

厚生労働科学研究委託事業(難治性疾患等実用化研究事業(免疫アレルギー疾患等実用化研究事業
移植医療技術開発研究分野))

「本邦における造血細胞移植一元化登録研究システム及び研究データ質管理システムの確立」
平成 26 年度 TRUMP 統計セミナー：応用編

日時： 平成 26 年 11 月 30 日（日） 10:00～15:30

場所： 名古屋大学大学院医学系研究科・医学部保健学科 東館 4 F 大講義室
〒461-0047 名古屋市東区大幸南一丁目 1 番 20 号

《プログラム》

9:30～ 受付、資料配布

10:00-10:30

- I. 我流 Stata の使い方：自分なりの方法を確立しよう
(島根大学医学部附属病院 腫瘍・血液内科 准教授 鈴木 律朗)

10:30-12:00

- II. TRUMP を用いた二次調査を計画する際に知っておきたい疫学的なこと
(九州大学大学院医学研究院 予防医学分野 教授 松尾 恵太郎)

12:00-13:00

昼食

13:00-14:15

- III. 多変量解析モデル構築の具体例
(日本造血細胞移植データセンター 熱田 由子)

14:15-15:30

- IV. 時間依存性変数の扱いの解説と演習
(自治医科大学附属さいたま医療センター 血液科 講田 淳也)

我流 Stata の使い方：自分なりの方法を確立しよう

2014年11月30日(日)
名古屋大学大幸キャンパス

我流Stataの使い方： 自分なりの方法を確立しよう

鶴根大学医学部附属病院
腫瘍センター 腫瘍・血液内科
鎌本律朗

昨年のセミナー



前回：2013.11.23. TRUMP統計セミナー @ 名古屋大学
昨年から基礎編と応用編に分けました。

昨年のセミナー



前回：2013.11.23. TRUMP統計セミナー @ 名古屋大学
昨年から基礎編と応用編に分けました。

飛行機人間 と グライダー人間

- 飛行機人間：
自分で空を飛ぶことができる
自分で問題を設定して、解決して行く
大学教育。
研究活動。
- グライダー人間：
空は飛べるが、助けが必要
自力で飛ぶことができない
他人が作った問題を解くのが上手
高校生(大学受験)まで

外山滋比古「思考の整理学」

TRUMPセミナー応用編にあたって

自分なりの「飛び方」を考えてください。

- 我々の言っていることが、必ずしも正しいとは限りません。
- ネットで調べるなどして、適切な手法か状況判断して下さい。 (自己責任)
- 質問・相談は歓迎しますが、高度な内容になると、回答者によって考え方方が異なります。
- 論文執筆になると、その方向性を左右することになります。

我流 Stata 使用法

Stata が少し使えるようになったら…

自分なりのメモ・使い方を、残しましょう。

ノートでもよいです。でも、紛失したり散逸したりするので、コンピューターのファイルがよいと思います。

記憶力に頼るのは、年齢的に危険です。

私は、Excelで自分が使えるようになったコマンドを管理して、日々追加しています。

コマンド	説明	例
exit	終了	exit
exit, saving	退出するときに、その状態を保存する	exit, saving
get help	ヘルプ	get help
get memory	メモリの確認	get memory
get memory off	メモリを開放する	get memory off
get search off	検索を停止する	get search off
get update off	更新を停止する	get update off
help	ヘルプ	help using <command>
help using	ヘルプ	help using <command>, search
help using <command>, search	ヘルプ	help using <command>, search
help using <command>, searchreplace	ヘルプ	help using <command>, searchreplace
help using <command>, searchreplace replace	ヘルプ	help using <command>, searchreplace replace
helps	ヘルプ	helps
helps using	ヘルプ	helps using <command>
helps using <command>, search	ヘルプ	helps using <command>, search
helps using <command>, searchreplace	ヘルプ	helps using <command>, searchreplace
helps using <command>, searchreplace replace	ヘルプ	helps using <command>, searchreplace replace
helps using <command>, searchreplace replace replace	ヘルプ	helps using <command>, searchreplace replace replace
helps using <command>, searchreplace replace replace detail	ヘルプ	helps using <command>, searchreplace replace replace detail

コマンド	説明	例
help	ヘルプ	help <command>
get help	ヘルプ	get help
get memory	メモリの確認	get memory
get search off	検索を停止する	get search off
get update off	更新を停止する	get update off
help	ヘルプ	help using <command>
help using	ヘルプ	help using <command>, search
help using <command>, search	ヘルプ	help using <command>, search
help using <command>, searchreplace	ヘルプ	help using <command>, searchreplace
help using <command>, searchreplace replace	ヘルプ	help using <command>, searchreplace replace
helps	ヘルプ	helps
helps using	ヘルプ	helps using <command>
helps using <command>, search	ヘルプ	helps using <command>, search
helps using <command>, searchreplace	ヘルプ	helps using <command>, searchreplace
helps using <command>, searchreplace replace	ヘルプ	helps using <command>, searchreplace replace
helps using <command>, searchreplace replace replace	ヘルプ	helps using <command>, searchreplace replace replace
helps using <command>, searchreplace replace replace detail	ヘルプ	helps using <command>, searchreplace replace replace detail

こっちの方が、TRUMP セミナー基礎編のテキストより、自分用にカスタマイズして使用することができます。

テキストを読んだりして、使った経験のあるコマンドを追加していくと良いと思います。

コマンドが増えれば、それだけ自分のスキルが向上したことが目で見て分かります。

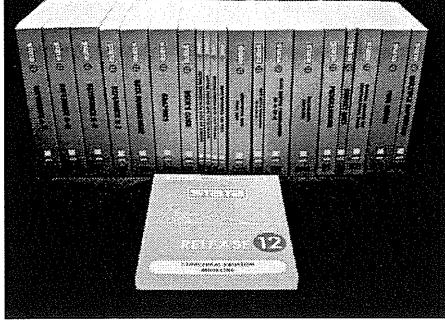
基本コマンド

- cii confidence interval immediately
- tabi tabulate immediately
- csi chi square immediately
- cci case control immediately
- iri incidence rate immediately
- sampsi sample size

help コマンドと、Stataのマニュアル

help コマンドと、Stataのマニュアル

Stataには、もともとマニュアルの本が附随。(有料)



ver. 13 では、紙のマニュアル本は廃止されました。

help コマンド

help + コマンド名で、コマンドの使い方が参照できます。

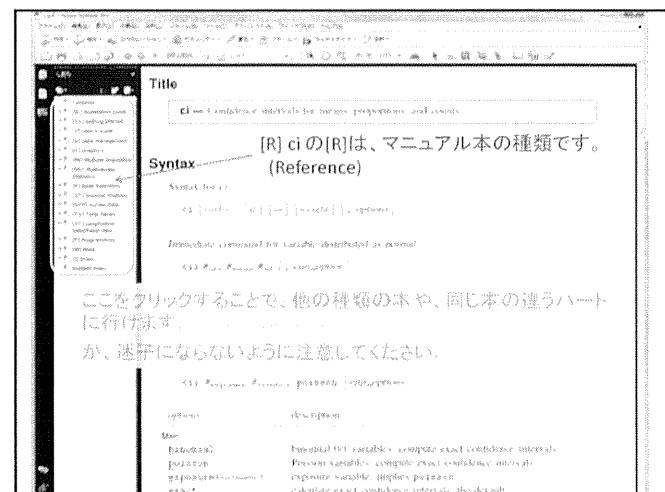
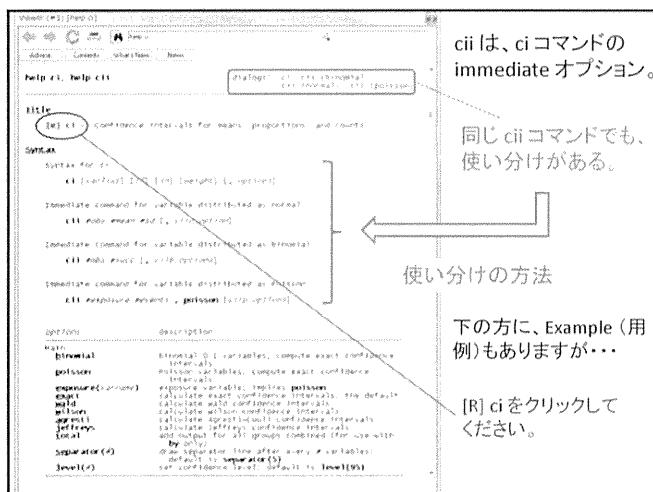
help cii と、タイプしてみてください。

```

STATA 11.2 Copyright 1985-2005 StataCorp LP
StataCorp LP, 4900 Stata Way
College Station, Texas 77845 USA
http://www.stata.com
979-695-4000 stata@stata.com
979-695-4001 (fax)

Single-user Stata perpetual license:
Serial number: 2011133810F
Extended to: R. Suzuki
Licensee info:
Notes:
1. E-line option or -set memory-> 50.00 MB allocated to data
2. E-line option or -set maxvar-> 5000 maximum variables
3. help sticek
4. help cii
5.

```



PDF版のhelp(または本)の方が、結果表示があるので分かりやすいです。

Ordinary confidence intervals

Example 1

Without the binomial or poisson option, `ci` produces "ordinary" confidence intervals, meaning those that are correct if the variable is distributed normally, and asymptotically correct for all other distributions satisfying the conditions of the central limit theorem.

use <http://www.statapress.com/datasets/ci11.dta>
1970 Automobile Data

コマンド行 → `ci mpg price`

Variable	Obs	Mean	Std. Err.	(95% Conf. Interval)
mpg	74	21.973	.6725441	19.959 - 22.63769
price	74	6165.257	342.8719	5826.914 - 6484.5

結果表示

コマンド行 → `ci mpg price, level(99)`

Variable	Obs	Mean	Std. Err.	(95% Conf. Interval)
mpg	74	21.973	.6725441	19.51849 - 23.07611
price	74	6165.257	342.8719	5256.405 - 2672.108

Immediate form

Example 7

We flip a coin 10 times, and it comes up heads only once. We are shocked and decide to obtain a 99% confidence interval for this coin.

```
cii 10 1, level(99)
```

Variable	Obs	Mean	Std. Err.	(95% Conf. Interval)
	10	.1	.0048683	.0006011 -.6442871

Example 6

We are reading a soon-to-be-published paper by a colleague. In it is a table showing the number of observations, mean and standard deviation of 1980 median family income for the Northeast and West. We correctly think that the paper would be much improved if it included the confidence intervals. The paper claims that for 106 cities in the Northeast, the average of median family income is \$19,369 with a standard deviation of \$4,379.

For the Northeast.

```
cii 106 19809 4379
```

Variable	Obs	Mean	Std. Err.	(95% Conf. Interval)
	106	19559	303.8763	18837.98 - 20160.07

For the West.

```
cii 256 22557 5000
```

Variable	Obs	Mean	Std. Err.	(95% Conf. Interval)
	256	22557	312.6875	21541.42 - 23472.78

sts graph のようなグラフコマンドだと、PDFでは図が入る。自分が描きたい図を、目で見て探せるので便利。

`. sts graph, by(drug) risktable`

The graph shows Kaplan-Meier survival estimates for two groups: drug = 0 (solid line) and drug = 1 (dashed line). The x-axis is 'analytic time' from 0 to 40, and the y-axis is 'survival' from 0.0 to 1.0. The legend is located at the bottom right of the plot area.

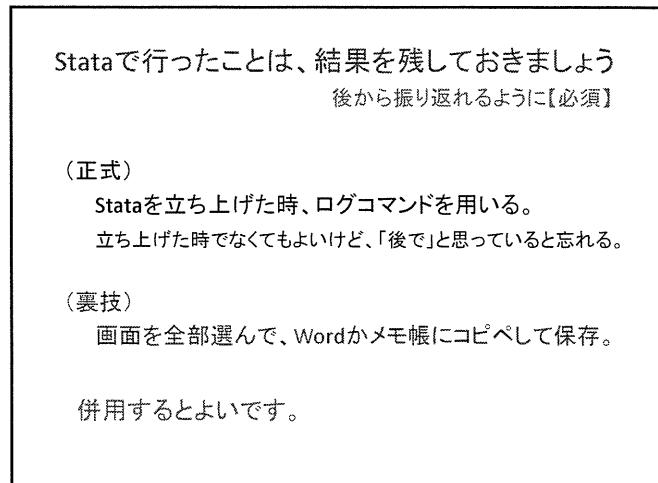
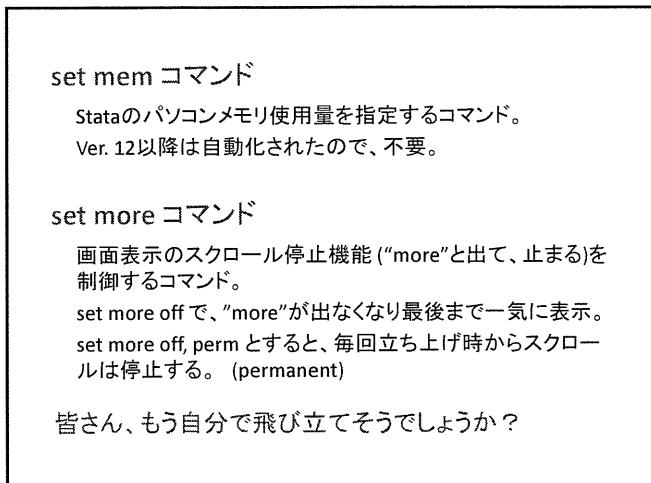
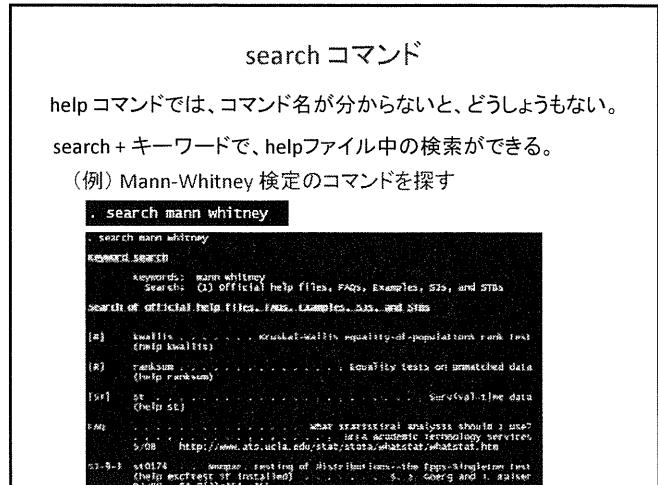
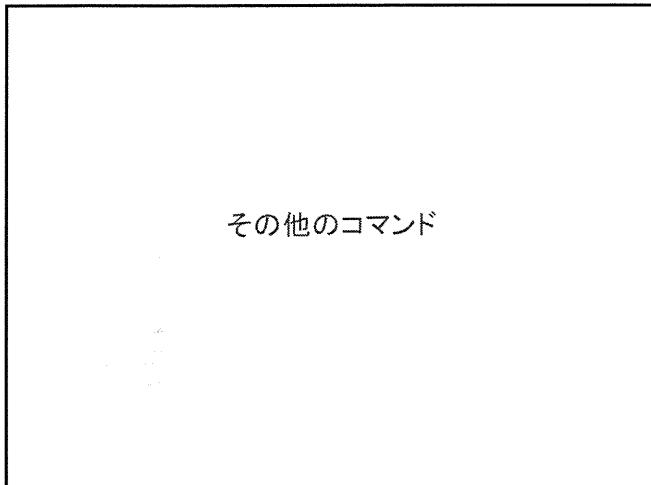
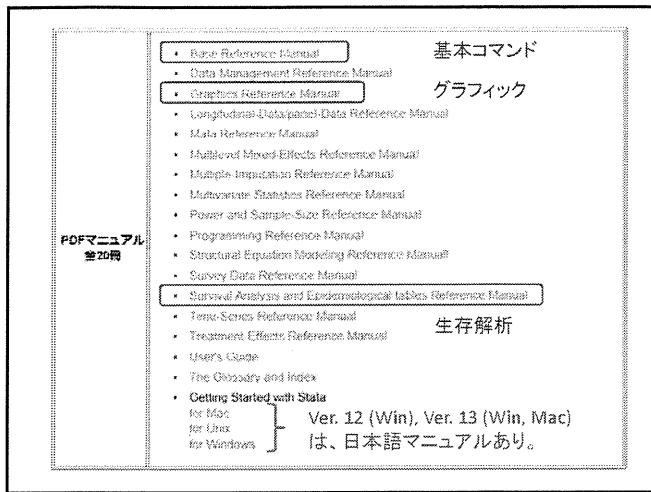
risktable オプションで、at risk 数を追加できる。

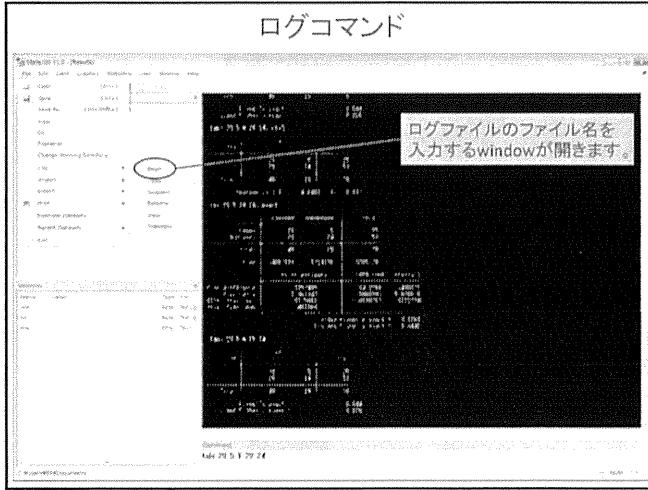
`. sts graph, by(drug) risktable legend(lg1 position(2) rows(2))`

The graph shows Kaplan-Meier survival estimates for two groups: drug = 0 (solid line) and drug = 1 (dashed line). The x-axis is 'analytic time' from 0 to 40, and the y-axis is 'survival' from 0.0 to 1.0. The legend is located in the top right corner of the plot area.

By default, row titles are placed on the left of the at-risk table and are right-justified. We can illustrate this by changing the text of the row titles to have an unequal length.

凡例 (legend) を移動させるオプション。

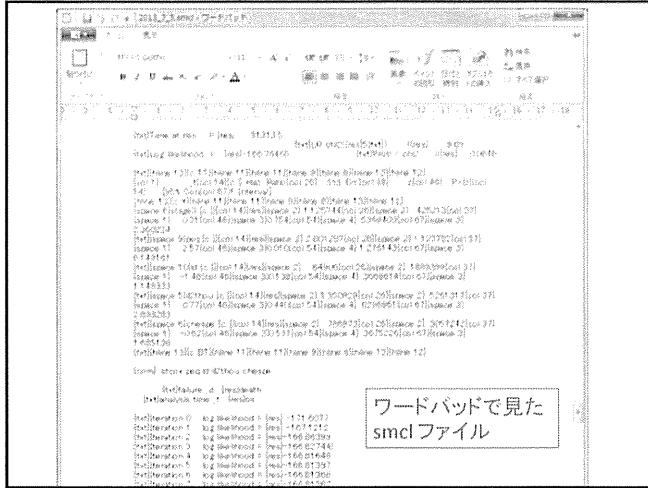




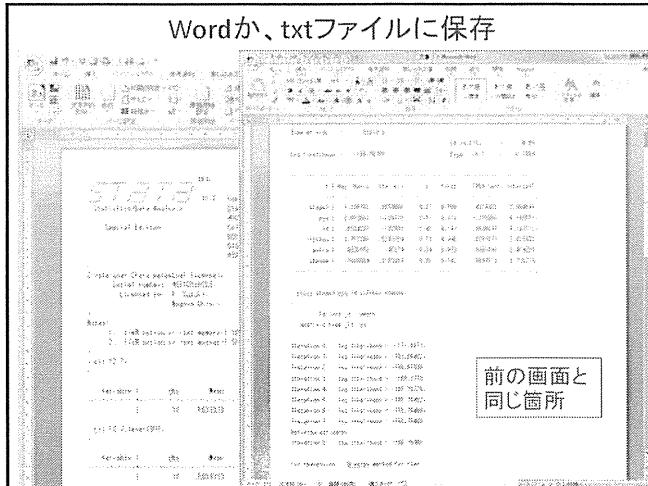
Stata ログファイルは .smcl ファイルになります

ダブルクリックすると、Stataが立ち上がって内容を見ることができますが、中身は text file です。

でも、通常の text reader では、Stata の描画コマンド (line など)が邪魔をして、かなり慣れないと理解不能です。



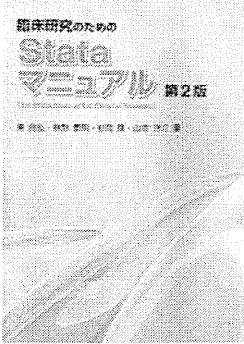
Stataでは、こう見える



テキスト形式でのログファイルの保存

- ※ 後からもできるので、気付いた時に対応可能。
- ※ Stataとは別画面で見られる。
- ※ Stataがインストールされていない別のパソコンで見る時には、分かりやすい。
- ※ コマンド行をStataにコピペして、動作を再現することも可能。
- ※ ただし、何百行も解析する時は、最初の結果が失われることもあるので、十分なバッファ量を確保するか、正式なログコマンドを使うべき。

参考文献



参考文献 (PDF)

あざやわかる Stata 入門

<http://www.sscn.jp/stata/pdf/001.pdf> (2013/09/06) 50 頁

はじめに

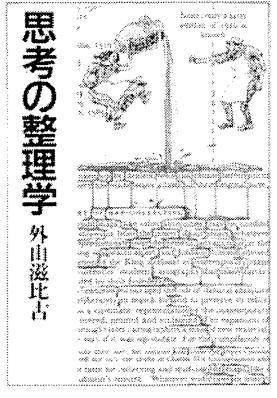
このページは 2004 年に大蔵大学で作成した講義資料を更新したものとされています。また、著者からの許可を得て、あくまで個人的な目的で複数枚複数部を複数人で共有する場合は問題ありません。
ただし、商業利用や複数枚複数部を複数人で共有する場合は、著作権法の規定により法律で禁じられています。
また、著作権法で禁じられております。

Stata 入門の入門

<http://www.sscn.jp/stata/pdf/002.pdf> (2013/09/06) 50 頁

はじめに

このページは 2004 年に大蔵大学で作成した講義資料を更新したものとされています。また、著者からの許可を得て、あくまで個人的な目的で複数枚複数部を複数人で共有する場合は問題ありません。
ただし、商業利用や複数枚複数部を複数人で共有する場合は、著作権法の規定により法律で禁じられています。
また、著作権法で禁じられております。



ちくま文庫 562円

今のが社会は、つよい学校信仰ともいべきものを持っている。

学校はグライダー人間の訓練所である。飛行機人間は作らない。グライダーの練習に、エンジンのついた飛行機などがまじっていては迷惑する。危険だ。学校では、ひっぱられるままに、どこへでもついて行く従順さが尊重される。勝手に飛び上がったりするのは規律違反。たちまちチェックされる。やがてそれぞれに、グライダーらしくなって卒業する。

グライダーとして一流である学生が、卒業間際になって論文を書くことになる。これは、これまでの勉強といさか勝手が違う。(中略) グライダーは途方にくれる。突如としてこれまでと違うことを要求されてもできるわけがない。グライダーとして優秀な学生ほどあわてる。

一般に、学校教育を受けた期間が長ければ長いほど、自力飛翔の能力は低下する。グライダーとして飛べるのに、危ない飛行機になりたくないのは当たり前である。

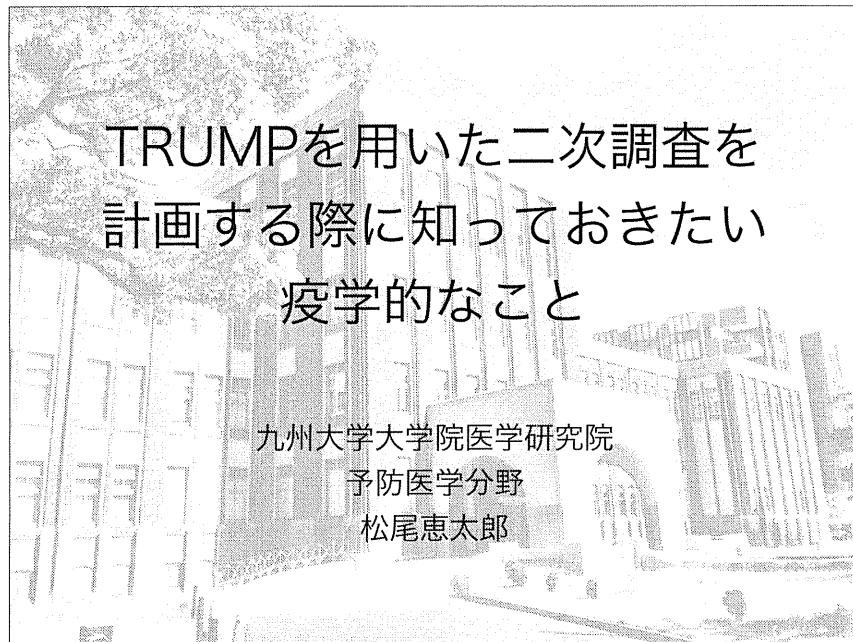
医師の社会は、つよい学会信仰ともいるべきものを持っている。

学会はグライダー人間の訓練所である。飛行機人間は作らない。グライダーの練習に、エンジンのついた飛行機などがまじっていては迷惑する。危険だ。学会では、ひっぱられるままに、どこへでもついて行く従順さが尊重される。勝手に飛び上がりたりするのは規律違反。たちまちチェックされる。やがてそれぞれに、グライダーらしくなって評議員になる。

グライダーとして一流である医師が、学位が必要になって論文を書くことになる。これは、これまでの診療といさか勝手が違う。(中略) グライダーは途方にくれる。突如としてこれまでと違う統計解析を要求されてもできるわけがない。臨床医として優秀な医師ほどあわてる。

一般に、臨床経験を受けた期間が長ければ長いほど、自力飛翔の能力は低下する。臨床マニュアル医として飛べるのに、危ない統計家もどき研究者になりたくないのは当たり前である。

TRUMP を用いた二次調査を計画する際に知っておきたい疫学的なこと



1

研究前の基本：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(Exposure)
 - アウトカム(Outcome)
 - 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 Covariates or Confounder）
- どの研究デザインを採用するか？

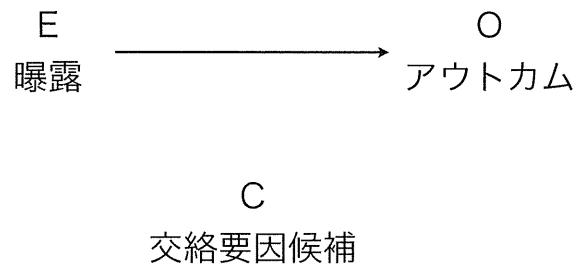
3

今日の話の大まかな流れ

- 研究前の基本：大まかな流れ
- 追加調査を必要とする場合の研究デザイン
- 基本研究デザイン
- 多変量解析の仕組
- 第3の要因
- おまけ
- 統計パッケージで解析する際の基本原則

2

E, O, and C



この構図に落とし込めてない状態では解析しない

4

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

5

例えは・・・

- 疑問：男女で血液悪性疾患の移植後の成績が違うか？
- 移植後の成績：OS, aGVHD, cGVHD、再発 etc…
- 手持ちの変数：男女、年齢、疾患名、graftの種類、適合度、移植前PS、前処置、GVHD予防法

6

例えは・・・ 続き

- まずはaGVHDから解析
- 手持ちの変数を全部評価
 - 単変量
 - 多変量
- 一杯結果がでたな・・・有意な因子が5つもある・・・、どれも面白そう
- さて次にOSでも同じことするか・・・
- 有意な結果が20個もある。あれ、私は何の研究をしてたっけ？

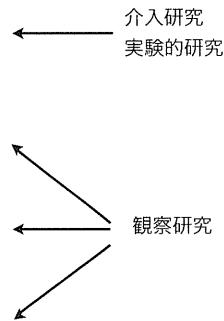
7

基本的な 研究デザイン

8

4つの代表的な研究デザイン

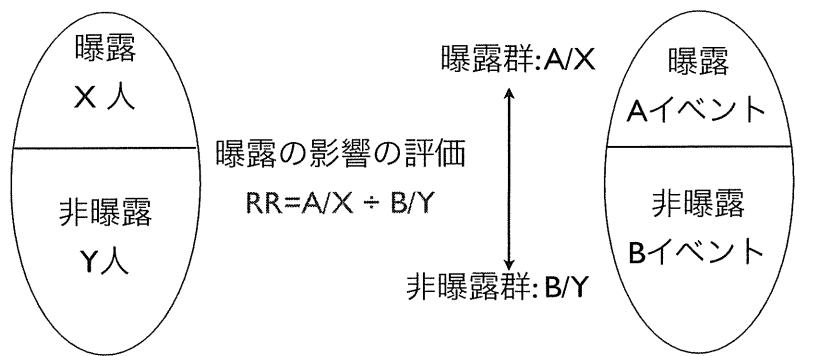
- ▶ 無作為割付試験
- ▶ コホート研究
- ▶ 症例対照研究
- ▶ 横断研究



9

9

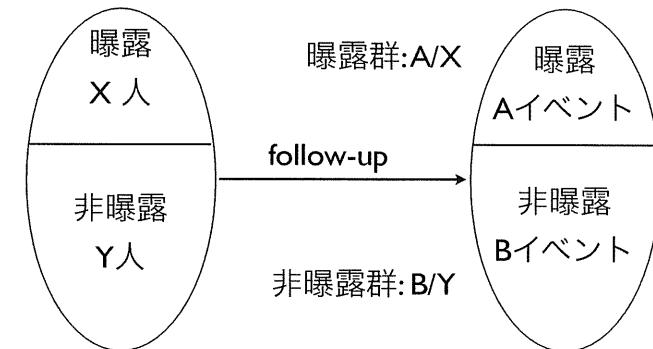
前向きコホート研究



11

前向きコホート研究

曝露の測定を最初にする

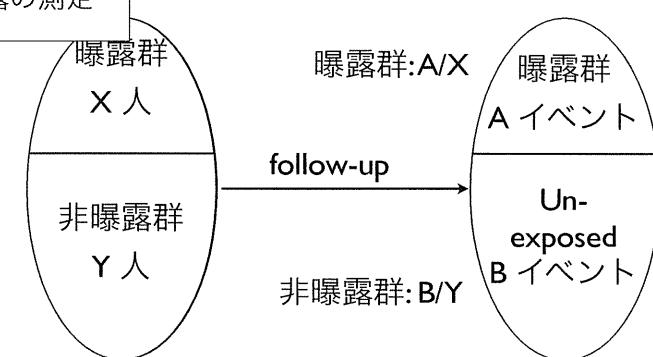


10

後向きコホート研究

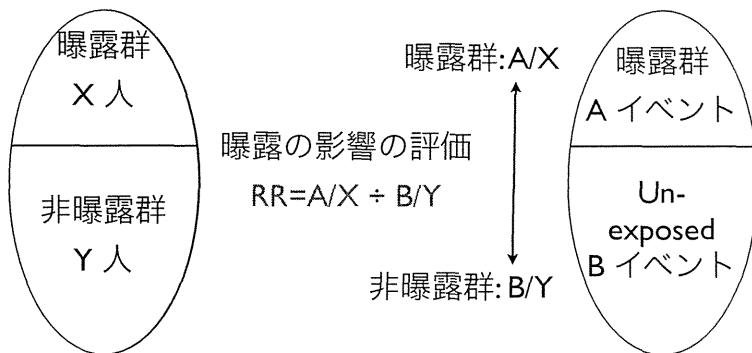
診療録レビュー等
・曝露の測定

既にあるリストに基づき、
「過去の曝露」を調べる



12

後向きコホート研究



13

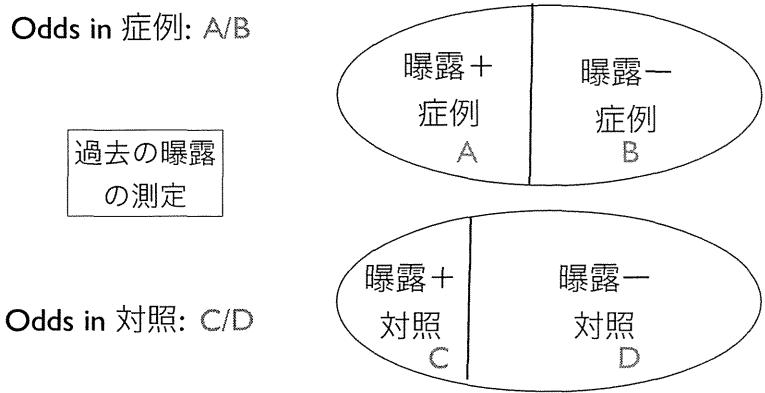
コホート研究(続き)

• 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？

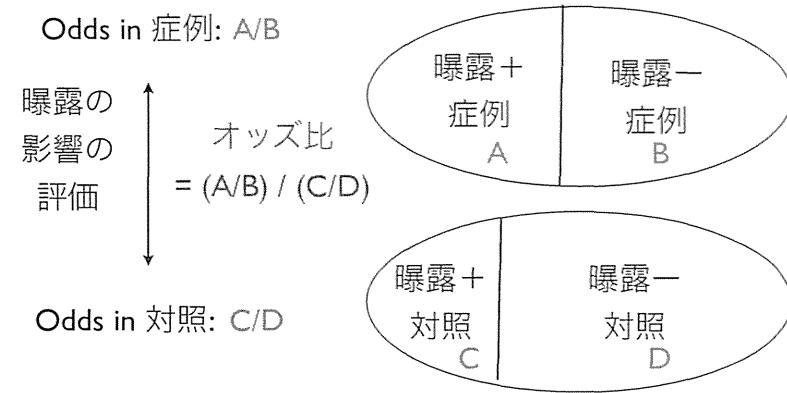
14

症例対照研究



15

症例対照研究



16

Case-control study (cont')

- Pro

- 前向き、後向きコホートよりも時間がかかるない

- 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）

17

TRUMPデータを用いた
リスク・予防要因探しは

コホート研究か症例対照研究

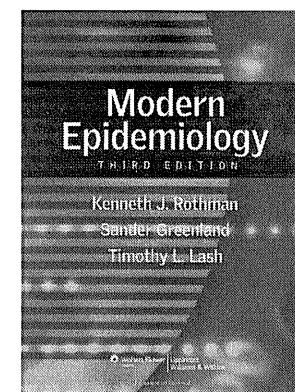
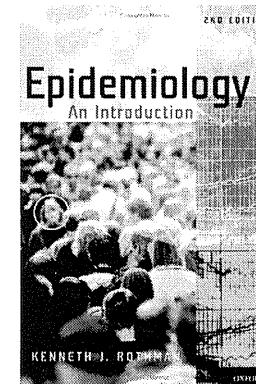
19

横断研究

この研究デザインは原因を云々する
デザインではない

18

疫学の学習に関して



Epidemiology: An Introduction.
by Rothman KJ.

Modern Epidemiology.
by Rothman KJ.

20

多変量解析

Cox modelを例に

21

Cox model、多変量解析

- $\lambda(t|x_1) = \lambda_0(t) \cdot \exp(b_1 X_1 + \dots + b_3 X_3)$
- $\log(\lambda(t|x_1)) = \log(\lambda_0(t)) + (b_1 X_1 + \dots + b_3 X_3)$

—

- $\log(\lambda(t|x_1=1)) = \log(\lambda_0(t)) + (b_1 \cdot 1 + \dots + b_3 X_3)$
- $\log(\lambda(t|x_1=0)) = \log(\lambda_0(t)) + (b_1 \cdot 0 + \dots + b_3 X_3)$

$$\log(\lambda(t|x_1=1)) - \log(\lambda(t|x_1=0)) = b_1$$

$$\log \frac{\lambda(t|x_1=1)}{\lambda(t|x_1=0)} = b_1 \rightarrow HR = \exp(b_1)$$

ハザード比

23

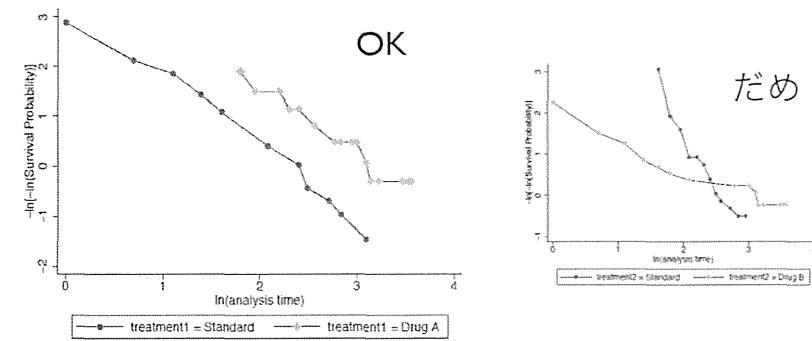
Cox model

- 回帰分析の一つ。式で表すと
 - $\lambda(t|x_1, \dots, x_n) = \lambda_0(t) \exp(b_1 X_1 + \dots + b_n X_n)$
- 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)

22

STATAでのproportional hazard assumptionの検証

- stphplot, strata(X_1)



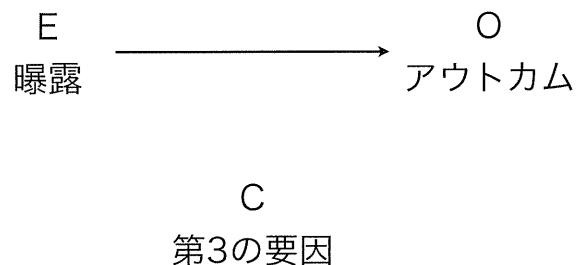
24

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についての有意性の検討をしてくれる

25

E, O, and C



大まかに4つのパターン

27

第3の要因

交絡、交互作用等

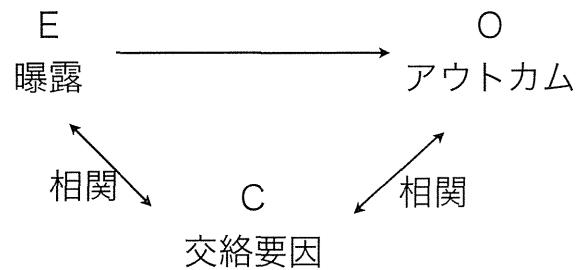
26

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

28

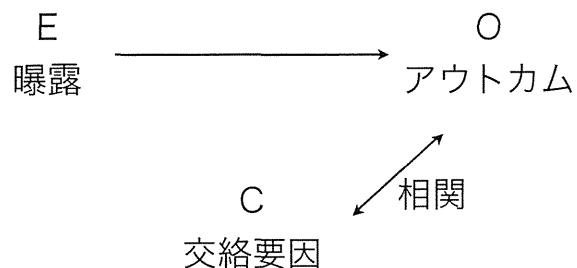
E, O, and C



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

29

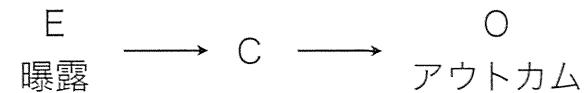
Eと相関してないC



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

31

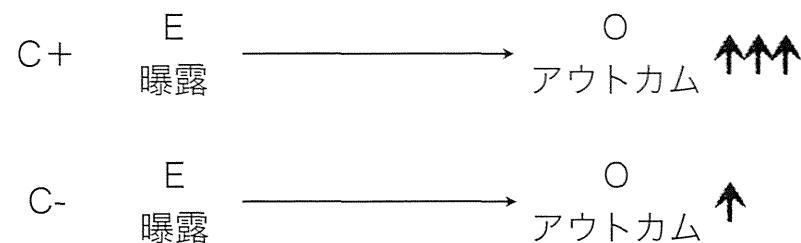
中間要因



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

30

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

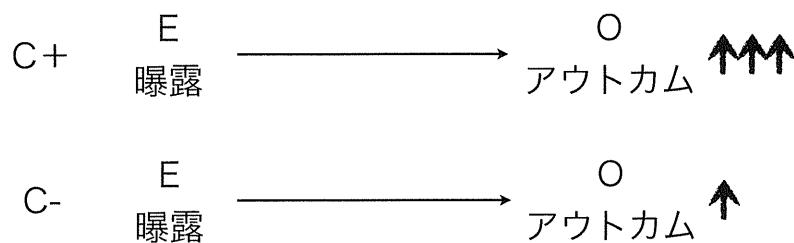
32

交絡要因を調整しないと

$E \leftrightarrow C$	$C \leftrightarrow O$	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	過小評価
負の相関	負の相関	過大評価

33

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

35

交絡の影響の排除法

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

34

交互作用がある時

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

$HR=5=1.5 \times 1.5 \times 2.22$ (交互作用ハザード比)

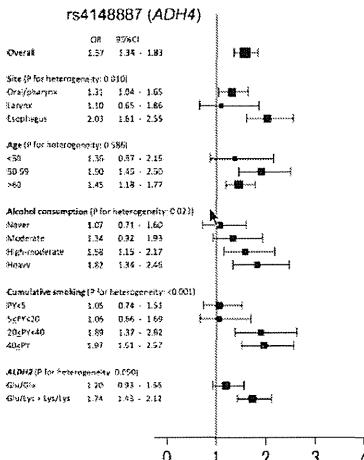
36

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

37

Forrest plot



39

- 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

38

第3の要因を意識した プレゼンテーション

- 大切な点
 - 研究の疑問を意識で来ているか？
 - 研究の疑問で上がっている曝露は何か？
 - それと、第3の要因は明確に分けてプレゼンする

40

良くある区別しない例

研究上の疑問：GANPという物質の免疫染色のレベルの予後への影響
第3の要因：閉経状態、腫瘍径、リンパ節転移、核型の4つ

Table 2: Uni- and multivariate analysis for breast cancer specific survival of KUH cohort.

Variables	Categories	N	Univariate			Multivariate		
			HR	95%CI	p-values	HR	95%CI	p-values
Menopause	post. vs. pre	123/253	1.10	0.50-2.44	0.808	Excluded	-	-
Tumor size	≥2cm vs. <2cm	162/214	3.26	1.43-7.41	0.005	1.14	0.41-3.16	0.799
Nodal status	Positive vs. Negative	168/204*1	3.84	1.74-8.50	0.001	2.93	1.13-7.60	0.027
Nuclear grade	2-3 vs. 1	168/204*2	2.97	1.30-6.79	0.01	1.19	0.89-1.59	0.229
GANP	IHC 2-3 vs. IHC 0-1	170/206	3.43	1.45-8.08	0.005	3.71	1.52-9.01	0.004

41

レビューに言われた、そこまで強気になれない

研究上の疑問：GANPという物質の免疫染色のレベルの予後への影響
第3の要因：閉経状態、腫瘍径、リンパ節転移、核型の4つ

Table 2: Uni- and multivariate analysis for breast cancer specific survival of KUH cohort.

Variables	Categories	N	Univariate			Multivariate		
			HR	95%CI	p-values	HR	95%CI	p-values
GANP	IHC 2-3 vs. IHC 0-1	170/206	3.43	1.45-8.08	0.005	3.71	1.52-9.01	0.004
Other covariates								
Menopause	post. vs. pre	123/253	1.10	0.50-2.44	0.808	Excluded	-	-
Tumor size	≥2cm vs. <2cm	162/214	3.26	1.43-7.41	0.005	1.14	0.41-3.16	0.799
Nodal status	Positive vs. Negative	168/204*1	3.84	1.74-8.50	0.001	2.93	1.13-7.60	0.027
Nuclear grade	2-3 vs. 1	168/204*2	2.97	1.30-6.79	0.01	1.19	0.89-1.59	0.229

43

究極これで良い

研究上の疑問：GANPという物質の免疫染色のレベルの予後への影響
第3の要因：閉経状態、腫瘍径、リンパ節転移、核型の4つ

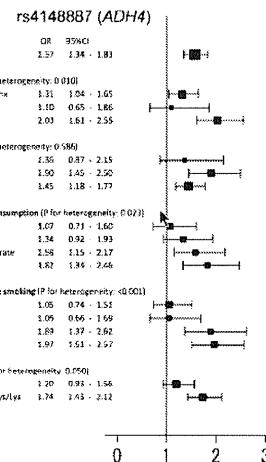
Table 2: Uni- and multivariate analysis for breast cancer specific survival of KUH cohort.

Variables	Categories	N	Univariate			Multivariate*1		
			HR	95%CI	p-values	HR	95%CI	p-values
GANP	IHC 2-3 vs. IHC 0-1	170/206	3.43	1.45-8.08	0.005	3.71	1.52-9.01	0.004

*1 Adjusted for xxx, yyy, zzz, and qqq.

42

再掲 Forrest plot



層別化解析を可視化

交互作用の有無が見やすい

44