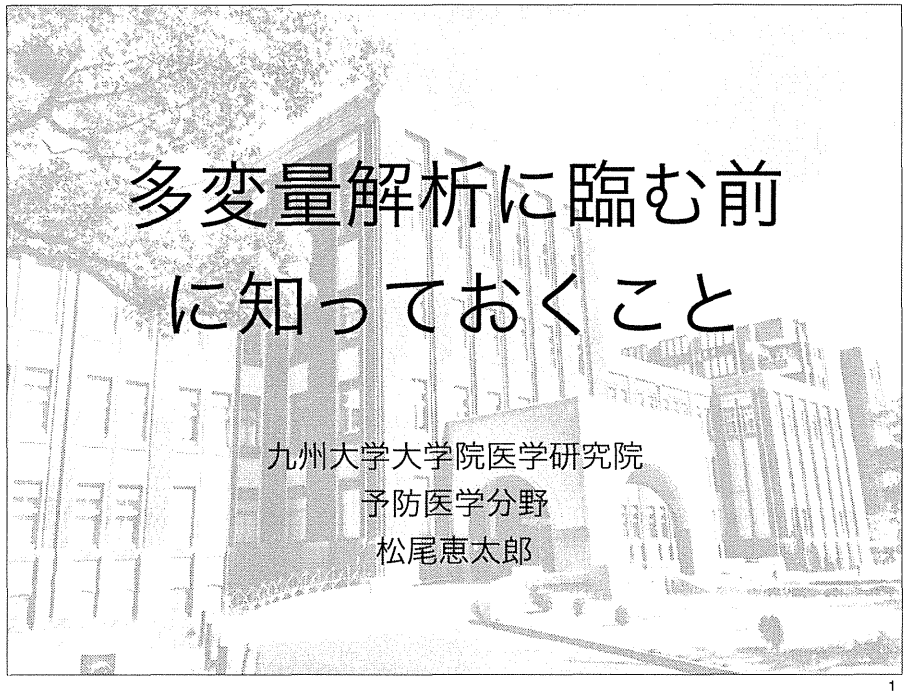


多変量解析に臨む前に知っておくこと

九州大学大学院医学研究院
予防医学分野
松尾恵太郎

本日の話の大まかな流れ

- 解析前の流れ
- 生存解析
- 多変量解析の仕組
- 第3の要因をどう取り扱うか？
- おまけ
  - モデルの検証
  - 統計パッケージで解析する際の基本原則
  - α、βエラー

解析前の流れ

大まかな流れ

- 研究で答える疑問をただ一つ設定
  - 対象者(Population)
  - 主たる検討要因(**E**xposure)
  - アウトカム(**O**utcome)
  - その他の関連する要因（交絡、交互作用の可能性のある要因 **C**ovariates or **C**onfounder）
- どの研究デザインを採用するか？

E, O, and C

E
曝露

O
アウトカム

C
交絡要因候補

この構図に落とし込めてない状態では解析しない

5

疑問の設定時

- TRUMPのようなデータでは、色々なことが出来てしまいます
- 失敗する研究は何がしたいかが絞れていない研究です

6

4つの代表的な研究デザイン

- 無作為割付試験
- コホート研究
- 症例対照研究
- 横断研究

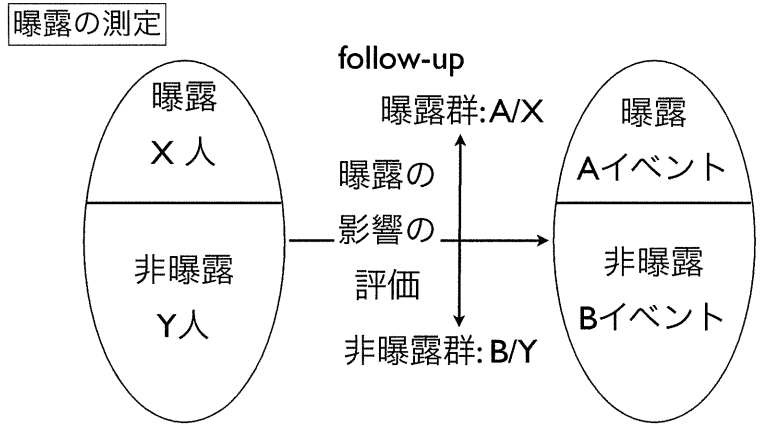
Intervention
Experimental

Observational

7

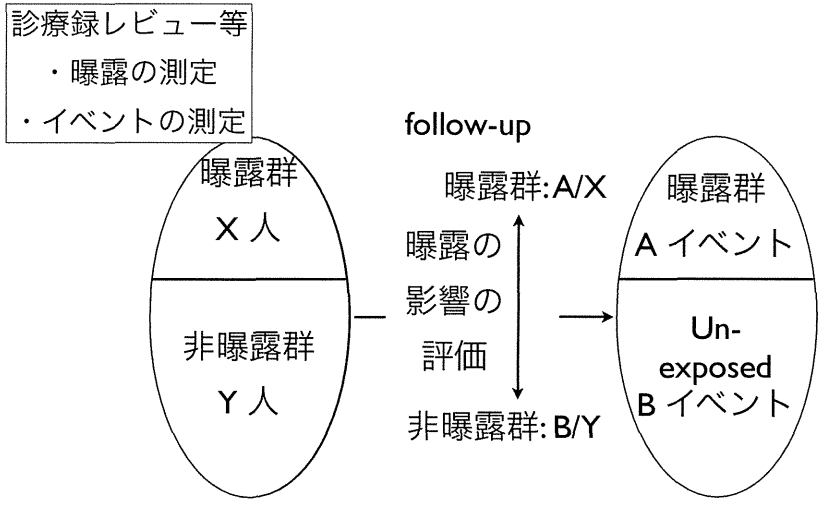
7

前向きコホート研究



8

後向きコホート研究



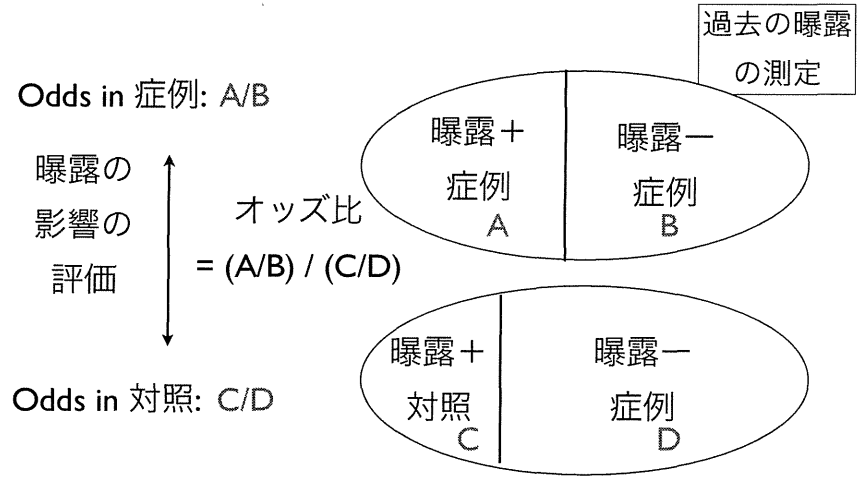
9

コホート研究 (続き)

- Pro
  - イベントがおきる前に曝露要因を測定している、つまり曝露要因に関するエラーが少ない（前向き）
  - 情報の拾い方などにエラーを起こす可能性がある（後向き）
- Con
  - イベント発生までに時間がかかる（前向きの場合）
  - 希なイベントは評価しがたい
- よく使う解析モデルは比例ハザードモデル
- TRUMPデータでやる生存解析は前向きコホート？後向きコホート？

10

症例対照研究



11

Case-control study (cont')

- Pro
  - 前向き、後向きコホートよりも時間がかからない
- Con
  - 曝露測定にバイアスの入る可能性
    - 曝露要因次第（遺伝子とかの場合関係ない）
  - 対照群設定の難しさ
- よく使う解析モデルはロジスティック回帰分析（マッチング有りの時は条件付きロジスティック回帰分析）

12

横断研究

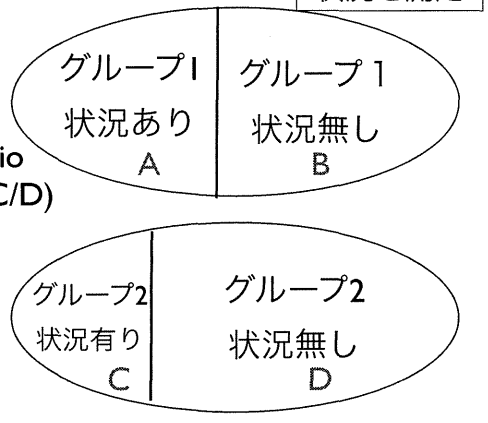
調査時点での状況を測定

Odds in グループ1: A/B

群間の状況の違いを評価

Odds ratio = (A/B) / (C/D)

Odds in グループ 2: C/D



この研究デザインは原因を云々するデザインではない

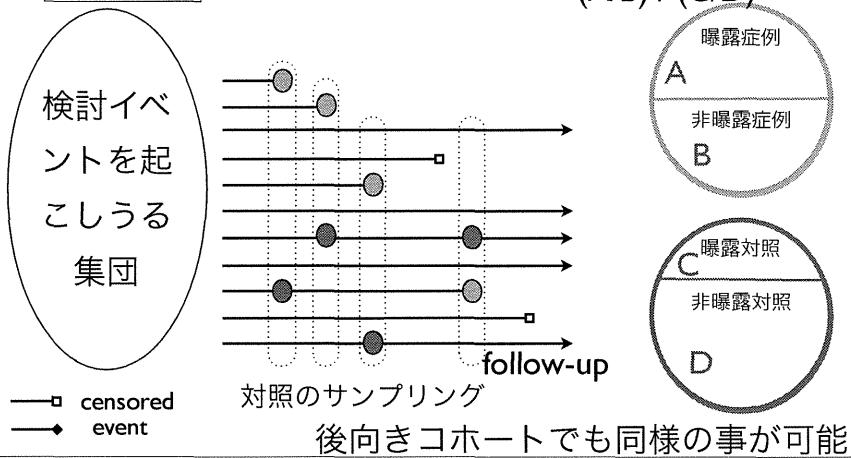
13

コホート内症例対照研究

Odds in 症例: A/B　オッズ比

曝露の測定

Odds in 対照: C/D　= (A/B) / (C/D)



後向きコホートでも同様の事が可能！

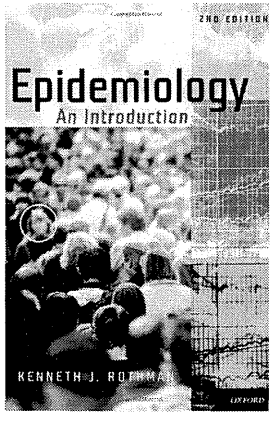
14

コホート内症例対照研究（続き）

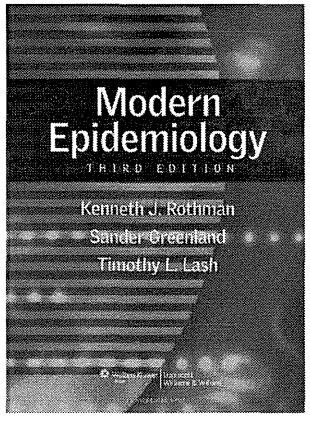
- コホート研究のメリットを享受できる
  - 曝露測定はコホートの枠組み
- 追加調査が必要な場合に全例を調査しなくても済む
  - 実施可能性を高める
- 解析は条件付きロジスティック回帰分析

15

疫学の学習に関して



Epidemiology: An Introduction.
by Rothman KJ.



Modern Epidemiology.
by Rothman KJ.

16

生存解析

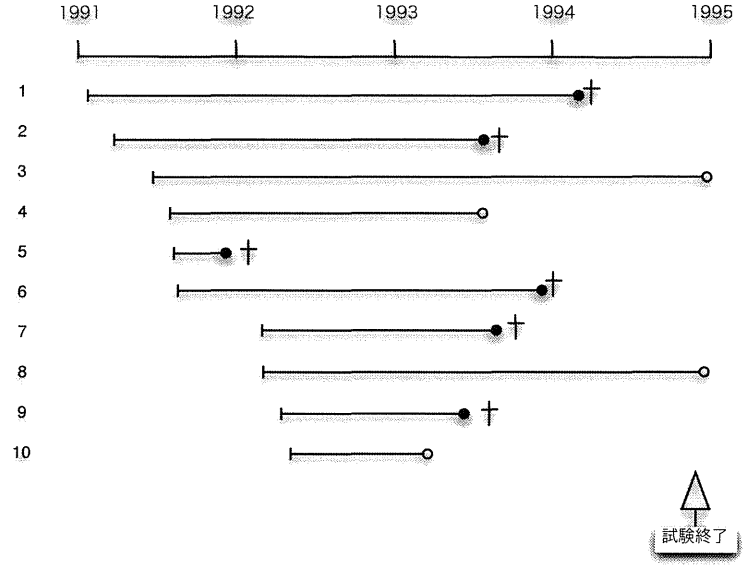
17

Survival analysis

||

Time-to-event analysis

18



19

$$\text{死亡率} = \frac{\text{死亡数（=6）}}{\text{全体（=10）}}$$

20

何が問題か？

- 観察期間の存在を無視している
- イベントの観察が行われなかった症例(中途打ち切り症例: censored case)の取扱い
  - at-risk for death でない人を分母に入れている

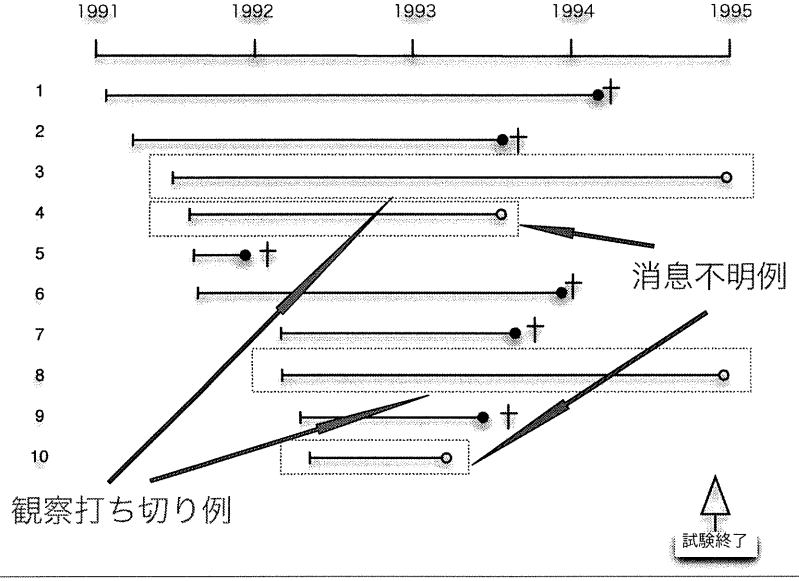
21

Censored case

- 中途打ち切り例
  - 消息不明例
  - 観察打ち切り例

22

エンドポイント：生存



23

二つの中途打ち切り例は同じか？

- 観察打ち切り例
- 消息不明例

No

消息不明例が多くなるような場合には解析結果の解釈に注意が必要

informative censoring

24

中途打ち切り例を検討できる代表的な解析手法

- Kaplan-Meier法
- Cox Proportional Hazard Model

25

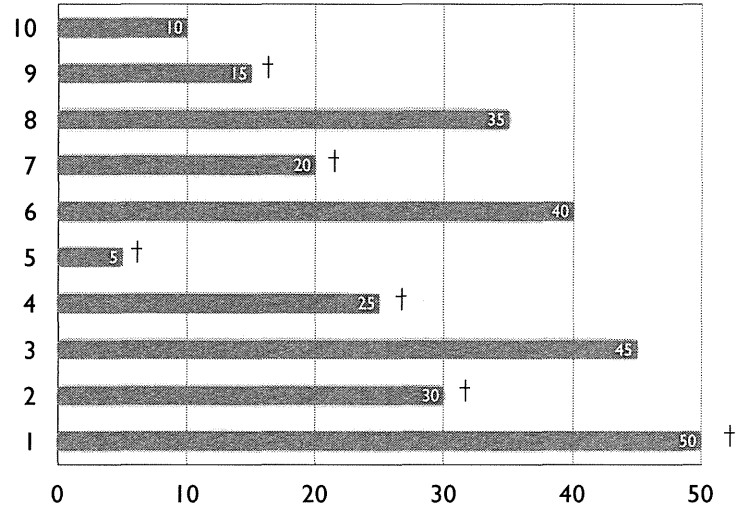
生存率

- Kaplan-Meier曲線による生存率
- 観察期間の中央値などの指標は提示されているか？
- 中途打ち切り例がちゃんとヒゲで分かるようになっているか？

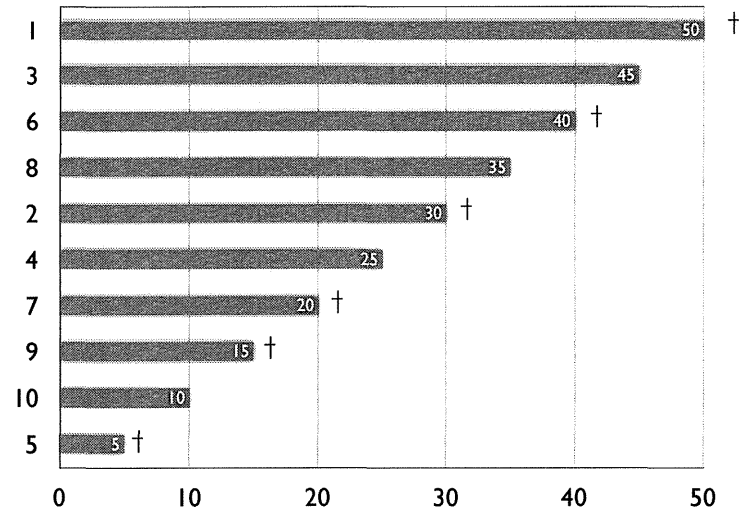
26

ID	Tx day	Last day	Survival	Months	Censored
1	1990.1	1994.2	Dead	50	
2	1991.3	1993.8	Dead	30	
3	1991.5	1995.1	Alive	45	観察打切
4	1991.7	1993.7	Alive	25	消息不明
5	1991.8	1991.12	Dead	5	
6	1991.8	1994.12	Dead	40	
7	1992.2	1993.9	Dead	20	
8	1992.2	1995.1	Alive	35	観察打切
9	1992.3	1993.5	Dead	15	
10	1992.4	1993.2	Alive	10	消息不明

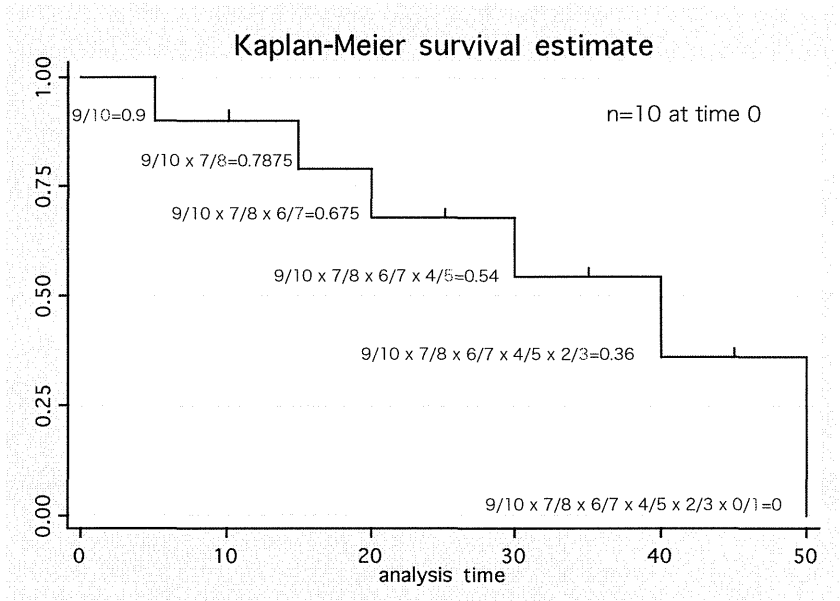
27



28



29



30

KM curveではある期間内のリスクを
掛け合わせていったものが
生存率の推計値として用いられている

↑

KMでは中途打ち切り例は
長く観察していれば、打ち切りに
なっていない症例と同様のイベント
発生率を示す、ということが前提

31

Cox model

- 回帰分析の一つ。式で表すと
  - $\lambda(t| x_1, ..., x_n)= \lambda_0 (t)\exp(b_1X_1+....+b_nX_n)$
- ハザードとは、「非常に微少な時間における死亡率」
- ハザードは検討対象変数X_1を除いて一定
  （proportional hazard assumption）

32

Cox model、多変量解析

- $\lambda(t|x_1,x_2,x_3)= \lambda(t_0)\exp(b_1X_1+... + b_3X_3)$
- $\log(\lambda(t|x_1,x_2,x_3))= \log(\lambda(t_0))+ (b_1x_1+... + b_3X_3))$
- $\log(\lambda(t|_{x1=1},x_2,x_3))= \log(\lambda(t_0))+ (b_1 \cdot 1 + ... + b_3X_3))$
- $\log(\lambda(t|_{x1=0},x_2,x_3))= \log(\lambda(t_0))+ (b_1 \cdot 0 + ... + b_3X_3))$

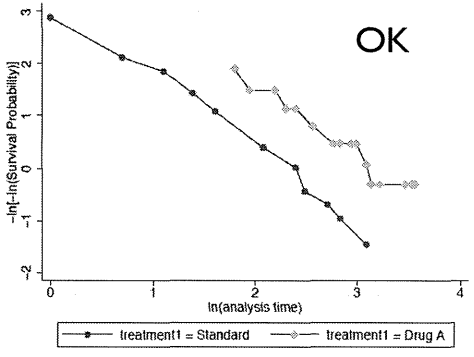
$\log(\lambda(t|_{x1=1},x_2,x_3))-\log(\lambda(t|_{x1=0},x_2,x_3)) = b_1$

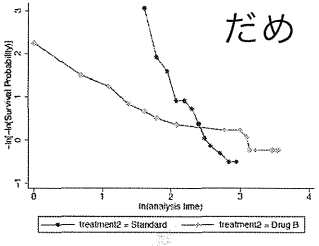
$$\log \frac{\lambda(t|_{x1=1},x_2,x_3)}{\lambda(t|_{x1=0},x_2,x_3)} = b_1 \longrightarrow HR=\exp(b_1)$$

ハザード比

STATAでのproportional hazard assumptionの検証

- stphplot, strata(X_1)





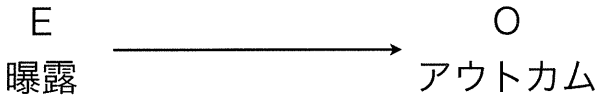
STATAでのproportional hazard assumptionの検証　その他

- stcox x1 x2 x3　を実施後に
- estat phtest, detail　とすると、
  - モデル全体並びに各変数ごとにPHAに関しての有意性の検討をしてくれる

第3の要因をどう取り扱うか？

交絡、交互作用等

E, O, and C



C
第3の要因

大まかに4つのパターン

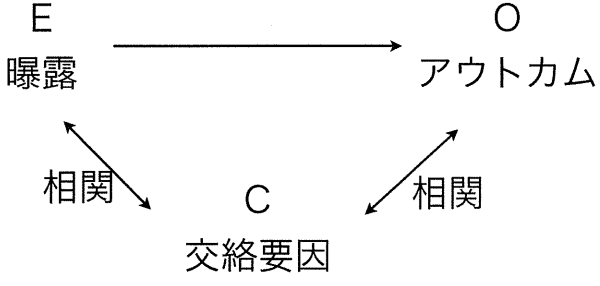
37

交絡要因
Confounder

- 条件
  - それ自体がアウトカムの発生と相関している
  - 検討対象曝露と相関している
  - 曝露とアウトカムの中間的な要因ではない

38

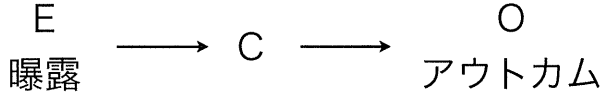
E, O, and C



EとOの関係は、Cの影響を考慮しないと正確に評価が出来ない

39

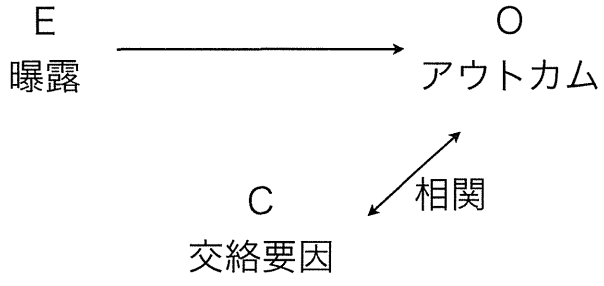
中間要因



CはEの結果として、EとOの間にあるため、これを解析に入れるとEの影響が正しく評価できない　＝　補正してはいけない

40

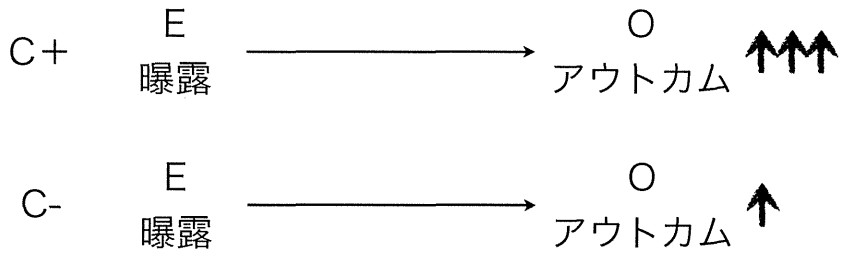
Eと相関してないC



EとOの関係は、Cの影響を考慮しなくても
評価できる　＝　補正しなくても良い

41

CはEのOへの影響に影響を与える（交互作用）



CはE->Oの関係を変える要因＝効果変容因子
補正してはいけない–＞層別化が必須

42

交絡要因を調整しないと

E <-> C	C <-> O	非調整の結果
正の相関	正の相関	過大評価
正の相関	負の相関	過小評価
負の相関	正の相関	過大評価
負の相関	負の相関	過小評価

43

交絡の影響の排除法

- 研究前に出来ること：比較集団間で交絡要因をマッチングする
- 研究後に出来ること：
  - 多変量解析による調整
  - 層別化解析　ーーー＞
    - 交互作用の検討も出来る

44

交互作用

C+　E 曝露 ⟶ O アウトカム　↑↑↑

C-　E 曝露 ⟶ O アウトカム　↑

CはE->Oの関係を変える要因＝効果変容因子
補正してはいけない–＞層別化が必須

45

交互作用がある時

	検討対象 曝露ー	検討対象 曝露＋
交互作用 要因ー	HR=1.0	HR=1.5
交互作用 要因＋	HR=1.5	HR=5

HR=5=1.5 x 1.5 x 2.22　交互作用ハザード比

46

STATAでの層別化・交互作用の検討

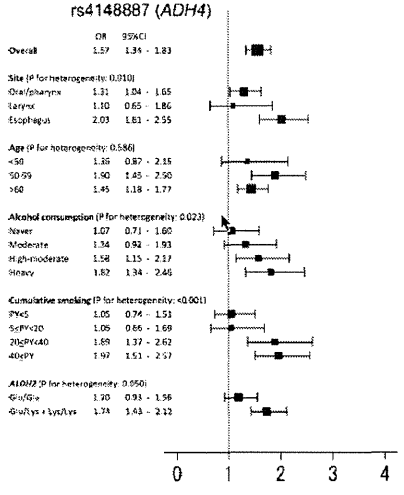
- 交絡要因の層別に検討する曝露のHRを推定
  - 同じようなＨＲが出ているか？
  - ＨＲが層別で大きな差があるとき
    - →交互作用の推定

47

- 例えば年齢50歳以上(age_50)でHLA mismatch(seromis6abdr)の影響が異なるかもしれない（交互作用があるかもしれない）と考えている
- 実例
- 交互作用の解釈に関して

48

Forrest plot



層別化解析を可視化

交互作用の有無が見やすい

おまけ１
多変量モデルの評価

ある変数を入れてモデルは良くなったか？

- Likelihood ratio test （尤度比検定）が一般的
  - stcox E cov1 cov2 cov3
  - estimates store m1　<-m1という名前で尤度を保存
  - stcox E cov1 cov2 cov3 cov4
  - estimates store m2　<-m2という名前で尤度を保存
  - lrtest m1 m2 <--LR test

おまけ２
統計パッケージで解析する際の基本原則

原則

- 元データは読み込むが上書きはしない
- 全ての解析をスクリプトでやる
  - 逐次解析をしない（しても良いが、最終解析は絶対逐次解析にしない）
  - 必ずログが残る解析にする
  - 再現性が保てない解析は解析でない

53

- set more off　（スクロールストップがかからなくなる）
- set logtype t　（ログをテキスト形式にする）
- log using XXXX,txt, replace
- use xxxx.dta, clear
- save work_xxxx.dta, clear
- describe　（データの場所、人数、変数とうの基本情報）
- codebook　（変数の詳細な情報がでる、毎回は不要）

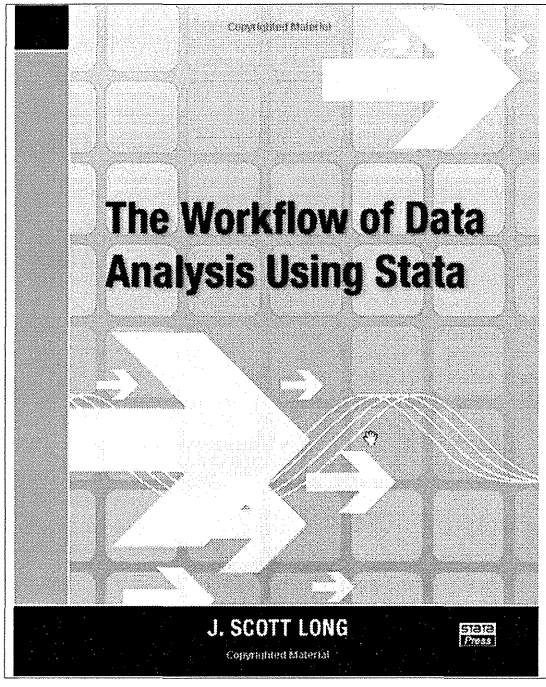
- /*for table 1*/
- tab xxxx yyyy, col row chi2 exact

- /*for table 2*/
- stset os, f(death==1)
- stsum

- sts graph, by(sex)
- graph save xxx.gph, replace

- log close

54

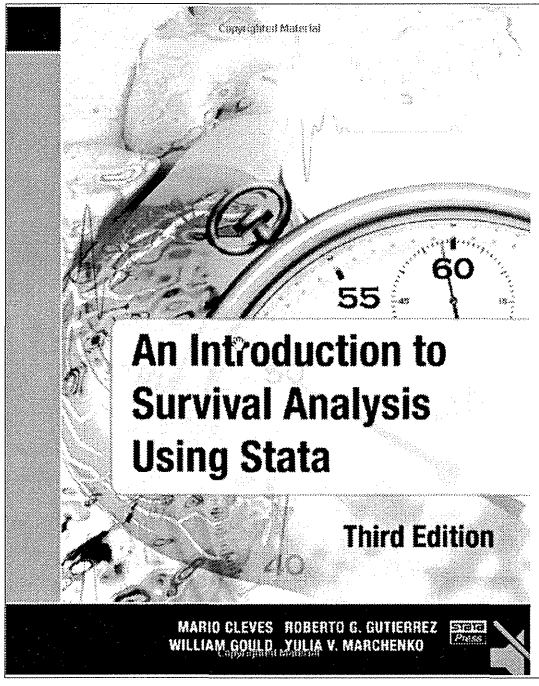


プロフェッショナルな解析をSTATAでする上での必携書

コンセプト自体はSTATA以外のパッケージでも使える

US$54.02
Amazon.com

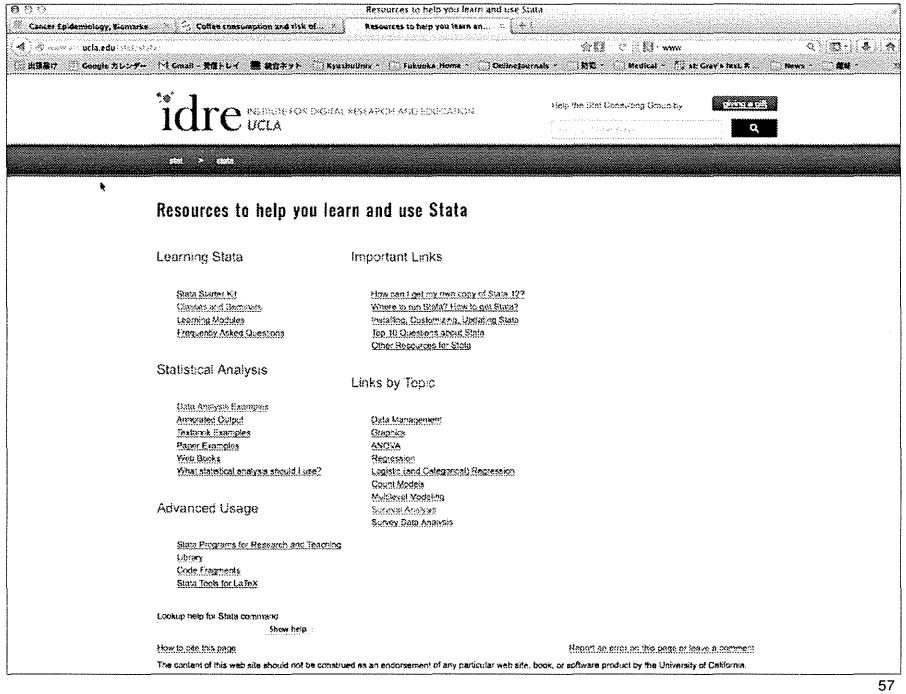
55



生存解析をSTATAでする上で参考になる

US$67.26
Amazon.com

56



57

おまけ３
α、βエラー

58

- 例）今持っているサイコロが、いかさまサイコロかどうかを実験で評価する
  - もしちゃんとしたサイコロなら、１がでる可能性は1/6である
  - これを検証するために実験を計画する

59

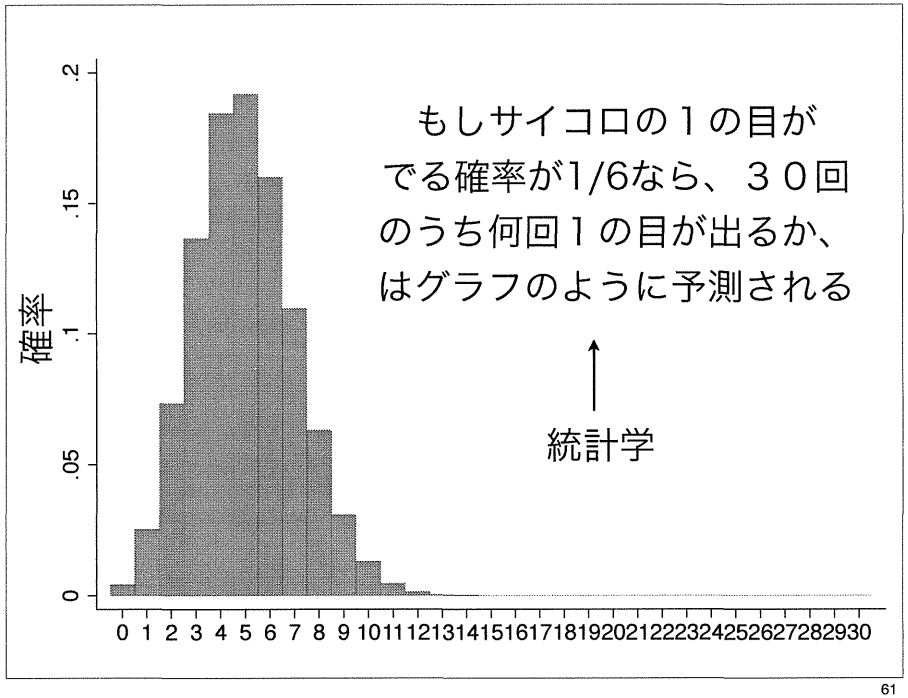
1.「１」の目がでる可能性は1/6である 仮説の設定

2. サイコロを30回振る　実験方法の規定

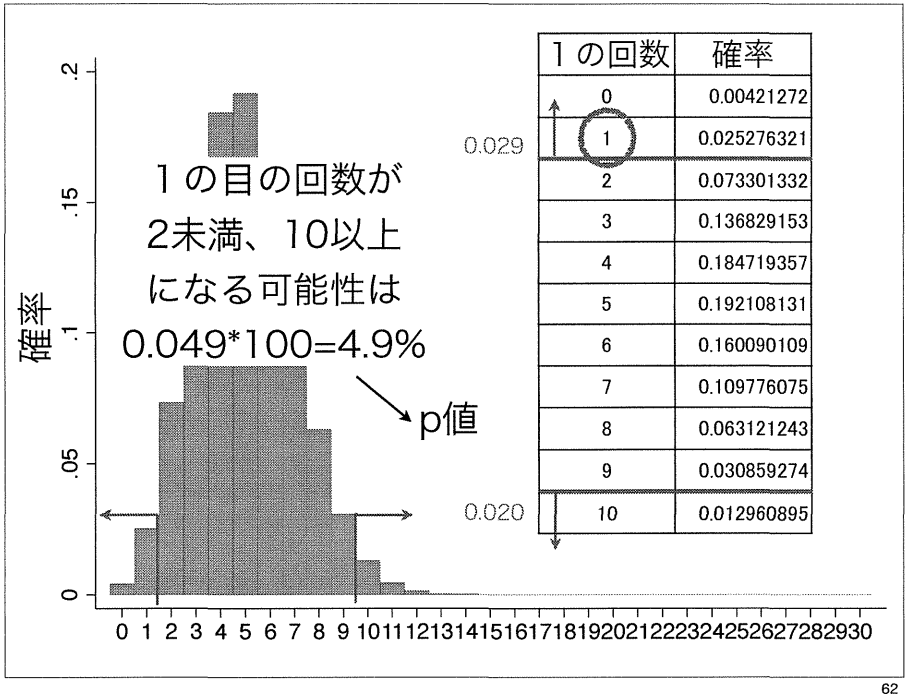
3.「１」の目が２回未満あるいは10回以上出た場合はいかさまサイコロと判定　判断基準の設定

これらを実験前にあらかじめ規定してから評価することが、実験により科学的評価をする、という事である。

60

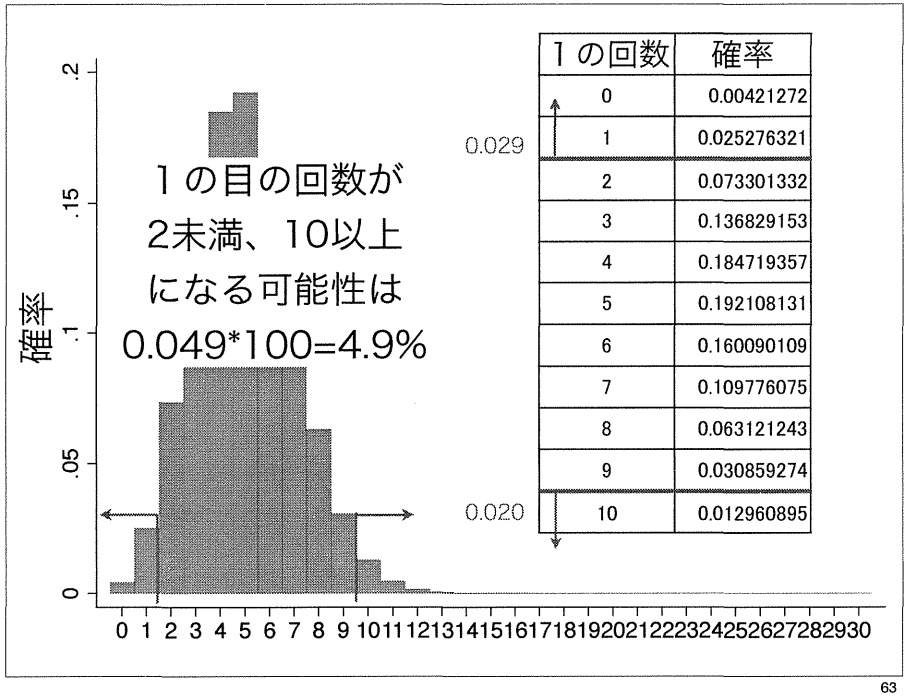


61



１の回数	確率
0	0.00421272
1	0.025276321
2	0.073301332
3	0.136829153
4	0.184719357
5	0.192108131
6	0.160090109
7	0.109776075
8	0.063121243
9	0.030859274
10	0.012960895

62



１の回数	確率
0	0.00421272
1	0.025276321
2	0.073301332
3	0.136829153
4	0.184719357
5	0.192108131
6	0.160090109
7	0.109776075
8	0.063121243
9	0.030859274
10	0.012960895

63

サイコロ

仮説の設定
↓
実験方法の規定
判断基準の設定
↓
実験を実施し
結果を測定
↓
結果を評価

・サイコロの確率は1/6

・30回振る
・「１」が2回未満、
10回以上でいかさま判定

「１」の目が１回
＝すごく低い確率
の事（p値）が起きている

どうもサイコロの
確率は１／６では
ないらしい

帰無仮説
対立仮説
サイコロの確率
は1/6でない

仮説の検定

64

Cox modelの場合

仮説の設定
↓
実験方法の規定
判断基準の設定
↓
実験を実施し
結果を測定
↓
結果を評価

・要因Aがあると
HRは1.0

帰無仮説

・800人の集団でHRを計算
・p値がαエラー(0.05)より低い場合には、帰無仮説を捨てて、対立仮説を採るぞ

対立仮説
HR != 1.0

・要因Aがある場合に
HR=1.3であった
p値=0.02だった

要因Aはイベントの発生を変える要因であった

65

帰無仮説が正しい時のエラー＝αエラー

仮説検定の結果

	HR=1と判断	HR != 1と判断
HR=1が真実 ＝帰無仮説正しい	正しい	誤り ＝αエラー

66

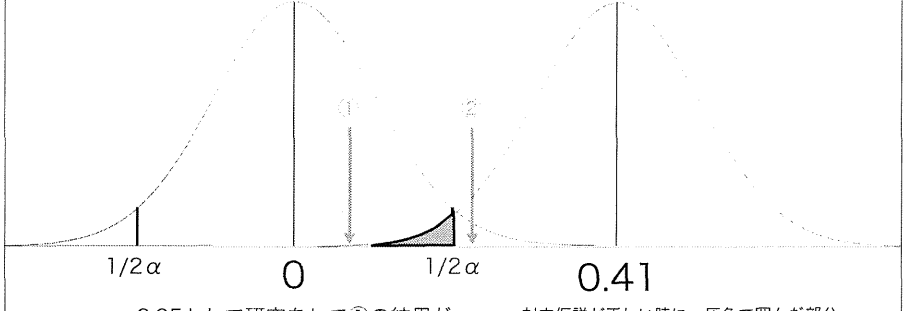
もう一つのエラー
βエラー

- αエラーは、帰無仮説が正しい時に、間違って帰無仮説が誤っているという判断をする確率
- βエラーは、対立仮説が正しい時に、間違って対立仮説を誤っているという判断をする確率
  - つまり、100-βは、対立仮説が正しい時に、それを正しいと判断する確率（検出力とかパワーと呼ばれる）

67

帰無仮説
HR=1
logHR=0

対立仮説
HR=1.5
logHR=0.41



α=0.05として研究をして①の結果だった場合、有意差無し。
②の結果の場合は有意である。

対立仮説が正しい時に、灰色で囲んだ部分（帰無仮説とα/2と被っている部分）の面積は、対立仮説が正しいのにうっかり対立仮説を否定して帰無仮説を選択してしまう確率＝βエラー

68

二つのエラー

仮説検定の結果

	帰無仮説が 正しいと**判断**	対立仮説が 正しいと**判断**
帰無仮説が本当 に正しい時	正しい	誤り ＝αエラー
対立仮説が本当 に正しい時	誤り ＝βエラー	正しい

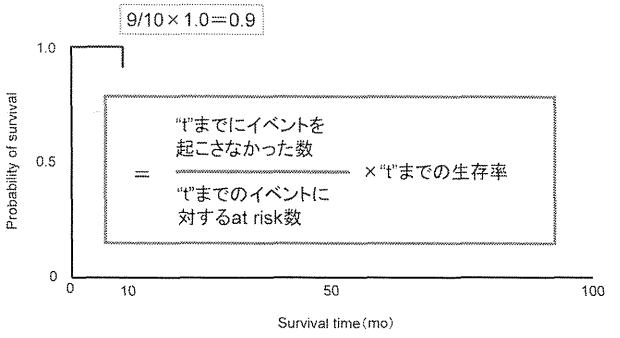
69

競合リスクイベントの扱いの解説と演習

TRUMP統計セミナー2013 応用編
2013年11月23日

競合リスクイベントの扱いの解説と演習

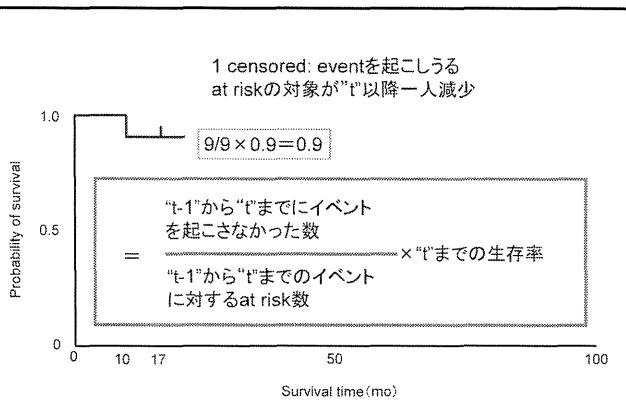
名古屋大学医学部
造血細胞移植情報管理・生物統計学
熱田　由子



Outcome dataの種類

1st type: "survival data"
- ランダムな時間におこる単純なイベント
- Overall survival (death), Relapse free survival (death or relapse)

2nd type: "competing risk data"
- あるイベントが起こることで、同じ対象における他のイベントが起こらなくなるイベント
- Relapse ⇔death without relapse



Outcome dataの種類

Kaplan-Meier法で生存曲線を描出

Cumulative incidence curveで描出すべき