北大SPH・統計解析の基礎④

2群の比較

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 連続量データの比較
- 対応のない場合
  - t検定
    - Student型と Welch型
  - Wilcoxon順位和検定
- 対応のある場合
  - 対応のある t 検定
  - Wilcoxon符号付き順位検定、McNamer検定

2

結果変数の型による分類

3

目的	連続尺度	分類尺度	時間イベント尺度
分布の記述	ヒストグラム、箱ヒゲ図、散布図	ヒストグラム、分割表	生存曲線（Kaplan-Meier法）
要約統計量	平均、分散、中央値、パーセント点、相関係数	頻度、一致度、相関係数	x年生存確率、中央生存期間
検定（単純）	t検定、分散分析、Wilcoxon検定	χ^2検定、Fisher正確検定	ログランク検定
検定（層別）	共分散分析	Mantel-Haenszel検定	層別ログランク検定
回帰モデル	重回帰分析	ロジスティック回帰分析	Cox回帰分析

3

血圧を下げる飲料！？

4

- 収縮期血圧が140mmHg以上
  (平均は144mmHg)であった
  大学生男性20名に飲料を摂取させた
- 30分後に血圧を測定したところ、
  平均収縮期血圧は125mmHgになった
- この飲料は収縮期血圧を下げる！

あなたは血圧の高い人にこの飲料をすすめますか？

4

のんだ、なおった、きいた？

5

- 同じ条件でこの飲料をのまなかったら？
  - 一息ついてリラックスしただけ？
- そもそも血圧の高い大学生男性って？
  - 直前に運動をしていた？？
- 平均は120mmHgだが、そのバラツキは？
- 血圧の測定器はどのようなもの？
  - 同じ機種を用いたか？
  - 人が聴診して測定したか？

5

コントロール(対照)の設定

6

- 飲料を摂取しないだけで、
  その他は同じ条件にしたグループをおく
- 血圧の変化量を
  飲料摂取群と非摂取群で比較

6

平均への回帰 regression to the mean

7

- ある基準以上という人だけ選んでみると、次の測定までに何もなされていなくても、ランダムなバラツキがゆえに平均値に結果が近づく現象
  - たまたま血圧の高かった人が対象者になったため、飲料とは関係なく30分後の血圧が下がっただけ？

7

対象集団

8

- 第1相試験では、健康な男性大学生が対象
  - それはヒトへの安全性をみるためでしょう
- この飲料を売るターゲットは中高年
  - 同じ高血圧でも大学生のそれとはわけが違う
  - 大学生と同じように効くとは限らない

8

測定の信頼性

9

- 血圧計の較正は難しい
  - 全く同じものを測定しても、血圧計によって値が異なる
- ましてや人が目で見ていたら・・
- そもそも血圧は個人内変動が激しい
- 測定回数は全員1回だけ？
  - 2回測定した平均をとった人もいた！？
- 1mmHg単位での信頼性はない！

9

コントロール群をおいた研究

10

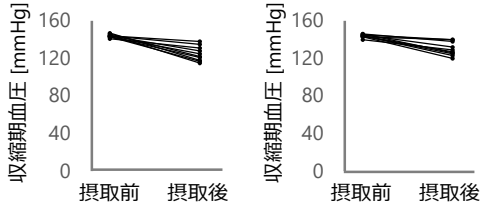
- 対象
  - 高血圧症と診断された、50歳以上の男女
- 方法：ランダム化
  - 飲料摂取群10名
  - 飲料非摂取群10名
    - 有効成分を除いた飲料（プラセボ）を摂取
- 測定
  - 同じ施設・測定器・実施者により安静時血圧を摂取前と1カ月の継続摂取後に各々3回測定し、その中央値を測定値に採用

10

血圧の変化

11

- 飲料摂取群
- 飲料非摂取群



研究仮説：
飲料摂取群は非摂取群より収縮期血圧が下がるか？

11

エンドポイント（評価項目）

12

- そのままの値
  - 1カ月後の測定値
- 絶対的変化量
  - (1カ月後の測定値) − (摂取前の測定値)
- 相対的変化量
  - $\frac{\text{(1カ月後の測定値)} - \text{(摂取前の測定値)}}{\text{(摂取前の測定値)}}$

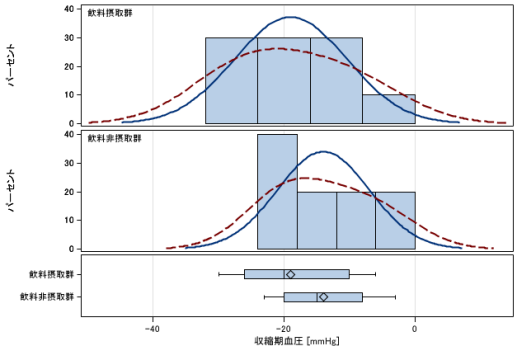
12

絶対的変化量を採用 13

- 飲料摂取群（10名）
  - -6, -9, -10, -16, -18, -22, -24, -26, -29, -30
  - 平均[SD]：-19.0 [8.6]
  - 中央値[四分位範囲]：-20 [-10, -26]
- 飲料非摂取群（10名）
  - -3, -5, -8, -11, -13, -17, -19, -20, -21, -23
  - 平均[SD]：-14.0 [7.1]
  - 中央値[四分位範囲]：-15 [-8, -20]

13

絶対的変化量の分布 14



14

分布の違い？ 15

- モーメントを用いた比較
  - 平均：位置
  - 分散：分布の広がり
  - 歪度：分布の歪み
  - 尖度：分布のすその重さ
- 中央値や分布関数を用いた比較
  - 中央値：位置
  - 分布関数：分布の広がり

15

位置の比較 16

- t 検定：平均値の比較
- Wilcoxonの順位和検定：中央値の比較
- 並べ替え検定：どちらも可能
- 分布の広がり、歪み、すその重さは、比較群間で同様と仮定
  - ランダム化によって介入前の分布は同じに
  - 介入によって分布の形状が変わることは考えづらい

16

ランダム化 17

- 20名を同じ確からしさで摂取群か非摂取群に割付
  - 合計10名ずつになるように
- ランダム化した結果、対象者の割り付けパターン数は・・？
  - ${}_{20}C_{10}$=184,756通り

17

並べ替え検定 18

- 帰無仮説：どちらの群でも変化量が同じ
  - 全割付パターンを計算すれば、帰無仮説の下、観察されうる結果がすべて判明
  - しかも逆のパターンが必ずあるので、「差がない」場合の結果の分布が分かる
    - AABABBというパターンがあれば、BBABAAというパターンも必ずある
- 例えば、変化量平均値の群間差を各割付パターンで求めてみよう

18

割付を並べ替え

19

ID	変化量	実際の割付	パターン①	パターン②
1	-6	摂取群	摂取群	摂取群
2	-9	摂取群	非摂取群	非摂取群
⋮	⋮	⋮	⋮	⋮
10	-30	摂取群	摂取群	非摂取群
11	-3	非摂取群	非摂取群	非摂取群
12	-5	非摂取群	摂取群	摂取群
⋮	⋮	⋮	⋮	⋮
20	-23	非摂取群	非摂取群	摂取群

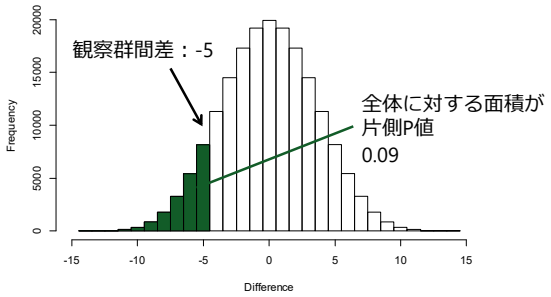
. . .

19

変化量平均値の群間差の分布

20

- 184,756通り計算



20

並べ替え検定に基づくP値

21

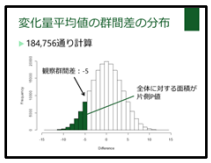
- ありうるパターンの何番目にいるか
  - 正確なP値
  - 二値(成功/失敗)なら手計算で可能(Fisher検定)
    - 変化量（アウトカム）が連続量だと、通り数が増えすぎて大変
  - 10例 vs 10例では18万通り以上計算せねば
    - 100例 vs 100例では約10^{29}通り
- 変化量の違いをみる指標には平均値のほか、中央値や分位点も可能
  - 検討したい仮説に応じて選べる

21

平均値の違いに注目

22

- 帰無仮説 $H_0: \mu_X = \mu_Y$
  - 摂取群平均値μ_Xと非摂取群の平均値μ_Yが同じ
- 対立仮説 $H_1: \mu_X < \mu_Y$
  - 摂取群平均値μ_Xのほうが、より減少している
- 平均値の群間差の分布さえ分かれば・・・



22

t分布

23

- 正規分布より少しすそが重い分布
  - 自由度が+∞の場合は正規分布に一致
- データが独立に同一の正規分布に従う場合、平均の差が従う正確な分布
  - データが正規分布に従っている必要はない
  - 別にデータが正規分布に従っていなくても、ランダム化試験のように分散が同じと仮定できるなら検定は妥当(valid)

23

Studentのt検定

24

- tが自由度$n_X + n_Y - 2$のt分布に従う
  - n_X, n_Y：各群の人数
  - $\bar{x}, \bar{y}$：各群の平均値

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(1/n_X + 1/n_Y\right)s^2}},$$

$$s^2 = \frac{\sum_i^{n_X}(x_i - \bar{x})^2 + \sum_i^{n_Y}(y_i - \bar{y})^2}{(n_X - 1) + (n_Y - 1)}$$

24

自由度 Degrees of Freedom

25

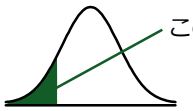
- 偏差平方和(データのバラツキの表現方法)
  - n個のデータ$x_1, \ldots, x_n$のバラツキは、それぞれ平均値からの偏差$(x_i - \bar{x})$の二乗和
$$\sum_i^n (x_i - \bar{x})^2$$
- 自由度
  - 偏差平方和が何個の独立な二乗和から成るか
  - 何個の二乗和が特定されれば、偏差平方和が判明するか
    - 平均が分かっているので、$n-1$個が独立な個数

25

飲料の例（t検定）

26

- 帰無仮説の下でt統計量を計算
$$t = \frac{-19-(-14)}{\sqrt{\left(\frac{1}{10}+\frac{1}{10}\right)\frac{664+448}{(10-1)+(10-1)}}} = -1.42$$
- 自由度18のt分布からP値を計算
  - 各群の自由度は9(=10-1)より、両群では18



この面積(片側P値)が0.09

2.5%より大きいので、
片側2.5%水準で有意でない

26

95%信頼区間

27

- 効果の大きさを含めた議論
- 真の平均値の差δ
$$t = \frac{(\bar{x}-\bar{y})-\delta}{\sqrt{\left({}^1\!/_{n_X} + {}^1\!/_{n_Y}\right)s^2}}$$
- 両側5%水準で有意とならないδの範囲
$$(\bar{x}-\bar{y}) \pm 2.10\sqrt{\left({}^1\!/_{n_X} + {}^1\!/_{n_Y}\right)s^2}$$

自由度18のt分布における
片側2.5%点

27

練習①　平均値の区間推定

28

- 飲料の例（スライド#13）において、収縮期血圧の平均変化の群間差の95%信頼区間をt分布を基に求めよ

28

飲料の例（平均値の区間推定）

29

$$-19-(-14) \pm 2.10\sqrt{\left(\frac{1}{10}+\frac{1}{10}\right)\frac{664+448}{18}}$$
$$= (-12.4, 2.4)$$

- 95%信頼区間に0を含んでいる
  - すなわち、片側2.5%(両側5%)で有意でない

29

信頼区間に注目

30

- 平均の差は-5
  - その信頼区間は(-12.4 , 2.4)
- -10mmHgくらい下がるならば臨床的な意義もあるかも
- 十分なサンプルサイズだった？

30

医学的有意差と統計的有意差 31

- 対象者数が非常に多い場合
  - わずかな治療効果(差)であっても統計的に有意
  - わずかな差が臨床的に重要?
- 対象者数が少ない場合
  - (非常に)大きな差でも統計的有意とはならない
  - 医学的に重要な差であれば、
    無視すべきものでなく、さらに検討すべき

31

(Studentの)t検定の仮定 32

- 分散が同じという仮定が必要
  - ランダム化試験では、リーズナブル
- 分散が多少違っていても、
  比較群間で人数が大体揃っていれば、
  検定はほぼ妥当（使ってよい）
  - αエラー率が、名目水準を上回らないこと
  - 観察研究では、めったに人数は揃わない

32

Welchのt検定 33

- 分散が同じという仮定が不要
  - 分散が異なる場合にも対応
  - s_X^2, s_Y^2：各群で求めた不偏分散

$$t_w = \frac{\bar{x} - \bar{y}}{\sqrt{{s_X^2}/{n_X} + {s_Y^2}/{n_Y}}}$$

  - Satterthwaiteの近似（式は省略）
    によって求めた自由度のt分布に従う

33

予備検定方式 34

- ものの本には、
  等分散性の検定を行い、
  有意差があればWelch型、
  なければStudent型　　とある
  - 予め検定を行い、次に行う検定方法を決める
- 研究デザインに応じて、
  検定方法は決めるべき

34

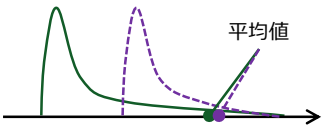
StudentかWelchか 35

- ランダム化比較試験であればStudent型
  - 割付比が1:1の場合
  - 検出力(1-βエラー)はStudent型が高い
    - 本当は差がある場合に有意差ありという確率
    - ランダム化する人（サンプルサイズ）を減らせる
- 観察研究はWelch型
  - 検定はどうでもよく、
    信頼区間を表示することを最優先
    - 少しでもマシな信頼区間を出すために、
      Welch型に基づく信頼区間を示すべき

35

分布形によらない検定 36

- t検定は元のデータが正規分布に従うとき
  最も効率の良い(=有意になりやすい)検定
  - せめて左右対称な分布形であるとよい
- 例えば極端に歪んだ分布
  - 平均値の解釈がわかりづらい
  - t検定では効率が悪い(=有意差を出しづらい)



36

ノンパラメトリックな検定 37

- データの確率分布を仮定しない
  - 対義語：パラメトリックな検定
- データの分布に注目 $F_X(u) = F_Y(u + \Delta)$
  - 帰無仮説$H_0: \Delta = 0$
    - 分布が重なっている
  - 対立仮説$H_1: \Delta \neq 0, (\Delta > 0, \Delta < 0)$
    - 分布がΔだけずれている
- 解釈は中央値の比較とすればよい

37

Wilcoxonの順位和検定 38

- Mann-WhitneyのU検定 と等価
- 全データの順位を割当
- 群の順位和U_Xを計算
  - U_Xの期待値 $n_X(n_X + n_Y + 1)/2$
  - U_Xの分散 $n_X n_Y(n_X + n_Y + 1)/12$
- 以下の検定統計量を計算

$$\frac{U_X - n_X(n_X + n_Y + 1)/2}{\sqrt{n_X n_Y(n_X + n_Y + 1)/12}}$$

38

飲料の例（Wilcoxonの順位和検定） 39

- 各データの順位を計算
  - 摂取群
    - -30, -29, -26, -24, -22, -18, -16, -10, -9, -6
    - 1, 2, 3, 4, 6, 10, 12, 15, 16, 18 (順位)
  - 非摂取群
    - -23, -21, -20, -19, -17, -13, -11, -8, -5, -3
    - 5, 7, 8, 9, 11, 13, 14, 17, 19, 20 (順位)
- 摂取群での順位和を計算
  - $U_X = 1 + 2 + \cdots + 18 = 87$

39

飲料の例（Wilcoxonの順位和検定） 40

- 検定統計量を計算

$$\frac{87 - 10(10 + 10 + 1)/2}{\sqrt{10 \cdot 10\ (10 + 10 + 1)/12}} = -1.36$$

- 検定統計量が標準正規分布に従うことを利用し、P値を計算
  - -1.36>-1.96より片側2.5%水準で有意差なし

40

t検定かWilcoxon検定か① 41

- 統計解析の医学的な解釈
  - 多くの状況では平均値の比較が解釈容易
    - 平均値に差がでた or 中央値に差がでた
    - 平均値の信頼区間 or 中央値の信頼区間
  - 平均値を議論することが無意味な場合
    - 分布が大きく歪んでいる
      - 平均値は外れ値の影響を受けやすい
    - 一部の患者のみ特異的に大きな値をとる状況
      - 閾値を定め、閾値を以上/未満を議論すべき

41

t検定かWilcoxon検定か② 42

- ものの本には、
  - ~~サンプルサイズが小さい時に、Wilcoxon検定~~
- サンプルサイズが極端に小さい時には?
  - 研究デザインからの逸脱は？
  - 仮説検定の問題か?（データの提示で十分?）
  - Wilcoxon検定は検出力が低い

42

2群の比較といっても…

43

- 同一患者さんに、時期をずらして、2つの治療法を施す臨床試験
  - クロスオーバー試験
- 治療しなければ回復が見込めない下での治療前後での患者さんの状態を比較
  - 手術による効果の検討
- 術式による手術成績を比較するため、背景のリスク要因が似た対象者同士をマッチングして比較

43

中心角膜厚データ（一部）

44

- 術後24週と2年で違いがあるか

ID	術後24週	術後2年
1	511	532
2	525	538
3	540	546
4	640	710
5	509	529
6	505	525
7	626	550
8	489	503
9	595	543
10	539	523
11	561	572

Kinoshita S, *et al. N Engl J Med.* 2018; 995-1003.

44

対応のあるt検定

45

- i番目の患者さんの結果X_{Ai}, X_{Bi}
  - 差 $d_i = X_{Ai} - X_{Bi}$
- 帰無仮説H_0：差の平均がゼロ
  - 差の平均値とその標準偏差　$\bar{d}, s_d$
  - 検定統計量tが自由度$n-1$のt分布に従う

$$t = \frac{\bar{d}}{\sqrt{s_d^2/n}}$$

45

練習②

46

- 中心角膜厚研究（スライド#44）において、術後24週と術後2年の間に統計的に有意な変化がみられるかを対応のあるt検定に基づき判断せよ

46

中心角膜厚の例

47

- 差の平均値とその標準偏差を計算
  - $\bar{d} = 2.82, s_d = 39.13$
- 検定統計量の計算
  - $t = \frac{2.82}{\sqrt{39.13^2/11}} = 0.24$
- 自由度10(=11-1)のt分布からP値を計算
  - 片側P値：0.41

47

Wilcoxonの符号付き順位検定

48

- 対応のあるデータに対するノンパラ検定
- 帰無仮説H_0：分布の中央がゼロ
  - ゼロを境に、正の値と負の値は同じバラツキ
- 差の絶対値$|d_i|$について、順位R_iを考える
  - ゼロは順位をつけない

48

順位和の計算

49

- $T_+ = \sum R_i$(正の値をとったデータ)
- $T_- = \sum R_i$(負の値をとったデータ)
- 帰無仮説の下では、$T_+ = T_-$
  - $E(T_+) = E(T_-) = n(n+1)/4$
  - $V(T_+) = V(T_-) = n(n+1)(2n+1)/24$
- 検定統計量が標準正規分布に従う

$$\frac{T_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

49

順位和の計算

50

ID	術後24週	術後2年	差	絶対値の順位	正符号の順位
1	511	532	21	8	8
2	525	538	13	3	3
3	540	546	6	1	1
4	640	710	70	10	10
5	509	529	20	6.5	6.5
6	505	525	20	6.5	6.5
7	626	550	-76	11	
8	489	503	14	4	4
9	595	543	-52	9	
10	539	523	-16	5	
11	561	572	11	2	2

- 検定統計量は、$\frac{(8+3+1+10+6.5+6.5+4+2)-11\cdot(11+1)/4}{\sqrt{11\cdot(11+1)(2\cdot11+1)/24}} = 0.71$
  - 片側P値=0.24

50

対応のない／ある群間比較

51

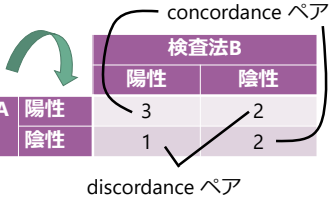
変数の種類	対応のないデータ	対応のあるデータ
連続量	• (2標本) t検定 • Wilcoxon順位和検定	• (1標本)t検定 • Wilcoxon符号付き順位検定
カテゴリカル	• χ^2検定 • Fisherの直接検定	• McNamer検定

51

対応のある2値アウトカム

52

患者ID	検査法A	検査法B
1	+ (陽性)	- (陰性)
2	+	+
3	+	+
4	-	-
5	-	+
6	+	-
7	-	-
8	+	+



52

McNamer検定

53

- Discordanceペアだけに注目

		検査法B	
		陽性	陰性
検査法A	陽性	a	b
	陰性	c	d

- 自由度1のカイ二乗検定を利用

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

53

まとめ

54

- 並べ替え検定、t検定、Wilcoxon順位和検定
  - 平均の比較をするなら、たいていt検定
  - ランダム化比較試験ではStudent型の検定、観察研究ではWelch型に基づく区間推定
- 対応のあるデータに対する検定
  - 同一対象者、マッチされたペアのように、背景情報を似せている（同じにする）場合

54