# 第11回　クロス集計(1)

- 今回はカテゴリ変数が2つ以上ある場合に，その関係をみる話に入ります
  - クロス集計の方法とクロス集計表の操作
  - 2つのカテゴリ変数が独立（無相関）であるという帰無仮説の検定
  - 第3の変数で層別化することによって交絡を制御する話
  - 2つのカテゴリ変数間の関連の程度の評価（次回）

# クロス集計表の作成

- 2つのカテゴリ変数をもつデータがあるとする
- （例）AGE（年齢），EXPOSURE（曝露の有無）とDISEASE（病気の有無）についての40人のデータ
- タブ区切りテキストファイル
  http://phi.med.gunma-u.ac.jp/medstat/s11.txt

| AGE | EXPOSURE | DISEASE |
|---|---|---|
| 69 | "YES" | "YES" |
| 54 | "YES" | "NO" |
| 76 | "YES" | "YES" |
| 44 | "YES" | "NO" |
| 50 | "YES" | "YES" |
| 70 | "YES" | "YES" |
| 40 | "YES" | "YES" |
| 54 | "YES" | "YES" |
| 50 | "YES" | "YES" |

# さまざまな集計の例

- データを読みこむ（Rを起動してプロンプト">"に下記を打つ）
```
x <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/s11.txt")
```
- データ構造を見て，ちゃんと読めたか確認
```
str(x)
```
- 曝露の有無(EXPOSURE)について集計
```
table(x$EXPOSURE) または xtabs(~EXPOSURE, data=x)
```
- 集計結果をEXCという名前のオブジェクトに付値
```
EXC <- table(x$EXPOSURE)
```
などとする
- 曝露ありの人についてだけ，病気の有無(DISEASE)を集計
```
table(x$DISEASE[x$EXPOSURE=="YES"])
```
または
```
table(subset(x,EXPOSURE=="YES")$DISEASE)
```
- この結果をEXDという名前のオブジェクトに付値
```
EXD <- table(x$DISEASE[x$EXPOSURE=="YES"])
```

# さまざまな集計の例(続き)

- 曝露なしの人についてだけ病気の有無(DISEASE)を集計しNEDに付値
  ```
  NED <- table(x$DISEASE[x$EXPOSURE=="NO"])
  ```
- これら２つの集計結果を行方向に結合するとクロス集計表になる
  ```
  rbind(EXD,NED)
  ```
- しかし実は最初からクロス集計はtable()でもxtabs()でも可能
  ```
  table(x$EXPOSURE,x$DISEASE)
  ```
  または
  ```
  xtabs(~EXPOSURE+DISEASE, data=x)
  ```
- 結果のクロス集計表は下記

| | DISEASE | |
|---|---|---|
| EXPOSURE | NO | YES |
| NO | 12 | 8 |
| YES | 4 | 16 |

- 最初からクロス集計表の人数がわかっていれば
  ```
  matrix(c(12,4,8,16),2,2)
  ```
  でもOK

# さまざまな集計の例（続き）

違う名前でもいい。行名，列名の順

- カテゴリに名前を付けて，オブジェクトTBLとして保存するには
  ```
  TBL <- matrix(c(12,4,8,16),2,2)
  ```
  としてから
  ```
  rownames(TBL) <- colnames(TBL) <- c("NO","YES")
  ```
  または
  ```
  dimnames(TBL) <- list(EXPOSURE=c("NO","YES"),DISEASE=c("NO","YES"))
  ```
- 上のようにすると，オブジェクトTBLは，下記xtabs結果と一致
  ```
  TBL <- xtabs(~EXPOSURE+DISEASE, data=x)
  ```
- 60歳未満の人だけ（AGE<60）についてクロス集計するなら
  ```
  xtabs(~EXPOSURE+DISEASE, data=subset(x,AGE<60))
  ```
- 60歳以上の人だけなら，同様に
  ```
  xtabs(~EXPOSURE+DISEASE, data=subset(x,AGE>=60))
  ```
- 10歳刻みで別々にクロス集計表を作るには
  ```
  x$AC <- cut(x$AGE,seq(min(x$AGE),max(x$AGE)+1,by=10),right=FALSE)
  xtabs(~EXPOSURE+DISEASE+AC, data=x)
  ```

# ３次元のクロス集計表（続き）

- 10歳刻みで別々にクロス集計表を作るには（再掲）
  ```
  x$AC <- cut(x$AGE,seq(min(x$AGE),max(x$AGE)+1,by=10),right=FALSE)
  ```
  xtabs(~EXPOSURE+DISEASE+AC, data=x) または
  table(x$EXPOSURE,x$DISEASE,x$AC)

- これは３次元のクロス集計表。
  ACの代わりに60歳未満／以上の２区分では，
  xtabs(~EXPOSURE+DISEASE+(AGE>=60), data=x)

- ２つのクロス集計表から合成するには，
  ```
  YTBL <- xtabs(~EXPOSURE+DISEASE, data=subset(x,AGE<60))
  ETBL <- xtabs(~EXPOSURE+DISEASE, data=subset(x,AGE>=60))
  T3 <- array(c(YTBL,ETBL),dim=c(2,2,2))
  ```
  行名，列名，表名が全部消えてしまうので，付け直すには
  ```
  dimnames(T3) <- list(E=c("N","Y"),D=c("N","Y"),AGE=c("<60",">=60"))
  ```

- ３次元クロス表の１枚である２次元クロス集計表を参照
  T3[,,1]やT3[,,"<60"]はYTBLと同値。T3[,,2]はETBLと同値。
  病気ありの人だけで曝露の有無と年齢のクロスはT3[,2,]でOK。
  T3[,,1]+T3[,,2]により年齢区分をしないクロス表が得られる
  QUIZ: 年齢区分をしないで曝露と病気のクロス表を得るには？

# クロス集計表をどのように分析する？

- 知りたいこと：2つのカテゴリ変数（曝露と病気）に関連があるか？
  - 「関連がない」帰無仮説の検定＝独立性の検定
    - カイ二乗検定（数学的に比率の差の検定と同値）
    - フィッシャーの直接確率
  - どの程度の関連があるか＝3つのアプローチ
    - 属性相関係数（やポリコリック相関係数）を調べる
    - 「比」曝露した人の**病気のなりやすさ**が曝露が無い人の何倍かを調べる
    - 「差」曝露した人の**病気のなりやすさ**が曝露が無い人よりどれだけ大きいかを調べる
- 注意点
  - 「病気のなりやすさ」をどうやって示すか？　「率と割合」
  - 曝露と病気の関連が歪められていないか？　「交絡」

次回やります

# 独立性の検定

- カイ二乗検定
  - クロス集計表の実際の数値を観測度数，関連が無かった場合に観察されるはずの人数を期待度数とし，適合度検定と同じく差の二乗を期待度数で割ったものを加えてカイ二乗値を計算する（自由度に注意）
  - 度数は離散値でカイ二乗分布は連続分布なため，Yatesの連続修正（度数の差に0.5を足したり引いたりする）によりカイ二乗分布の近似をよくする
  - 通常は，chisq.test(クロス集計表)でOK。
- フィッシャーの直接確率
  - 周辺度数が決まっているときのすべてのあらゆる組み合わせを考え，観察されている表よりも出現確率が低い表の出現確率の総和をとって有意確率を得る
  - 通常はfisher.test(クロス集計表)でOK。

# 前出の例(s11.txt)では？

- やりたいことの分解
  - データを読む
    **x <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/s11.txt)**
  - 曝露の有無と病気の有無のクロス集計表を計算・表示
    **(TBL <- xtabs(~EXPOSURE+DISEASE,data=x))**
  - 曝露の有無と病気の有無に関係が無いという帰無仮説の検定
    **chisq.test(TBL)**または**fisher.test(TBL)**
- 期待度数が小さすぎるセルがあると**chisq.test()**は警告が出る。現在ではコンピュータは充分速いので，常に**fisher.test()**で**OK**

```
> chisq.test(TBL)

        Pearson's Chi-squared test with Yates' con
        correction

data:  TBL
X-squared = 5.1042, df = 1, p-value = 0.02387
```

```
> fisher.test(TBL)

        Fisher's Exact Test for Count Data

data:  TBL
p-value = 0.02248
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.226738 32.733453
sample estimates:
odds ratio
  5.707765
```

この例では，どちらでも，ほぼ同じ有意確率が得られる

この部分の説明は次回

# 属性相関係数

- ファイ係数(φ)
  - 曝露の有無，発症の有無を1/0で表した相関係数
  - π1=曝露群の有病割合，π2=非曝露群の有病割合，θ1=病気ありの人の曝露割合，θ2=病気なしの人の曝露割合として，
    φ=√((π1-π2)(θ1-θ2))
  - ２×２に限らず一般のk×m分割表について計算可能
  - カイ二乗統計量χ2と総人数nを用いると√(χ2/n)
- ピアソンのコンティンジェンシー係数C
  - ファイ係数からカテゴリ数の影響を除去したもの
  - C=√(φ2/(1+φ2))
- クラメールのV
  - ファイ係数の取りうる値の範囲を0から1にしたもの
  - kとmの小さな方をtとして，V=φ/√(t-1)
- vcdライブラリのassocstats()関数で計算できる

# 属性相関係数の計算例

- **s11.txt**で曝露と病気についての属性相関係数は？

```
# データを読む
x <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/s11.txt")
# 集計
(TBL <- xtabs(~EXPOSURE+DISEASE,data=x))
# vcdライブラリを読み込む
library(vcd)
# 属性相関係数の計算
assocstats(TBL)
```

```
> assocstats(TBL)
                    X^2 df  P(> X^2)
Likelihood Ratio 6.9044  1 0.0085985
Pearson          6.6667  1 0.0098233

Phi-Coefficient   : 0.408
Contingency Coeff.: 0.378
Cramer's V        : 0.408
```

Yatesの連続性の修正がされていないカイ二乗検定の結果

ファイ係数

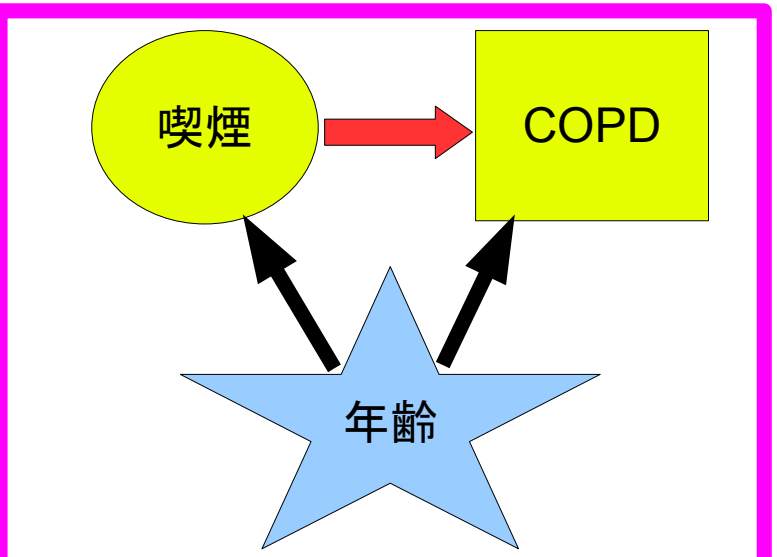
コンティンジェンシー係数

クラメールのV

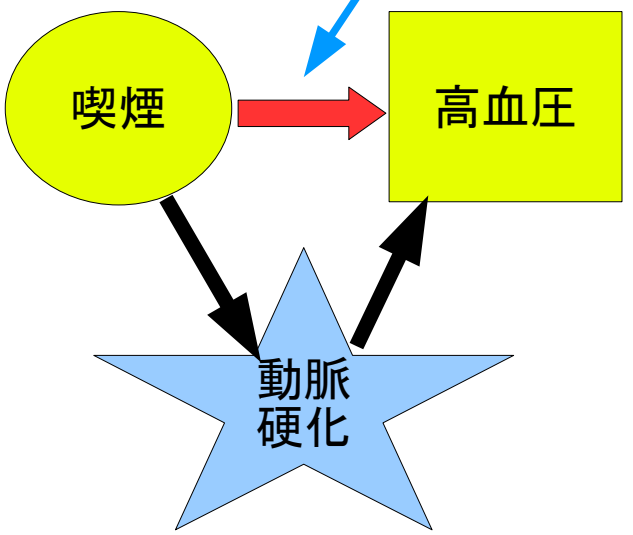
# 交絡とは何か？

- 原因への曝露と結果である病気との因果関係を歪めるバイアスの1つ
- 交絡因子の3つの条件
  - 曝露と関係している
  - 病気と関係している
  - 曝露の結果ではない（因果パスの中間にはない）

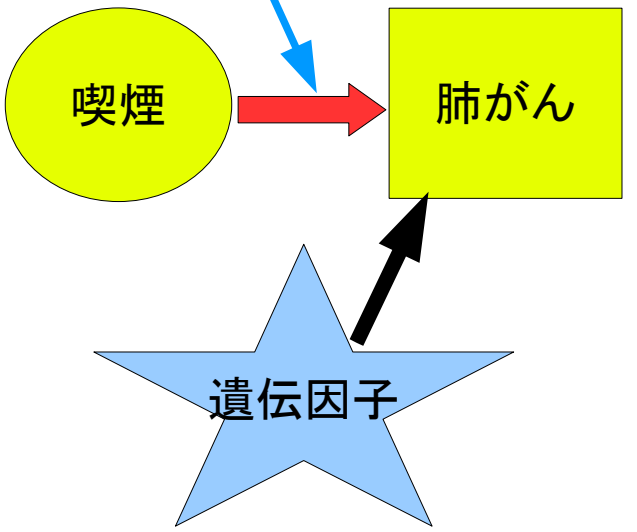
**交絡因子でなければ，制御しなくても因果関係は歪まないので，2次元クロス集計で分析可能**




喫煙への曝露とCOPD発症の因果関係において**年齢は交絡因子**




こういう因果パスがあったとすると，動脈硬化は**交絡因子ではない**




曝露と関係していなければ**交絡因子でない**

# 交絡の検討方法

- 限定による解析
  - ある要因曝露と病気の関係が高齢者でだけ見られる場合は，広い年齢層のデータを一緒にしてしまうと関連がマスクされるので，対象を高齢者に「限定」して解析
- 層別解析
  - 上の例で，若い人も調べるが高齢者とは年齢層別に解析することにすれば，高齢者での関連がマスクされないだけでなく，もし若い人に別の関連性が潜んでいても見いだせる
- 層別のクロス集計表の併合による解析
  - どの層でも同じ方向に関連がある，といいたいとき
    マンテル＝ヘンツェルの要約カイ二乗検定
    **mantelhaen.test(**３次元の集計表**)**
  - ただし３次元の交互作用がないことを確認するため
    **Woolf**の検定で有意でないことを要確認
    **library(vcd); woolf_test(**３次元の集計表**)**
- ロジスティック回帰分析など多変量解析（第**14**回参照）

# 交絡が疑われる時の解析例(s11.txt)

- データは既に**x**に読めているとする
- **60歳以上に限定**するには
```
fisher.test(xtabs(~EXPOSURE+DISEASE,
data=subset(x,AGE>=60))
```
- **60歳以上か未満かで層別に解析**するには
```
T3 <- xtabs(~EXPOSURE+DISEASE+(AGE>=60),data=x)
fisher.test(T3[,,1])
fisher.test(T3[,,2])
```
- クロス表を併合して解析するなら
  `fisher.test(T3[,,1]+T3[,,2])`や
  `chisq.test(T3[,,1]+T3[,,2])`の結果と
  `mantelhaen.test(T3)`の結果を比較 → このデータの場合，p値が大きく異なるので，層別因子にした「年齢」は交絡である可能性が高い
  ただし
```
library(vcd)
woolf_test(T3)
```
  で有意確率が大きく３次の交互作用が有意でないことが前提

# 付．反復測定の一致度について（詳細は省略するのでテキスト参照）

- 対象者に同じ検査や質問紙調査を反復測定あるいは別の検査者や評価者による測定が実施された場合に測定の信頼性（test-retest reliabilityやinter-rater reliability）を調べたいとき
- 形はクロス集計となるが，帰無仮説「関連が無い」では論理的におかしい。つまり関連はあって当然。偶然よりも一致度が高くなって初めて意味があるので，帰無仮説は「偶然の一致と差が無い」
  →カイ二乗検定やフィッシャーの検定は使えない
  - カッパ統計量の計算と有意性検定（fmsbライブラリのKappa.test)，拡張マクネマー検定(mcnemar.test)などの方法がある