H28.11.29

# 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１回：疫学と統計の基礎

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/

# コーネル大学　疫学の講義への参加

## “Advanced Methods in Epidemiology”

## by Yrjö Gröhn教授（College of Veterinary Medicine)

日程：8月24日～12月4日までの毎週月、水、金。基本的に、朝8:30から9:20まで。
50分/回×3回/週×14-15週

教科書：Kleinbaumら ”Epidemiologic Research” (Wiley, 1982, KKM) など

前提：基礎疫学の受講、統計学の受講

このセミナーでやります

内容：疫学全般（研究方法、疫学で用いる指標、交絡因子、層化解析等）
から、ロジスティック回帰、生存時間分析、ポアソン回帰、
また、それらの拡張（repeated observation、clusteringの考慮）まで。

対象：修士課程の学生（獣医に限らない）

時間があればやります

講義の様子

出席者：獣医分野の学生（7人程度）、
Gröhn先生の研究室のポスドク2人とインターンシップ生１人、
Gröhn先生が副指導教員をしている統計学科のPhD candidateの学生１人。

宿題等：課題が11回出た。教科書の演習問題や、Gröhn先生の論文を元にしたデータ解析
（SASを利用）。提出期限は、出題から、2週間程度（課題による）。
他にも、Quiz（US letterサイズの用紙2枚程度の問題に、その場で回答し提出）が数回あった。

# 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 (12/19 or 20)：　交絡因子とその調整方法（仮）
第3回 (1/23 or 30の週)：　統計ソフトRの基礎（仮）
第4回 (2/20 or 27の週)：　ロジスティック回帰（仮）＋α？

# このセミナーで取り上げる解析対象と応用

このセミナーでは、以下の対象を想定した**解析方法を紹介します**。

解析対象

- **ヒトや動物の疾病等**　例：がん、感染症、周産期病
  感染症だけでなく、生活習慣病等も対象

応用

- **解析方法は、他の様々な対象に適用可能**
  - 植物の病気
  - 農作物の収量・品質
    “リスク因子（規定因子）”: 気温、水環境、肥料、場所
    “疾病”: 収量、品質

# 第1回目　疫学と統計の基礎

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

疫学の基礎

- 研究方法

  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仮説検定、信頼区間

  罹患率（incidence ratio）、有病率（prevalence）

  リスク比、オッズ比

統計の基礎

- 仮説検定（カイ二乗検定）、信頼区間

# 第2回目　交絡因子とその調整方法

予定:

- 交絡因子
  　定義と例
- 調整方法①　マッチング
  　カイ二乗検定？
- 調整方法②　層化解析（Stratification）
  　マンテル・ヘンツェル検定

# 第3回目　統計ソフトRの基礎

予定:

- Rを電卓として使う
- 四則演算
- Rにおける「変数」の扱いを理解する
- データを読み込む
- 記述統計学（平均や分散を求める、作図する）
- 推測統計学（信頼区間を求める、検定をする）

# 第4回目　ロジスティック回帰（＋α?）

予定：

- 回帰分析の種類
  単回帰と重回帰、ロジスティック（重）回帰
- ロジスティック回帰の解釈
- ロジスティック回帰の発展？
- ロジスティック回帰の実例

# 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

- 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

# 今日の目標と内容

目標:

コホート研究（追跡）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。

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

# はじめに

疫学とは何か

有名な疫学研究

疫学の主なトピック

リスク因子の同定、分割表、検定

# 疫学とは何か

【目的】

**疾病の頻度と分布および規定因子を明らかにして、適切な対策の樹立に必要な資料を提示する**

【定義に含まれる要素】

1. 感染症の研究
2. 疾病自然史の研究
   がん：健康→前がん状態→早期→進行→末期→死亡
3. 疾病（健康障害）の頻度と分布に関する研究
4. 疾病の頻度と分布に影響を与える要因（リスク因子）の研究
5. 人間集団を扱う研究
6. 予防医学や公衆衛生の基礎科学

『疫学マニュアル』第１章

# 有名な疫学研究 1854-55年

## ジョン・スノウによるコレラの研究 （wikiより）

コレラのイギリス侵入（1831年10月）当時、コレラは空気感染すると考えられており恐れられていた。しかしスノウは同じ流行地域でも患者が出る家は飛び飛びである等の知見を得て空気感染説に疑問を持ち、**「汚染された水を飲むとコレラになる」という「経口感染仮説」を立て、疫学的調査と防疫活動を行った**。

ブロード・ストリート事件
1854年8月、コレラ患者が多量発生したロンドンのブロード街にて患者発生状況の調査を行い、ある井戸が汚染源と推測、あてはまらない事例について調査を行い、**「汚染された井戸水を飲んでいる人は罹る」**と結論した。行政がこれに従い問題の井戸を閉鎖したため、流行の蔓延を防ぐ事が出来た。

水道会社給水範囲とコレラ患者発生との関係の調査
ロンドンの水道会社はテムズ川から取水していたが、当時のテムズ川は汚濁がひどく衛生的とは言えなかった。スノウは患者発生マップと各水道会社の給水地域との比較照合を行い、**特定の水道会社の給水地域においてコレラ患者が多発している**ことを突き止めた。同社の取水口は糞尿投棄の影響を受ける位置にあったという。
これは1883年にロベルト・コッホが**コレラ菌を発見する30年前**の事であった。

# 疫学の主なトピック

- 疫学で用いる指標　　本セミナーでやります
- 研究方法
- 標本抽出
- 誤差・偏り（選択、情報、交絡バイアス）とその制御
- 因果関係の判定
- スクリーニング
- 疫学に必要な統計手法　　Gröhn先生の講義の（後半の）主な対象
- サーベイランス
- 感染症の疫学
- 特定分野の疫学
- リスクアセスメント
- 疾病の経済評価
- 疫学資料
- 疫学研究と倫理

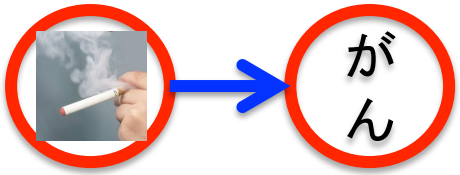
『疫学マニュアル』、『獣医疫学』

# リスク因子の同定（疫学の目的の1つ）

（本セミナーのテーマ）

暴露（Exposure）と疾病（Disease）の関係は？

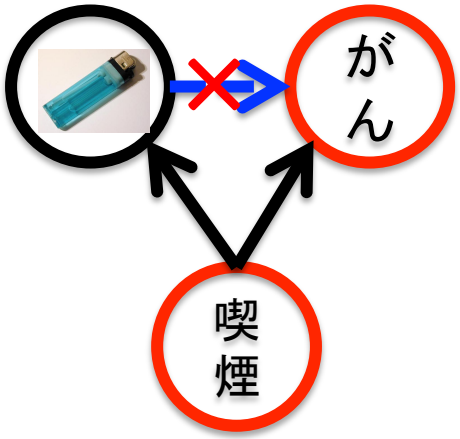
暴露：特定の状態のこと。例：毎日30分散歩する




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

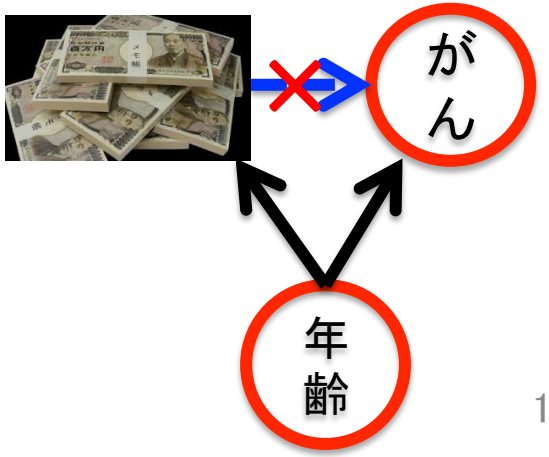



例2：ライター所持は肺がんのリスク因子か？　→　No..

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！　⇒　次回やります

# リスク因子同定のための表（2×2分割表）

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

今日のセミナーの前提

- １つの暴露と疾病の発生を調べる
  つまり、交絡因子となりうる暴露（変数）は無視する（次回以降は考慮します）

重要!

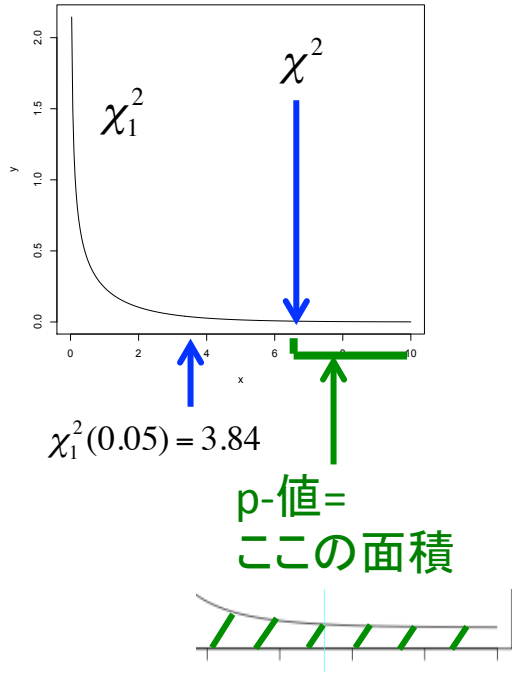
# リスク因子同定のための検定：カイ二乗検定

- 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> $\chi^2 = \dfrac{n(ad-bc)^2}{n_1\, n_0\, m_1\, m_0}$

大きいほど暴露効果あり！

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

～ $\chi_1^2$ ：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

## 研究方法

コホート研究（追跡研究）

症例対照研究

## 疫学で用いる指標

罹患率、有病率

リスク比、オッズ比

# 研究方法 Study Design

1. 観察研究 observational study

   記述的 descriptive

   分析的 analytical

      横断 cross-sectional （一時点の有病率等）
      生態学的 ecological （集団の相関解析等）
      コホート cohort（追跡or前向き。これから疾病に罹患する）
      症例対照 case/control（後ろ向き。既に罹患している）

2. 介入研究 intervention study

   臨床試験 clinical trialなど

# 疫学で用いる指標 Epidemiological Measures

基本的な指標

- 罹患率 incidence rate
- 累積罹患率 cumulative incidence
- 有病率 prevalence
- 死亡率 mortality
- 致命率 fatality

暴露と疾病の関連性の指標

- 相対危険 relative risk or リスク比 risk ratio
- オッズ比 odds ratio
- 生存率

# コホート（cohort、follow-up）研究

内容
- 一定期間、**疾病にかかり得る集団を追跡**し、疾病の発生を調べる
- これからデータをとる or これから疾病に罹患する！

種類
- Fixed cohort：　研究期間中に集団は変わらない（理想的。レア）
- Dynamic cohort（dynamic population）：
  研究期間中に集団からの出入りがある（実用的。農場等）

特徴
- 疾病の発生前（！）に、暴露の状態を知ることが出来る！！　　暴露⇒疾病

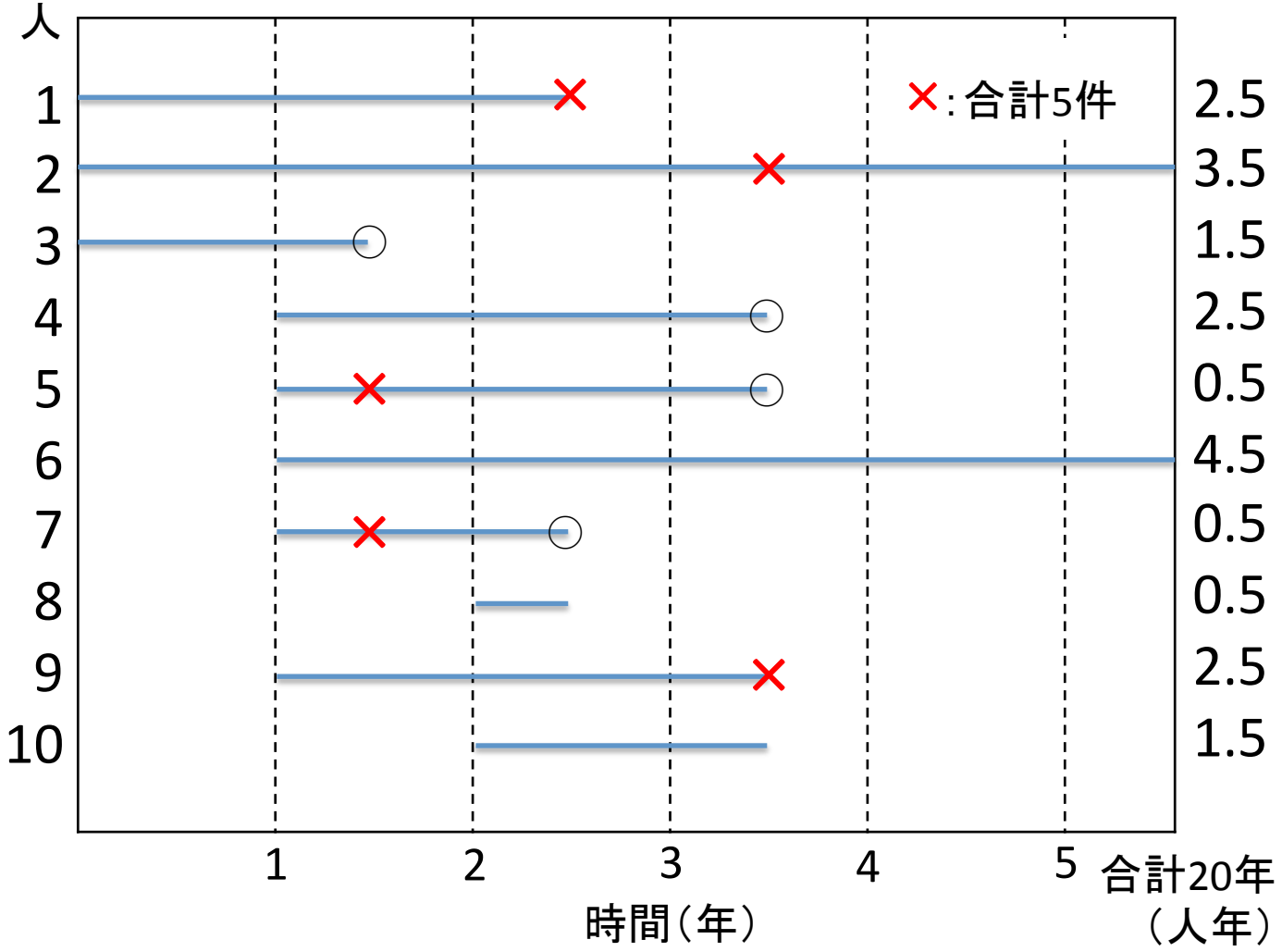
欠点
- コスト、時間がかかる
- レアな疾病の研究には向かない（研究期間中に疾病が発生しないかも。。）

仮定：
研究開始時には集団の各メンバーは健康（disease free）であるとする

重要!

# 今日のポイント：疾病のタイミング！

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

×：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

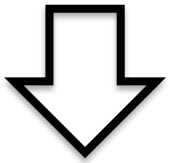
平均して、6人を5年半追跡

研究人数=10人
脱落者=8人
10-8/2=6

重要!

# 今日のポイント：疾病のタイミング！

1. いつ何を調べたのか
2. (1)点で見たのか、(2)期間で見たのか
3. (1)追跡したのか、（今から疾病発生）
   (2)過去にさかのぼったのか（既に疾病発生）



2. (1): 有病率、 (2): 罹患率、累積罹患率、死亡率

3. (1): follow-up研究（コホート研究）
   (2): case/control研究（症例対照研究）

# 罹患率（incidence rate、incidence density）と累積罹患率（cumulative incidence)

- 罹患率（IR）$= \dfrac{\text{新規発生件数}}{\text{疾病にかかり得る期間の合計}}$

例： $IR = \dfrac{5}{2.5+3.5+1.5+2.5+0.5+4.5+0.5+0.5+2.5+1.5} = \dfrac{5}{20}$ 件/年

= 5件/20年 = 25件/100年 = 0.25件/1年 = 1件/4年

- 累積罹患率（CI）$= \dfrac{\text{新規発生件数}}{\text{疾病にかかり得る人数（平均等）}}$

例： $CI = \dfrac{5}{10\text{人} \times 5.5\text{年}} = \dfrac{9}{100}$ = 9%（年間） ＊fixed cohortを仮定

or $\dfrac{5}{(10-8/2)\text{人} \times 5.5\text{年}} = \dfrac{15}{100}$ = 15%（年間）

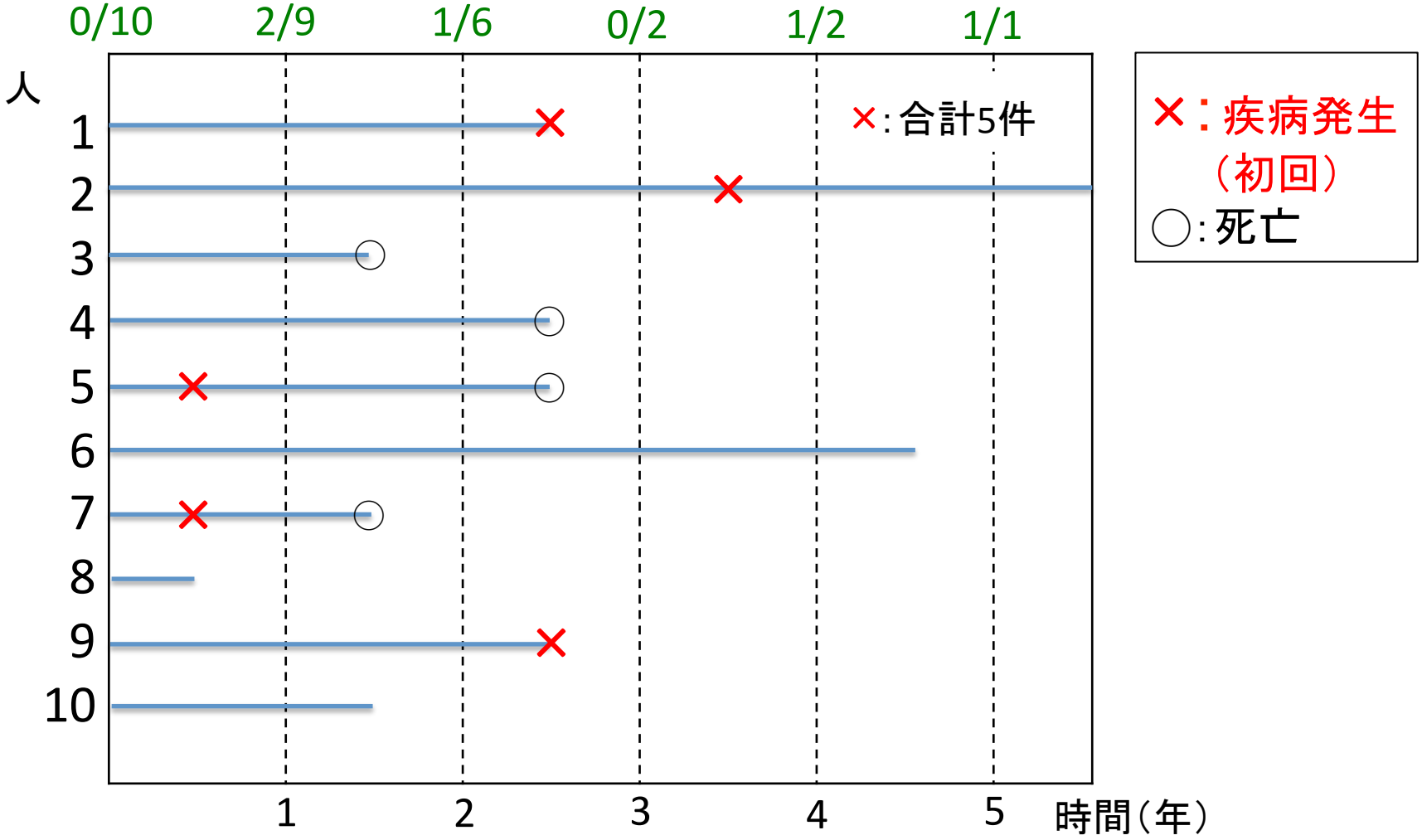
8人：研究期間内に脱落した人数　　平均して6人を5年半追跡

# 有病率（Prevalence）or 点有病率

期間有病率もある

- 有病率 $= \dfrac{\text{罹患者数}}{\text{観察者数}}$

簡単のため、全ての人の研究開始時点は同じとする。
研究期間内に罹患者は回復しないとする。



# 症例対照（case/control）研究

よく用いられる

内容

- 既に疾病が発生している集団と、発生していない健康な集団を比較する

  例：　病院に行って、カルテを見て、疾病集団と健康集団の暴露状態を比較する

特徴

- 既に疾病に罹患している！
- コスト、時間の負担が少ない！
- レアな疾病にも適用できる

欠点

- 疾病前の暴露の状態を知ることが出来ない。。

  つまり、暴露が疾病の原因か、結果かはわからない

# 暴露効果の指標（リスク比やオッズ比）と統計的推測（検定と信頼区間）

# 暴露効果の指標と統計的推測
# 準備1: 目的

Cohort研究（dynamic、fixed）、case-control研究において以下を紹介する

1. 暴露効果があるかどうかの指標
2. 暴露効果があるかどうかの検定
3. 1.の信頼区間

1と3: 暴露効果があるかどうかが数値でわかる

例: オッズ比OR = 3 > 1（効果がありそう）

case 1. ORの95%信頼区間 = [2.0, 4.0] ⇒ OR>1（有意）

case 2. ORの95%信頼区間 = [0.5, 5.5] ⇒ OR>1でない

2: 暴露効果があるかどうかがp値でわかる

例: カイ二乗の検定のp値 = 0.02 < 0.05 ⇒ 暴露効果あり

# 暴露効果の指標と統計的推測
# 準備2：データレイアウト

変更点！

- Dynamic Cohort

| | 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) | $L_1$ | $L_0$ | $L=L_1+L_0$ |

- Fixed Cohort
  or 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 | $m_0=c+d$ |
| | $n_1$ | $n_0$ | |

$c=n_1-a$
$d=n_0-b$

| | 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 | $n_1$ | $n_0$ | $n=n_1+n_0$ |

# 暴露効果の指標と統計的推測
# 準備3:リスクとオッズ、それらの比

| | 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 |
| | $n_1$ | $n_0$ | |

- リスク:罹患数とそうでない被験者数の比

  E群 リスク = a/c　　$\overline{E}$群 リスク = b/d

  リスクの比= $\dfrac{a/c}{b/d} = \dfrac{ad}{bc}$

- オッズ:事象が起こる確率と起こらない確率の比

  D群 暴露のオッズ = $(a/m_1) / (b/m_1) = a/b$

  $\overline{D}$群 暴露のオッズ = $(c/m_0) / (d/m_0) = c/d$

- オッズ比:2つのオッズの比

  D群と$\overline{D}$群での暴露のオッズ比 = $\dfrac{a/b}{c/d} = \dfrac{ad}{bc}$

# 暴露効果の指標と統計的推測
# 準備4：仮説検定

帰無仮説$H_0$：暴露効果あり
対立仮説$H_a$：暴露効果なし

<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）

参考：　巻末の補足資料1（検定統計量とその分布：t検定の場合）

# 暴露効果の指標：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) | $L_1$ | $L_0$ | $L=L_1+L_0$ |

罹患率：Incidence rate　$IR_1=\frac{a}{L_1}$（暴露あり）；　$IR0=\frac{b}{L_0}$（暴露なし）

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

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

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　$IR_1>IR_0 \Leftrightarrow IRR>1$

一部参考

# 暴露効果の推測：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 | m1=a+b |
| Population Time (Person・Year) | $L_1$ | $L_0$ | $L=L_1+L_0$ |

A: E and Dとなる人数（確率変数）
$p_0=L_1/(L_1+L_0)$, $q_0=L_0/(L_1+L_0)$

帰無仮説H0：暴露効果なし、のもとで、$A \sim Bin(m_1,p_0)$となるので、

$$Pr(A \geqq a \mid H_0) = \Sigma_{j=1}^{m1} C_j^{m1} p_0^{\ j} q_0^{\ m1-j}$$ （計算できる）

$C_j^{m1}$：$m_1$個からj個取り出す組み合せの数

しかし、面倒なので、large sample test（近似）を使うと、

**$E(A)=m_1 p_0$**, $Var(A)=m_1 p_0 q_0$ より、 $A \sim N(m_1 p_0, m_1 p_0 q_0)$なので、

（二項分布の期待値と分散）

検定統計量 $Z = \dfrac{A - m_1 p_0}{\sqrt{m_1 p_0 q_0}} \sim N(0,1)$ （帰無仮説のもとで）

or $\chi^2 = Z^2 \sim \chi_1^{\ 2}$

一部参考

# 暴露効果の推測：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) | $L_1$ | $L_0$ | $L=L_1+L_0$ |

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

A: E and Dとなる人数（確率変数）

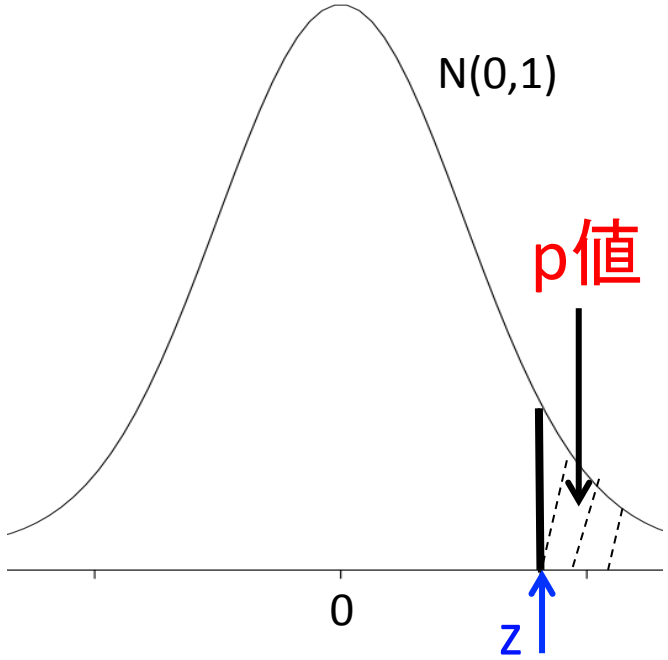
検定統計量 $Z=\dfrac{A-m_1 p_0}{\sqrt{m_1 p_0 q_0}} \sim N(0,1)$ （帰無仮説のもとで）

or $\chi^2=Z^2 \sim \chi_1^2$

実際のデータ（表）から、

$z=\dfrac{a-m_1 p_0}{\sqrt{m_1 p_0 q_0}}$ の値を求めて、

平均0、分散1の正規分布（右図）のどこにあるかを調べればいい！
（p値が求まる）



# 暴露効果の指標：fixed 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 at risk | $n_1$ | $n_0$ | $n=n_1+n_0$ |

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

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

リスクオッズ比！：odds ratio

$$ROR=\frac{CI_1/(1-CI_1)}{CI_0/(1-CI_0)}=\frac{ad}{bc}$$

$c=n_1-1$
$d=n_0-b$

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

一部参考

# 暴露効果の推測：fixed 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 at risk | $n_1$ | $n_0$ | $n=n_1+n_0$ |

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

$CI=\frac{n_1CI_1+n_0CI_0}{n_1+n_0}=\frac{m_1}{n}$（combined）

帰無仮説$H_0$：$CI_1=CI_0$
対立仮説$H_a$：$CI_1>CI_0$

比率の差の検定！

検定統計量 $Z=\frac{(CI_1-CI_0)-0}{\sqrt{CI(1-CI)(1/n_1+1/n_0)}}=\frac{\sqrt{n}\,(ad-bc)}{\sqrt{n_1\,n_0\,m_1\,m_0}}\sim N(0,1)$（under $H_0$）

$c=n_1-a$
$d=n_0-b$
$m_0=c+d$

$\Leftrightarrow\quad \chi^2=Z^2=\frac{n(ad-bc)^2}{n_1\,n_0\,m_1\,m_0}\sim \chi_1^2$（under $H_0$）

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

# 暴露効果の指標と推測：case-controlの場合

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

Fixed cohortの場合と同じ

- 累積罹患率（CI）
- リスク比（RR）
- リスク差（RD）
- リスクオッズ比（ROR）

検定統計量

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

（under $H_0$）

同じでない

- 暴露オッズ比 EOR = $\frac{a/b}{c/d} = \frac{ad}{bc}$

Exposure

# 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 = $\theta \exp\left\{\pm 1.96\sqrt{(\log\theta)^2/\chi_{MH}{}^2}\right\}$

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

# 例題：レアな疾病の暴露効果の評価

| | E：暴露あり | $\overline{E}$：暴露なし | 合計 |
|---|---|---|---|
| D：疾病あり | 28 (a) | 22 (b) | 50 ($m_1$) |
| $\overline{D}$：疾病なし | 20 (c) | 30 (d) | 50 ($m_0$) |
| | 48 ($n_1$) | 52 ($n_0$) | 100 (n) |

KKM 例題15.1

問題設定（case-control研究）

- 与えられた母集団において、過去5年間の新規の疾病罹患者（50人）について暴露の有無を調べた。また、同数の非罹患者を同じ母集団からサンプリングし、同様に暴露の有無を調べた。
- この暴露が疾病の罹患に寄与しているかを検証する。

- EOR = ad/bc = 28・30/(22・20) = 1.91 > 1 より、暴露効果がありそう

- $\chi_{MH}^2 = \dfrac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0} = \dfrac{(100-1)(28\cdot 30-22\cdot 20)^2}{48\cdot 52\cdot 50\cdot 50} = 2.54 \quad \sim \chi_1^2$より

  p=0.055≒0.05。したがって、有意差については”ボーダーライン”

# 例題：レアな疾病の暴露効果の評価

| | E：暴露あり | $\overline{E}$：暴露なし | 合計 |
|---|---|---|---|
| D：疾病あり | 28 (a) | 22 (b) | 50 ($m_1$) |
| $\overline{D}$：疾病なし | 20 (c) | 30 (d) | 50 ($m_0$) |
| | 48 ($n_1$) | 52 ($n_0$) | 100 (n) |

KKM 例題15.1

- EOR= 1.91、$\chi_{MH}^2$ = 2.54、p=0.055 ≒ 0.05
- EORの95%信頼区間 = $EOR \exp\left\{\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right\}$

$= 1.91 \exp\left\{\pm 1.96\sqrt{\frac{1}{28}\ \frac{1}{22}\ \frac{1}{20}\ \frac{1}{30}}\right\}$

= [0.863, 4.229] （large sample）

- または、$= EOR^{1 \pm 1.96/\sqrt{\chi_{MH}^2}} = 1.91^{1 \pm 1.96/\sqrt{2.54}}$ = [0.862, 4.233]
- どちらの場合も有意でない（信頼区間に1を含む）（検定ベース）

# 演習

1. 疫学指標1:
   罹患率、累積罹患率、有病率の計算

2. 疫学指標2と推測:
   リスク比、オッズ比の計算と統計的推測

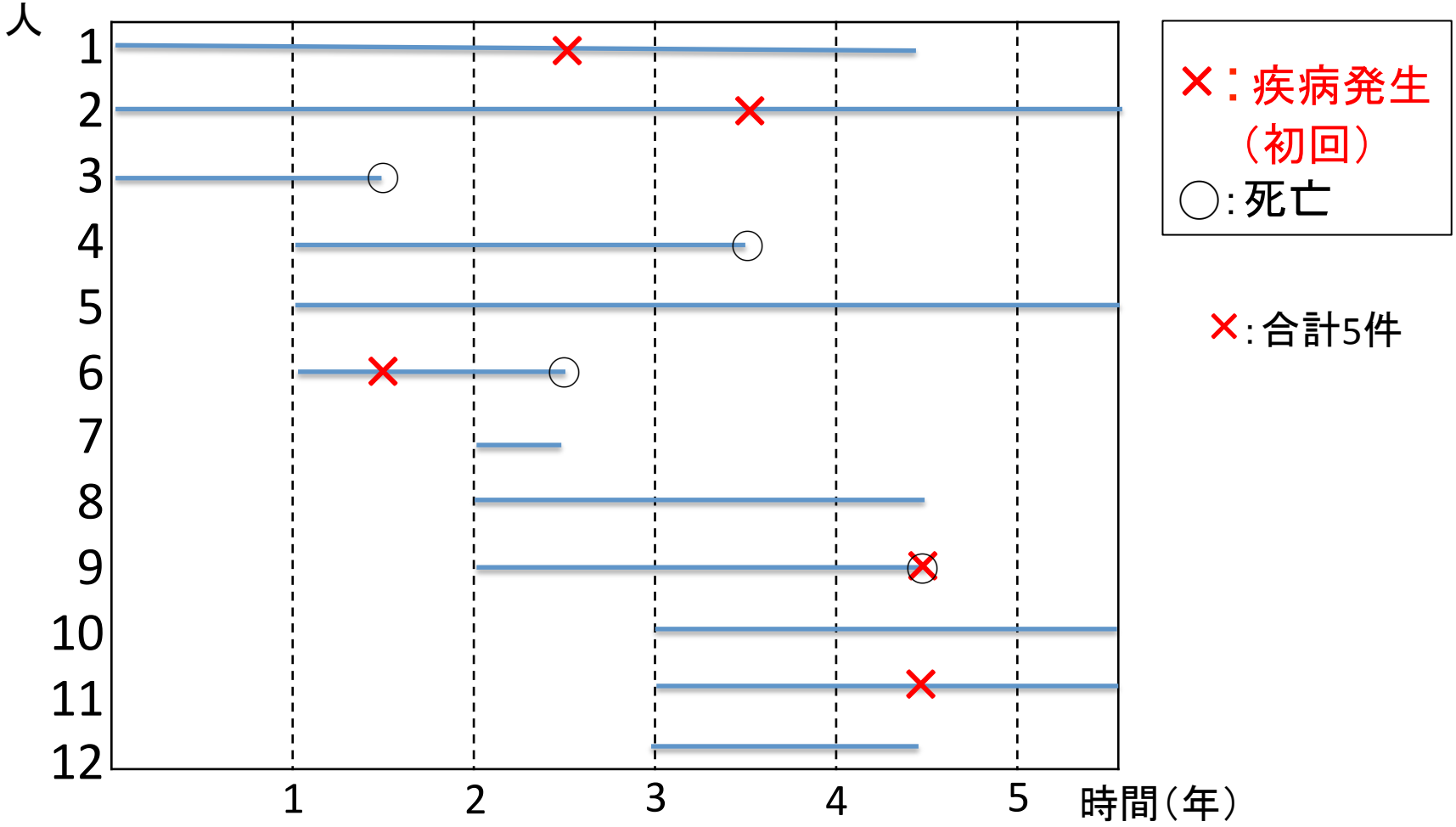
エクセルファイル:

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

# 演習1: 疫学指標1

KKM Ex. 6.1 [改]

5年半の12人のコホート研究（下記）について、
罹患率と各年（0,1,2,…,5年）における有病率を求めて下さい。
ただし、1度罹患した個体は研究期間内には回復しないとする。



# 演習2: 疫学指標2と統計的推測

| | E: 暴露あり | $\overline{E}$: 暴露なし | 合計 |
|---|---|---|---|
| D | 70 (a) | 40 (b) | 110 ($m_1$) |
| $\overline{D}$ | 42 (c) | 58 (d) | 100 ($m_0$) |
| | 112($n_1$) | 98 ($n_0$) | 210 (n) |

| | E: 暴露あり | $\overline{E}$: 暴露なし | 合計 |
|---|---|---|---|
| D | 105 (a) | 60 (b) | 165 ($m_1$) |
| $\overline{D}$ | 63 (c) | 87 (d) | 150 ($m_0$) |
| | 168($n_1$) | 147($n_0$) | 315 (n) |

上記のcase-control研究の結果について、以下をそれぞれ求め、暴露効果があるのかどうか、また、データ数が結果に与える影響を考察して下さい。

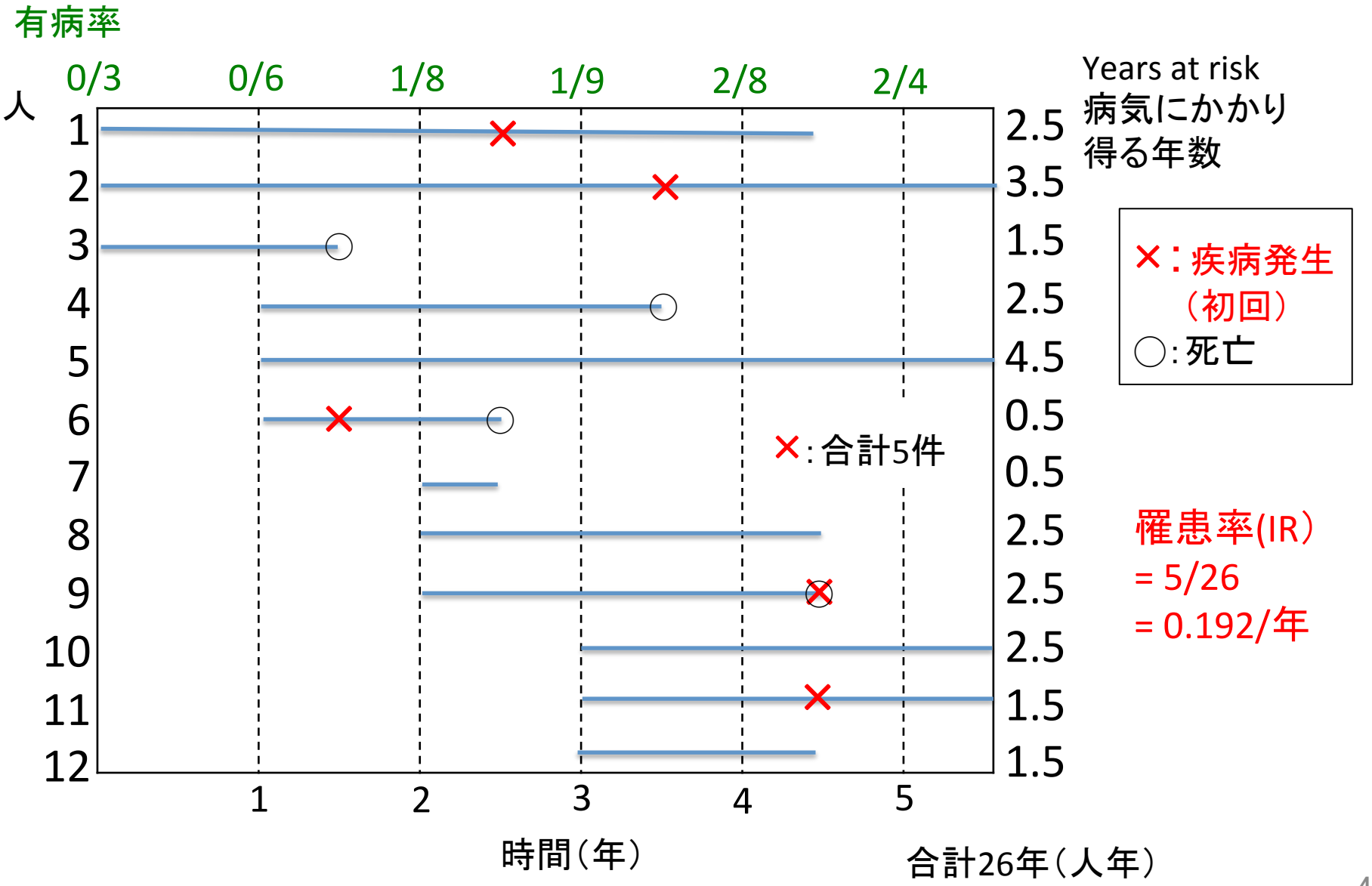
- 暴露オッズ比EOR
- カイ二乗統計量$\chi_{MH}^2$（MHタイプ）
- $\chi_{MH}^2$から求めたp値

Excelでp値を求める関数：CHISQ.DIST

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

# [解答] 演習1: 疫学指標1

KKM Ex. 6.1 [改]



# [解答] 演習2: 疫学指標2と統計的推測

| | E:暴露あり | $\overline{E}$:暴露なし | 合計 |
|---|---|---|---|
| D | 70 (a) | 40 (b) | 110 ($m_1$) |
| $\overline{D}$ | 42 (c) | 58 (d) | 100 ($m_0$) |
| | 112($n_1$) | 98 ($n_0$) | 210 (n) |

- EOR = ad/bc
  = 70・58/(40・42) = 2.41
- $\chi_{MH}^2 = \dfrac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0}$
  $= \dfrac{(210-1)(70\cdot 58-40\cdot 42)^2}{112\cdot 98\cdot 110\cdot 100}$
  = 9.80
- p値 = 0.00174

| | E:暴露あり | $\overline{E}$:暴露なし | 合計 |
|---|---|---|---|
| D | 105 (a) | 60 (b) | 165 ($m_1$) |
| $\overline{D}$ | 63 (c) | 87 (d) | 150 ($m_0$) |
| | 168($n_1$) | 147($n_0$) | 315 (n) |

- EOR = ad/bc
  = 105・87/(60・63) = 2.41
- $\chi_{MH}^2 = \dfrac{(n-1)(ad-bc)^2}{n_1\, n_0\, m_1\, m_0}$
  $= \dfrac{(315-1)(105\cdot 87-60\cdot 63)^2}{168\cdot 147\cdot 165\cdot 150}$
  = 14.73
- p値 = 0.00012

有意な暴露効果がある（p<0.01）　　（有意な影響が出やすい）

EORは同じだが、データ数が多い方がp値が低くなる！

# 今日の目標と内容

目標:

コホート研究（追跡）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。

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

# お願い：Rのインストール

- 第３回目（１月下旬予定）に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

- 次回（12月中）に確認します（？）

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

1. 検定統計量とその分布（t検定の場合）
2. データの種類に応じた2因子の関連性の評価

# 1. 検定統計量とその分布（t検定の場合）

データ　平均値　分散

- 1群：　$X_1,\cdots,X_m$　　$\overline{X}$　　$S_X^2$
- 2群：　$Y_1,\cdots,Y_n$　　$\overline{Y}$　　$S_Y^2$

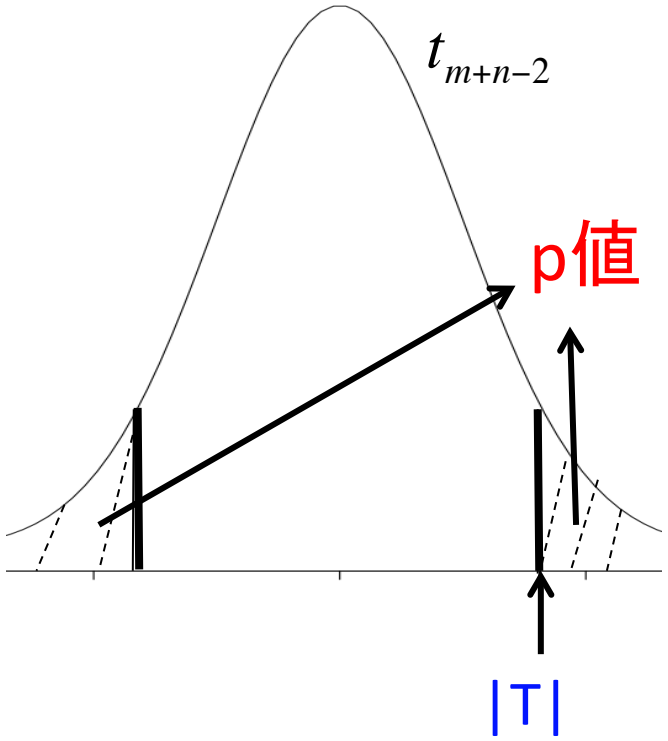
両群の分散は等しいとする

- 帰無仮説$H_0$：平均値は等しい
- 対立仮説$H_a$：平均値は異なる

- 検定統計量と分布

$$T=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{m+n}{mn(m+n-2)}\left(mS_X^2+nS_Y^2\right)}}$$

$\sim t_{m+n-2}$　（under $H_0$）



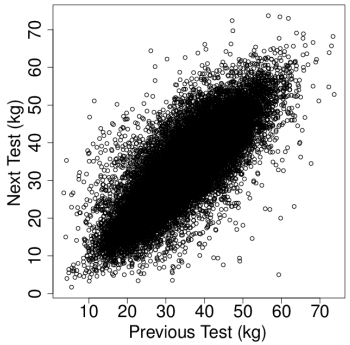
# 2. データの種類に応じた2因子の関連性の評価

## データの種類

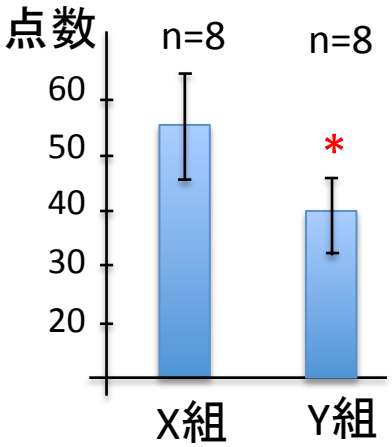
1. 量的（数値）：身長、体重、年齢、、、
2. 質的（数値でない）：性別、品種、暴露や疾病の有無、、

## 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 |