北大・統計解析の基礎②

データの記述と統計的推測

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 医学データの生成過程
  - 記述統計と推測統計
  - 反事実アウトカム
- 「バラツキ」の分解と誤差の確率変数による定式化
- 確率分布
- 統計量と標本分布理論
  - 大数の法則と中心極限定理
  - 統計数値表

2

手元にあるデータをどう活用するか

3

- 記述統計
  - どのようにデータが得られたかを明らかに
  - データのタイプに応じた要約
    - 連続量、カテゴリカル、生存時間
- 推測統計
  - 想定する源泉集団において、曝露や治療とアウトカムの関連を検討
  - 統計的検定、推定を利用
    - 治療(曝露)効果の検証や推定

3

データを集めました！①

4

- ある健診データでの身長(cm)

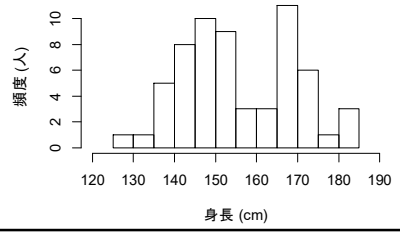
157.0、170.2、164.6、167.6、168.9、
167.6、167.6、168.9、170.9、180.3、
167.1、160.8、170.2、168.4、165.1、
168.4、172.2、182.9、165.9、163.8、
180.3、168.9、173.5、176.5、171.5、
・・・

4

ヒストグラム (histogram)

5

- データのバラツキ状態を可視化
  - 外れ値が存在するか
  - データの分布が多峰性を示すか

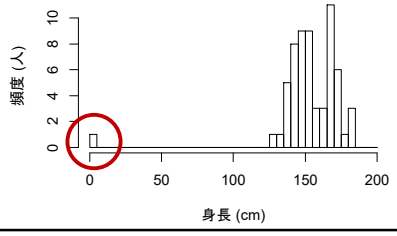


5

ヒストグラム (histogram)

6

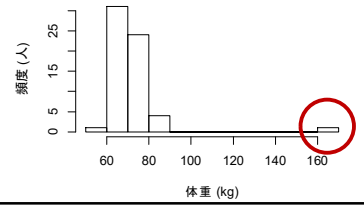
- 外れ値がみられる
  - 145(cm)を誤って1.45(m)と入力
  - データの確認に有効



6

ヒストグラム (histogram) 7

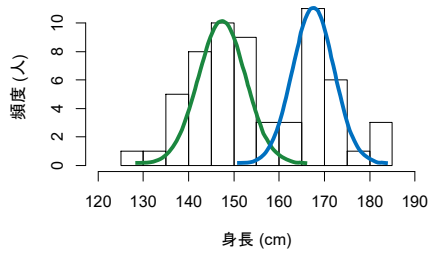
- 166(kg)と入力されている
  - 外れ値とは限らない
  - 医学データの場合、
    外れ値の統計的基準による判定は不向き



7

ヒストグラム (histogram) 8

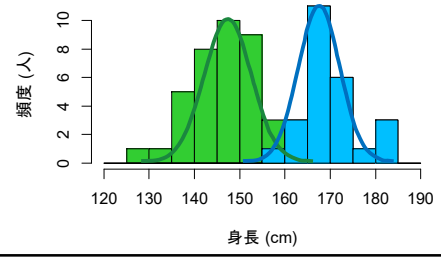
- 二峰性を示しているよう・・・
  - 偶然なのか、理由があるのかを検討



8

ヒストグラム (histogram) 9

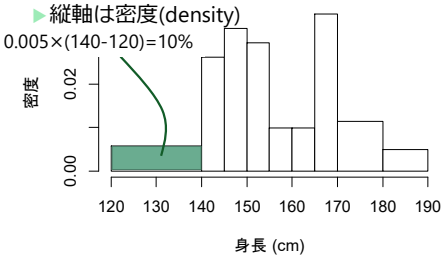
- 15歳以下のグループ(緑色)と
  18歳以上のグループ(青色)が混合(mixture)



9

ヒストグラム (histogram) 10

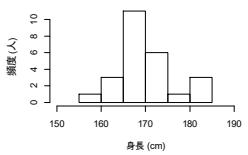
- 面積が割合となるように図示
  - 横幅は等間隔でなくてもよい
  - 縦軸は密度(density)



10

記述統計学 11

- 標本データを要約し、特性を把握
  - 一部の情報を失う
  - 散布図、ヒストグラム、箱ヒゲ図
  - 中央値、範囲、四分位範囲、
    平均値、標準偏差、歪度、尖度・・・



中央値：168.9
四分位範囲：167.1-171.5

11

位置の指標 12

中央値 median

- データを順に並べ替え、
  真ん中の値
- 外れ値に対し頑健
- 数学的な性質は
  あまりよくない

平均値 mean

- 全てのデータを足し、
  データの個数で除した値
- 外れ値に対し頑健でない
- 数学的な性質はよい

1, 2, 3, 4, 50

中央値：3　　平均値：12

12

バラツキの指標 (中央値とセットで使うもの)

13

- 範囲 range
  - (最大値) − (最小値)
  - 医学論文では最小値と最大値をそのまま記述
- 四分位範囲 inter-quartile range
  - 上側四分位(75%)点と下側四分位(25%)点を用いた範囲

1, 2, 3, 4, 50

範囲：1-50　　四分位範囲：2-4

13

バラツキの指標 (平均値とセットで使うもの)

14

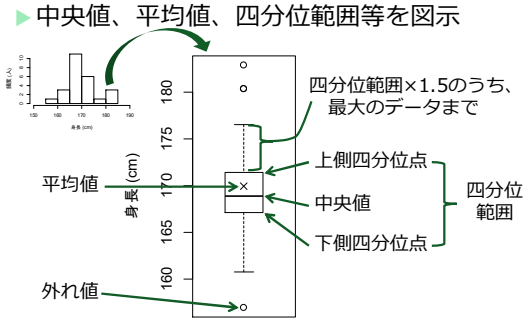
- 標準偏差 Standard Deviation
  - 不偏分散
    - 各データと平均値との差（偏差）の2乗和を(データの個数)-1 で除したもの
    - $\frac{(1-12)^2+(2-12)^2+(3-12)^2+(4-12)^2+(50-1\ )^2}{5-1} = 452.5$
  - 不偏分散の平方根が標準偏差
    - $\sqrt{452.5} \approx 21.3$
- 暗に正規分布を仮定した指標
  - 医学データでは正規分布に従うことはまれ
  - ほぼ、データを記述する場面では不向き

14

箱ヒゲ図 box-whisker plot

15

- 中央値、平均値、四分位範囲等を図示



15

変動係数、歪度と尖度

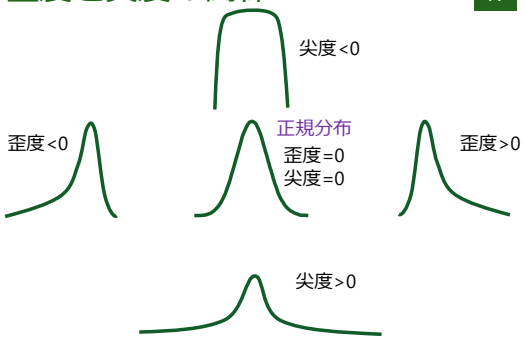
16

- 変動係数 coefficient of variation,CV； σ/μ
  - バラツキが平均を単位にしてどの程度大きいか
  - 測定精度の表現として%表示することも
- 歪度 skewness
- 尖度 kurtosis
  - 正規分布を基準にして考える
  - 平均まわりでの尖りの強さでなく、分布のすそに関する重さを表す
  - 平均、分散とあわせてモーメントに分類

16

歪度と尖度の関係

17

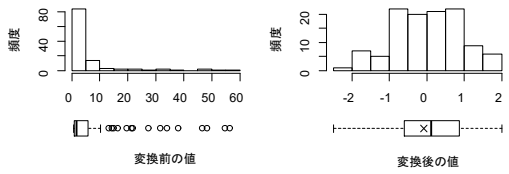


17

変数変換

18

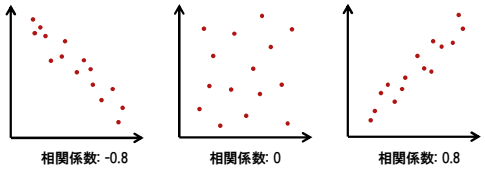
- 歪んだ分布である場合
  - 分布の歪みをとりたい
  - 正規分布、せめて左右対称な分布に近づけたい
    - Box-Cox変換



18

2つのデータの関係に注目 19

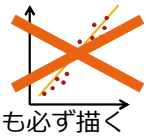
- 散布図: 縦軸、横軸にそれぞれ変数を配置した図
- 相関係数: 強さと方向を-1から1をとる値で代表
  - 0： 2つの変数間に相関なし
  - 1： 2つの変数に正の相関
  - -1 : 2つの変数に負の相関



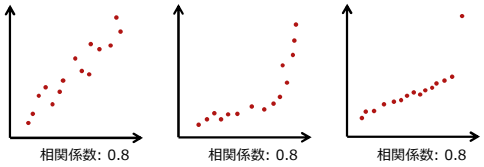
19

散布図と相関係数の注意 20

- 相関に注目する場合、回帰直線は描かない
  - 相関と回帰は別の解析



- 相関係数だけでなく散布図も必ず描く
  - 以下の散布図はすべて同じ相関係数



20

カテゴリカル(二値,多値)データ 21

- 治療(曝露)の有無、進行度ステージ(I, II, III, IV)、疾患の有無
- 分割表による要約
  - 人数と曝露群別に求めた割合を表記

治療	進行度ステージ				合計
	I	II	III	IV	
新治療	2 [4%]	5 [10%]	23 [46%]	20 [40%]	50
標準治療	4 [8%]	4 [8%]	24 [48%]	18 [36%]	50

曝露	疾病発生		合計
	あり	なし	
あり	12 [20%]	48 [80%]	60
なし	16 [10%]	144 [90%]	160

21

割合、率、比 22

- 割合 proportion
  - 全体に占める程度
  - 0から1をとる指標
- 率 rate
  - 単位時間あたりの発生数
  - 0から∞(無限大)をとり、1/単位時間 が単位
- 比 ratio
  - 同じ単位をもつ2指標の相対的な大きさ
  - 0から∞(無限大)をとり、単位はなし

22

効果の指標 23

- 治療(曝露)効果の方向や大きさの表現

指標	差の指標	比の指標
リスク、割合	リスク差	リスク比
オッズ		オッズ比
率	率差	率比
ハザード		ハザード比

23

割合に関する効果の指標① 24

- リスク差 risk difference, リスク比 risk ratio
  - 疾病発生割合(リスク)の群間比較
  - 直感的な解釈
  - NNT (Number Needed to Treat): -1/(リスク差)
    - 何名治療すれば、1人の疾病発生を抑えられるか
- 発生オッズ比 incidence odds ratio
  - 疾病発生オッズ(発生数/非発生数)の群間比較
  - 数学的によい性質

24

割合に関する効果の指標②

25

曝露	疾病発生		合計
	あり	なし	
あり	12 [20%]	48 [80%]	60
なし	16 [10%]	144 [90%]	160

- リスク差: $\frac{12}{60} - \frac{16}{160} = 0.1$
- リスク比: $\frac{12}{60} / \frac{16}{160} = 2.0$
- オッズ比: $\frac{12}{48} / \frac{16}{144} = 2.25$

- 曝露により、疾病発生リスクが
- 10%増加
- 2倍に増加
- 上昇(オッズ比：2.3倍)

25

発生率に関する効果の指標①

26

- 観察人時間 observed person-time
  - のべ何単位時間だけ観察したか
    - 40人を5年、20人を6年観察した場合、320人年
    - 発生数を観察人時間で除したものが発生率
- 発生率差 incidence rate difference
- 発生率比 incidence rate ratio

26

発生率に関する効果の指標②

27

曝露	疾病発生数	観察人年	発生率 [/年]
あり	12	320	0.0375
なし	16	800	0.0200

- 発生率差
  - $^{12}/_{320} - {}^{16}/_{800} = 0.0375 - 0.0200 = 0.0175$ [/年]
  - 曝露により1年あたりの発生率が0.0175増加
- 発生率比
  - $\frac{^{12}/_{320}}{^{16}/_{800}} = \frac{0.0375}{0.0200} \approx 1.88$
  - 曝露により1年あたりの発生率が1.88倍に増加

27

発生率とハザード

28

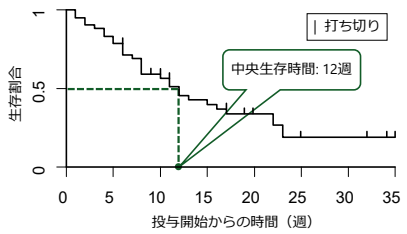
- 発生率は観察人時間あたりの疾病発生数
  - カウントデータ
- よく似た指標に、ハザード (hazard / hazard rate)
  - 直前まで観察である下で、その次の瞬間における発生しやすさ
  - 1/単位時間 が単位であり、解釈は発生率とほぼ同様
    - 効果の指標としてハザード比
  - 生存時間データ
    - 打ち切り(censoring)を含むデータ
    - ある時点まで生存は確認されているが、その時点以降に存在するイベント時点が正確に分からない

28

カプラン・マイヤー曲線

29

- 各時点における生存割合 (1-(疾病発生割合))を階段状にプロット



29

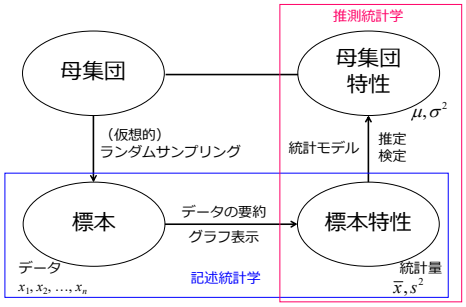
推測統計学

30

- 集団全体を調べなくとも、その特性を標本から確率的に推測
  - 部分から全体への推測
- 頻度流統計学 v.s. ベイズ流統計学
  - 今日は頻度流統計学を紹介
    - 多くの数理統計学の教科書で記述
    - 医学データへの適用例が豊富
- 統計的推測 Statistical Inference
  - 推定、検定の２種類に大別

30

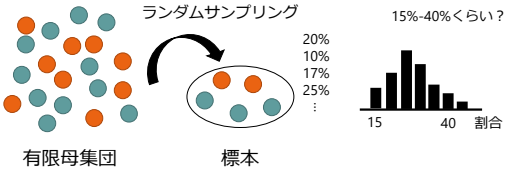
記述統計と推測統計の役割 31



31

母集団からのランダムサンプリング 32

- 有限母集団の特性を推測するには、ランダムサンプリングした標本を用いた
  - 選挙の出口調査



32

統計学を医学研究へ応用 33

- がん登録データベースを用いて、患者の予後に影響を与える因子を検討
- ある睡眠薬について催奇性の有無を調べるケースコントロール研究
- 有限母集団からのランダムサンプリング？？

33

仮想的無限母集団による正当化 34

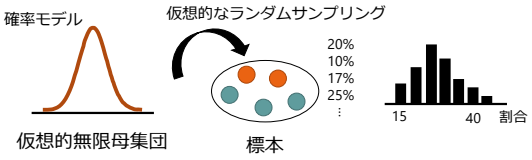
- 仮想的無限母集団
  - がん患者、睡眠薬を処方される人
  - 現在のみならず、過去や将来の患者も含める
- ランダムサンプリングを行った
  - 手元にあるデータは仮想的無限母集団からの標本とみなす
- 有限母集団に対する統計理論が適用可能

時に、この正当化は無理な場合がある

34

確率モデルから得られた標本 35

- 知りたい仮説を確率モデルで表す
  - がん患者の予後因子と死亡の因果関係
  - 睡眠薬が奇形児をもたらす影響
- 母集団は、研究対象（標本）の特徴から操作的に作られるもの



35

この治療は効くんですか？ 36

- 医学的要請が高いもののひとつ
  - 製薬会社は新薬を上市するために治験を行う
  - 研究者は新規治療法の有効性を確認するため、医師主導臨床試験を行う
- 治療法をランダムに被験者に割り付け、2群での有効性を比較する
  - ランダム化、ランダム割付
  - ランダムサンプリングなどしていない

36

内的妥当性 37

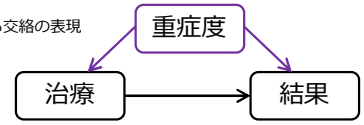
- 研究対象集団で調べたいものが正しく調べられているか
- 4つの妥当性が必要
  - 比較の妥当性
    - そもそも比較群はよく似ている集団か？
  - 追跡の妥当性
  - 測定の妥当性
  - 解析の妥当性

37

比較の妥当性 38

- 交絡(confounding)バイアスによって比較の妥当性が欠けてしまう
  - 治療法Aは軽症な人ほど受けやすい
  - 治療法Bは重症な人ほど受けやすい
  - 治療法の比較では、治療効果をみるのか、重症度の違いによる影響をみるのか不明

DAGによる交絡の表現



38

ランダム割り付けの意義 39

- 全員が治療 A を受けた場合の結果
  - 治療 A を実際に受けた半分の参加者の結果で代用
- 全員が治療 B を受けた場合の結果
  - 治療 B を実際に受けた半分の参加者の結果で代用

これらの妥当性は、
治療法のランダム割り付けが保証

39

ある個人に対する潜在的な結果 40

	治療法	
	A	B
タイプ1	治る	治る
タイプ2	治る	治らない
タイプ3	治らない	治る
タイプ4	治らない	治らない

40

反事実結果変数モデル 41

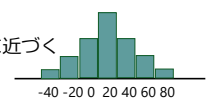
counterfactual outcome model

- 仮想的な10名の対象者の治療に対する潜在的な結果

	A	B	
タイプ1	＋	＋	2名
タイプ2	＋	－	3名
タイプ3	－	＋	1名
タイプ4	－	－	4名

推定したい真の群間差
＝（5 - 3）/10 = 0.2（20%）

  - ランダム割り付けのパターンは、${}_{10}C_5$ = 252通り
- 群間差の分布は？
  - 252通りの群間差の計算
  - 真の治療効果を中心に、対象者を増やせば正規分布に近づく



41

頻度流統計学を適用 42

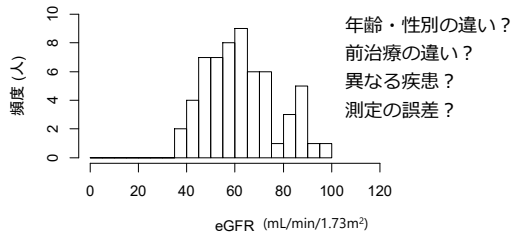
- 有限母集団からのランダムサンプリング
- 仮想的無限母集団からの仮想的なランダムサンプリング
  - 仮想的無限母集団に確率モデルを仮定
- 内的妥当性を保証するためのランダム化に基づく比較
  - 反事実結果変数モデルに基づく平均治療効果に関する統計的推測

42

データを集めました！②

43

- 腎摘出術を受ける患者の術前eGFR
  （推定糸球体濾過量）



年齢・性別の違い？
前治療の違い？
異なる疾患？
測定の誤差？

Isotani S, et al. *Clin Exp Nephrol*. 2015

43

バラツキ variation

44

- 同じようなものを測定したはずなのに、値が異なってしまうこと
  - 統計で問題にするのはこの「バラツキ」
- 頻度流統計学では

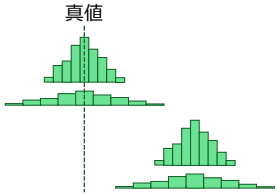
測定値 = 真値 + バイアス + 誤差
measurement　true value　bias　error

44

測定値の正しさ

45

- 正確度 accuracy
  - 真値に一致しているか、ズレていないか
- 精度 precision
  - 真値の周りに集まっているか



正確度	精度
○	○
○	×
×	○
×	×

45

バイアスの特定と制御

46

- 研究結果と知りたい真値とのズレ
  - 選択バイアス、情報バイアス、交絡、・・・
  - 研究者(医師)の主観、測定器の違い、・・・
- 研究デザインと解析モデルにおいて、バイアスを制御することを目指す
  - 疫学研究はバイアスとの戦い
  - 臨床試験では、ランダム化によって、制御できないバイアス要因を期待的にゼロに
    - これらを誤差に転化してしまう

46

誤差の確率変数による定式化

47

変数Xが次の条件を満たすとき、
これを確率変数という

1. Xはいろいろな値をとり得るが、とり得る値の範囲は定まっている
2. Xは、ある時点が過ぎると値が確定するがそれまでは値が不確定である
3. Xのとり得る値についての確率分布は定まっている

吉村功ら,医学・薬学・健康の統計学,サイエンティスト社.2009

47

(例)治療が成功するか X

48

- 条件1
  - Xのとり得る値は1(成功),0(失敗)のいずれか
- 条件2
  - Xの値は、治療するまで不確定
- 条件3
  - Xがそれぞれの値をとる確率は$1/2$ずつ

$$\Pr(X = x) = 1/2\,, x = 0,1$$

48

どうやって真値を調べるか？ 49

- この治療が成功する割合は50%だ
  - ラットでも、イヌでも、サルでも・・・
- 実験で確かめるしかない！
  - 5人に治療を行って、
    成功した人数を調べてみよう

49

5人に治療して成功する人数 K 50

- これも確率変数

$$K = X_1 + X_2 + X_3 + X_4 + X_5$$

  - 1人目について治療が成功するか　X_1
  - 2人目について治療が成功するか　X_2
  - ・・・
- 各対象者が治療成功するか$X(=0,1)$は
  確率pのベルヌーイ分布に従う

$$\Pr(X = x) = p^x(1-p)^{1-x}$$

  - 母平均p 、母分散$p(1-p)$

50

Kの平均や分散は？ 51

- 平均値
  - 各対象者が治療成功するかの平均値を足す
    - $E(K) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) = 5p$
- 分散
  - 各対象者が治療成功するかの分散を足す
    - $V(K) = V(X_1) + V(X_2) + V(X_3) + V(X_4) + V(X_5) = 5p(1-p)$
    - 分散の加法性

51

治療実施をさぼった・・ 52

- 1人目の結果を、
  2,3,4,5人目の結果としてコピペした
  - 平均値
    - 1人目が治療成功するかの平均値を5倍
      - $E(K) = 5 \times E(X_1)$
  - 分散
    - 1人目が治療成功するかの分散を5^2倍
      - $V(K) = 5^2 \times V(X_1) = 25V(X_1)$

52

確率変数の線形変換 53

- 一般化すると、
  平均と分散の特徴は以下の通り
  - $E(aX + bY) = aE(X) + bE(Y)$
  - $V(aX + bY) = a^2V(X) + b^2V(Y)$
    - a, b：定数、X, Y：確率変数

53

治療成功確率pを実験から確認 54

- 5人に治療して成功する確率$\bar{X}$は、

$$\bar{X} = \frac{K}{5} = \frac{1}{5}X_1 + \cdots + \frac{1}{5}X_5$$

  - 平均や分散は
    - $E(\bar{X}) = \frac{1}{5}E(X_1) + \cdots + \frac{1}{5}E(X_5) = p$
    - $V(\bar{X}) = \frac{1}{5^2}V(X_1) + \cdots + \frac{1}{5^2}V(X_5) = \frac{1}{5}p(1-p)$

54

確率分布 55

- 標本空間 sample space
  - 確率変数のとり得る値の全体、集合
- 確率変数 random variable
  - 標本空間上で確率分布が定まっているときに、その分布に従って実現値を出す変数
- 確率変数を特徴づけるもの
  - 確率分布 probability distribution

55

確率分布の特徴づけ 56

- 確率関数、確率密度関数
- 分布関数
- モーメント（積率）
  - 原点周り、平均周り
  - 平均、分散、歪度、尖度
- 確率母関数、積率母関数、特性関数
- キュムラント

56

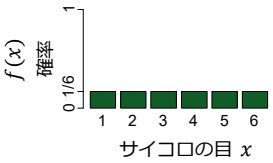
離散型 / 連続型確率関数 57

- 離散型確率関数
  - 二値(陽性/陰性)や整数値のようなとびとびの値をとる
  - 確率分布は確率関数によって特徴づけ
- 連続型確率関数
  - とり得る範囲であらゆる値をとる
  - 確率分布は確率密度関数によって特徴づけ

57

確率関数(離散型確率変数の場合) 58

- 確率変数Xがとる値に対して、その値がとる確率を考えることができる
- $f(x) = \Pr(X = x)$をxの関数とみたもの
- 例：サイコロの目の確率関数



$$f(x) = \begin{cases} 1/6, & x = 1,2,\dots,6 \\ 0, & \text{それ以外} \end{cases}$$

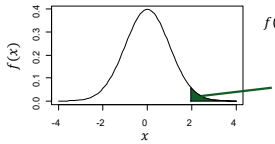
58

確率密度関数(連続型確率変数の場合) 59

- 区間$(a, b]$のどれかの値が実現する確率

$$\Pr(a < X \le b) = \int_a^b f(x)dx$$

- 横軸をx、縦軸を密度関数$f(x)$で図示
- 例：標準正規分布(平均0,分散1^2の正規分布)



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\Pr(1.96 < X < +\infty) = \int_{1.96}^{+\infty} f(x)dx = 0.025$$

59

分布族と母数 60

- 分布族 family of distribution
  - 性質の似た確率分布をグループ分け
- 母数 parameter
  - 分布を一意に指定する役割をもつ変数
  - 母数空間：母数のとり得る値の範囲

分布族	母数
正規分布	平均μ、分散σ^2
二項分布	サイズn、出現率p

60

代表的な分布(族) 61

離散型分布
- 二項分布
- 多項分布
- ポアソン分布
- 負の二項分布
- ベータ二項分布
- 超幾何分布
- 幾何分布

連続型分布
- 正規分布
- カイ二乗分布
- t 分布
- F 分布
- 一様分布
- 指数分布
- ワイブル分布
- ガンマ分布

61

二項分布 $Bin(n,p)$ 62

- ある治療の治癒確率がpの場合、n人の患者に治療した際の治癒した人数X
- 確率関数　$\Pr(X=x) = {}_n\mathrm{C}_x p^x (1-p)^{n-x}$
  - 標本空間　$\{0,1,\dots,n\}$
  - 母数空間　nは正の整数、$p: 0 \le p \le 1$
- 平均 np、分散$np(1-p)$

62

ポアソン分布 $Poisson(\lambda)$ 63

- ハザード（ある瞬間の治癒率）がλの場合、十分に患者がいる下で治癒した人数X
  - ただし治癒発生は、まれである
  - 2項分布にて$np=\lambda$とおき、$n \to \infty$としたときの極限（少数の法則 law of small numbers）
- 確率関数　$\Pr(X=x) = \lambda^x e^{-\lambda}/x!$
  - 標本空間　$\{0,1,\dots,n\}$
  - 母数空間　$\lambda > 0$
- 平均λ、分散λ

63

正規分布 $N(\mu,\sigma^2)$ 64

- 十分に症例数が大きいとき、(ほとんどの)医学データの平均値が従う
  - この性質はあとで解説
- 確率密度関数　$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - 標本空間　$(-\infty,\infty)$
  - 母数空間　$-\infty < \mu < \infty, 0 < \sigma < \infty$
- 平均μ、分散σ^2（標準偏差σ）
  - $N(0,1^2)$としたものを標準正規分布とよぶ

64

カイ二乗分布 $\chi^2(k)$ 65

- 確率変数$X_1, X_2, \dots, X_n$が同一かつ独立に$N(0,1^2)$に従う場合、

$$Y = \sum_{i=1}^{k} X_i^2$$

は自由度kのカイ二乗分布に従う

65

t 分布 $t(m)$ 66

- 確率変数$X_1, X_2, \dots, X_n$が同一かつ独立に$N(\mu,\sigma^2)$に従う場合、

$$T = \frac{\bar{X}-\mu}{\sqrt{s^2/n}}$$

は自由度$m(=n-1)$の t 分布に従う
- $\bar{X} = \sum_i^n X_i/n$
- $s^2 = {}^1/_{n-1}\sum_i^n (X_i-\bar{X})^2$
  - 不偏分散という

66

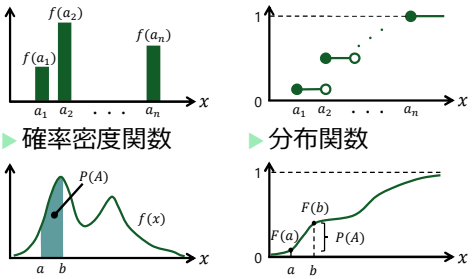
分布関数 $F(x)$ 67

- 標本空間で1つの値xより左側にある確率
  - 区間$(-\infty, x]$の確率
  - 離散型・連続型確率変数が統一的に扱える
- 離散型分布:$\sum_{u \le x} \Pr(X = u)$
- 連続型分布:$\int_{-\infty}^{x} f(u)du$
  - $F(x)$は単調増加で$F(-\infty) = 0, F(\infty) = 1$
  - $F(x)$は右連続：$\lim_{x \to a+0} F(x) = F(a)$

67

確率(密度)関数と分布関数 68

- 確率関数
- 分布関数



- 確率密度関数
- 分布関数

68

パーセント点 69

- 分布関数をある値αにする横軸xの値
  - 統計数値表にまとめられる
  - この逆を統計数値表にまとめることもある
- パーセント点 percentile
  - 下側100α%点：$F(x) = \alpha$となるxの値
  - 上側100α%点：$F(x) = 1 - \alpha$となるxの値
  - 両側100α%点：$F(x) = 1 - {}^{\alpha}/_{2}$となる$x$の値

69

モーメント（積率） 70

- 統計数値表は個別的すぎる
  - 母数が多いと表が大きくなりすぎる
- 分布の形状や位置といった性質を知りたい
  - 原点まわりのk次のモーメント

$$\mu_k = \int x^k f(x)dx, \sum_x x^k \Pr(X = x)$$

  - 平均まわりのk次のモーメント

$$\nu_k = \int (x - \mu_1)^k f(x)dx, \sum_x (x - \mu_1)^k \Pr(X = x)$$

70

平均、分散、標準偏差 71

- 平均：原点まわりの1次のモーメントμ_1
  - 慣習的にμであらわす
- 分散：平均まわりの2次のモーメントν_2
  - 慣習的にσ^2であらわす
  - $\sigma^2 = \mu_2 - \mu_1^2$
- 標準偏差：分散の平方根 σ
  - 変数や平均と同じ次元であり、比較しやすい

71

統計量 72

- 知りたいパラメータの推定するために、個々のデータを要約したもの
  - 治療成功確率を推定するために、（パラメータ）
    5人に治療を行い成功割合（平均）を求める（要約）
- データを集めて計算した平均値は統計量の代表例

72

標本分布

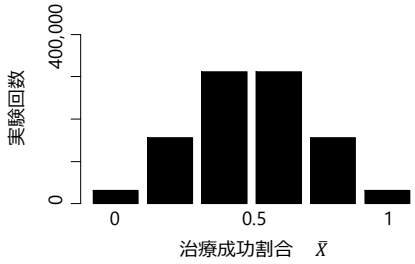
73

- 統計量自体の分布
  - 対象者ごとに治療結果がばらつくように、結果をまとめた統計量もばらつく
- $p = 0.5$として、n人に治療した場合の、成功する割合$\bar{X}$の分布を確認
  - n人に治療して$\bar{X}$を計算する実験を100万回やってみた

73

$n = 5$の場合

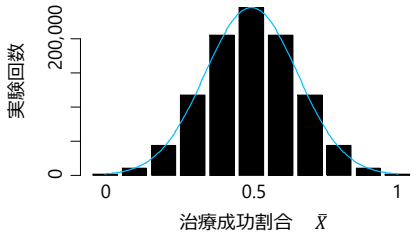
74



74

$n = 10$の場合

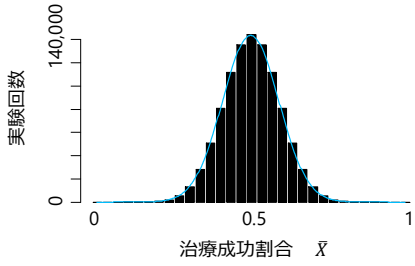
75



75

$n = 30$の場合

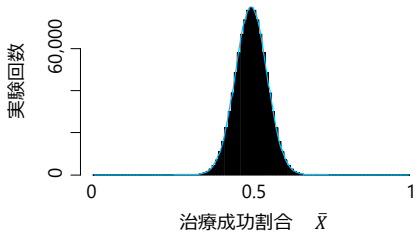
76



76

$n = 100$の場合

77



77

標本分布の導出

78

- もとの確率変数が従う分布が分かれば、標本分布は計算可能
  - 極めて限定的な状況でのみ正確に解ける
  - コンピュータシミュレーションによる数値計算
- もとの確率変数が従う分布が不明でも、標本分布は正規分布にて近似（漸近論）
  - nが十分に大きい場合の特徴
  - 平均と分散のみ計算
  - 大数の法則と中心極限定理で証明

78

大数の法則 79

- 平均値μをもつ確率分布からの独立な確率変数$X_1, X_2, \ldots, X_n$の標本平均$\bar{X}$はμに収束
  - 任意に小さい正の数ε

$$\lim_{n\to\infty} Pr(|\bar{X} - \mu| > \varepsilon) \to 0$$

79

中心極限定理 80

- 独立な確率変数$X_1, X_2, \ldots, X_n$の重み付き和(例えば標本平均)は、正規分布に収束
  - $X_1, X_2, \ldots, X_n$が平均μ、分散σ^2の独立同一分布に従う場合、その標本平均$\bar{X}$について

$$\frac{(\bar{X} - \mu)}{\sqrt{\sigma^2/n}}$$

標準化統計量という

標準誤差という Standard Error

が標準正規分布に従う

80

大数の法則と中心極限定理を認めれば 81

- 1回の研究結果(統計量)と仮説が、どの程度異なるかを定量的に評価できる
- 仮説検定 hypothesis testing
  - ある仮説が正しいと仮定した場合に、研究結果が観測されることはどの程度(順位)まれであるかを確率で表現
  - 意思決定に利用
- 区間推定 interval estimation
  - 仮説検定で「まれ」と判断されない仮説の範囲

81

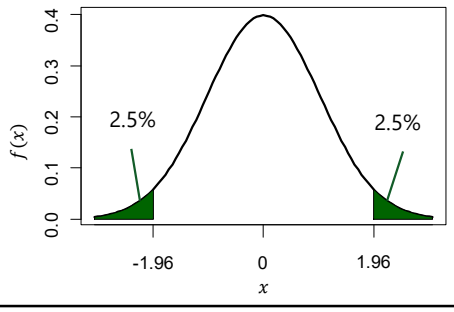
例：抗がん剤の反応割合 82

- がん患者50名に抗がん剤を投与したところ、反応した(CR+PR)者が20名であった
  - 20/50=40%の反応割合
- この反応割合の信頼度は？
  - 信頼区間 confidence interval/limit
- 大数の法則と中心極限定理より、

$$\frac{\hat{p} - \mu}{SE(\hat{p})} \sim N(0,1^2), SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$$

82

標準正規分布 83



83

割合の95%信頼区間 84

- 正規近似による信頼区間
  - 下式を真値μに関して解く

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

  - 今の例では、$0.4 \pm 1.96 \cdot \sqrt{\frac{0.4 \times 0.6}{50}} \approx (0.26, 0.54)$

84

同じ40%でも

85

- 人数によって信頼区間は異なる
  - 得られる情報量の違い
- 同じ40%でも・・
  - 4/10　(0.10,0.70)
  - 20/50　(0.26,0.54)
  - 40/100　(0.30,0.50)
  - 400/1000　(0.37,0.43)

85

90% or 99%信頼区間は？

86

- 95%信頼区間を示すことが多い（慣習）
- 50人中20人が反応した場合
  - 90%信頼区間：　(0.29, 0.51)
    - $\frac{\hat{p}-\mu}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \pm 1.64$を解いた
  - 95%信頼区間：　(0.26, 0.54)
  - 99%信頼区間：　(0.22, 0.58)
    - $\frac{\hat{p}-\mu}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \pm 2.58$を解いた

86

統計数値表

87

- 確率とパーセント点の対応をまとめた表
  - 上側確率0.05に対応するパーセント点uは？
    - このuを用いれば、　%信頼区間が分かる
  - パーセント点0.84に対応する上側確率は？

87

まとめ①

88

- 記述統計と効果の指標
  - ヒストグラム、箱ひげ図、平均値と中央値
  - 分割表による集計
    - リスク差、リスク比、オッズ比
    - 率差、率比
  - カプラン・マイヤー法とハザード比
- 頻度流統計学の適用場面
  - 疫学データのような無限母集団からの仮想的ランダムサンプリング
  - ランダム化研究のような内的妥当性の確保

88

まとめ②

89

- 真値、バイアス、誤差への分解
  - バイアスは制御を目指す
  - 制御できない分は誤差へ転化
  - 確率変数による誤差の定式化
- 標本分布理論と漸近論
  - 大数の法則と中心極限定理により標本平均は真値を中心とした正規分布に収束

89