北大・統計解析の基礎③

推定と検定

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 統計的仮説検定
  - カイ二乗検定
- 区間推定
- P値廃止運動
- サンプルサイズ設計

2

炎上したが・・・

5

- 統計的に話を整理する

1. 対戦相手と実力が互角という仮定
   - 勝利する確率が0.5
2. 対戦結果は29勝0敗
3. 1.の仮定の下で、29勝0敗が起こることは
   どのくらいまれなことかを計算
   - 確率ではなくp値と言えば炎上しなかった！？

5

統計的仮説検定

6

- 科学的方法である背理法を導入

1. 「対戦相手との実力に差はない」と仮定
   - 帰無仮説という
   - 群間比較の場面では「比較群間に差がない」
2. 仮定が偽であることをいう
   - P値が事前に定めた有意水準より低い
3. 「実力に差がある」と結論づけ
   - 対立仮説を受容

6

帰無仮説と対立仮説

7

- 帰無仮説 H_0（否定したい仮説）
  - 勝率は○○である
  - 比較群間で効果の大きさに違いがない
  - 曝露とアウトカムは無関係だ　　…etc.
- 対立仮説 H_1
  - 帰無仮説と逆の内容
- 観察データが帰無仮説に反するかに注目
  - 対立仮説に反するかには注目しない

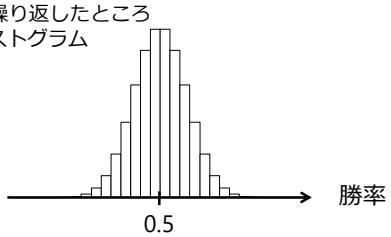
7

帰無仮説が正しい場合

8

- 同様の研究を繰り返したならば、
  観察される勝率は、
  0.5を中心として、左右対称に分布

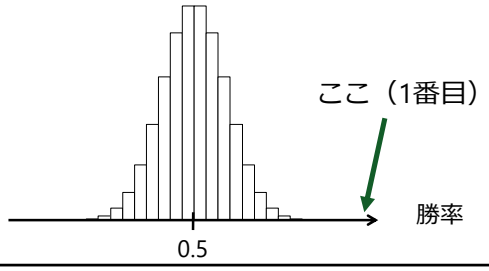
「29試合行う」を繰り返したところ
得られた勝率のヒストグラム



8

観察結果はどこにある？ 9

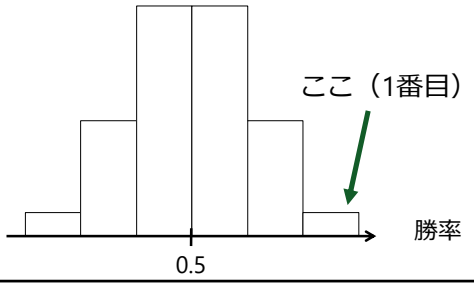
- 勝率100%（29勝0敗）は、
  ヒストグラムのどこに位置するのか？



9

たとえば5連勝だったら？ 10

- まだありえそう



10

同じ「1番目」にまれな出来事 11

- 試合数に応じて、「まれ」っぷりが違う
- 帰無仮説の下で、ありえるパターン数を分母において考える
  - 勝率が0.5だとして、5試合行うと、
    対戦結果は2^5=32通り考えられ、
    そのうち1番まれな結果が得られた
  - 勝率が0.5だとして、29試合行うと、
    対戦結果は2^{29}=536870912通り考えられ、
    そのうち1番まれな結果が得られた

11

どちらがよりまれ？ 12

- 勝率が0.5である下で、
  - 29勝1敗
    - 勝率 97%
    - 全パターンのうち、2番目にまれ
  - 5勝0敗
    - 勝率 100%
    - 全パターンのうち、1番目にまれ

12

p値 13

- 帰無仮説の下で、
  考えられうる全ての結果パターンに対し、
  観察結果が得られた順位
  - 順位を0から1の間で基準化
- 事前に定めた有意水準より低ければ、
  帰無仮説は正しくないと意思決定する
  - 有意水準は、通常片側2.5%

13

正規近似を用いた検定 14

- 以下の検定統計量Zが
  標準正規分布に従うことを利用
  - 平均0、分散1^2の正規分布

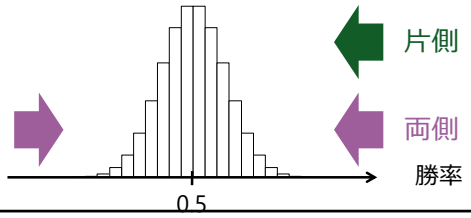
$$\mathrm{Z} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1^2)$$

ただし、$\hat{p}$は観察された勝率
p_0は帰無仮説の下での勝率
nは試合数（サンプルサイズ）

14

片側p値？両側p値？ 15

- 勝率が高い方からみるか、
  勝率が低い方からみるものも加えるか
  - 通常、両側p値は片側p値の2倍
  - 両側p値で報告することが多い



15

片側検定と両側検定 16

- 言いたい対立仮説に応じて選ぶべき
- 片側検定
  - 帰無仮説：勝率は0.5である
    対立仮説：勝率は0.5より高い
- 両側検定
  - 帰無仮説：勝率は0.5である
    対立仮説：勝率は0.5より高い、もしくは低い

16

脚気論争 17

- 長期航海において脚気患者が続出
  - 食事内容を変更し、同一航路で訓練航海

食事	脚気の発生		合計
	あり	なし	
洋食	14 (4.2%)	319	333
米食	169 (44.9%)	207	376

- 洋食にすれば脚気は減る？

17

割合の95%信頼区間を計算 18

- 洋食
  - $\frac{14}{333} \pm 1.96\sqrt{\frac{\frac{14}{333}\times\frac{333-14}{333}}{333}} \approx (0.020, 0.064)$
- 米食
  - $\frac{169}{376} \pm 1.96\sqrt{\frac{\frac{169}{376}\times\frac{376-169}{376}}{376}} \approx (0.399, 0.500)$
- どうやら差はありそう

18

統計的仮説検定を導入 19

- 帰無仮説H_0：
  食事によって脚気発生割合は変わらない
- 対立仮説H_1：
  洋食は米食より脚気発生割合が低い
  - 片側検定を用いる
- 脚気発生割合の群間差に注目

19

仮説をパラメータ化 20

- 洋食での脚気発生割合p_1
- 米食での脚気発生割合p_2
- 帰無仮説$H_0: p_1 = p_2$
  - 差($\Delta = p_1 - p_2$)で書き直せば、$H_0: \Delta = 0$
- 対立仮説$H_1: p_1 < p_2$
  - 差($\Delta = p_1 - p_2$)で書き直せば、$H_1: \Delta < 0$

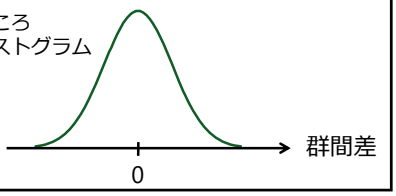
20

帰無仮説が正しい下で

21

- 同様の研究を繰り返したならば、
  観察される群間差は、
  ゼロを中心として、左右対称に分布
  - サンプルサイズ大きくなるにつれ、
    正規分布に近づく

研究を繰り返したところ
得られた群間差のヒストグラム



21

群間差が従う正規分布

22

- 帰無仮説が正しい下で、群間差について
  - 平均は0
  - 分散は$\frac{t}{N}\cdot\frac{N-t}{N}\cdot\left(\frac{1}{n}+\frac{1}{m}\right)$

比較群	疾病発生		合計
	あり	なし	
試験群	a	b	n
対照群	c	d	m
合計	t	$N-t$	N

22

割合の差の検定

23

- 帰無仮説の下で、
  以下の検定統計量Zが標準正規分布に従う

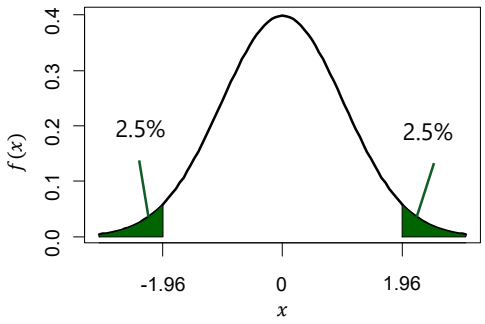
$$Z=\frac{(\widehat{p_1}-\widehat{p_2})-0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}}\sim N(0,1^2)$$

ただし、$\widehat{p_1}=\frac{a}{n},\widehat{p_2}=\frac{c}{m},\hat{p}=\frac{t}{N}$

23

標準正規分布

24



24

カイ二乗検定

25

- 検定統計量Zを2乗して整理すると

$$Z^2=\chi^2=\frac{N(ad-bc)^2}{nmt(N-t)}\sim\chi_1^2$$

  - 自由度1のカイ二乗分布
  - 割合の差の検定と全く同じ
- 2×2分割表以外の分割表にも拡張可能

25

脚気論争の例

26

- 割合の差の検定
  - $Z=\frac{\left(\frac{14}{333}-\frac{169}{376}\right)-0}{\sqrt{\frac{183}{709}\cdot\frac{526}{709}\left(\frac{1}{333}+\frac{1}{376}\right)}}=-375.74<-1.96$
- 片側2.5%有意水準で有意差あり
  - 「洋食は米食より脚気発生割合が低い」

26

この例では？

27

比較群	疾病発生		合計
	あり	なし	
試験群	10 (50%)	10	20
対照群	4 (20%)	16	20
合計	14	26	40

- 割合の差の検定：$Z = 1.66 < 1.96$
  - 有意差なし
- 「群間に差がなかった」とはいえない

27

仮説検定での2つのエラー

28

- αエラー (type-I エラー,第一種の過誤)
  - 本当は差がないのに有意差ありという誤り
    - 消費者リスク
    - (**A**watenbo)あわてんぼうさんの間違い
- βエラー (type-II エラー,第二種の過誤)
  - 本当は差があるのに有意差を出せない誤り
    - 生産者リスク
    - (**B**onyari)ぼんやり者の間違い

28

αエラーとβエラー

29

検定結果	母集団（真）	
	差がない	差がある
有意差なし	正しい	第2種の過誤 β エラー
有意差あり	第1種の過誤 α エラー	正しい

有意水準で大きさを決める

サンプルサイズを増やすことで軽減

29

有意差なし ≠ 差がない

30

- 本当に差がなかった
- βエラーによって、
  本当は差があるのに有意差なしかも
  - どちらが正しいかはデータから判定不能
  - 背理法の理屈からも、
    帰無仮説が正しいとはいえない

30

統計的仮説検定のポイント

31

- 二者択一の意思決定
  - 差がある or 差があるか分からない
- 有意差がある場合、
  差があることを主張する強力な方法
  - 背景には背理法の理屈
- 有意差がない場合、解釈が難しい
  - 差がない、といってはならない

31

効果や影響の大きさ

32

- 検定では、差があることしか分からない
- 脚気割合の場合は、信頼区間を求めた

$$\frac{\hat{p}-\mu}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \pm 1.96$$

- 群間差でも信頼区間を求めたい
  - 推定値の精度を議論できる

32

脚気発生割合の場合

33

信頼区間

- $\frac{\hat{p}-\mu}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \pm 1.96$ をμについて解いた

p値の計算

- $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1^2)$
  - 事前に定めたp_0が正しいか

信頼区間は、検定で有意にならない範囲

33

群間差の95%信頼区間

34

信頼係数という

- 割合の差の検定をベースに、以下の式のμを解く

$$\frac{(\widehat{p_1}-\widehat{p_2})-\mu}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} = \pm 1.96$$

34

脚気論争の例

35

$$(\widehat{p_1}-\widehat{p_2}) \pm 1.96\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}$$

$$\left(\frac{14}{333}-\frac{169}{376}\right) \pm 1.96\sqrt{\frac{183}{709}\cdot\frac{526}{709}\left(\frac{1}{333}+\frac{1}{376}\right)}$$

- (-0.343, -0.472)

35

脚気論争例の解釈

36

- (-0.343, -0.472)
  - 洋食にすると脚気発生割合が30%下がる！
    - ウソ
  - 洋食にすると脚気発生割合が40%下がる！
    - 否定できない
  - 洋食にすると脚気発生割合が下がらない！
    - ウソ
- 差の信頼区間が0を含まない
  ⇔ 検定で有意差あり

36

信頼区間の解釈

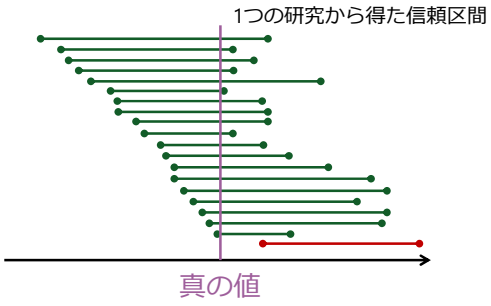
37

- 厳密な解釈
  - 同じような研究を繰り返した場合、研究ごとに信頼区間を推定
  - それら信頼区間100通りあたり、95通りは真の値を含む
- 実際の解釈
  - 今回の研究データから効果の大きさとしてありうるであろう範囲

37

仮想的に研究を繰り返し

38

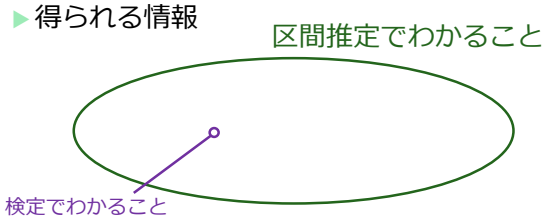


38

区間推定のメリット

39

- 単に差があることをいうだけでなく、<u>どの程度の効果</u>であるのかを議論できる
  - 研究の精度もコミの議論
- 得られる情報



39

Scientists rise up against statistical significance

40

Amrhein V, Greenland S, McShane B. Nature. 2019;567:305-307.

40

P値の誤用入門①

41

- 死亡群と生存群では年齢が異なった
- P>0.05だったから両群は同じ
- 年齢の違いではP=0.04、
  性別の違いではP=0.02ゆえ、
  性別のほうが与える影響が大きい

41

P値の誤用入門②

42

- 有意差がつかないと論文にならないから
  Fisher検定をカイ二乗検定に変えた
- 有意差があったから、
  グループ間には間違いなく差があるんだ

42

P値廃止運動

43

- あまりに科学界で検定を濫用している
- いっそ、P値を示すことを禁止しては
- 医学研究では臨床試験における
  アウトカム比較でのみ検定を使用すべき

43

サンプルサイズ設計

44

- 研究の位置づけによって異なる
- 介入研究
  - 探索的な要素が強いか、検証的な意味合いか
    - パイロット研究か否か
- 観察研究
  - 仮説創出か、なぞる目的か、
  - 因果関係の探索か、予測なのか

44

適切なサンプルサイズ 45

- サンプルサイズをただ大きくすればよいか？
  - 実際の研究では、多くの人数を集めることは時間的、金銭的に困難
  - わずかな差でも有意となってしまい、そのような差に臨床的意味があるのか？
- サンプルサイズが小さいとどうなるか？
  - 臨床的に意味のある差があるのに、有意であるという結果が得られにくくなる
    - 検出力不足

45

サンプルサイズ設計のやり方 46

- 実現可能性に基づく設計
- 検定ベースの設計
  - 例：期待する治療効果がみられた場合に、検出力80%を満たすような最小例数
- 精度ベースの設計
  - 例：期待する反応割合の下で、95%信頼区間幅が30%になるような例数

46

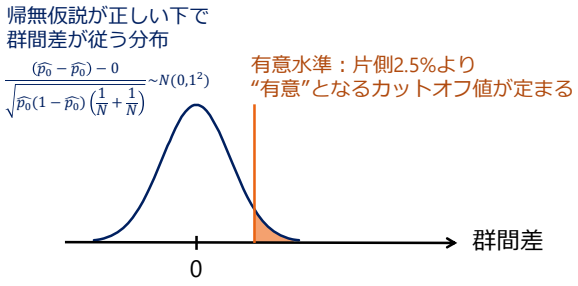
検定ベースの例数設計（二値）① 47

- 対照群の反応割合 p_0
- 期待する試験群の反応割合 p_1
- 割合の差の検定（カイ二乗検定）を実施
  - 群間差 $\Delta = p_1 - p_0$
- 有意水準：片側0.025
- 検出力：0.8
- 割り付け比：均等（1対1）
- 片群のサンプルサイズ N

47

検定ベースの例数設計（二値）② 48

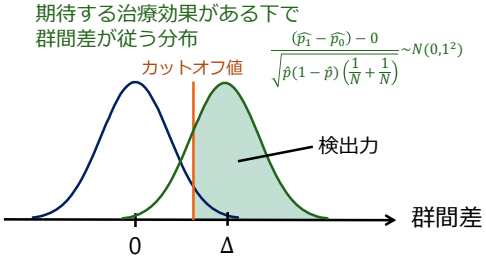
- 帰無仮説が正しい下で群間差が従う分布



48

検定ベースの例数設計（二値）③ 49

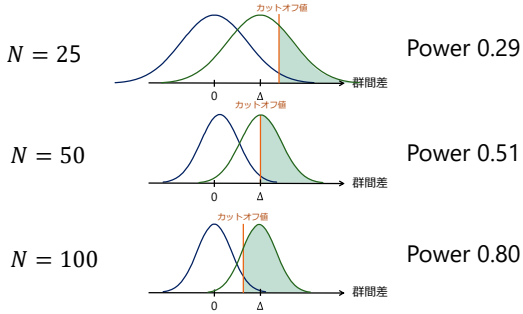
- 期待する治療効果がある下で有意差がつく確率（検出力）を求める



49

検定ベースの例数設計（二値）④ 50

- 検出力が80%以上となる最小のNを探索



50

例数設計に必要な情報 51

- 検定方法、有意水準、検出力
- 効果の大きさとバラツキ
  - 連続量：群間差、群間差の標準偏差
  - 二値：両群の割合
  - 生存時間：ハザード比、対照群での生存率＋登録期間、最低追跡期間

51

サンプルサイズ設計の実際 52

- たいていは専用ソフトを用いる
  - SAS、JMP、nQuery、PASS、、、（有料）
  - SWOGのホームページ（無料）
- 複雑な仮説の場合はシミュレーション
  - 群間差がない下で、n人分のデータを発生し、検定を行い、有意な割合が有意水準以下
  - 期待する効果の下で、n人分のデータを発生、検定を行い、有意な割合が検出力以上
    - このnをいろいろ変えて検討

52

まとめ 53

- 検定と推定
  - 検定は差があることをいうためには強力
  - 推定は効果の大きさを含めたより多くの議論
- カイ二乗検定
  - カテゴリカルデータの群間比較

53