第12回　クロス集計(2)

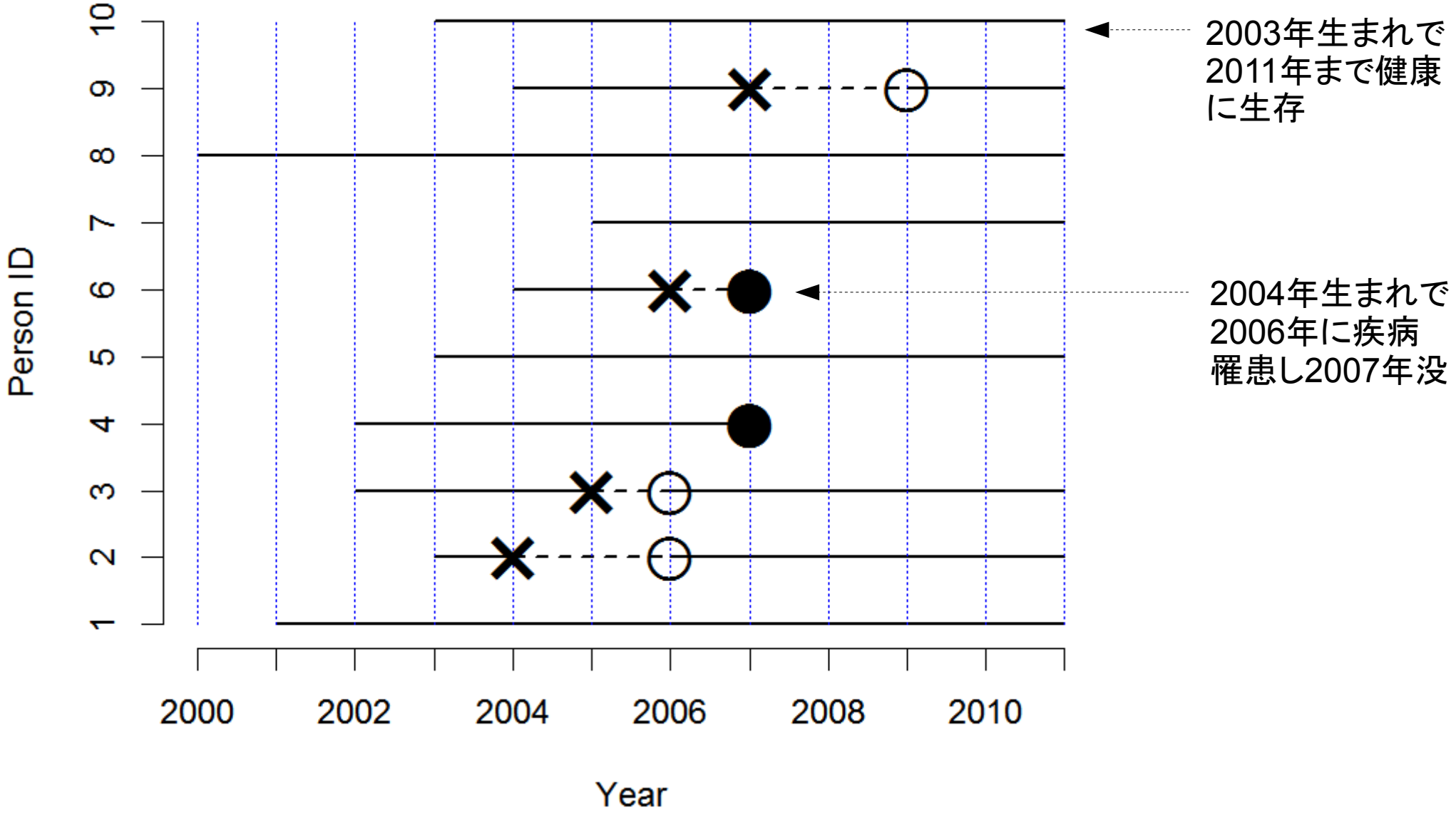
- 今回もカテゴリ変数が2つ以上ある場合に，その関係をみる話の続きです
  - クロス集計の方法とクロス集計表の操作（復習）
  - 2つのカテゴリ変数間の関連の程度の評価
＝疾病の有無と曝露の有無はどれくらい関連？
    - 疾病量をどうやって把握するのか
      - 割合とリスクと発生率（罹患率，死亡率）
    - 疾病と曝露の関連の程度をどうやって測るのか
      - 差か比か
    - 検定，推定，p値関数
      - 疫学的にはp値関数がベストとされるが，検定は便利
  - （時間があれば）スクリーニングの評価

クロス集計と操作の方法（復習）

- MASSライブラリのsurveyデータフレームについて，性別（変数Sex）と利き手（変数W.Hnd）のクロス集計をしてクロス集計表をTABというオブジェクトに付値すると同時に中身を表示するには，
  - **library(MASS)**
    (TAB <- xtabs(~Sex+W.Hnd, data=survey))
    # TAB <- matrix(c(7,10,110,108),2,2)
- クロス集計表TABの各セルの参照は下記
  - TAB[1,1] # SexがFemaleでW.HndがLeftの人数
  - # TAB[”Female”,”Left”]も同様
  - TAB[,2] # W.HndがRightの男女別人数ベクトル
  - TAB[”Male”,] # SexがMaleの利き手別人数

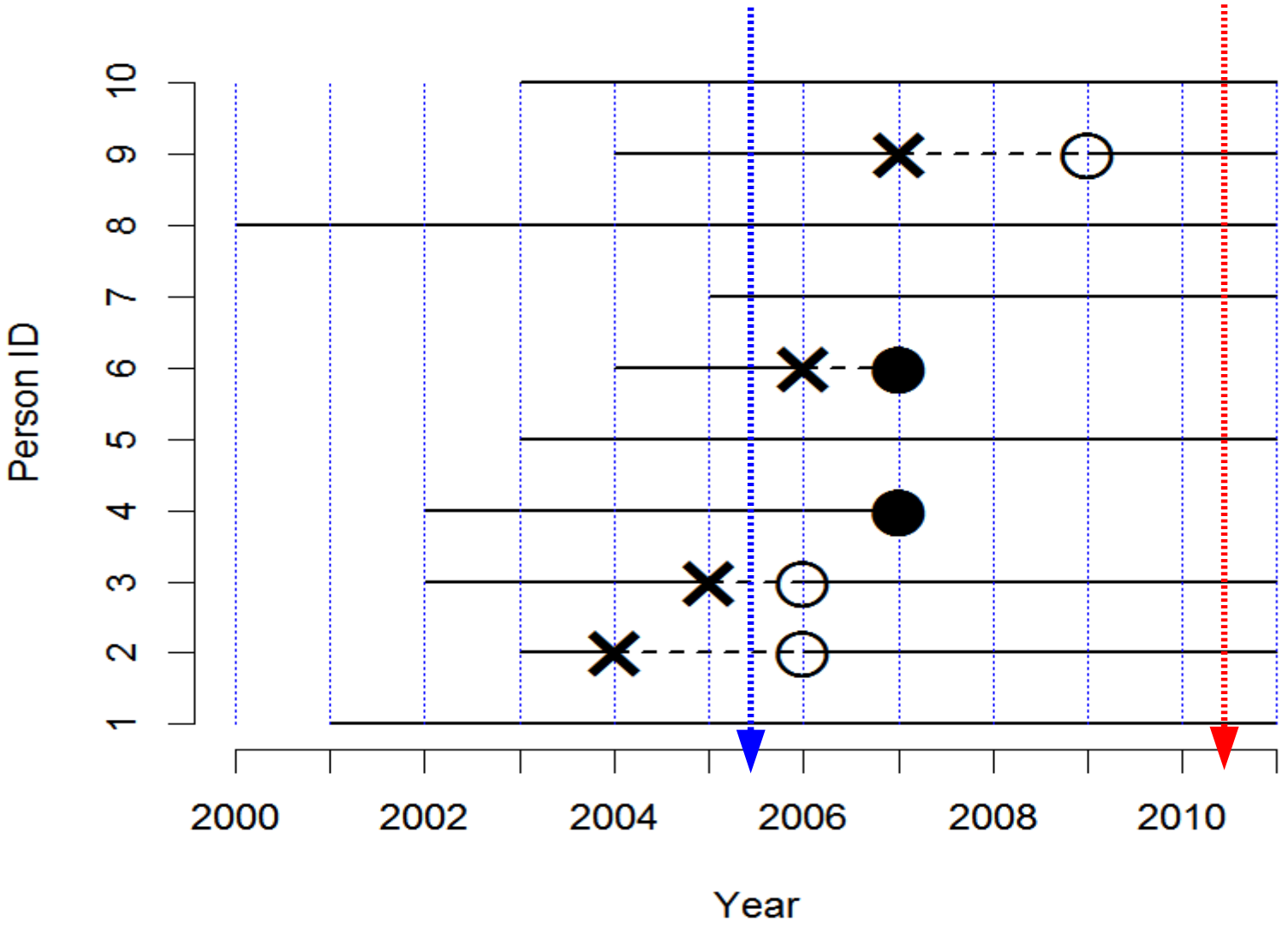
疾病量の把握

- 疾病の発生について，実際の状況を考えてみよう。下図は1本の線が1人を示し，横軸は年次，実線が健康（左端が出生），破線が疾病にかかっている状態，×が疾病発生，○が治癒，●が死亡とする。



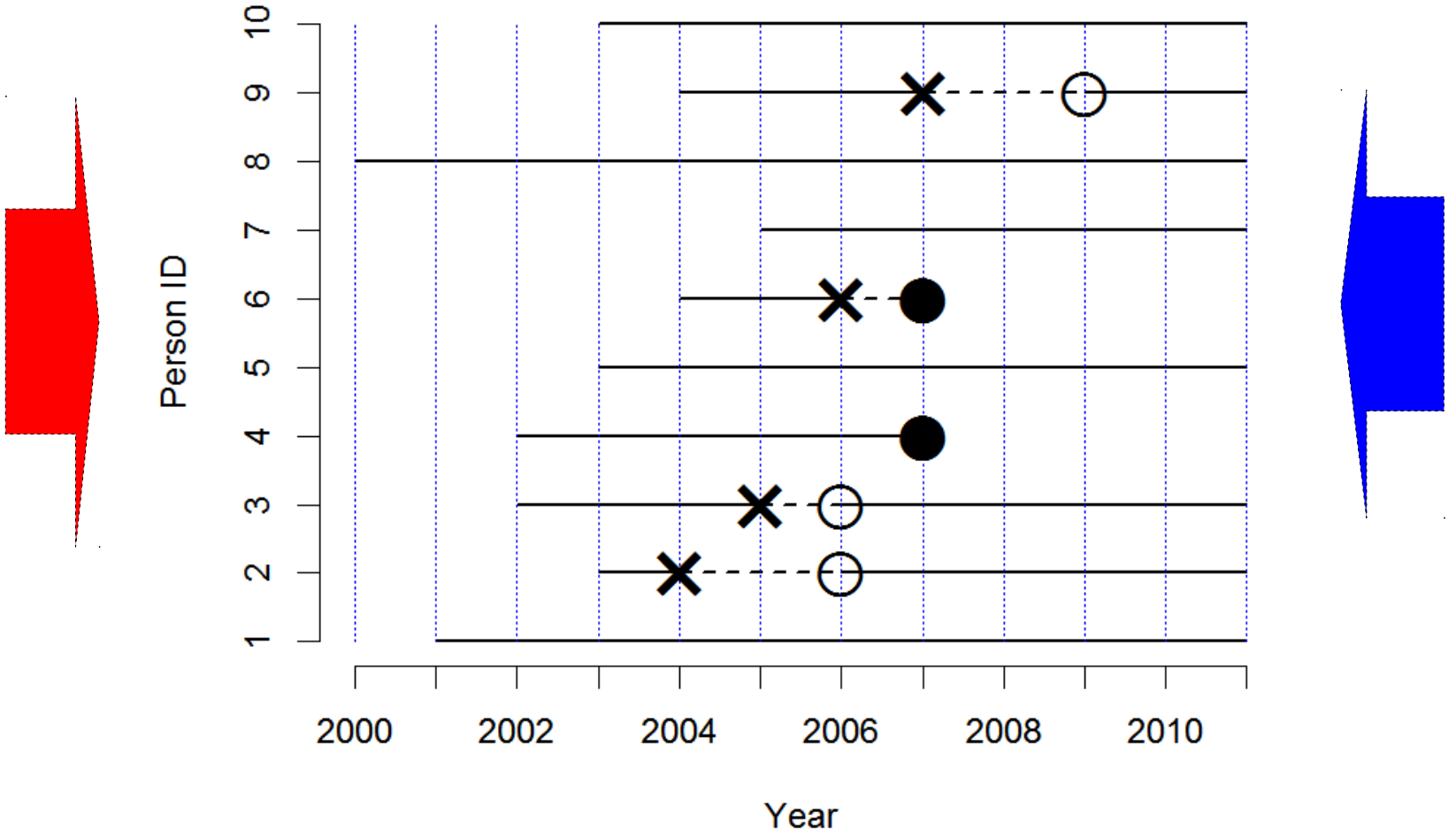
断面研究による有病割合の把握

- 2005年にこの集団の調査をしたら，10人中2人が病気なので疾病量としての有病割合は0.2（疾病オッズは2/8で0.25）といえる。←調査は簡単
- でも，2010年に調査をしたら0になってしまう。←正確さや代表性は弱い



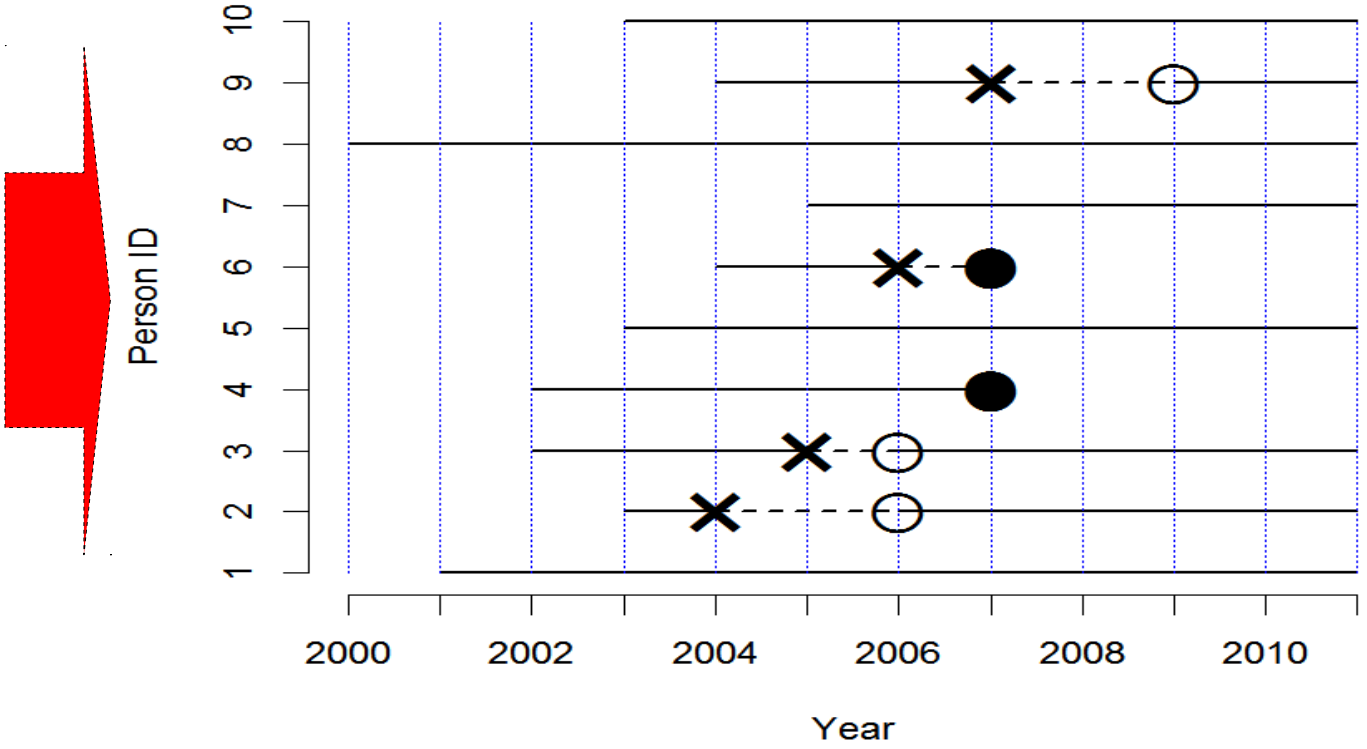
累積罹患率(リスク)の把握

- 2011年に生存者8名の親からの聞き取りで疾病罹患経験を調べた場合，最近10年間にこの疾病に罹った人は3人なので，10年間のリスクは3/8=0.375←調査は簡単だが，死亡例が脱落するし年齢もばらばら
- 2000年からの11年間のコホート研究ではリスクは4/10=0.4。しかし生後1年以内のリスクは1/10=0.1←観察期間によってリスクは変わる



コホート研究による罹患率の把握

- 2000年からの11年間のコホート研究のデータをフルに生かすには，分母を人数でなく人年にする方がいい。
- 一生に一度しか罹患しない疾病なら，罹患後（治癒後も含む）は感受性がない。死亡しても感受性がない。感受性のある（＝population at riskに属している）観察期間の合計を分母とし，発症数を分子にすると，罹患率が求められる。（1／年）という単位になる。
- 何度も罹患する疾病の場合，年初に感受性のある人口から1年間に発生する患者，と考え，年央人口を期待人年として分母に使い，1年間に発生する患者数を割って罹患率を推定することができる（通常は10万人当たりなどにする）。

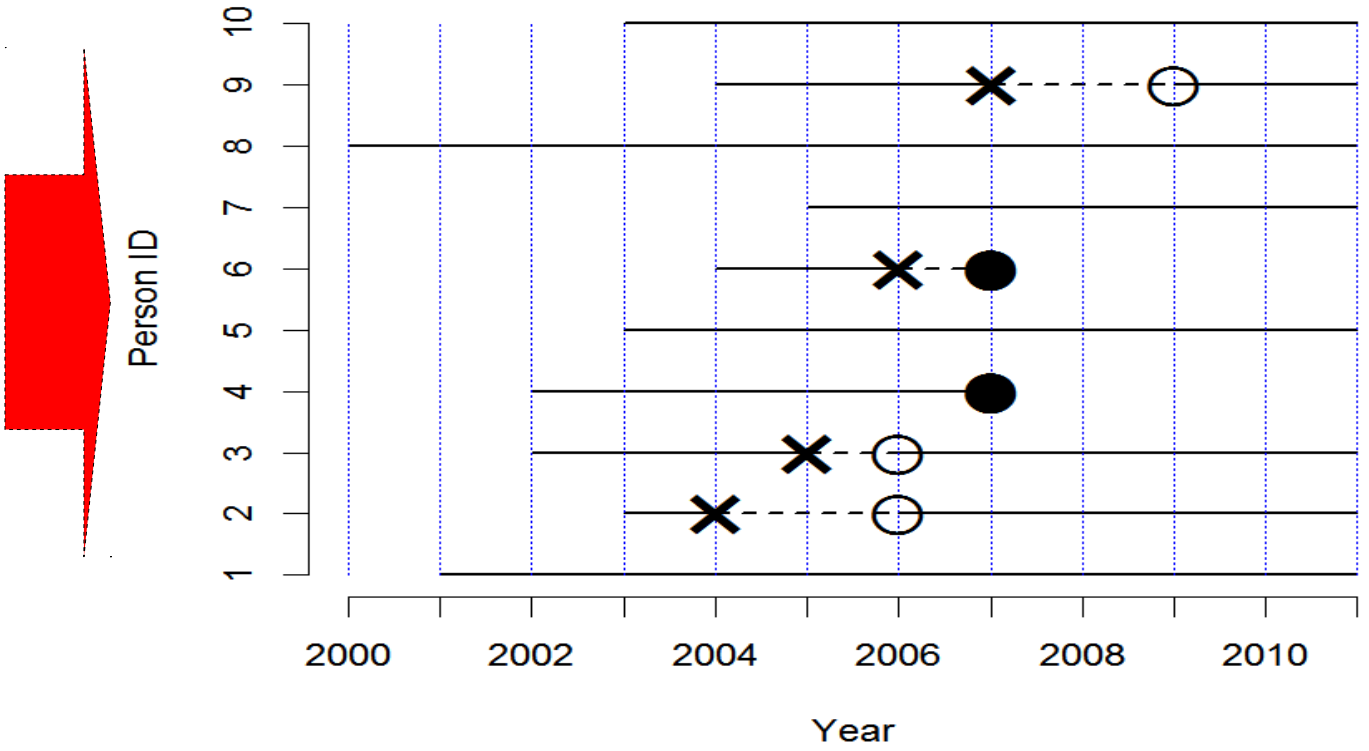


左の図で疾病×の罹患率は？

答えは次ページ

コホート研究による罹患率の把握

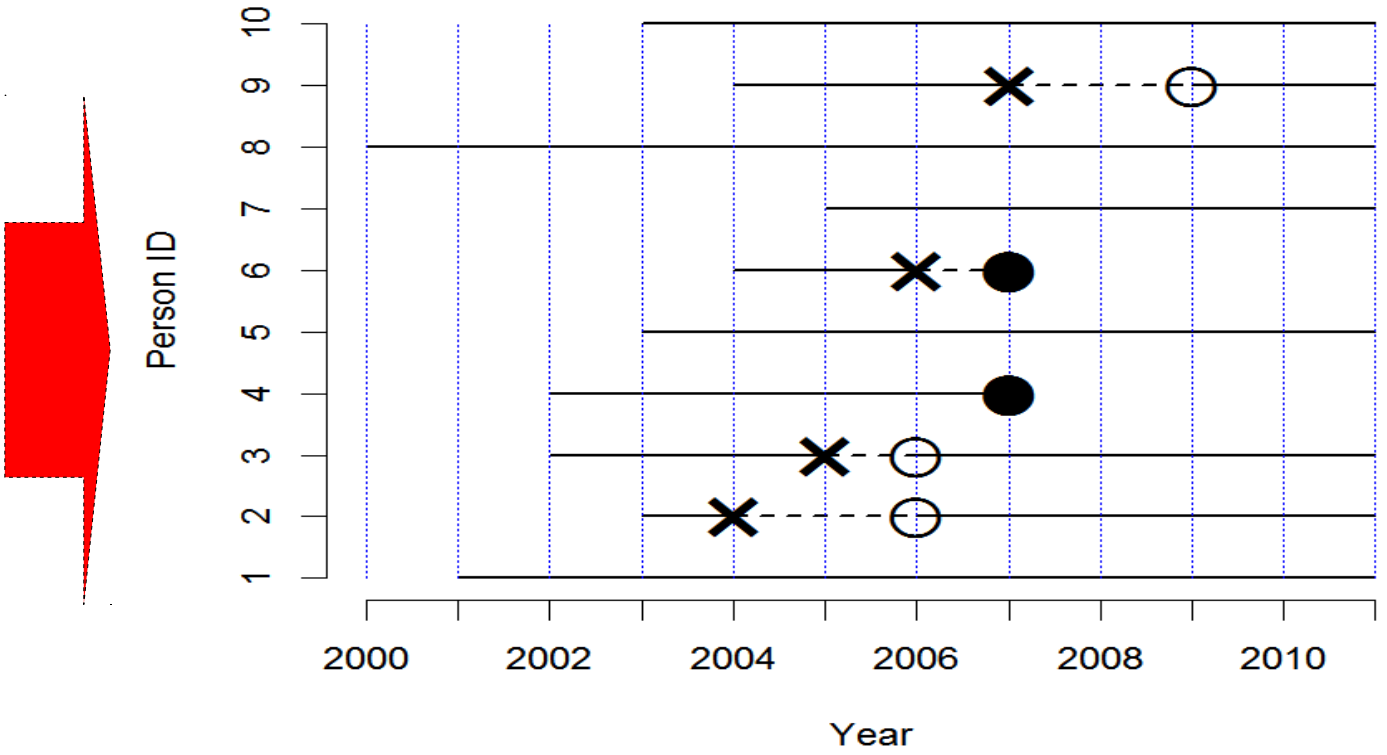
- 2000年からの11年間のコホート研究のデータをフルに生かすには，分母を人数でなく人年にする方がいい。
- 一生に一度しか罹患しない疾病なら，罹患後（治癒後も含む）は感受性がない。死亡しても感受性がない。感受性のある（＝population at riskに属している）観察期間の合計を分母とし，発症数を分子にすると，罹患率が求められる。（1／年）という単位になる。
- 何度も罹患する疾病の場合，年初に感受性のある人口から1年間に発生する患者，と考え，年央人口を期待人年として分母に使い，1年間に発生する患者数を割って罹患率を推定することができる（通常は10万人当たりなどにする）。



観察人年
8
3
11
6
2
8
6
3
1
10
合計58人年
4/58=0.069（／年）

死亡率の把握

- 観察のエンドポイントを疾病罹患でなく死亡にすると，死亡率が計算できる。
- 死亡は一生に一度しか罹患しない疾病と同様に考えることができる。
- 年初に感受性のある人口（＝まだ生きている人口）から1年間に発生する死亡，と考え，年央人口を期待人年として分母に使い，1年間に発生する死亡数を割って死亡率を推定することができる（通常は総死亡なら人口1000人当たり，死因別なら10万人当たりなどにする）。
  →**これだと2007年調査だけ0.2／年となり，他の年は0とばらつく。**



観察人年
8→8
3→7
11→11
6→6
2→3
8→8
6→6
3→9
1→8
10→10
合計58→76人年
2/76=0.026（／年）

（参考）個人史の図を描画するコード（関心がある人は後で見てください

```
> B1 <- B <- c(1, 3, 2, 2, 3, 4, 5, 0, 4, 3) + 2000
> DIS <- c(NA, 4, 5, NA, NA, 6, NA, NA, 7, NA) + 2000
> REC <- c(NA, 6, 6, NA, NA, NA, NA, NA, 9, NA) + 2000
> DIE <- c(NA, NA, NA, 7, NA, 7, NA, NA, NA, NA) + 2000
> E1 <- ifelse(is.na(DIS), ifelse(is.na(DIE), 2011, DIE), DIS)
> B2 <- ifelse(!is.na(DIS), DIS, NA)
> E2 <- ifelse(!is.na(DIS), ifelse(!is.na(REC), REC,
  ifelse(!is.na(DIE), DIE, 2011)), NA)
> B3 <- ifelse(!is.na(REC), REC, NA)
> E3 <- ifelse(!is.na(REC) & !is.na(DIE), DIE, 2011)
> par(mar=c(4,4,1,1),cex=2)
> plot(c(2000, 2011), c(1, 10), type="n", frame.plot=FALSE,
  xlab="Year", ylab="Person ID", axes=FALSE)
> axis(1, 2000:2011, 2000:2011)
> axis(2, 1:10, 1:10)
> segments(2000:2011,1,2000:2011,10,lwd=1,lty=3,col="blue")
> segments(B1, 1:10, E1, 1:10, lty=1, lwd=2)
> segments(B2, 1:10, E2, 1:10, lty=2, lwd=2)
> segments(B3, 1:10, E3, 1:10, lty=1, lwd=2)
> points(DIS,1:10,pch="×",cex=3)
> points(REC,1:10,pch="○",cex=3)
> points(DIE,1:10,pch="●",cex=3)
```

線分を描く（まず観察開始から疾病発生まで，次に疾病発生から治癒まで，最後に死亡まで）

イベントをプロットする（疾病発生，治癒，死亡の順）

疾病と曝露の関連の程度をみる＝曝露の有無間で疾病量を比べる

- 典型的な比べ方：
  差（絶対比較）または比（相対比較）
- 差と比のどちらがいいということはなく，目的次第で使い分ける。
- 疾病量ごとに異なる
  - リスクを出したとき＝リスク差かリスク比
  - 罹患率を出したとき＝罹患率差か罹患率比
  - 死亡率を出したとき＝死亡率差か死亡率比
  - 有病割合を出したとき＝有病割合比……ではない！オッズ比を用いる

絶対比較＝リスク差，罹患率差＝寄与危険（超過危険）ともいう

- （仮の例）送電線の近くに住む10万人（電磁波曝露群）を5年間観察して白血病発症者が毎年2人ずつの計10人，送電線から離れたところに住む10万人（対照群）を同じく5年間観察して白血病発症者が毎年1人ずつの計5人だったとする。また居住場所が送電線に近いかどうか以外の条件は2群間で差がないとする。
- リスク差は10/100000-5/100000=5/100000 (=5e-5)
- 罹患率差は
  10/(100000+99998+99996+99994+99992)
  -5/(100000+99999+99998+99997+99996)
  ≒0.0000100006（／年）
- いずれにせよ，希な疾病ならわずかな違いに見える。

（参考）曝露が予防や治療の場合のリスク差や罹患率差や再発率差

- （仮の例）胃がん治癒後に再発予防のために薬剤Aを定期服用する100人（予防群）では5年以内の再発が毎年１人の計５人，薬剤Aを定期服用しない100人（対照群）では5年以内の再発が毎年２人の計10人だったとする。なお，２群間には薬剤Aの服用有無以外に違いはないとする。こうした臨床疫学では，通常そのためすべての患者に対しインフォームドコンセントを得てランダムに2群を割り付ける（さらに，事前にその計画の倫理審査を通る必要あり）
- リスク差は5/100-10/100=–5/100。つまり薬剤Aの定期服用により，５年間の再発リスクは5/100小さくなる。逆数をとった20は，１人の再発を防ぐために必要な服用人数なので，NNT (Number Needed to Treat)と呼ばれる。臨床疫学の重要な概念。
- 再発率差は5/(100+99+98+97+96)-10/(100+98+96+94+92) =–0.01063（／年）。薬剤Aを100人が定期服用することで毎年約1人の再発を予防できることを意味する。
- 差は絶対比較なので，リスクや率が大きいと大きくなる。

相対比較（１）＝リスク比，罹患率比

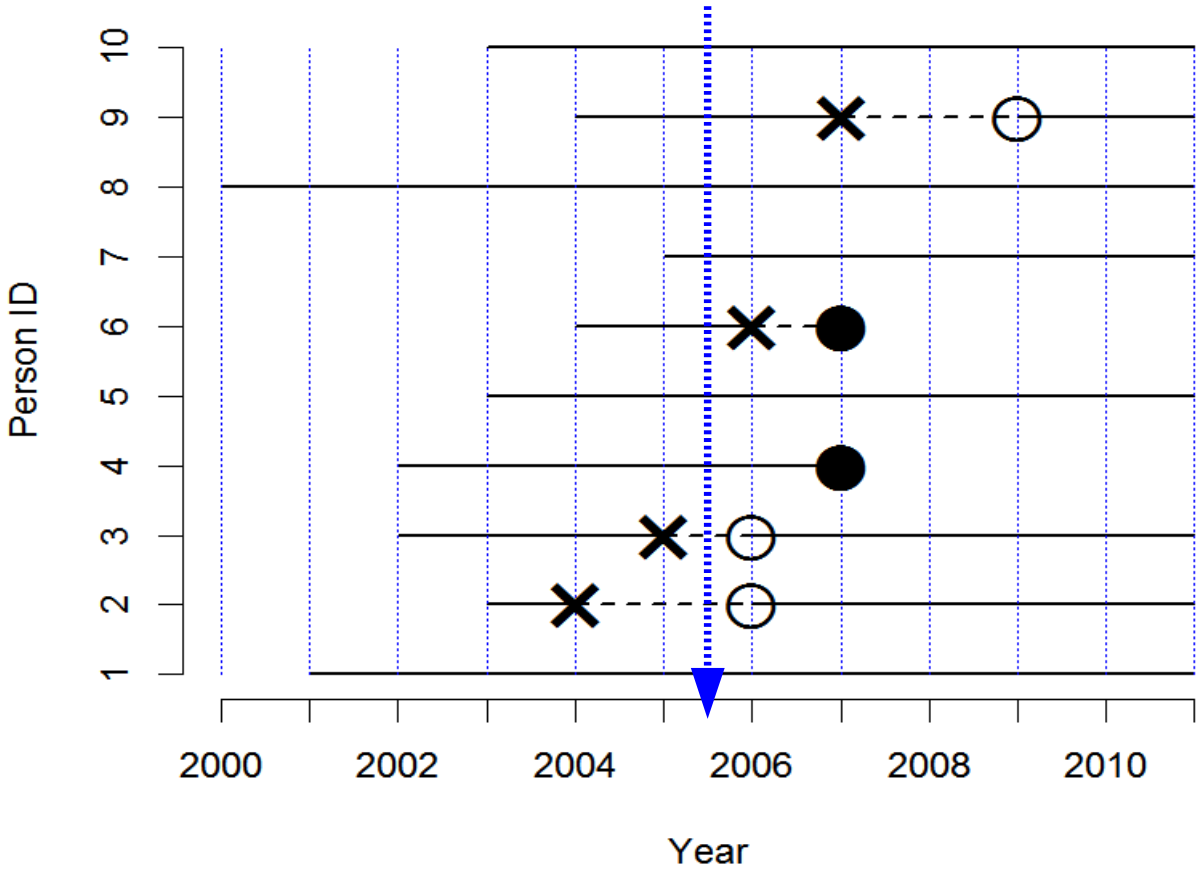
- （同じ例）送電線の近くに住む10万人（電磁波曝露群）を５年間観察して白血病発症者が毎年２人ずつの計10人，送電線から離れたところに住む10万人（対照群）を同じく５年間観察して白血病発症者が毎年１人ずつの計５人だったとする。また居住場所が送電線に近いかどうか以外の条件は2群間で差がないとする。
- リスク比は(10/100000)/(5/100000)=2
- 罹患率比は
  $\frac{10/(100000+99998+99996+99994+99992)}{5/(100000+99999+99998+99997+99996)} \fallingdotseq 2$
- どちらでも「電磁波曝露により白血病リスクが約2倍になった」といえる→統計学的有意性は？

リスク比の信頼区間と検定

- （同じ例）送電線の近くに住む10万人（電磁波曝露群）を5年間観察して白血病発症者が毎年2人ずつの計10人，送電線から離れたところに住む10万人（対照群）を同じく5年間観察して白血病発症者が毎年１人ずつの計5人だったとする。また居住場所が送電線に近いかどうか以外の条件は2群間で差がないとする。
- 前述の通り，リスク比は2だが，信頼区間の計算や，「リスクに差がない＝リスク比が1」という帰無仮説の検定ができる。
- 差についても，信頼区間の計算や「差がゼロ」という帰無仮説の検定はできる。ここでは省略する。
- 簡単なのは
  library(fmsb)
  riskratio(10, 5, 100000, 100000)
  とする方法。検定結果のp値だけみるよりp値関数推奨
  pvalueplot(matrix(c(10,5,99990,99995),2))

相対比較(2)＝オッズ比

- 断面研究では，有病割合の差や比を見ることはあまりない。2005年調査時に偶数番号の人が喫煙者，奇数番号の人が非喫煙者だったとすると，喫煙という要因あり群の疾病オッズも，要因なし群の疾病オッズも1/4



＊症例対照研究でもオッズ比を求める。2番と3番の人が患者として見つかったとき，1番と4～10番の人が対照群として選ばれたという状況なら，患者群の曝露オッズも対照群の曝露オッズも1
◎この例ではオッズの比は，いずれにせよ1

オッズ比の信頼区間と検定

- 断面研究でも症例対照研究でも，最初に曝露群と非曝露群が決まっているわけではないのでリスクや罹患率は出せないし，その比較もできない
- オッズ比はロジスティック回帰分析でも得られる。その場合は，共変量の影響が調整済み。
- （例）M市の成人から年齢分布に応じてランダムサンプルした200人について，喫煙の有無と持続的な咳の有無を質問紙調査した結果，次のクロス集計表が得られたとする

持続的咳	有	無
喫煙有	50	50
喫煙無	10	90

- オッズ比は，(50/50)/(10/90)=(50/10)/(50/90)=9
  この値はコホート研究におけるリスク比や罹患率比の近似値と考えることができる→喫煙者は非喫煙者の9倍持続的咳症状あり
- 信頼区間の推定や「オッズ比が1」という帰無仮説の検定は，

```
fisher.test(matrix(c(50,10,50,90),2)) # 最尤推定による
library(fmsb); oddsratio(50,10,50,90) # 正規近似による
pvalueplot(matrix(c(50,10,50,90),2),plot.OR=TRUE,xrange=c(0,20))
```

オッズ比の計算例

- MASSライブラリのsurveyデータフレームについて，性別（変数Sex）と利き手（変数W.Hnd）のクロス集計を行い，女性は男性の何倍左利きの可能性が高いかをオッズ比により検討せよ
- やりたいことの分解
  - MASSライブラリの読み込み
  - クロス集計表Xの作成
  - オッズ比計算
- Rのコードは下記（約0.69倍で5%水準で有意差なし）

```
library(MASS)
X <- xtabs(~Sex+W.Hnd, data=survey)
fisher.test(X) # または下記3行
library(fmsb)
oddsratio(X[1,1],X[2,1],X[1,2],X[2,2])
pvalueplot(X, plot.OR=TRUE)
```

スクリーニングの評価（1）

- 2×2クロス集計表のもう1つの使い方として，スクリーニングにおける検査法の評価がある

疾病	有	無
検査陽性	a	b
検査陰性	c	d

- この検査法の性能評価は，通常下記が高いほどよい
  - 感度(sensitivity)＝a/(a+c)
  - 特異度(specificity)=d/(b+d)
  - 陽性的中度(positive predictive value)=a/(a+b)
  - 陽性尤度比＝感度／（1－特異度）＝(a/(a+c))/(b/(b+d))
  - 陰性的中度(negative predictive value)=d/(c+d)
- 感度と特異度は有病割合の影響を受けないが，有病割合が低い希な疾患だと陽性的中度は低くなる

スクリーニングの評価（2）

- 検査方法の評価では，信頼性も重要。検査再検査信頼性(test-retest-reliability)は前回触れたκ係数など
- ROC (Receiver Operating Curve)による評価
  - 最適カットオフ値を決める
  - 複数の検査法の性能を比較する（カットオフ値が可変なとき，感度と特異度はトレードオフの関係になるので，片方では比較できない）
  - （例）精神科医がうつと診断した人6人とうつでないと診断した人4人について，うつ質問紙調査のスコア（10点～30点を取り得る）があるとき，何点をそこより上を陽性と判定するためのカットオフ値に決めるのが最適か？　を考える。
    ＊カットオフ値が9点なら全員が陽性なので感度1，特異度0
    ＊カットオフ値が30点なら全員が陰性なので感度0，特異度1
    →途中のどこかに感度も特異度もそこそこ高くなる最適カットオフ値があるはず→縦軸に感度，横軸に1－特異度をとってカットオフ値を変えて曲線で結ぶROCの最も左上に近い点が最適カットオフ。それが左上に近いほど高性能。
  - 時間があれば，library(fmsb); example(roc)を試されたい。
- 詳しくは3年公衆衛生学講義資料を参照のこと。