効果量としての意義も認められ、よく利用されている。またメタ分析の手法も確立している。この2×2で定義されているORを高次元に拡張し、そのORが望ましい不変性を持つように定義する。これまで拡張の提案はあるが、不変性が十分でない。ORはその分散の推定を元に、通常のオッズ比と同じくメタ分析を行うこともできる。拡張されたオッズ比が1となるのは、行（列）が線形従属な時である。実用上、2×2の場合とほぼ同様の処理ができ、関連する検定よりも良い基準とできる。

1. 緒言

オッズ比（Odds ratio. 以下OR）は従来から疫学･医学･薬学などで良く利用されていた。近年では、効果量としての価値も認められるようになり、英語教育等でも注目されている。

ORを高次元の行列に対して定義するには、2×2小行列群のORの集合を考えるのが主流であり［1, 2, 5］、その可視化の工夫もされ［6, 7, 8］、代数幾何学的考察もある［9］。一個の量としては、Goodman-Kruskal's γ から $OR_G=(1+\gamma)/(1-\gamma)$ とする［3, 4］などの提案がある。しかし、それらの量は本来のORの満たす性質の一部は持っているものの、重要な不変性が失われているなどの欠点もある。このため、2×2以外の複数の行列が与えられた時、その群の異質性や同質性を判断しようとすると、解りやすい不変量がなかった。

もしORと同様な不変性を持つ量を定めることができるならば、一群の行列が与えられた場合、その量のMantel-Haenszel型estimateや、メタ分析の手法での推定値が、その群の性質を表現することになるであろう。

以下において、正方行列の場合を主体として、ORの高次元への拡張を提案し、その性質を調べ、実際の適用例を解説する。また全般的理解の為に、交代群の不変式の一般論から始める。

2. 交代群の不変式

Vを体K上のn次元ベクトル空間とし、その座標関数の集合 $\{x_1, x_2, \ldots, x_n\}$ をXとする。Xには、n次対称群Snが添え字の置換として作用する。即ち、置換 $\sigma \in Sn$ に対して、X上の置換 $x_i \longmapsto x_{\sigma(i)}$ が引き起こされる。

このSnのX上への作用は、自然に多項式環K[X]上に誘導される。そのSnの作用での不変式環は、対称式の集合Sである。Snの元は偶置換と奇置換にわかれ、偶置換全体の集合Anは Snのindex 2の（正規）部分群をなす。置換は符号 sign (σ) をもち、これは群準同型を与える。

$$sign : Sn \longrightarrow \{\pm 1\}$$

偶置換は符号1, 奇置換は符号－1である。Anの不変式であって、Snの不変式ではないものを交代式とよぶ。Sと交代式の全体は再び環をなし、半不変式の環と呼ばれる。

交代式Δが存在し、半不変式＝S＋ΔSとなることが知られている。Kの標数が2でない場合は、Δとして基本差積をとることができる。

$$\Delta = \Pi i<j \ (x_i - x_j)$$

右辺を展開し、係数が1である単項式の和をΔp,

係数が－1となる単項式の和をΔnとすれば、

$\Delta = \Delta p - \Delta n$

となる。Kの標数が2の場合は、Δは対称式であり、半不変式＝S＋ΔpS＝S ＋ΔnSとなる。

以下Kの標数≠2とする。

次に、n×n行列M＝(x_{ij}) を考える。第二添え字に置換を含む単項式

$m(\sigma)=\Pi^{n}_{i=1}\ x_{i\sigma(i)}$

は、i行σ(i）列成分の積であり、各行各列から一か所づつ採択している。この第二添え字をSn全体に及ぼした和、

$perm(M) = \Sigma\ \sigma\in Sn\ m(\sigma)$

は、Mのpermanentと呼ばれ、グラフ理論で重要である。また、置換の符号をつけた和

$det(M) = \Sigma\ \sigma\in Sn\ sign(\sigma)m(\sigma)$

は、Mの行列式（determinant）である。

定義 1 detp. detn

$detp(M) = \Sigma\ \sigma\in An\ m(\sigma)$,

$detn(M) = \Sigma\ \sigma\in Sn\setminus An\ m(\sigma)$,

と定義する。

命題 1 detp, detnいずれもAn不変式である。

$det(M) = detp(M) - detn(M)$,

$perm(M)= detp(M) + detn(M)$,

となる。

M^{σ}を奇置換σに従いMの行を入れ替えた行列とすれば（列でも同様）次式が成立する。

$detp(M^{\sigma})=detn(M)$, $detn(M^{\sigma})=detp(M)$.

3. 高次元ORと類似不変量

体K上のn×n行列全体をM(n, K）と記す。以下、定義と命題を列挙する。命題の証明は容易であるので略す。

定義 2 OR

Dom={M∈M(n,K); $detp(M)\neq 0$,$detn(M)\neq 0$}

と置く。

OR: Dom⟶K＼{0}を

$OR(M) = detp(M)/detn(M)$

により定義する。

命題 2 ORは次の性質を満たす.

1. $OR(cM)=OR(M)$, $0\neq c\in K$.
2. M^{T}をMの転置行列とすれば $OR(M^{T}) = OR(M)$.
3. M'を、Mの2つの行、あるいは2つの列を入れ替えた行列とすれば$OR(M')=1/OR(M)$.
4. $OR^{-1}(1) = Dom\cap\{M: det(M)=0\}$.

また、Yule's Qの拡張が定義できる。Q(M) はGoodman-Kruskal's γの類似である。

定義 3 Q

$Q(M) = (OR(M)-1)/(OR(M) + 1)$

と定義する。ORの定義より次式が成立する。

$Q(M)= det(M)/perm(M)$.

拡張されたORは、加法・乗法と0でない除法について閉じているKの部分集合上に定義できる。例えば、0以上の実数R_{+}においても定義できる。それが通常のORの拡張となっている。

ORが1となるのは、det(M)＝0となる場合、つまり、Mがmulti-colinear（Mの行、また列が一次従属）な時である。

以下、K＝R_{+}とする。$-1<Q(M)<1$である。

ファイ係数の拡張は次の式で与えられる。下記において、colSums(), rowSums() はそれぞれ列和と行和ベクトル、prod(v) はベクトルvの成分すべての積を意味する。また、sqrt() は平方根である。

定義 4 Φ

$$\Phi(M) = \frac{det(M)}{sqrt(prod(colSums(M),\ rowSums(M)))}$$

対数オッズ比も同様に定義する。

定義 5 logOR

logOR: Dom⟶Rを次式で定める。

$logOR(M) = log(OR(M))$

命題3 関数Q, Φ, logORのいずれかをTとする。このときTは次の性質を満たす。

1. $T(cM)= T(M)$, $0\neq c\in R_{+}$.
2. M^{T}をMの転置行列とすれば $T(M^{T}) =T(M)$.
3. M'を、Mの2つの行、あるいは2つの列を入れ替えた行列とすれば $T(M')= -T(M)$.

従って、

$T(M^{\sigma})=\mathrm{sign}(\sigma)T(M),\ \sigma\in Sn.$

4. $T^{-1}(0)=\mathrm{Dom}\cap\{M:\ \det(M)=0\}.$

命題 3.1 によれば、T は（そして $\log(OR_G)$ も）n^2-1 次元射影空間上の関数となる。さらに、行番号の置換と列番号の置換として、Sn × Sn が Dom 上に作用しており、命題 3.3 により T は sign × sign の符号変化を受ける。その作用により、T は An × An 不変である。

命題 4

M"を、M の一つの行（又は一つの列）を c 倍（c ≠ 0）した行列とする。T＝OR, Q, logOR について T(M")＝T(M) が成立する。

本来の OR は勿論命題 4 を満たすので、その拡張である関数も同様であることが好ましい。

ファイ係数はこの性質を満たさず、拡張である Φ でも一般に Φ(M")≠Φ(M) となる。

Agresti の OR_G は命題 4 を満たさず、命題 2.3 も満たさない。従って、$\log(OR_G)$ は命題 3.3 を満たさない。これは、ORG また Goodman-Kruscal's γ が、元々順序因子のクロス集計を考慮して定義されていることから生じている。

4. logOR の性質。同質性

4.1 同質性と推定値

複数の行列が与えられた時、その同質性もしくは異質性を判断する必要が生じることがある。例えば、複数の治験例での OR を統合して、その全体のシステムでの OR の推定値を求めたい場合などである。以下行列は非負整数成分とする。

$G=\{M_1, M_2, \cdots, M_k;\ Mi \text{ in Dom}\}$

とする。このとき、G の同質性や異質性を判定し、同質ならば不変量を統合する手法として、Mantel-Haenszel 検定とその estimate、replicated G-test、又メタ分析での各種モデルが知られている。検定の場合、異質と判定する p の閾値は .05 に拘らず、分散分析の交互作用と同じく、分野によって定めるのが適切であろう。

Mantel-Haenszel estimate は、行列によって成分の値の大きさに偏りがあっても、多少緩和して加重平均する簡易な目安であり、高次元の OR に直接的に拡張可能である。ここで sum(M) は行列 M の成分の和を意味する。

定義 6 Mantel-Haenszel type estimate.

$$MHe(G)=\frac{\Sigma_i\ \mathrm{detp}(Mi)/\mathrm{sum}(Mi)^{n-1}}{\Sigma_i\ \mathrm{detn}(Mi)/\mathrm{sum}(Mi)^{n-1}}$$

メタ分析での OR 統合の手法を適用するには、logOR の分散（標準誤差 se の 2 乗相当であるが分散とよぶ習慣である）の推定値が必要となる。M が 2×2 行列の場合は、各成分の逆数和である次式がよく知られている。

$v(M)=1/m_{11}+1/m_{12}+1/m_{21}+1/m_{22}$

高次元での v 相当を理論的に求めることも可能ではあろうが、シミュレーションで容易に推定することができるので、実用上問題ない。そこで、$(\log OR(Mi), v(Mi))\ i=1,..,n$ をメタ分析に引き渡せば、モデルに従った推定値を得る。その結果を通行の Forest 図によって表現する（R の metfor パッケージ等）ことも可能である。

4.2 実例

簡単な例を挙げる。ここここでは後述する近傍作成によって得た 3 つの行列を用いる。

$$Mg=\{M_1, M_2, M_3\},$$

$$M_1=\begin{bmatrix}2&5&8\\7&8&5\\11&7&12\end{bmatrix},\quad M_2=\begin{bmatrix}4&3&8\\7&7&6\\9&10&11\end{bmatrix},\quad M_3=\begin{bmatrix}3&4&8\\7&6&7\\10&10&10\end{bmatrix}.$$

Mg に対する、replicated G-test（以下 replG）と、Cochran-Mantel-Haenszel 検定（以下 CMH）の結果は下記の通りである。実際はより多くの情報を出力するが、ここでは略記した。末尾の vib は、後述の近傍シミュレーションによる分散（se）推定値で

表 1 replG, CMH の結果

	G	df	p	VG	OR	logOR	vib
1	5.306	4	.257	.194	.719	-.329	0.980
2	2.220	4	.695	.130	1.056	1.055	1.078
3	2.020	4	.732	.125	1.052	1.050	1.080
total	9.546	12	.656		.927	-.076	
pooled	6.661	4	.155	.131	.947	-.055	1.019
heterog	2.885	8	.941				

Cochran-Mantel-Haenszel test

M^2 = 6.528, df = 4, p-value = 0.163

ある。また、VG は、G 値から求めたクラメール V である（通常の V とほぼ等しい）。pooled は、the pooled matrix であり、3 つの行列の単純和である。

CMH の結果は、replG の pooled matrix に対する結果と近いものである。replG の heterog（異質性検定 heterogeniety G-value であり、p< 閾値の場合に異質と判定する）からして、同質性が保証され、MHe＝0.927 が有効である。その値を total の OR の箇所に記入している。

一方、メタ分析（Random-effects model. Restrictid maximum-likelihood estimator）の結果は図 2 の通りである。なお、数値は OR とその信頼区間として与えているが、棒グラフのスケールは logOR に従って配置している。OR=1 の位置に点線を付している。その前後の目盛りは $1/\sqrt{2}, \sqrt{2}$ である。対数尺なので、一定比が等間隔となる。

Meta 分析の OR 統合値は 0.945（信頼区間はその右 [0.680,1.312]）であり、replG における OR(pooled)＝0.947 や MHe(G)＝0.927 と近い値であるが、いつもそうなるわけではなく、かなり異なることもある。ここでは信頼区間が 1 を跨ぐ。

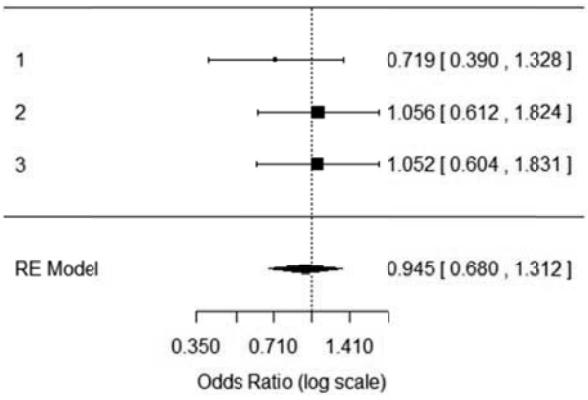

図２　meta 分析の結果の Forest 図

いずれにせよ、これらの数値が整合していることがわかる。

5. 計算の手続

5. 1　全体の手続

この手法を、大量の行列を含む G に適用することは無く、せいぜい 20 個以内であろう。また、サイズ n も実用上 7 まで、あるいは 5 までと思えばよい。Likert scale 7 段階の 2 つの変数のクロスを取ったとして 7×7 である。しかし、両端 {1, 2} {6, 7} を統合すれば 5×5 でよい。

従って、det や perm の計算におけるオーバーヘッドはそれほど深刻ではない。必要な detp, detn は直接再帰的に計算することにすればよい。もしも大規模ならば、det と perm の既存の高速近似計算から detp, detn を求めることになると思われる。

R を用いた素朴な detp, detn の計算は付録に掲げる。通常の計算では有効数字は 15 桁程度であり加算でも桁落ちが心配されるが、R では多倍長が簡単に行えるので、detp, detn を例えば 30 桁の整数として中途計算を行い、除算して OR を求める段階で通常の数値にすればよい。その実例も付録に含めた。

メタ分析の Forest 図の出力には、簡単な wrapper を設計すればよい。このため、replG のスクリプトにおいて、行列の近傍を生成し、logOR の分散を求めるようにして、その結果を metafor::rma に引き渡す。

5. 2　近傍データの作成とその性質

以下、M の成分は非負整数とする。

行列 M の近傍（neighborhood）の生成にあたっては、まず M をクロス表と見て、Untable(M) によって、クロスを取った元の 2 つの質的変数を再現する。それらの長さは N＝sum(M) である。この質的変数の組から N 個のサンプリングを nrep 回行い、各々を table すれば M に近い行列 nrep 個を得る。その nrep 個の行列に対して logOR を求め、標本分布をみて、分散 v(M) を求めることができる。状況確認には nrep＝100 で行い、本番では nrep＝1000, 2000 程度に取ればよい。

M の成分に 0 が含まれる場合、この方法ではサンプルのその成分は常に 0 であることに注意されたい。必要があれば、強制的に 1 を入れる。

この近傍データは十分によく分布するであろうか。またこのとき logOR はどのように分布しているであろう。これらを確認するために、専用のスクリプトを作成した。

そのスクリプトでは、行列空間を n^2 次元のベクトル空間とみて、元の行列 M と近傍の行列の群 G(M) を配置し、主成分分析を行う。その PC1-PC2 主成分スコア散布図に、95% 許容楕円（信頼楕円）を付加し、かつ PC1 への射影の分布の density と rug を配置する。さらに logOR の分

布の density を平均を 0 にそろえて上書きする。標準的な凡例を左上に、また右上にデータフレーム名と nrep を記載する。

図 3 は ee の近傍での 1000 個の分布である。

$$ee = \begin{bmatrix} 20 & 3 & 3 \\ 3 & 20 & 3 \\ 3 & 3 & 20 \end{bmatrix}$$

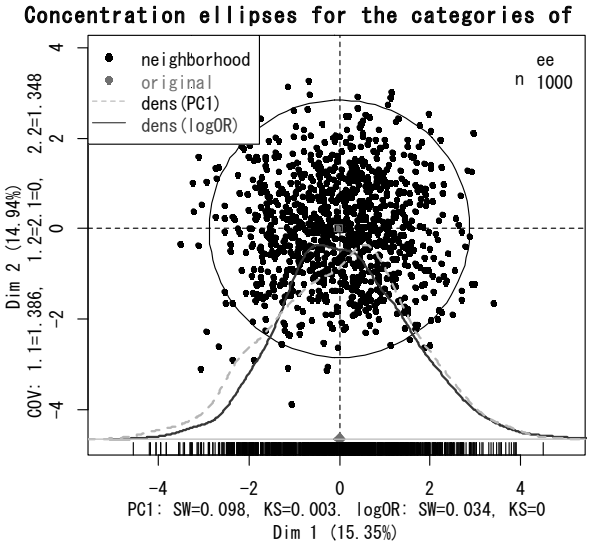

図３　ee の近傍行列と logOR の分布

元の ee は実際中央（原点）に来ている。まさに近傍に相応しく分布しているのがわかる。

図 3 の灰色の破線が PC1 方向への射影の density であり、実線が logOR の分布の density である。この場合はよく重なっている。logOR の分布は尖度が大きくなくなることもある。

この ee の近傍 10 個に対してメタ分析を適用した結果の一例は図 4 のとおりである。この例の場合、MHe＝15.935 であった。また、replG の異質性検定は p＝0.646 となり同質と判断する。

CMH : M^2＝686.66, df＝4, p-value < 2.2e-16 となるがこれは G(pooled)＝658.648, df＝4 と同程度の内容であり、役に立つ情報とはいえない。

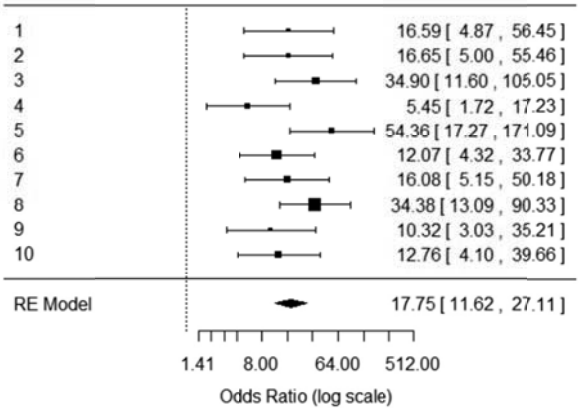

図４　ee の近傍 10 個のメタ分析

6. 矩形行列への拡張

もとより、本稿の結果を r×c 行列（r < c）に拡張することは可能である。すべての r × r 小行列 N＝cCr 個、または代表として連続する r 列からなる c-r+1 個の r×r 小行列に対する OR の集合を対象とすればよい。この場合、余次元 1 の代数的集合 L を除いた r×c 行列全体の集合からの写像となる。

OR ： $M(r,c;\ K)\setminus L \longrightarrow K^N$ 又は K^{c-r+1}

特に r＝2 の場合は、通行の方式と一致する。

さらにこの OR の集合を一つの代表値にするには MHe あるいは、メタ分析による推定値を用いればよい。

一例として、Agresti (2013) p.395, Table10.5 の 4×6 行列を取り上げる。その列 ABCDEF の連続する 4 列づつを取った 3 つの 4×4 行列の OR とメタ分析による統合は図 5 の通りである。

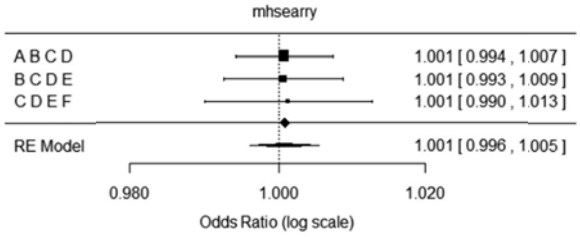

図５　Cross-Classification of Mental Health Status and Socioeconomic Status.[1] Table10.5.

7. 結言

高次元の OR を、十分な不変性を持つように定義した。CMH が役に立つ情報を提供しないのに比較して、効果量として十分な機能を果たすと認められる。今後より精密な性質の検討を行う。

参考文献

［1］Agresti A.(2013). Categorical Data Analysis. 3rd edition. Wiely, New York.

［2］Agresti A.(1984). Analysis of Ordinal Categorical Data. Wiely, New York.

［3］Agresti A.(1980). Generalized odds ratios for ordinal data. *Biometrics* 36:59-67.

［4］Edwardes MD & E.Baltzan (2000). The generalization of the odds ratio, risk ratio and risk difference to r × k tables. *Stat Med*. 19(14):1901-14.

［5］Subramanyam K.(1988). Analysis of Odds Ratios in 2 × n Ordinal Contingency Tables. *J. Mult. Anal*. 27:478-493.

［6］De Rooij M. and Anderson, C.J.(2007). Visualizing, summarizing and comparing odds ratio structures, *Methodology* 3:139-48.

［7］D'Ambra, L., Camminatiello I. and Sarnacchiaro P. (2013) Generalized log odds ratio analysis for the association in two-way contingency table, Electric Book "Advances in Latent Variables", Eds. Brentari E., Carpita M., Vitae Pensiero, Milan, Italy, ISBN978 88 343 25568.

［8］Sarnacchiaro P., Gallo M. and D' Ambra L.(2014). Three-way decomposition of weighted log-odds ratio for customer satisfaction analysis, Innovation and Society 2013 Conference, IES 2013. *Procedia Economics and Finance* 17:30-38.

［9］Slovkovic A.B. and Fienberg S.E.(2010). Algebraic geometriy of 2x2 contigency tables, "Algebraic and Geometric Methods in Statistics", Eds. Gibilisco P., Riccomago E., Rogantin M.P., Wynn H.P., Cambridge University Press, Cambridge. pp.63-82.

附録A.

A.1. detp, detn の計算

パッケージ Rmpfr によって多倍長計算を行う。関数 mpfr（k, 100） は、k を 100bit, 2^100 ≒ 10^30.1 で取り扱うという意味であり、整数で 30 桁が確保される。出力 pn により、

pn[1] = detp, pn[2] = detn を受け取った後で

as.numeric(pn[1]/pn[2])

として、通常の数に変換する。

```
library(Rmpfr)  # 事前に外部で実行しておく
rcsvm.det <- function(x)  # x: matrix
{
 n <- nrow(x)
 pn <- c(mpfr(0,100),mpfr(0,100))
 if (n==2) {
  pn[1] <- mpfr(x[1,1],100)*mpfr(x[2,2],100)
  pn[2] <- mpfr(x[1,2],100)*mpfr(x[2,1],100)
  return(pn)
 } else {
  indx <- rep(c(1,2),n)
  for (cnt in 1:n){
     pn[1] <- pn[1] + mpfr(x[1,cnt],100)*
          Recall(x[-1, -cnt])[indx[cnt]]
     pn[2] <- pn[2] + mpfr(x[1,cnt],100)*
          Recall(x[-1, -cnt])[indx[cnt+1]]
  }
  return(pn)
 }
}
```

ここで Recall は R における再帰的呼び出しであり、これによって、全体の関数名を変更しても、コーディングを変更する必要が無い。上記の関数は行列式の小行列展開そのものを素朴にコーディングしたものであり、n が増加すると時間がかかる。通常の数で実行する場合、上記のコードから mpfr(,100) の文字を削除すればよい。

A.2. Untable について

Untable は、クロス集計表から元のデータを復元する。例えば下記の行列 mat に対して DescTools::Untable(mat) とすれば右のとおり 2 因子変数の data.frame が生成される。

```
> mat
  A B
a 1 3
b 2 4
```

```
> table(Untable(mat))
   Var2
Var1 A B
   a 1 3
   b 2 4
```

```
> Untable(mat)
   Var1 Var2
1     a    A
2     b    A
3     b    A
4     a    B
5     a    B
6     a    B
7     b    B
8     b    B
9     b    B
10    b    B
```

A.3. Agresti の Genelarized Odds Ratio. OR_G

行列 M（r 行 c 列）に対し

$$OR_G = P/Q$$

$$P = \sum_{i<j}\sum_{s<t} m_{is}\, m_{jt},\ \#\ \text{concordant pairs}$$

$$Q = \sum_{i<j}\sum_{s>t} m_{is}\, m_{jt}.\ \#\ \text{discordant pairs}$$

と定義される。P, Q はいずれも $rC_2 \cdot cC_2$ 個の和からなる。

x <- DescTools::ConDisPairs(M) と置けば OR_G = xC/xD　である。

A.4. Forest 図実例

実例として、Agresti[1] の Table 8.9, 8.21, 11.19, 12.3 の行列群、また 13.10 を 3 通りに見た場合、さらに Agresti[3] の p.61 の例を図 A に提示する。

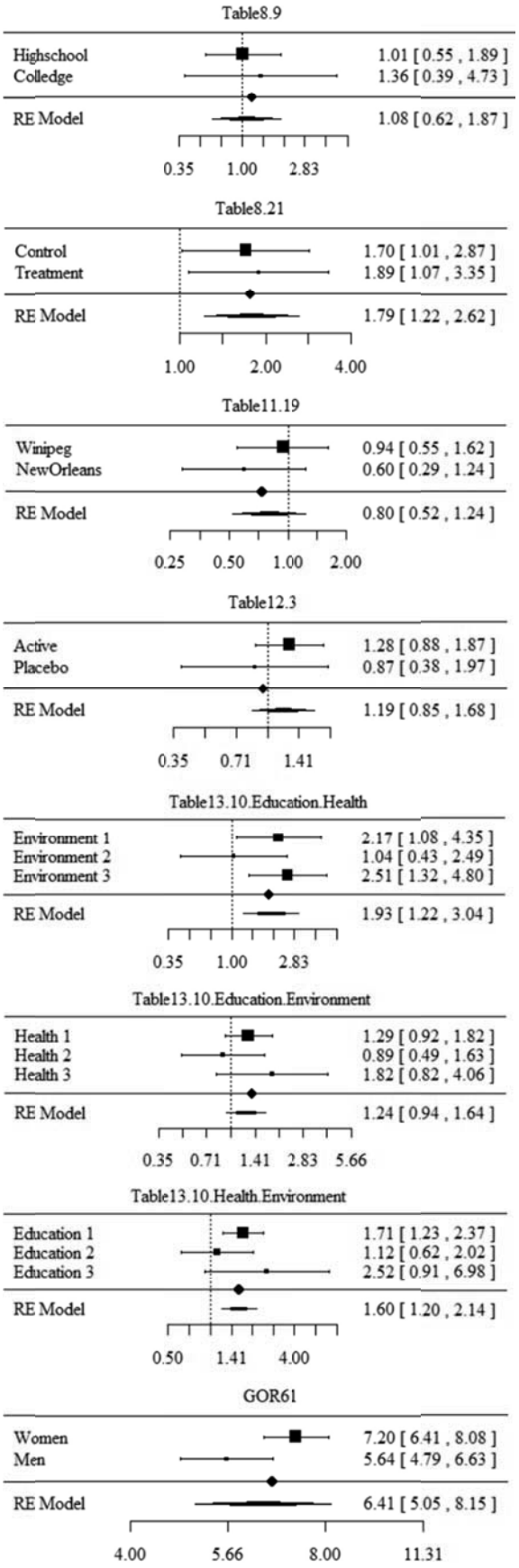

図 A　Agresti の書籍・論文にある実例