

医学統計勉強会

第5回 比率と分割表

帝京大学臨床研究センター（TARC）・帝京大学大学院公衆衛生学研究科 共催

帝京大学大学院公衆衛生学研究科

宮田 敏

カテゴリカル変数 (categorical variable) は、カテゴリーの集まりから構成される測定尺度の一つ。

- **名義変数** (nominal variable)：順序を持たないカテゴリ変数
- **順序変数** (ordinal variable)：順序を持つカテゴリ変数

カテゴリカルデータの要約:

数量的要約

- **度数** (frequency)：カテゴリに属する観測値の個数
- **相対度数** (relative frequency, %)：度数／総観測値数

必ず**度数 (%)** の形で提示する。変数によって欠測値の数が高くなるとき、片方だけでは各水準の相対的な大きさをうまく表現できない。

視覚的要約

(相対) 度数分布表を**棒グラフ**で表す。 **ヒストグラム**。

比率と分割表

疾患の発症率など、物事の頻度 (frequency) を議論する際、以下の三つの概念を使い分ける。

- **比 (ratio)** : A, B ($\neq 0$) が存在するとき, A/B を比という. A と B は互いを含まない.
 - 例 : 性比. $BMI = \text{体重} / \text{身長}^2$
- **割合 (proportion)** : A, B ($\neq 0$) が存在し, 分子 A が分母 B に含まれるとき, A/B を割合という. $0 \leq \text{割合} \leq 1$.
比率 \equiv 割合
- **率 (rate)** : 単位時間あたりのイベントの発生割合.
 - 率 = イベント発生件数 / 延べ観察時間

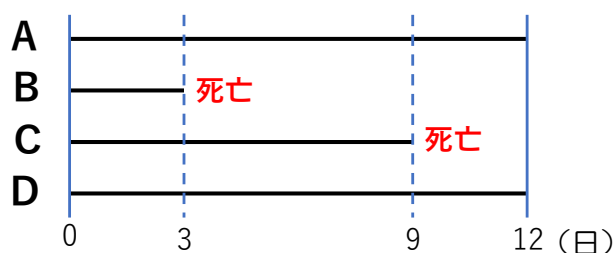
2024/11/13

医学統計勉強会 第5回 比率と分割表

3

割合と率 :

例 : 4匹のマウスの生存時間を観察した。



死亡割合 : イベント数 / 標本数 = $2/4 = 0.5$

死亡率 (人日法) : イベント数 / 延べ観察時間
 $= 2 / (12 + 3 + 9 + 12) = 2/36 = 0.0556$

「率」は、時間単位によって値が変わることに注意。

2024/11/13

医学統計勉強会 第5回 比率と分割表

4

比率の比較の尺度

二つのカテゴリカル変数の、水準（レベル）の組み合わせの頻度は**分割表 (contingency table)** の形にまとめられる。

	(-)	(+)
reference	a	b
intervention	c	d

(1) リスク差 RD (risk difference) : $p_1 - p_0 = \frac{a}{a+b} - \frac{c}{c+d}$
寄与危険度(attributable risk) (比較臨床試験では有効率の差)

(2) リスク比 RR (risk ratio) : $p_1/p_0 = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$
相対危険度(relative risk)

(3) オッズ比 OR (odds ratio) : $\frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{ad}{bc}$

2024/11/13

医学統計勉強会 第5回 比率と分割表

5

実際の計算方法

問題：フラミンガム研究の結果の一部

健常な白人男性を18年間追跡し、研究開始時の収縮期血圧とその後18年間での冠状動脈性疾患(CHD)の発生との関係を45-59歳に限ってまとめた結果が次の表の形にまとめられている。

		18年間でのCHD発生		
		なし	あり	計
収縮期 血圧	165mmHg未満	385	107	492
	166mmHg以上	57	39	96
計		442	146	242

2024/11/13

医学統計勉強会 第5回 比率と分割表

6

相対危険度 : $RR = P_1/P_0 = (39/(57+39)) / (107/(385+107)) = 0.406 / 0.218 = 1.86$

→したがって、収縮期血圧が165mmHg以上の人は18年間のCHDリスクが1.86倍高くなると解釈される。

寄与危険度 : $AR = P_1 - P_0 = (39/(57+39)) - (107/(385+107)) = 0.406 - 0.218 = 0.188$

→したがって、収縮期血圧が165mmHg以上の人は18年間のCHDリスクが18.8%増加すると解釈される。あるいは収縮期血圧が165mmHg以上の人が100人いると18年間のCHD発生が18.8人増加すると解釈される。

オッズ比 : $OR = (P_1/(1-P_1))/(P_2/(1-P_2)) = (385(39))/(107(57)) = 2.462$

→ $OR > 1$ なら、イベントのリスク増。 $OR < 1$ なら、リスク減。

一般に、イベントの発生確率が p である母集団から、 n 個のサンプルを無作為に抽出したとき、そのイベントが x 回出現する確率は**二項分布** (binomial distribution), $B(n, p)$ で与えられる。

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \dots, n$$

$$\hat{p} = \frac{X}{n} \sim N \left(p, \sqrt{\frac{p(1-p)}{n}} \right)$$

$$\text{Confidence interval : } \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Clopper-Pearson正確信頼区間

前ページの信頼区間と仮説検定はよく用いられるが、標本数が十分大きいときに適用される近似を用いている。（通常 $np > 5$ 程度）また、 $\hat{p} = 0$ あるいは 1 となったとき、信頼区間が構成出来ない。

Clopper-Pearson正確信頼区間

$$\left(\frac{x}{(n-x+1)F_1 + x}, \frac{(x+1)F_2}{(x+1)F_2 + n-x} \right)$$

ただし、

$F_1 = F(\alpha/2, 2(n-x+1), 2x) : df1=2(n-x+1), df2=2x$ の F 分布の上側 $100(\alpha/2)\%$ 点。
 $F_2 = F(\alpha/2, 2(x+1), 2(n-x)) : df1=2(x+1), df2=2(n-x)$ の F 分布の上側 $100(\alpha/2)\%$ 点。

小標本の下での比率の検定

標本数が少ない、あるいは比率の推定値が 0, 1 に極端に近いときは、信頼区間の場合と同様に仮説検定でも正規近似に基づく方法は使えない。
二項分布に基づく正確検定を行う必要がある。

例： $x = 0, n = 10, \hat{p} = 0/10 = 0$

```
> binom.test(0, 10)
```

```
Exact binomial test
```

```
data: 0 and 10
```

```
number of successes = 0, number of trials = 10, p-value = 0.001953
```

```
alternative hypothesis: true probability of success is not equal to 0.5
```

```
95 percent confidence interval:
```

```
0.0000000 0.3084971
```

```
sample estimates:
```

```
probability of success
```

```
0
```

フラミンガム研究：収縮期血圧165mmHg未満群のCHD発症率

正規近似による
信頼区間

```
> prop.test(107, (107 + 385))

1-sample proportions test with continuity correction

data: 107 out of (107 + 385), null probability 0.5
X-squared = 155.95, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1823444 0.2571139
sample estimates:
      p 
0.2174797

> binom.test(107, (107 + 385))

Exact binomial test

data: 107 and (107 + 385)
number of successes = 107, number of trials = 492, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1818018 0.2565924
sample estimates:
probability of success
      0.2174797
```

Clopper and Pearson
正確信頼区間

$$X_1 \sim B(n_1, p_1), X_2 \sim B(n_2, p_2)$$

Risk Difference (RD) : $(p_1 - p_2)$

$$(\hat{p}_1 - \hat{p}_2) \sim N \left((p_1 - p_2), \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right)$$
$$\text{CI} : \left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

Risk Ratio (RR) : p_1/p_2

$$\log(\hat{\text{RR}}) = \log(\hat{p}_1/\hat{p}_2) \sim N \left(\log(\text{RR}), \frac{1}{a} + \frac{1}{b} - \frac{1}{c} - \frac{1}{d} \right)$$
$$\text{CI} : \exp \left(\log(\text{RR}) \pm z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{c} - \frac{1}{d}} \right)$$

$X_1 \sim B(n_1, p_1), X_2 \sim B(n_2, p_2)$

Odds Ratio (OR) : $(p_1(1-p_1))/ (p_2(1-p_2))$

$\log(\hat{OR}) = \log((\hat{p}_1(1 - \hat{p}_1))/(\hat{p}_2(1 - \hat{p}_2))) \sim N \left(\log(OR), \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)$

CI : $\exp \left(\log(\hat{OR}) \pm z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right)$

Notice the expected structure of the data to be given to 'epitab':

Exposure	Disease	
	No (ref)	Yes
Level 1 (ref)	a	b
Level 2	c	d

```
> #install.packages("epitools")
> library(epitools)
>
> mtx <- matrix(c(385, 107, 57, 39), nrow=2, ncol=2, byrow=TRUE,
+               dimnames=list(c("SBP.le.165", "SBP.ge.165"),
+                               c("nonCHD", "CHD")))
> print(mtx)
      nonCHD CHD
SBP.le.165 385 107
SBP.ge.165  57  39
> epitab(mtx, method="riskratio")
$tab
      nonCHD      p0 CHD      p1 riskratio      lower      upper      p.value
SBP.le.165 385 0.7825203 107 0.2174797 1.0000000      NA      NA      NA
SBP.ge.165  57 0.5937500  39 0.4062500 1.867991 1.391835 2.507042 0.0001671871

$measure
[1] "wald"

$conf.level
[1] 0.95

$pvalue
[1] "fisher.exact"

> epitab(mtx, method="oddsratio")
$tab
      nonCHD      p0 CHD      p1 oddsratio      lower      upper      p.value
SBP.le.165 385 0.8710407 107 0.7328767 1.0000000      NA      NA      NA
SBP.ge.165  57 0.1289593  39 0.2671233 2.461879 1.553853 3.900529 0.0001671871

$measure
[1] "wald"

$conf.level
[1] 0.95

$pvalue
[1] "fisher.exact"
```

イベント率（人年法）の信頼区間と検定

イベント率＝イベント発生件数/延べ観察時間

```
> ##Examples from Rothman 1998, p. 238
> bc <- c(Unexposed = 15, Exposed = 41)
> pyears <- c(Unexposed = 19017, Exposed = 28010)
> dd <- matrix(c(41,15,28010,19017),2,2)
> dimnames(dd) <- list(Exposure=c("Yes","No"), Outcome=c("BC","PYears"))
>
> rateratio.wald(dd)
$data
      Outcome
Exposure BC PYears
Yes    41 28010
No     15 19017
Total 56 47027

$measure
      rate ratio with 95% C.I.
Exposure estimate  lower  upper
Yes 1.0000000      NA      NA
No 0.5388632 0.2982792 0.9734956

$p.value
      two-sided
Exposure midp.exact  wald
Yes      NA      NA
No 0.03545742 0.03736289
```

イベント率の比と95%信頼区間

イベント率の比較の検定の p-value
Midp-exact法とWald法の二つがある

2024/11/13

医学統計勉強会 第5回 比率と分割表

15

カテゴリデータの要約

分割表 (Contingency table), クロス集計表 (cross tabulation table) : 2種類のカテゴリデータの水準の組み合わせごとに、度数を求めた表.

(例)

2×2分割表

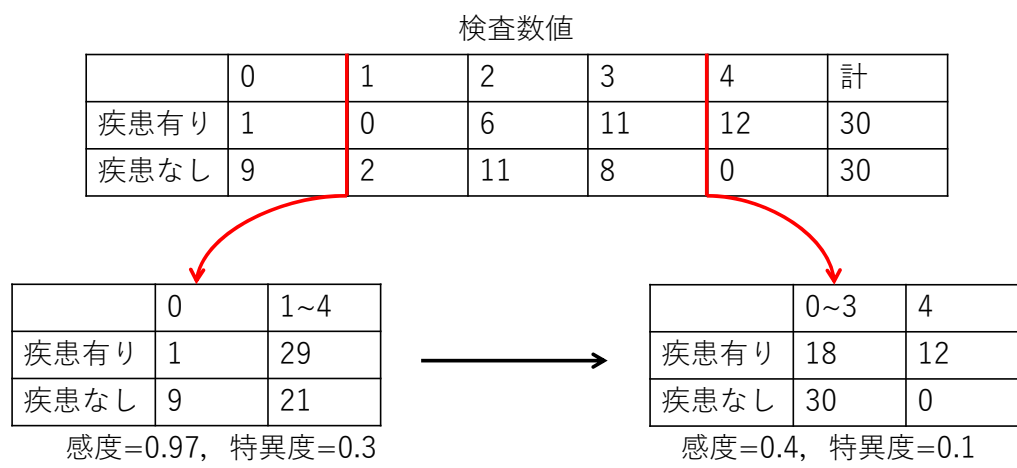
	疾患有り	疾患なし
要因陽性	a	b
要因陰性	c	d

2024/11/13

医学統計勉強会 第5回 比率と分割表

16

検査数値が、三段階以上もしくは実数値をとる場合、カットポイントの
とる場所により、複数の 2×2 分割表にデータをまとめることが出来る。
検査数値の順序に従って分割表を作り、感度、特異度が求める。

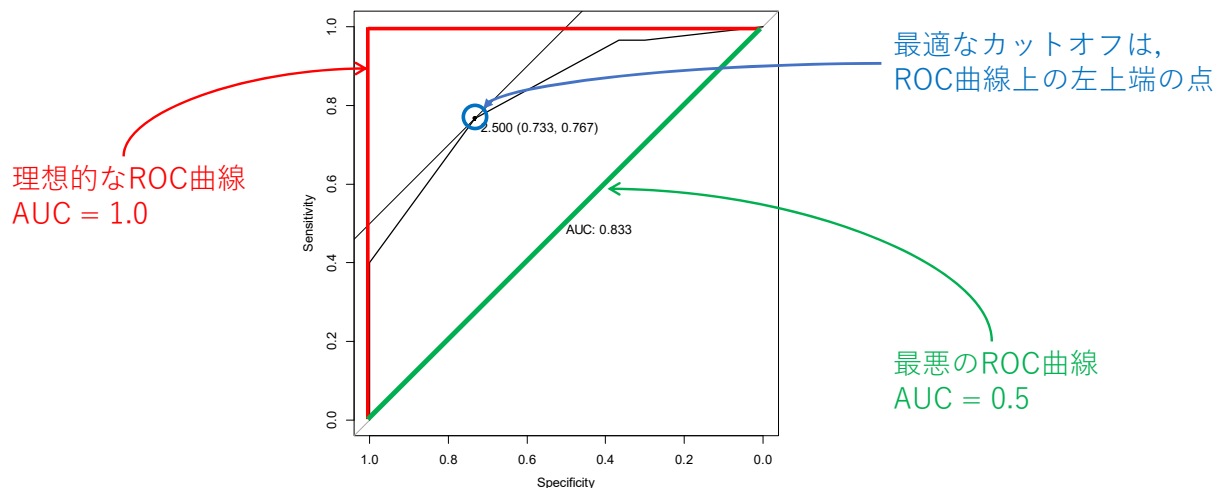


柳川 堯, 木 由布子 (著)「バイオ統計の基礎—医薬統計入門」近代科学社 (2010/02)

前項のように、検査数値の可能なカットポイントの順に、一連の分割表を作り感度、特異度が計算出来る。縦軸に感度、横軸に (1-特異度) をとったグラフを**ROC曲線** (Receiver Operating Characteristic curve) と呼び、ROC曲線の下を面積を **AUC** (Area Under Curve) あるいは**C統計量** (C statistic) と呼ぶ。

感度は 1 に近く、特異度は 1 に近い = (1 - 特異度) は 0 に近いことが望ましいわけだから、AUCは大きいほうが望ましい。AUCは $0.5 \leq AUC \leq 1.0$ となることが示される。

- 感度，特異度共に大きい(1.0に近い)方が望ましい。
- ROC曲線の左上端の点に対応する検査数値が，最適な分岐点。
- ロジスティック回帰の適合度の評価にも用いる。



要因の有無により疾患を予測するとき，その精度を測るために以下の概念を定義する。

- **正答率 (accuracy rate)** : $(a + d) / (a + b + c + d)$
- **感度 (sensitivity)** : 疾患有りのうち，陽性と判断される割合. $a / (a + c)$
疾患有りなのに陰性と判断される割合を**偽陰性率 (false negative rate)**と呼ぶ。
- **特異度 (specificity)** : 疾患なしのうち，陰性と判断される割合. $d / (b + d)$
疾患なしなのに陽性と判断される割合を**偽陽性率 (false positive rate)**と呼ぶ。
- **陽性的中率 (positive predictive rate)** : 陽性のうち疾患有りの割合.
 $a / (a + b)$
- **陰性的中率 (negative predictive rate)** : 陰性のうち疾患なしの割合.
 $d / (c + d)$

```

> ## 分割表: 正答率、感度、特異度、陽性的中率、陰性的中率 ##
>
> coords(tbl.roc, x="best", ret=c("accuracy", "1-sensitivity", "sensitivity",
+ "1-specificity", "specificity", "ppv", "npv"))
      accuracy 1-sensitivity sensitivity 1-specificity specificity
threshold    0.75      0.2333333    0.7666667    0.2666667    0.7333333
      ppv      npv
threshold 0.7419355 0.7586207

```

分割表の検定（独立性の検定）

	6カ月以内死亡	6カ月以上生存
牛乳抗体陽性	29	80
牛乳抗体陰性	10	94

$$\hat{p}_1 = 29/(29 + 80) = 0.266$$

$$\hat{p}_2 = 10/(10 + 94) = 0.0962$$

帰無仮説 $H_0: p_1 = p_2$ 母比率が一定.

対立仮説 $H_1: p_1 \neq p_2$ 母比率が異なる.

もし、二つの変数に**関連がなければ**、グループによらず（=v陽性でも陰性でも），母比率（= 6カ月以内に死亡する確率）は**一定**のはず ⇒

帰無仮説 H_0

関連があれば、母比率が**異なる** ⇒ **対立仮説** H_1

2×2分割表に限り、**片側仮説**が可能. $H_1: p_1 > (<) p_2$

分割表の検定（独立性の検定） 続き

χ^2 検定 (chi-squared test) : サンプル数が多いとき, 検定統計量の分布が χ^2 分布で近似されることを利用した検定. 分割表の度数は, 最低 5 は必要とされる. 必ず Yates の連続補正 (Yates's continuity correction) を行う.

Fisher の直接法 (Fisher's exact test) : サンプル数によらず, 正確な p -value を計算できる検定.

例：高血圧を合併した安定期慢性心不全患者に対するオルメサルタンの有効性に関する薬物介入臨床試験（SUPPORT試験）

	Olmesartan (+)	Olmesartan (-)
β -Blocker (+)	405	416
β -Blocker (-)	173	152

Fisher の直接法 : $p = 0.2388$

χ^2 検定 (Yates の連続補正あり) : $p = 0.2606$

χ^2 検定 (Yates の連続補正なし) : $p = 0.2339$

分割表の検定の結果の提示

例：

各群の総数を明示する

Name	Statin=0 (n=2933)	Statin=1 (n=1803)	p-value
HxSmoke01	1243 (42.4%)	864 (47.9%)	0
HxHFad	1616 (55.1%)	853 (47.3%)	0
HxHT	2200 (75%)	1472 (81.6%)	0
indx.DM	866 (29.5%)	786 (43.6%)	0
Cancer	380 (13%)	174 (9.7%)	0.001

カウントとパーセントを
両方提示する。

検定のp値（Fisher's exact test
が基本）を、必ず明示する。

対応のあるデータに対する比率の検定

同じサンプルから、対照群と処理群のデータを対応のあるデータ（1対1マッチングデータ）として取られる場合の有効率の比較を考える。

	control	treatment
Item 1	+	+
Item 2	+	-
Item 3	-	-
Item n	+	-



	Treatment (+)	Treatment (-)
Control (+)	a	b
Control (-)	c	d

(i, j) セルの確率を P_{ij} とすると、分割表全体の確率は以下ようになる。

	Treatment (+)	Treatment (-)
Control (+)	P_{11}	P_{12}
Control (-)	P_{21}	P_{22}

対照群の有効率は $P_{\text{cnt}} = P_{11} + P_{12}$ 処置群の有効率は $P_{\text{trt}} = P_{11} + P_{21}$ であるから、検定する仮説は以下の通り。

$$H_0 : P_{\text{cnt}} = P_{\text{trt}} \text{ vs. } H_1 : P_{\text{cnt}} \neq P_{\text{trt}} \Leftrightarrow H_0 : P_{12} = P_{21} \text{ vs. } H_1 : P_{12} \neq P_{21}$$

すなわち、(1, 1), (2, 2) セルは無視してよい。 $\theta = P_{12} / (P_{12} + P_{21})$ とすれば、

$$H_0 : \theta = 1/2 \text{ vs. } H_1 : \theta \neq 1/2$$

対応のあるデータに対する比率の検定

χ^2 分布を用いてこの検定を行う方法を、**McNemar検定**と呼ぶ。
McNemar検定は近似検定であり、連続補正が行われる。**二項分布**を用いて、正確な p 値を求める検定も可能である（こちらが望ましい）。

例：2004年と2008年のアメリカ大統領選挙における一般住民調査（男性）における投票行動

	2008 年 民主党	2008 年 共和党
2004 年 民主党	175	16
2004 年 共和党	54	204

Agresti, A. 2013. *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: Wiley. P. 414, Table 11.1

```

> tbl2
      2008 election
2004 election Democrat Republican
Democrat      175      16
Republican     54     188
> binom.test(16, (16 + 54))# 二項分布による正確検定

Exact binomial test

data: 16 and (16 + 54)
number of successes = 16, number of trials = 70, p-value = 5.854e-06
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.136657 0.344475
sample estimates:
probability of success
      0.2285714

> mcnemar.test(tbl2)# McNemar検定 (連続補正あり)

McNemar's Chi-squared test with continuity correction

data: tbl2
McNemar's chi-squared = 19.5571, df = 1, p-value = 9.764e-06

> mcnemar.test(tbl2)# McNemar検定 (連続補正なし)

McNemar's Chi-squared test with continuity correction

data: tbl2
McNemar's chi-squared = 19.5571, df = 1, p-value = 9.764e-06

```

シンプソンのパラドックス

分割表で考える2つの要因の双方に影響を与える因子を、**交絡因子 (confounding factor)** と呼ぶ。交絡因子を無視して検定を行うと、本来存在した関連が見えなくなったり、その逆が起こることがある。

いま、A, B 2つの要因の間の分割表を考えているとする。第3の要因Zの値によって、標本が二つの層に分けられたとする。それぞれの分割表は以下の通りとする。

Z1

	B	Not B
A	80	16
Not A	160	160

$$\text{Odds ratio} = (80/16) / (160/160) = \mathbf{5}$$

Z2

	B	Not B
A	50	452
Not A	10	452

$$\text{Odds ratio} = (50/452) / (10/452) = \mathbf{5}$$

交絡因子Zを無視して、二つの分割表を統合してしまうと、

	B	Not B
A	130	468
Not A	170	612

$$\text{Odds ratio} = ((80+50)/(16+452))/((160+10)/(160+452)) = \mathbf{1}$$

この分割表のオッズ比を計算すると、**オッズ比=1**となる。交絡因子Zで層別しておけばオッズ比=5、すなわち要因A有りのほうが5倍オッズが高かったことが分かるが、交絡因子を無視したためにリスクが全く見えなくなっている。

(参照：柳川堯「離散多変量データの解析」(1986) 共立出版)

⇒ **Mantel-Haenszel検定**

分割表で考える A, B 二つの要因の、双方に影響を与える第3の要因 C が**交絡因子**であった。

要因 A が原因、要因 B が結果である時、因果関係の連鎖の途中にある要因 D は、交絡因子にはなりえないことに注意する。



何が交絡因子になるかは、統計学だけでは答えられない。研究対象に関する科学的知識が必要。

交絡因子が判明した場合，層ごとに分割表を作り，それらを統合して検定を行う．この検定を，**Mantel-Haenszel検定**という．

```
> tbl <- array(c(80,160,16,160, 50, 10, 452, 452), dim=c(2,2,2))
> tbl
, , 1
      [,1] [,2]
[1,]    80    16
[2,]   160   160
, , 2
      [,1] [,2]
[1,]    50   452
[2,]    10   452

> mantelhaen.test(tbl)

Mantel-Haenszel chi-squared test with continuity correction

data:  tbl
Mantel-Haenszel X-squared = 57.2109, df = 1, p-value = 3.915e-14
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 3.205444 7.799232
sample estimates:
common odds ratio
5
```