2014年9月26日
第62回日本心臓病学会学術集会
モーニングレクチャー「医学統計の基礎」
於：仙台国際センター　第9会場

医学統計の基礎

東北大学大学院医学系研究科
循環器内科学分野

宮田　敏
miyata@cardio.med.tohoku.ac.jp

日本心臓病学会
COI 開示

東北大学大学院医学系研究科循環器内科学　宮田　敏

演題発表に関連し、開示すべきCOI関係にある企業などはありません。

Agenda:

1. データの準備　　—データが手に入ったら、無条件にやるべきこと—
2. データの提示　　—Table 1のまとめ方—
3. 解析結果の提示　—統計解析ソフトからの出力のどこを提示すべきか—
4. 統計家から医療者へのメッセージ

1. データの準備　　—データが手に入ったら、無条件にやるべきこと—

データが手に入った時、すぐ解析に取り掛かりたいのは人情です。

しかし、慌ててはいけません!!

まずは、データの確認とクリーニング。

1. データの準備

1-1. 元データの取り扱い

i. データの形は長方形

- 第一行目に変数名。全角文字は避ける方が無難。
- グラフ、解析結果などを張り付けない。別ファイルで保存。
- データの形は、長方形になるはず。

systemID	hospitalID	sex	age	height	bodyweight
4	1185645	1	64	173	75.4
11	3329388	1	69	164	72
12	4022624	1	78	155.2	47.2
14	4402536	1	83	159.1	60
22	4862866	2	73	147.6	40.5

1. データの準備

1-1. 元データの取り扱い

ii. 元データは絶対に改変しない。

- 解析の過程で、変数を変換したり、新しい変数を定義することがある。
- 新しく作ったデータを、元データに上書きしない。
- データを改変したら、新しいファイル名で保存。

- 元データを改変すると、元データが何であるか分からなくなる。元データが分からなくなれば、意図せざるデータのねつ造まであと一歩。

1. データの準備

1-1. 元データの取り扱い

iii. 患者さんの個人情報は記載しない。

- 残念ながら、いまだに氏名、カルテ番号など、患者さん個人を特定できる情報が付いたままのデータを見かける。
- 個人情報は、データ解析の立場からは無意味。
- 個人情報が漏えいすれば、研究は中止、研究者の辞表が何枚か必要。被害者には、お詫びの仕様がない。

- データを受け取ったら、個人情報はすぐに匿名化もしくは削除。

1. データの準備

1-1. 元データの取り扱い

iv. 解析記録の保存。

- 患者さんを診察すれば、医師がカルテに記録するのは当然。実験をすれば、実験ノートに記録するのは常識。統計解析の記録を残すのも、それと同じ。

- 元データと解析の記録を見れば、第三者が解析を再現できる程度の記録が必要。
  - 解析の再現性
  - 備忘録　　「三日後の自分は遠い親戚。一週間後の自分は赤の他人」
  - 出来れば、プログラムを書いて解析する。

1. データの準備

1-2. データ入手時にすべきこと：入力ミス、異常値の発見

表計算ソフトのフィルター機能が便利

- データの範囲： 本来正の値をとるはずが、負の値をとる。小数点の間違いで、体重35kgが3.5kgになる、等。
- 全角文字と半角文字の混在：“w” と “w”など。
- 質的変数の数字表記： 男性→1, 女性→2など。男性→M, 女性→Fのように書き直す。
- 異常な値の検出：“3.14” と “3,14” など。
- 欠測値の数： 欠測値の数が想定より多い場合、データが正常に認識されていないことがある。

2. データの提示 —Table 1のまとめ方—

得られたデータは、サンプルの群ごとに分類し基本統計量（平均、分散、その他）を計算して提示する。（Table 1)

i. 連続数（実数）の提示

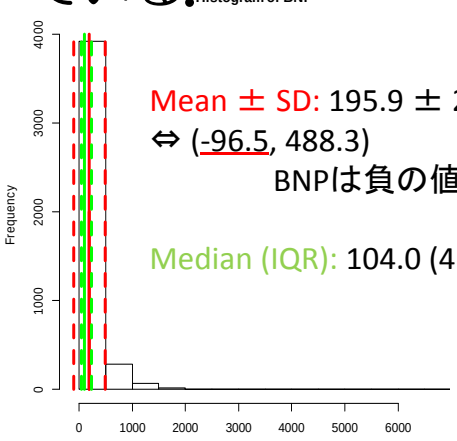
- Mean ± SD (Standard Deviation)
- Mean ± SE (Standard Error, SD/√N)
- Median (IQR, Inter quartile range)

第一選択は、Mean ± SD（データ全体の分布に興味があるとき）または、Mean ± SE（群の比較に興味があるとき）

Mean ± SDの上限、下限を実際に計算して、その変数が通常とる範囲を外れるときは、Median (IQR) を選択。

Mean ± SD (Standard deviation): 平均(Mean)を中心にMean ± SDの範囲に、データ全体の60～70%が分布している。

Median (interquartile range, IQR): 中央値(Median)を中心に，IQRの範囲にデータ全体の50%が分布している。



Mean ± SD: 195.9 ± 292.4
⇔ (-96.5, 488.3)
BNPは負の値をとらない

Median (IQR): 104.0 (41.3, 238.0)

Mean ± SDは，不合理な値(データの範囲を逸脱)をとることがある.

分布が歪んでいるときは，Median (IQR) が第一選択.

2. データの提示 —Table 1のまとめ方—

ii. 離散数(カウント)の提示

- 度数(個数、frequency)とパーセントを両方提示する。

よく、度数のみ、あるいはパーセントのみを提示した論文を見かけるが、お勧めしない。必ず両方出す。

グループ間の比較のための検定(必ずp値を記載する)

i. 連続数：Mean ± SE⇒ Welch's t-test(不等分散)
Median (IQR) ⇒ Mann-Whitney test, Wilcoxon rank sum test

ii. 離散数：Fisher's exact test(フィッシャーの直接法) or χ^2検定 第一選択はフィッシャーの直接法。χ^2検定は、すでに歴史的役割を終えている。(私見です)

3. 解析結果の提示　—統計解析ソフトからの出力のどこを提示すべきか—

医学統計で用いられるモデルの多くは、Outcomeを共変量の一次式 $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ で説明するもの。
提示すべきものは、以下の通り。

- 各パラメター β 、オッズ比、ハザード比などの推定値、信頼区間、有意確率（p値）
- モデル全体の適合度
- その他

医学統計で用いられる多変量モデル

- 線形回帰モデル

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

実数

- ロジスティック回帰モデル

$$\log\left(p/(1-p)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

ロジット、対数オッズ

- Cox比例ハザードモデル

$$\lambda(t \mid x_1, \ldots, x_k) = \lambda_0(t)\exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right)$$

ハザード $\lim_{\Delta t \to 0} \frac{P(t \le T < t+\Delta t \mid T \ge t)}{\Delta t}$

- 線形回帰モデル

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

実数

- 各パラメター β の推定値、信頼区間、有意確率（p値）

例：

	Estimate	CI	Pr(>|t|)
groupTrt	-0.371	(-1.025, 0.283)	0.249

- モデル全体の適合度
  - 決定係数 (coefficient of determination, R^2)
    応答変数の変動のうち、回帰で説明された部分の割合

Multiple R-squared: 0.07308

  - 分散分析表 (Analysis of Variance Table) のp値

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \text{ vs. } H_1 : \text{not } H_0$$

F-statistic: 1.419 on 1 and 18 DF, p-value: 0.249　H_0を棄却したくない

- 線形回帰モデル（続き）

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

実数

- 回帰モデルにおいて「応答変数の変動は、説明変数の一次式で近似される」という線形性の仮定が、もっとも重要。事前に散布図を描いて確認する。
- 決定係数はどの程度以上なら良い?

もちろん、R^2は大きいに越したことはない。R^2が小さいということは、yに影響する、今のモデルに含まれない変動要因が存在する、ということ。

もし、R^2が小さくてもなお有意な説明変数があれば、上記のlimitationの下で、回帰は有効である。

- ロジスティック回帰モデル

$$\log\left(p/(1-p)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

ロジット、対数オッズ

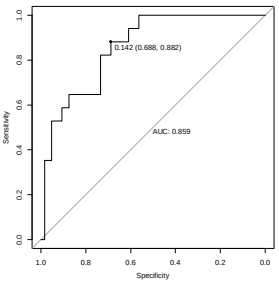
- オッズ比 e^{β} の推定値、信頼区間、有意確率（p値）

例：

	Odds ratio	CI	Pr(>|z|)
Start	0.813	(0.705, 0.924)	0.00229

- モデル全体の適合度
  - ROC曲線のAUC

イベントの予測確率と、実際のイベントの有無でROC曲線を描く。



  - Hosmer-Lemeshowの適合度検定

H_0を棄却したくない

参考文献:内田「SPSSによるロジスティック回帰分析」（平成23年）オーム社, p. 212

- Cox比例ハザードモデル

$$\lambda\left(t \mid x_1, \ldots, x_k\right) = \lambda_0(t)\exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right)$$

ハザード $\lim_{\Delta t \to 0} \frac{P(t \le T < t+\Delta t \mid T \ge t)}{\Delta t}$

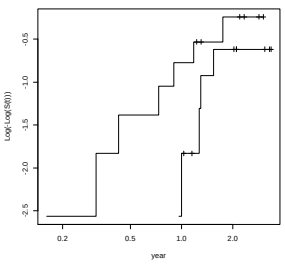
- ハザード比 e^{β} の推定値、信頼区間、有意確率（p値）

例：

	Hazard ratio	CI	Pr(>|z|)
rx	0.551	(0.174, 1.74)	0.31

- 比例ハザード性の確認
  - 補対数-対数プロット

ある共変量に対して比例ハザード性が成り立つなら、その共変量の層ごとに求めた log(-log(S(t))) は平行になる。



  - Schoenfield残差を用いた検定

共変量ごとに比例ハザード性の検定。

- Cox比例ハザードモデル（続き）

$$\lambda(t \mid x_1, \ldots, x_k) = \lambda_0(t)\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

ハザード $\lim_{\Delta t \to 0} \frac{P(t \le T < t+\Delta t \mid T \ge t)}{\Delta t}$

- モデル全体の適合度
  - ワルド検定 (Wald test)
  - 尤度比検定 (likelihood ratio test)
  - スコア検定 (Score test)

```
Likelihood ratio test= 1.05  on 1 df,   p=0.3052
Wald test            = 1.03  on 1 df,   p=0.3096
Score (logrank) test = 1.06  on 1 df,   p=0.3026
```

通常三通り出力されるが、どれを用いてもよい。私の好みは、ワルド検定である。

4. 統計家から医療者へのメッセージ

- 統計学者を怖がらない。
  - データ解析の相談を受けて、いやな気持になる統計学者はいない。（忙しい人は、いるかもしれない）
  - 初対面、他学部の研究者でも、事情を説明して相談してみよう。
- 「すごく基本的な質問で、申し訳ないのですが．．．」という前置きは、いらない。
  - 「基本的でつまらない問題」に出会ったことがない。
  - 出会ったことのないデータは、常においしい。