H28.12.20

疫学統計セミナー

疫学と統計の基礎からロジスティック回帰

第2回：交絡因子とその調整方法

担当：　茅野光範
グローバルアグロメディシン研究センター
獣医学研究部門

メール：kayano@、内線5521

セミナー資料：　http://www.obihiro.ac.jp/~kayano/epi-stat/

このセミナーについて

内容：　疫学と統計を復習し、交絡因子とその調整方法、
　　ロジスティック回帰等を紹介する

目標：　交絡因子調整の検定やロジスティック回帰を理解し、
　　R等で実行できるようになる！

ポイント：<u>疾病の規定要因（リスク因子）を正しく同定する</u>

日時（予定）：　毎月下旬月曜or火曜の午後5時から1.5時間程度

スケジュール（予定）：　全４回

第1回 (11/28)：　疫学と統計の基礎
第2回 (12/20)：　交絡因子とその調整方法
第3回 (1/24?)：　統計ソフトRの基礎と応用
第4回 (2/21?)：　ロジスティック回帰（仮）＋α？

今日の内容：交絡因子とその調整方法

- 前回の復習
  目標・内容、 リスク因子の同定と交絡因子の影響
  研究方法に応じたデータレイアウト
  暴露効果の指標と推測 （交絡の影響は無視）

- 交絡の調整方法1 研究方法で調整する （考え方の紹介）
  (1) 因子範囲の制限
  (2) マッチング

- 交絡の調整方法2 解析方法で調整する
  (1) 層化解析（Stratification）：マンテル・ヘンツェル検定
  (2) 回帰分析 ← 第4回目にやります

確認

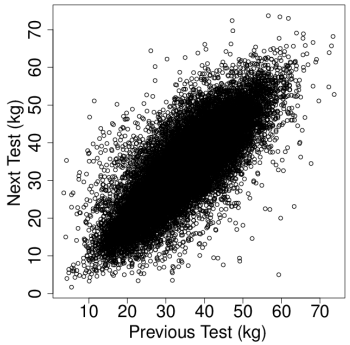
データの種類に応じた2因子の関連性の評価

データの種類

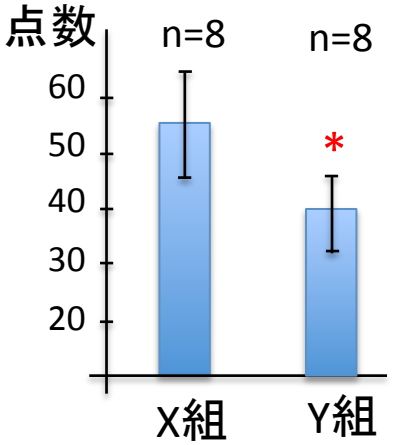
1. 量的（数値）：身長、体重、年齢、、、
2. 質的（数値でない）：性別、品種、暴露や疾病の有無、、

関連性の評価

1. 量的vs量的：　相関係数、回帰分析
2. 量的vs質的：　t検定、分散分析、Tukeyの方法
3. 質的vs質的：　カイ二乗検定、マンテル-ヘンツェル検定



ホルスタイン種2,499頭、4,391回の泌乳の約3万ペアの乳量のプロット(r = 0.820)



$p<0.05$

	E	$\overline{E}$	合計
D	a	b	m_1
$\overline{D}$	c	d	m_0
合計	n_1	n_0	n

前回の復習

目標と内容
リスク因子の同定と交絡因子の影響
研究方法に応じたデータレイアウト
暴露効果の指標と推測

復習

前回の目標と内容

目標：コホート研究（追跡、dynamic/fixed）と症例対照研究（case/control）において、暴露が疾病に関与しているかどうかを検証（検定）する。（ただし、交絡は無視する）

内容：

- はじめに
  疫学とは何か、有名な疫学研究、トピック、リスク因子の同定
- 研究方法（研究デザイン）と疾病のタイミング
  コホート研究（follow-up研究）、症例対照研究
- 疫学で用いられる指標と統計的推測
  罹患率（incidence ratio）、有病率（prevalence）
  リスク比、オッズ比、カイ二乗検定、信頼区間

復習

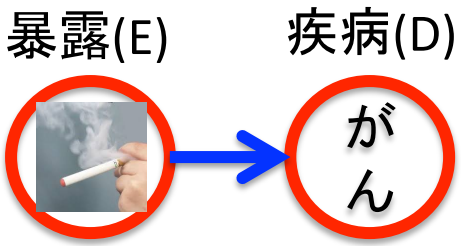
リスク因子の同定と交絡因子の影響

（本セミナーのテーマ）

暴露（Exposure）と疾病（Disease）の関係は？

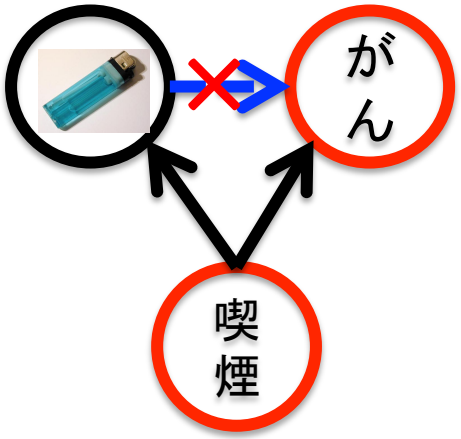
暴露：特定の状態のこと。例：毎日30分散歩する

例1：喫煙は肺がんのリスク因子か？　→　Yes!



交絡因子(Confounder)

EのDへの影響をゆがめてしまう要因

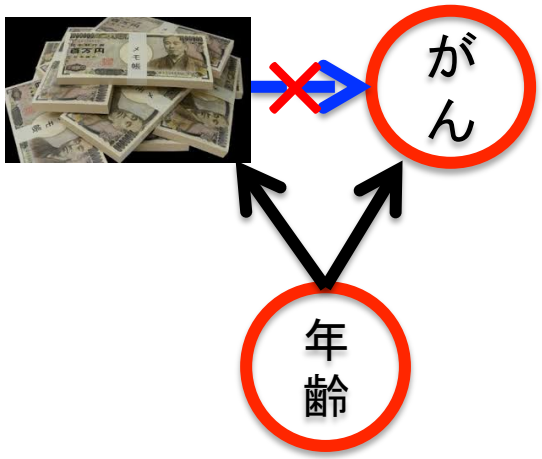


例2：ライター所持は肺がんのリスク因子か？　→　No..

喫煙が交絡因子

例3：年収は肺がんのリスク因子か？　→　No...

年齢が交絡因子



通常、年齢と性別は交絡因子になる。

調整（補正）する必要がある！　⇒　今日やります

復習

リスク因子同定のための表（2×2分割表）

	例：喫煙 E：暴露あり （E=1）	例：非喫煙 $\overline{E}$：暴露なし （E=0）	合計
D：疾病あり （D=1）	a	b	$m_1=a+b$
$\overline{D}$：疾病なし （D=0）	c	d	$m_0=c+d$
合計	$n_1=a+c$	$n_0=b+d$	$n=n_1+n_0$ $=m_1+m_0$ $=a+b+c+d$

この行（$\overline{D}$）は、
研究方法によっては
他の項目になる

a, b, c, d：
対応する人数

復習

リスク因子同定のための表（データレイアウト）

- Fixed Cohort
  or case-control

変更点！

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a	b	$m_1=a+b$
$\overline{D}$：疾病なし	c	d	$m_0=c+d$
	n_1	n_0	$n=n_1+n_0$ $=m_1+m_0$ $=a+b+c+d$

a, b, c, d：対応する人数

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a	b	$m_1=a+b$
Population at risk	n_1	n_0	$n=n_1+n_0$

- Dynamic Cohort

	E：暴露あり	$\underline{E}$：暴露なし	合計
D：疾病あり	a	b	$m_1=a+b$
Population Time (Person・Year)	L_1	L_0	$L=L_1+L_0$

$p_0=L_1/(L_1+L_0)$
$q_0=L_0/(L_1+L_0)$

復習

暴露効果の指標と統計的推測
準備：仮説検定

帰無仮説H_0：暴露効果なし
対立仮説H_a：暴露効果あり

<u>p値を求めるために必要なこと</u>

1. 検定統計量を決める
2. 帰無仮説H_0のもとで、検定統計量の分布を求める
3. データから求めた検定統計量の値が、
   2で求めた分布のどこにあるか（どのくらい外れているか？）
   で、p値が求まる！　（統計ソフトの内部でやっていること）

重要!

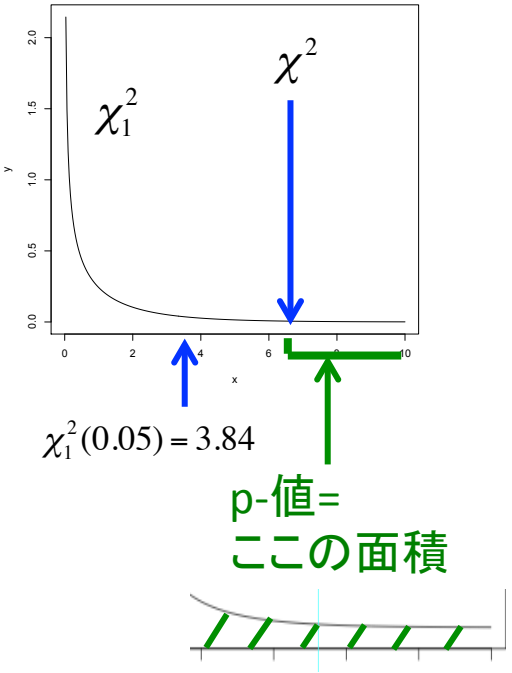
リスク因子同定のための検定:カイ二乗検定

- 帰無仮説H_0:暴露効果なし
- 対立仮説H_a:暴露効果あり
- <u>検定統計量</u> $\chi_{MH}^2 = \dfrac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0}$

大きいほど暴露効果あり！

オッズ比=ad/bcとad-bcの値は比例する
オッズ比=1のとき、ad-bc=0となる

$\sim \chi_1^2$:自由度1の<u>カイ二乗分布</u>



	E:暴露あり (E=1)	E:暴露なし ($\bar{E}$=0)	合計
D:疾病あり (D=1)	a	b	m_1=a+b
D:疾病なし ($\bar{D}$=0)	c	d	m_0=c+d
合計	n_1=a+c	n_0=b+d	$n=n_1+n_0$ $=m_1+m_0$ =a+b+c+d

a, b, c, d:対応する人数

復習

暴露効果の指標と推測：fixed cohortやcase/control

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a	b	$m_1=a+b$
$\overline{D}$：疾病なし	c	d	$m_0=c+d$
	n_1	n_0	

累積罹患率：
Cumulative incidence
罹患リスク (incidence risk)

$$CI_1 = \frac{a}{n_1}\ (\text{暴露あり});\quad CI_0 = \frac{b}{n_0}\ (\text{暴露なし})$$

リスク比：
risk ratio

$$RR = \frac{CI_1}{CI_0} = \frac{a/n_1}{b/n_0}$$

リスク差：
difference

$$RD = CI_1 - CI_0 = \frac{a}{n_1} - \frac{b}{n_0}$$

オッズ比：
odds ratio

$$OR = \frac{ad}{bc} \left(= \frac{CI_1/(1-CI_1)}{CI_0/(1-CI_0)} = \frac{a/b}{c/d} \right)$$

帰無仮説H_0：暴露効果なし　$CI_1=CI_0 \Leftrightarrow RR=1 \Leftrightarrow RD=0 \Leftrightarrow OR=1$

検定統計量 $\chi_{MH}^2 = \dfrac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0} \sim \chi_1^2$　(under H_0)

復習 暴露効果の指標と推測：dynamic cohortの場合

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a	b	$m_1=a+b$
Population Time (Person・Year)	L_1	L_0	$L=L_1+L_0$

$p_0=L_1/(L_1+L_0)$
$q_0=L_0/(L_1+L_0)$

罹患率：Incidence rate　$IR_1=\dfrac{a}{L_1}$（暴露あり）；$IR_0=\dfrac{b}{L_0}$（暴露なし）

罹患率比：ratio　$IRR=\dfrac{IR_1}{IR_0}=\dfrac{a/L_1}{b/L_0}=\dfrac{a\,L_0}{b\,L_1}$

罹患率差：difference　$IRD=IR_1-IR_0=\dfrac{a}{L_1}-\dfrac{b}{L_0}$

帰無仮説H0：暴露効果なし　$IR_1=IR_0 \Leftrightarrow IRR=1 \Leftrightarrow IRD=0$
対立仮説Ha：暴露効果あり　$IR_1>IR_0 \Leftrightarrow IRR>1$

検定統計量 $Z=\dfrac{a-m_1 p_0}{\sqrt{m_1 p_0 q_0}} \sim N(0,1)$　（under H_0）

or　$\chi^2=Z^2 \sim \chi_1^2$　（under H_0）

復習

暴露効果指標の信頼区間

テイラー展開による近似（large sample）

- リスク比RRの95%信頼区間 = $RR \ \exp\left\{\pm 1.96\sqrt{\frac{(1-CI_1)\,(1-CI_0)}{n_1CI_1\ n_0CI_1}}\right\}$
- オッズ比ORの95%信頼区間 = $OR \ \exp\left\{\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right\}$

検定ベース

- 差θの95%信頼区間 = $\theta \pm 1.96\sqrt{\theta^2/\chi_{MH}^2} = \theta\,(1 \pm 1.96/\sqrt{\chi_{MH}^2})$
- 比θの95%信頼区間 = $\theta \exp\left\{\pm 1.96\sqrt{(\log\theta)^2/\chi_{MH}^2}\right\}$

$$= \theta^{1 \pm 1.96/\sqrt{\chi_{MH}^2}}$$

交絡因子が無い（無視できる）

⇒　リスク因子の同定は難しくない

交絡因子がある

⇒　その影響を調整・補正する！

交絡の調整方法1：研究方法で調整する

(1) 因子範囲の制限

(2) マッチング

交絡因子とその調整方法1:
研究方法で調整する

(1) <u>因子範囲の制限</u>

<u>方法</u>:交絡因子の範囲を制限する(例:高齢のみの解析)

<u>利点</u>:コントロール可能、便利、安い、簡単

<u>欠点</u>:一般的な推測が出来ない、制限範囲内で交絡が残る可能性がある

(その場合は、層化解析を行うのが良い)

(2) <u>マッチング</u>

<u>方法</u>:交絡因子毎(例:性別、年齢、体重)に患者と対照をマッチングさせて、対応のある患者群・対照群を設定する。

a) 1人の患者に対して1人の対照をマッチングさせる

b) 1人の患者に対して複数の対照をマッチングさせる

<u>利点</u>:スマートな方法、効率がいい

<u>欠点</u>:コスト・手間がかかる、柔軟性に乏しい

<u>注意</u>:データレイアウトもオッズ比の計算も、<u>**使う検定**</u>も他とは異なる!

<u>**対応のあるカイ二乗検定(McNemar検定)**</u>

参考:KKM Ch.18や『医学への統計学』(第3版)p108-111, 130-132

交絡の調整方法2：解析方法で調整する

層化解析（Stratification analysis）
マンテル・ヘンツェル検定（Mantel-Haenszel test）

Motivated data：
冠状動脈疾患のリスク因子の同定

問：体内のカテコールアミン（ストレスで増加する化学物質）レベル（CAT：High/Low）はその後7年間の冠状動脈疾患の罹患（CHD：yes/no）に影響しているか？

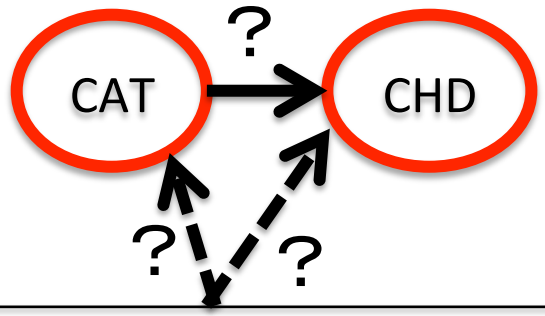
データ

	High CAT	Lo CAT	合計
CHD	27	44	71
No CHD	95	443	538
	122	487	609

リスク比RR= CI_1/ CI_0
= (27/122) / (44/487) = 2.45

χ_{MH}^2 = 16.22 ($p<0.001$)

RR>1！（significant）
⇔CATはCHDのリスク因子！



他の観測変数
AGE：年齢（数値）
ECG：心電図異常の有無（yes/no）
CHL：血清コレステロール値（数値）
DBP：血圧（数値）
QTI：ケトレー指数（=BMI、数値）
SMK：喫煙状況（喫煙/非喫煙）
SES：経済レベル（数値。12(high) – 84(low)）
OCC：職種（農家/農家でない）
HPT：高血圧（high/low）

層化解析（Stratification）

患者群と対照群を交絡因子の暴露状況（例：年齢high/low）について同一の層に分け、暴露効果の有無を調べる。

	High CAT	Lo CAT	合計
CHD	27	44	71
No CHD	95	443	538
	122	487	609

リスク比RR= 2.45

$\chi_{MH}^2 = 16.22$ ($p<0.001$)

年齢で分ける（55歳未満・以上）

<55歳

	High CAT	Lo CAT
CHD	4	24
No CHD	21	309

RR=2.22, $\chi_{MH}^2 = 2.49$ ($p=0.06$)

≧55歳

	High CAT	Lo CAT
CHD	23	20
No CHD	74	134

RR=1.83, $\chi_{MH}^2 = 4.80$ ($p=0.01$)

層化解析の条件

1. 各層（Stratum）で十分なn数がある
2. コントロール因子を適切に選べる
3. コントロール因子を適切にカテゴリカル化出来る（カテゴリカル化に意味がある、交絡の影響が残らない）

層化解析：データレイアウト

- Fixed Cohort
  or case-control

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a_g	b_g	$m_{1g}=a_g+b_g$
$\overline{D}$：疾病なし	c_g	d_g	$m_{0g}=c_g+d_g$
	n_{1g}	n_{0g}	$n_g=n_{1g}+n_{0g}$ $=m_{1g}+m_{0g}$ $=a_g+b_g+c_g+d_g$

変更点！　g=1,2,…,G（層番号）

- Dynamic Cohort

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a_g	b_g	$m_{1g}=a_g+b_g$
Population Time (Person・Year)	L_{1g}	L_{0g}	$L_g=L_{1g}+L_{0g}$
			$p_{0g}=L_{1g}/(L_{1g}+L_{0g})$ $q_{0g}=L_{0g}/(L_{1g}+L_{0g})$

⇒ カイ二乗検定を行いたい、オッズ比を求めたい！
=マンテル・ヘンツェル検定！

重要!

マンテル・ヘンツェル検定とオッズ比等

（case/control、fixed cohort等の場合）

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a_g	b_g	$m_{1g}=a_g+b_g$
$\overline{D}$：疾病なし	c_g	d_g	$m_{0g}=c_g+d_g$
	n_{1g}	n_{0g}	n_g

g=1,2,…,G
（層番号）

検定統計量

$$\chi^2_{MHS} = \frac{\left(\sum_{g=1}^{G}\frac{a_g d_g - b_g c_g}{n_g}\right)^2}{\sum_{g=1}^{G}\frac{n_{1g}n_{0g}m_{1g}m_{0g}}{(n_g-1)n_g^2}} \sim \chi^2_1$$

（under H_0）

層化しない場合

$$\chi_{MH}^2 = \frac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0} \sim \chi_1^2$$

$$OR = \frac{ad}{bc} \qquad RR = \frac{a/n_1}{b/n_0} = \frac{a\, n_0}{b\, n_1}$$

調整オッズ比

$$mOR = \frac{\sum_{g=1}^{G}\frac{a_g d_g}{n_g}}{\sum_{g=1}^{G}\frac{b_g c_g}{n_g}} \left(= \frac{\sum_{g=1}^{G} W_g OR_g}{\sum_{g=1}^{G} W_g} \right)$$

各層のオッズ比
$OR_g = a_g d_g/(b_g c_g)$
の重み付き平均！

$W_g = b_g c_g/n_g$

（If $b_g c_g \neq 0$）

リスク比

$$mRR = \frac{\sum_{g=1}^{G}\frac{a_g n_{0g}}{n_g}}{\sum_{g=1}^{G}\frac{b_g n_{1g}}{n_g}}$$

マンテル・ヘンツェル検定の統計量のイメージ

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a_g	b_g	$m_{1g}=a_g+b_g$
$\overline{D}$：疾病なし	c_g	d_g	$m_{0g}=c_g+d_g$
	n_{1g}	n_{0g}	n_g

g=1,2,…,G
（層番号）

マンテル・ヘンツェル検定

$$\chi^2_{MHS}=\frac{\left(\sum_{g=1}^{G}\frac{a_g d_g-b_g c_g}{n_g}\right)^2}{\sum_{g=1}^{G}\frac{n_{1g}n_{0g}m_{1g}m_{0g}}{(n_g-1)n_g^2}}$$

層化しない場合

$$\chi^2_{MH}=\frac{(n-1)(ad-bc)^2}{n_1n_0m_1m_0}$$

$$=\frac{(n-1)\left(\frac{ad-bc}{n}\right)^2}{\frac{n_1n_0m_1m_0}{n^2}}=\frac{\left(\frac{ad-bc}{n}\right)^2}{\frac{n_1n_0m_1m_0}{(n-1)n^2}}$$

層化しない場合の統計量χ_{MH}^2の
分子と分母を、層ごとに足し合わせている！

参考

検定統計量と分布の導出
（Case/control, fixed cohortの場合）

Dynamic cohortの場合も同様

a=Σa_gの期待値、分散をそれぞれE(A)、Var(A)とすると以下が成り立つ

$$\chi^2_{MHS} = \frac{\left(\boxed{a - E(A)}\right)^2}{Var(A)} \sim \chi^2_1 \quad \text{(under } H_0\text{)}$$

←標準正規分布とカイ二乗分布の性質

ここで、 $E(A) = \sum_{g=1}^{G} E(A_g) = \sum_{g=1}^{G} \frac{n_{1g} m_{1g}}{n_g}$ $\quad Var(A) = \sum_{g=1}^{G} Var(A_g) = \sum_{g=1}^{G} \frac{n_{1g} n_{0g} m_{1g} m_{0g}}{(n_g - 1) n_g^2}$ 、

また、 $\boxed{a - E(A)} = \sum_{g=1}^{G} a_g - \sum_{g=1}^{G} \frac{n_{1g} m_{1g}}{n_g} = \sum_{g=1}^{G} \frac{a_g n_g}{n_g} - \sum_{g=1}^{G} \frac{n_{1g} m_{1g}}{n_g} = \sum_{g=1}^{G} \left(\frac{a_g (n_{1g} + n_{0g})}{n_g} - \frac{n_{1g}(a_g + b_g)}{n_g} \right)$

$$= \sum_{g=1}^{G} \left(\frac{a_g n_{0g} - n_{1g} b_g}{n_g} \right) = \sum_{g=1}^{G} \left(\frac{a_g (b_g + d_g) - (a_g + c_g) b_g}{n_g} \right) = \sum_{g=1}^{G} \frac{a_g d_g - b_g c_g}{n_g}$$

より、結局、

$$\chi^2_{MHS} = \frac{\left(\sum_{g=1}^{G} \frac{a_g d_g - b_g c_g}{n_g} \right)^2}{\sum_{g=1}^{G} \frac{n_{1g} n_{0g} m_{1g} m_{0g}}{(n_g - 1) n_g^2}} \sim \chi^2_1 \quad \text{(under } H_0\text{)}$$

となる。

参考

調整オッズ比mORの95%信頼区間（検定ベース）

95%信頼区間 = $\left[mOR^{1-1.96/\chi_{MHS}}, mOR^{1+1.96/\chi_{MHS}} \right]$

例：mOR=1.89, χ_{MHS}^2 = 4.15のとき、χ_{MHS} = 2.037より、

mORの95%信頼区間は、

下限 = $1.89^{1-1.96/2.037}$=1.02

上限 = $1.89^{1+1.96/2.037}$=3.49

となる。

マンテル・ヘンツェル検定とリスク比

（dynamic cohortの場合）

	E：暴露あり	$\overline{E}$：暴露なし	合計
D：疾病あり	a_g	b_g	$m_{1g}=a_g+b_g$
Population Time (Person・Year)	L_{1g}	L_{0g}	$L_g=L_{1g}+L_{0g}$

g=1,2,…,G（層番号）

$p_{0g}=L_{1g}/(L_{1g}+L_{0g})$
$q_{0g}=L_{0g}/(L_{1g}+L_{0g})$

検定統計量

$$\chi^2_{MHS}=\frac{\left(a-\sum_{g=1}^{G}m_{1g}p_{0g}\right)^2}{\sum_{g=1}^{G}m_{1g}p_{0g}q_{0g}}\sim\chi^2_1$$

$a=\Sigma a_g$

（under H_0）

層化しない場合

$$\chi_{MH}^2=\frac{(a-m_1p_0)^2}{m_1p_0q_0}\sim\chi_1^2$$

$$IRR=\frac{a/L_1}{b/L_0}=\frac{a\,L_0}{b\,L_1}$$

罹患率比

$$mIRR=\frac{\sum_{g=1}^{G}\frac{a_gL_{0g}}{L_g}}{\sum_{g=1}^{G}\frac{b_gL_{1g}}{L_g}}\left(=\frac{\sum_{g=1}^{G}W_gIRR_g}{\sum_{g=1}^{G}W_g}\right)$$

$W_g=b_gL_{1g}/L_g$

各層の罹患率比
$IRR_g=a_gL_{0g}/(b_gL_{1g})$
の重み付き平均！

Rothman&Boice (1979)

例1：CAT-CHD association　データレイアウト

年齢（AGE）と心電図異常の有無（ECG）で層化解析を行う（G=4）

層1 (N.S)　<55歳、ECG=0

RR_1=2.01

	High CAT	Lo CAT	
CHD	1	17	18
No CHD	7	257	264
	8	274	282

層2 (N.S)　<55歳、ECG=1

RR_2=1.49

	High CAT	Lo CAT	
CHD	3	7	10
No CHD	14	52	66
	17	59	76

層3 (p=0.05)　≧55歳、ECG=0

RR_3=1.88

	High CAT	Lo CAT	
CHD	9	15	24
No CHD	30	107	137
	39	122	161

層4 (N.S.)　≧55歳、ECG=1

RR_4=1.54

	High CAT	Lo CAT	
CHD	14	5	19
No CHD	44	27	71
	58	32	90

Crude Data

	High CAT	Lo CAT	合計
CHD	27	44	71
No CHD	95	443	538
	122	487	609

cRR=2.45
（p<0.001）

例1： CAT-CHD association
マンテル・ヘンツェル検定、リスク比とオッズ比

$$\chi^2_{MHS} = \frac{\left(\boxed{\frac{1\cdot 257-17\cdot 7}{282}}+\frac{3\cdot 52-7\cdot 14}{76}+\frac{9\cdot 107-15\cdot 30}{161}+\frac{14\cdot 27-5\cdot 44}{90}\right)^2}{\boxed{\frac{8\cdot 274\cdot 18\cdot 264}{281\cdot 282^2}}+\frac{17\cdot 59\cdot 10\cdot 66}{75\cdot 76^2}+\frac{39\cdot 122\cdot 24\cdot 137}{160\cdot 161^2}+\frac{58\cdot 32\cdot 19\cdot 71}{89\cdot 90^2}} = \underline{4.15}$$

(p=0.02!)

$$mRR = \frac{\boxed{1\cdot\frac{274}{282}}+3\cdot\frac{59}{76}+9\cdot\frac{122}{161}+14\cdot\frac{32}{90}}{\boxed{17\cdot\frac{8}{282}}+7\cdot\frac{17}{76}+15\cdot\frac{39}{161}+5\cdot\frac{58}{90}} = 1.70$$

層1 (N.S) <55歳、ECG=0

RR_1=2.01

	High CAT	Lo CAT	
CHD	1	17	18
No CHD	7	257	264
	8	274	282

$$mOR = \frac{\boxed{1\cdot\frac{257}{282}}+3\cdot\frac{52}{76}+9\cdot\frac{107}{161}+14\cdot\frac{27}{90}}{\boxed{17\cdot\frac{7}{282}}+7\cdot\frac{17}{76}+15\cdot\frac{30}{161}+5\cdot\frac{44}{90}} = 1.89$$

mORの95%信頼区間
=[1.02,3.49]

例2：肥満による死亡リスクの評価（dynamic cohort）

450人の白人女性（53-74歳。肥満：150人、肥満でない：300人←研究開始時に決定）を8年間（1960-1967）追跡した。それぞれ少なくとも1年間追跡した。研究期間中に105人が死亡した。研究対象は60-75歳とした（1960年に58歳の人は～5年間追跡可能）

層1 (N.S) Ages 60-64

IRR_1=1.81

	Obese	Non obese	
Deaths	7	9	16
Person-Year	234.5	544.5	779

層2 (N.S) Ages 65-69

IRR_2=1.68

	Obese	Non obese	
Deaths	11	11	22
Person-Year	264.5	444.5	709

層3 (N.S.) Ages 70-74

IRR_3=1.54

	Obese	Non obese	
Deaths	12	16	28
Person-Year	200	410	610

Crude (p=0.02)

IRR=1.67

	Obese	Non obese	
Deaths	30	36	66
Person-Year	699	1399	2098

KKM 例17.4 (page 337)

例2：肥満による死亡リスクの評価（dynamic cohort）

$$E(A)=\sum_{g=1}^{G} m_{1g}p_{0g}=\frac{16\cdot 234.5}{799}+\frac{22\cdot 264.5}{709}+\frac{28\cdot 200}{610}=22.204$$

$$Var(A)=\sum_{g=1}^{G} m_{1g}p_{0g}q_{0g}=\frac{16\cdot 234.5\cdot 544.5}{799^2}+\frac{22\cdot 264.5\cdot 444.5}{709^2}+\frac{28\cdot 200\cdot 410}{610^2}=14.6825$$

より、

$$\chi^2_{MHS}=\frac{\left(a-E(A)\right)^2}{Var(A)}=\frac{(30-22.2040)^2}{14.6825}=4.14$$ (p=0.02)

肥満は死亡リスクの１つ(p=0.02)

層1 (N.S)	Ages 60-64		
IRR_1=1.81	Obese	Non obese	
Deaths	7	9	16
Person-Year	234.5	544.5	779

層2 (N.S)	Ages 65-69		
IRR_2=1.68	Obese	Non obese	
	11	11	22
	264.5	444.5	709

層3 (N.S.)	Ages 70-74		
IRR_3=1.54	Obese	Non obese	
	12	16	28
	200	410	610

演習

1.症例対照研究における層化解析
2. fixedコホート研究における層化解析

エクセルファイル：

http://www.obihiro.ac.jp/~kayano/epi-stat/

演習1: 症例対照研究における層化解析

	層1　お菓子をよく食べる			
		E	not E	合計
		歯磨きをする	しない	
虫歯あり	D	13	32	45
なし	not D	7	14	21
	合計	20	46	66

層2　お菓子を食べない			
	E	not E	合計
	歯磨きをする	しない	
D	25	8	33
not D	63	22	85
合計	88	30	118

Crude			
	E	not E	合計
	歯磨きをする	しない	
D	38	40	78
not D	70	36	106
合計	108	76	184

184人の小児の症例対照研究結果（上表）について、

暴露E（歯磨きをよくする）の疾病D（虫歯あり）への効果を検証して下さい。

（1） 各層とCrudeデータのそれぞれにおいて、罹患リスクIR、オッズ比RRを求め、カイ二乗検定を行う

（2） 調整したオッズ比mORを求め、マンテル・ヘンツェル検定を行う。

Excelでカイ二乗検定のp値を求める関数：CHISQ.DIST

使い方：p値 = 1 – CHISQ.DIST（χ_{MH}^2,1,TRUE）

『医学への統計学』（第3版）
Page226 例題13.3

演習2: fixedコホート研究における層化解析

層1　女性、年齢≦20

	E	not E	合計
D	4	30	34
not D	10	251	261
合計	14	281	295

層2　女性、年齢>20

	E	not E	合計
D	5	7	12
not D	18	61	79
合計	23	68	91

層3　男性、年齢≦20

	E	not E	合計
D	23	29	52
not D	27	102	129
合計	50	131	181

層4　男性、年齢>20

	E	not E	合計
D	19	5	24
not D	36	14	50
合計	55	19	74

Crude

	E	not E	合計
D	51	71	122
not D	91	428	519
合計	142	499	641

641人のfixedコホート研究結果（上表）について、
暴露Eの疾病Dへの効果を検証して下さい。

（1）各層とCrudeデータのそれぞれにおいて、
罹患リスクIR、リスク比RRを求め、カイ二乗検定を行う

（2）調整したリスク比mRRを求め、over allなカイ二乗検定を行う

Excelでカイ二乗検定のp値を求める関数：CHISQ.DIST

使い方：p値 = 1 – CHISQ.DIST（χ_{MH}^2,1,TRUE）　[(1)の場合]

[解答例] 演習1: 症例対照研究における層化解析

層1　お菓子をよく食べる

		E	not E	合計
		歯磨きをする	しない	
虫歯あり	D	13	32	45
なし	not D	7	14	21
	合計	20	46	66

IR1=13/20=0.650（E）

IR0=32/46=0.696（not E）

OR=13・14/(32・7)

= 0.813

χ_{MH}^2 = 0.1319（p=0.717）

層2　お菓子を食べない

		E	not E	合計
		歯磨きをする	しない	
	D	25	8	33
	not D	63	22	85
	合計	88	30	118

IR1=25/88=0.284

IR0= 8/30=0.267

OR=25・22/(8・63)

=1.091

χ_{MH}^2 = 0.033（p=0.855）

Crude

		E	not E	合計
		歯磨きをする	しない	
	D	38	40	78
	not D	70	36	106
	合計	108	76	184

IR1=38/108=0.352

IR0=40/76 =0.526

OR=38・36/(40・70)

=0.489

χ_{MH}^2 = 5.530（p=0.019）

調整オッズ比mOR=0.968

χ_{MHS}^2 =0.008 (p=0.929!)

E:「歯磨きをする」のD:「虫歯あり」への
暴露効果があるとは言えない。
Crude解析（右上）では、「お菓子をよく食べる」が交絡となった見せかけの暴露効果が出ている。

『医学への統計学』（第3版）Page226 例題13.3とは
各層のカイ二乗値の値が異なります。例題13.3では、その計算に別の式を使っています

[解答例] 演習2: fixedコホート研究における層化解析

層1　女性、年齢≦20

	E	not E	合計
D	4	30	34
not D	10	251	261
合計	14	281	295

IR1=0.286 (E)

IR0=0.107 (not E)

RR=2.676

χ_{MH}^2=4.174

p= 0.041

層2　女性、年齢>20

	E	not E	合計
D	5	7	12
not D	18	61	79
合計	23	68	91

IR1=0.217

IR0=0.103

RR=2.112

χ_{MH}^2=1.945

p= 0.163

層3　男性、年齢≦20

	E	not E	合計
D	23	29	52
not D	27	102	129
合計	50	131	181

IR1=0.460

IR0=0.221

RR=2.078

χ_{MH}^2=10.010

p= 0.002

層4　男性、年齢>20

	E	not E	合計
D	19	5	24
not D	36	14	50
合計	55	19	74

IR1=0.345

IR0=0.263

RR=1.313

χ_{MH}^2=0.431

p= 0.512

調整リスク比mRR=1.948

χ_{MHS}^2=14.364

p=0.0002!

性別と年齢で調整をしても、
暴露効果あり(p<0.001)

Crude

	E	not E	合計
D	51	71	122
not D	91	428	519
合計	142	499	641

IR1=0.359

IR0=0.142

RR=2.524

χ_{MH}^2=33.690

p= 6×10^{-9}

今日の内容：交絡因子とその調整方法

- 前回の復習
  目標・内容、 リスク因子の同定と交絡因子の影響
  研究方法に応じたデータレイアウト
  暴露効果の指標と推測 （交絡の影響は無視）

- 交絡の調整方法1 研究方法で調整する （紹介のみ）
  (1) 因子範囲の制限
  (2) マッチング： 対応のあるカイ二乗検定（McNemar検定）

- 交絡の調整方法2 解析方法で調整する
  (1) 層化解析（Stratification）：マンテル・ヘンツェル検定
  (2) 回帰分析 ← 第4回目にやります

このセミナーについて

内容：　疫学と統計を復習し、交絡因子とその調整方法、
　　ロジスティック回帰等を紹介する

目標：　交絡因子調整の検定やロジスティック回帰を理解し、
　　R等で実行できるようになる！

ポイント：<u>疾病の規定要因（リスク因子）を正しく同定する</u>

日時（予定）：　毎月下旬月曜or火曜の午後5時から1.5時間程度

スケジュール（予定）：　全4回

第1回 (11/28)：　疫学と統計の基礎
第2回 (12/20)：　交絡因子とその調整方法
次回：第3回 (1/24?)：　統計ソフトRの基礎と応用
第4回 (2/21?)：　ロジスティック回帰（仮）＋α？

お願い：Rのインストール

- 次回（１月下旬）はRを使います
- それまでにRをインストールしておいて下さい

Rダウンロードリンク

- Windows: https://cran.ism.ac.jp/bin/windows/base/
  https://cran.ism.ac.jp/bin/windows/base/R-3.3.2-win.exeを
  クリックして、実行ファイルをダウンロード⇒実行、で、手順に従う
- Mac https://cran.ism.ac.jp/bin/macosx/　上と同じように

参考

http://www.okadajp.org/RWiki/?R%20%E3%81%AE%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%BC%E3%83%AB#p7074c04

補足資料

Dynamic Cohortの例

疫学の教科書・参考書（追加あり）

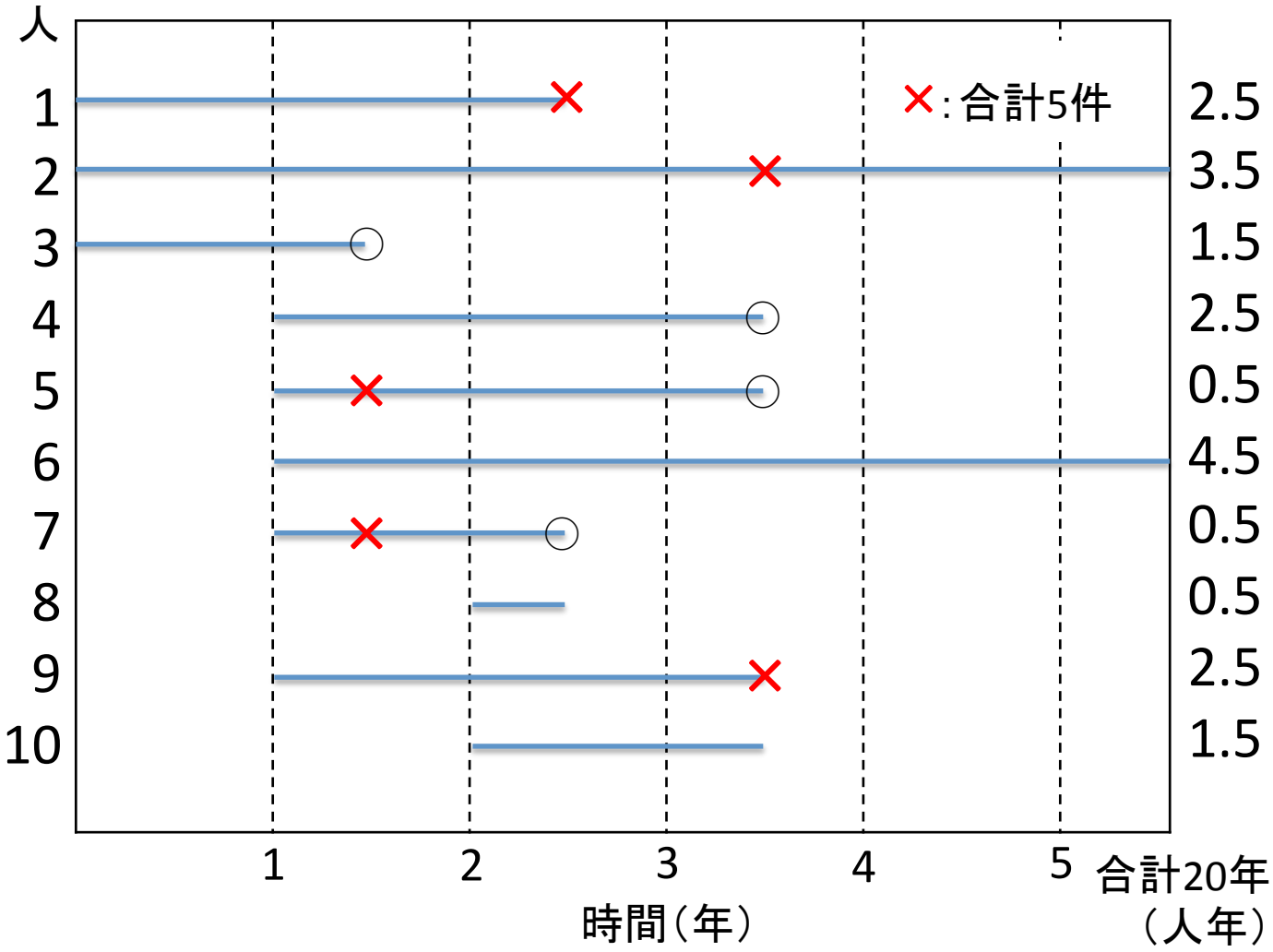
カイ二乗検定と検定統計量

$$\chi^2 = \frac{n(ad-bc)^2}{n_1\, n_0\, m_1\, m_0} \quad \text{or} \quad \chi_{MH}^2 = \frac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0}$$

復習

Dynamic Cohortの例

例：10人の被験者の5年半の追跡（コホート）研究。Dynamic population
研究開始時は全員健康（disease-free、その病気にかかっていない）で、
今後その病気にかかり得るとする。



Years at risk
病気にかかり
得る年数

2.5
3.5
1.5
2.5
0.5
4.5
0.5
0.5
2.5
1.5

×：疾病発生
（初回）
○：死亡

・10人を調べて
20年で5件発生
・各時点での有病率

平均して、6人を5年半追跡

研究人数=10人
脱落者=8人
10-8/2=6

疫学の教科書・参考書1

- **Kleinbaum, Kupper, Morgenstern『Epidemiologic Research』(Wiley, 1982,KKM)**
  Gröhn先生が講義で使われていた教科書。ロジスティック回帰まで網羅。
  実例も式も豊富。
- **柳川『疫学マニュアル』(第7版, 南山堂、2012)**
  オススメです。式も出てきますが、見やすくまとまっています。
  ロジスティック回帰・Cox回帰(生存時間分析)まで網羅。
- **Dohoo et al.『Veterinary Epidemiologic Research』(2nd Ed., VER Inc, 2009, 865 pages..)**
  (厚い＆重いけど)オススメです。最新の疫学手法をカバーしている。
  Gröhn先生の講義で扱った手法はほとんど載っている。
- Pfeiffer『獣医疫学へのファーストステップ』(緑書房, 2012)
- Pfeiffer『Veterinary Epidemiology: An Introduction』(Wiley, 2010)
  はじめに手に取りやすい。読みやすい(基本的な考え方を学べる)。
- 日本疫学会『はじめて学ぶやさしい疫学』(第2版、南江堂, 2010)
- 中村『楽しい疫学』(第3版、医学書院, 2012)
- 獣医疫学会編『獣医疫学』第2版(近代出版, 2011)
  Pfeiffer本の次に or 一緒に。

疫学の教科書・参考書2

- 丹後俊郎『医学への統計学』（第3版）　追加
  統計の基礎～交絡因子の調整、生存時間解析等まで網羅されている

- Allison『Survival Analysis using the SAS system』(SAS Institute Inc., 1995）
- Rothman, Greeland『Modern Epidemiology』（Lippincott-Raven Publishers, 1998）
- Hosmer, Lemeshow『Applied Logistic Regression』（Wiley, 1989）
- Stokes et al.『Categorical data analysis using the SAS system』(SAS Institute Inc. 1995)

Gröhn先生の講義の参考書

カイ二乗検定と検定統計量：検定の仕組み

データと期待度数の差を求めて、それが十分大きいかを調べる

データ

	E：暴露あり（E=1）	$\overline{E}$：暴露なし（E=0）	合計
D	a	b	m_1
$\overline{D}$	c	d	m_0
合計	n_1	n_0	n

期待度数

暴露効果なしの場合の表

	E：暴露あり（E=1）	E：暴露なし（E=$\overline{0}$）	合計
D	$a^*=n_1 \cdot m_1/n$	$b^*=n_0 \cdot m_1/n$	m_1
$\overline{D}$	$c^*=n_1 \cdot m_0/n$	$d^*=n_0 \cdot m_0/n$	m_0
合計	n_1	n_0	n

- **検定統計量**

$$\chi^2 = \frac{(a-a^*)^2}{a^*} + \frac{(b-b^*)^2}{b^*} + \frac{(c-c^*)^2}{c^*} + \frac{(d-d^*)^2}{d^*}$$

エクセルでやる場合、a*,b*,c*,d*を手動で求めてから、CHISQ.TEST関数を使って、p値が求められる

カイ二乗検定と検定統計量：統計量の変形

データ

	E：暴露あり（E=1）	$\overline{E}$：暴露なし（E=0）	合計
D	a	b	m_1
$\overline{D}$	c	d	m_0
合計	n_1	n_0	n

期待度数

暴露効果なしの場合の表

	E：暴露あり（E=1）	E：暴露なし（E=$\overline{0}$）	合計
D	$a^*=n_1 \cdot m_1/n$	$b^*=n_0 \cdot m_1/n$	m_1
$\overline{D}$	$c^*=n_1 \cdot m_0/n$	$d^*=n_0 \cdot m_0/n$	m_0
合計	n_1	n_0	n

- **検定統計量**

$$\chi^2 = \frac{\boxed{(a-a^*)^2}}{a^*} + \frac{(b-b^*)^2}{b^*} + \frac{(c-c^*)^2}{c^*} + \frac{(d-d^*)^2}{d^*}$$

$$\boxed{a-a^*} = a - \frac{m_1 n_1}{n} = \frac{\boxed{an - m_1 n_1}}{n} = \frac{ad-bc}{n}$$

$$\boxed{an - m_1 n_1} = a(n_1+n_0) - m_1 n_1 = (a-m_1)n_1 + an_0 = -b(a+c) + a(b+d) = ad - bc$$

$$\therefore \frac{(a-a^*)^2}{a^*} = \frac{(ad-bc)^2}{m_1 n_1 n}$$

カイ二乗検定と検定統計量：統計量の変形

つづき

- **<u>検定統計量</u>** $$\chi^2 = \frac{(a-a^*)^2}{a^*} + \frac{(b-b^*)^2}{b^*} + \frac{(c-c^*)^2}{c^*} + \frac{(d-d^*)^2}{d^*}$$

同様に、 $$\frac{(a-a^*)^2}{a^*} = \frac{(ad-bc)^2}{m_1 n_1 n} \qquad \frac{(b-b^*)^2}{b^*} = \frac{(ad-bc)^2}{m_1 n_0 n}$$

$$\frac{(c-c^*)^2}{c^*} = \frac{(ad-bc)^2}{m_0 n_1 n} \qquad \frac{(d-d^*)^2}{d^*} = \frac{(ad-bc)^2}{m_0 n_0 n}$$

$$\chi^2 = \frac{(ad-bc)^2}{m_1 n_1 n} + \frac{(ad-bc)^2}{m_1 n_0 n} + \frac{(ad-bc)^2}{m_0 n_1 n} + \frac{(ad-bc)^2}{m_0 n_0 n}$$

$$= \frac{(ad-bc)^2}{n_1 n_0 m_1 m_0 n}\left(m_0 n_0 + m_0 n_1 + m_1 n_0 + m_1 n_1\right)$$

$$\boxed{= \frac{n(ad-bc)^2}{n_1 n_0 m_1 m_0}} \quad \text{or} \quad \boxed{\chi_{MH}^2 = \frac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0}}$$

$$m_0 n_0 + m_0 n_1 + m_1 n_0 + m_1 n_1$$
$$= m_0(n_0 + n_1) + m_1(n_0 + n_1)$$
$$= (m_0 + m_1)(n_0 + n_1)$$
$$= n^2$$