2021/1/20　医療統計学（３年）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

- 記述統計
  - 連続量、カテゴリカル、生存時間
- 「バラツキ」の分解
  - ランダム化による誤差への転化
- 統計量と標本分布理論
  - 大数の法則と中心極限定理
- 信頼区間の構成

2

データを集めました！①

3

- ある健診データでの身長(cm)

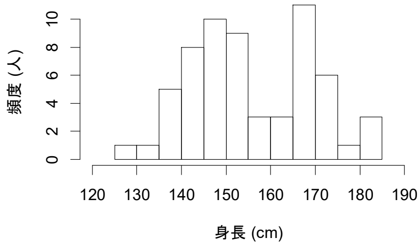
157.0、170.2、164.6、167.6、168.9、
167.6、167.6、168.9、170.9、180.3、
167.1、160.8、170.2、168.4、165.1、
168.4、172.2、182.9、165.9、163.8、
180.3、168.9、173.5、176.5、171.5

3

ヒストグラム (histogram)

4

- 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

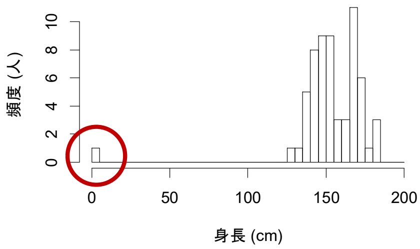


4

ヒストグラム (histogram)

5

- 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

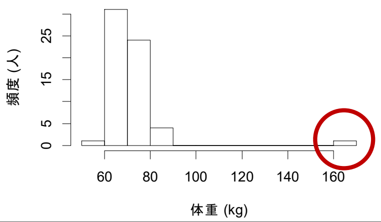


5

ヒストグラム (histogram)

6

- 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

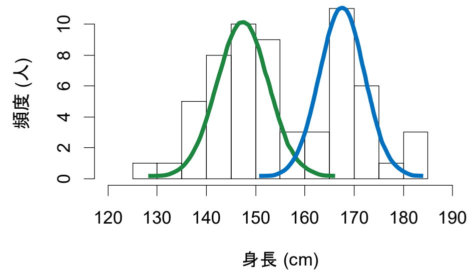


6

ヒストグラム (histogram)

7

- 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

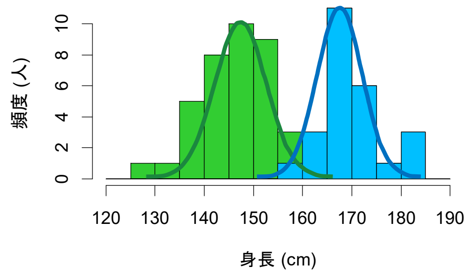


7

ヒストグラム (histogram)

8

- 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)



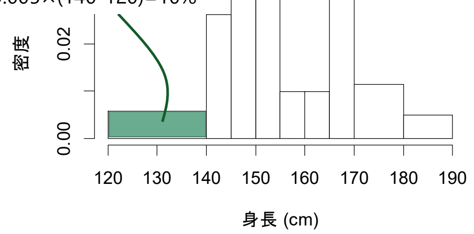
8

ヒストグラム (histogram)

9

- 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)

0.005×(140-120)=10%

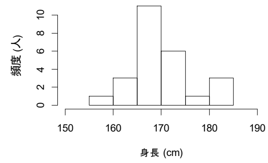


9

記述統計学

10

- 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：152.4
四分位範囲：145.5-167.6

10

位置の指標

11

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

11

バラツキの指標 (中央値とセットで使うもの)

12

- 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

12

バラツキの指標 (平均値とセットで使うもの)

13

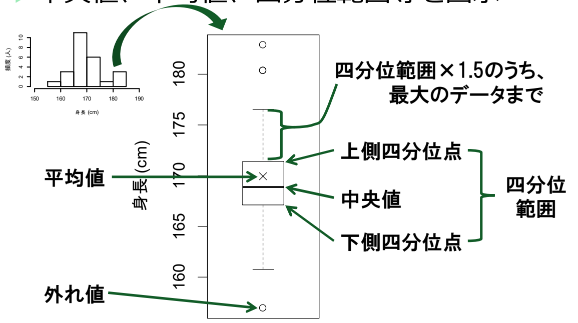
- 標準偏差 Standard Deviation
  - 不偏分散
    - 各データと平均値との差（偏差）の2乗和を(データの個数)-1 で除したもの
    - $\frac{(1-12)^2+(2-12)^2+(3-12)^2+(4-12)^2+(50-1\ \ )^2}{5-1}=452.5$
  - 不偏分散の平方根が標準偏差
    - $\sqrt{452.5}\approx 21.3$
- 暗に正規分布を仮定した指標
  - 医学データでは正規分布に従うことはまれ
  - ほぼ、データを記述する場面では不向き

13

箱ヒゲ図 box-whisker plot

14

- 中央値、平均値、四分位範囲等を図示



14

変動係数、歪度と尖度

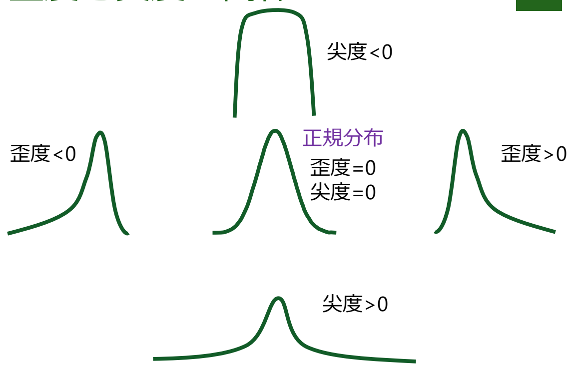
15

- 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

15

歪度と尖度の関係

16

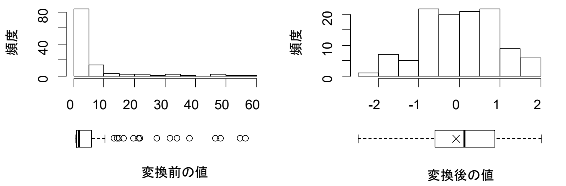


16

変数変換

17

- 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
  - 群間比較を行う際に、群内バラツキを揃えたい

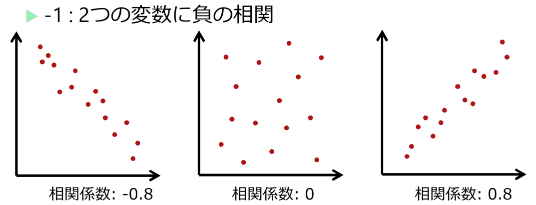


17

2つのデータの関係に注目

18

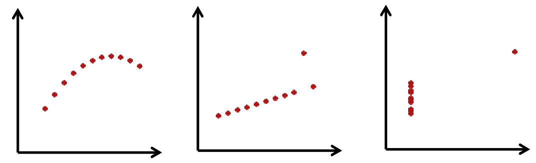
- 散布図: 縦軸、横軸にそれぞれ変数を配置した図
- 相関係数: 強さと方向を-1から1をとる値で代表
  - 0：2つの変数間に相関なし
  - 1：2つの変数に正の相関
  - -1 : 2つの変数に負の相関



18

相関係数は0.816 (Anscombeの例)

19



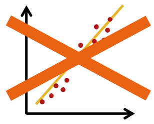
- 相関係数だけでなく散布図も必ず描く
  - 直線性があることを確認

19

相関 ≠ 回帰

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：2つのデータを対等に扱う
  - 縦軸と横軸を入れ替えても、同じ相関係数
- 回帰：結果(縦軸)を原因(横軸)で予測
  - 縦軸と横軸を入れ替えても、回帰直線は対称に入れ替わるわけではない

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

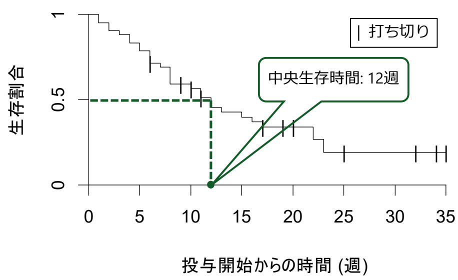
曝露	疾病発生		合計
	あり	なし	
あり	12 [20%]	48 [80%]	60
なし	16 [10%]	144 [90%]	160

21

カプラン・マイヤー曲線

22

- 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

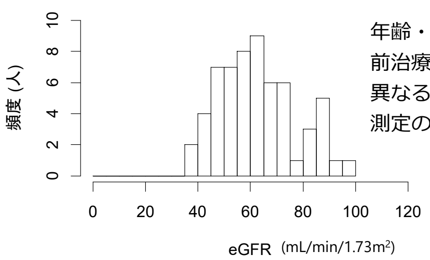


22

データを集めました！②

23

- 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

23

バラツキ variation

24

- 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 (measurement) ＝ 真値 (true value) ＋ バイアス (bias) ＋ 誤差 (error)

腎摘出術を受ける患者の真のeGFR

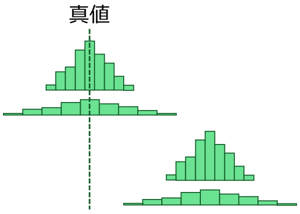
年齢・性別による影響
前治療による影響
⋮

24

測定値の正しさ

25

- 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
○	○
○	×
×	○
×	×

25

バイアスの特定と制御

26

- 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、交絡バイアスを期待的にゼロに
    - これらを誤差に転化してしまう

26

バイアスと誤差の評価

27

バイアス

- バラツキの原因として特定
  - 医学的に有益な情報？
  - 測定値から引けば、正確度が向上（バラツキの制御）
- 制御を目指す
  - できない分は誤差に

誤差

- 確率変数としてモデル化
  - バラツキの原因が特定できない分
  - 制御ができない or あえてしない分
- 原因が分かればバイアスに転化可能

トレードオフの関係

27

誤差の確率変数による定式化

28

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

28

(例)治療が成功するか X

29

- 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

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

29

どうやって真値を調べるか？

30

- 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

30

5人に治療して成功する人数 K

31

- 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$p(1-p)$

31

Kの平均や分散は？

32

- 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

32

治療実施をさぼった・・

33

- 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
      - $V(K) = 5^2 \times V(X_1) = 25V(X_1)$

33

確率変数の線形変換

34

- 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

34

治療成功確率pを実験から確認

35

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

35

統計量

36

- 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

36

標本分布

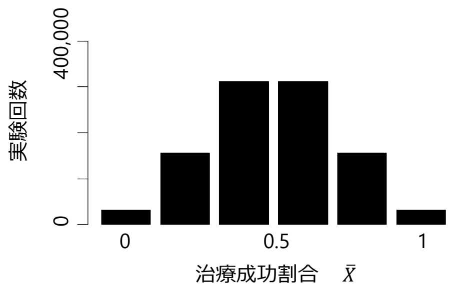
37

- 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

37

$n = 5$の場合

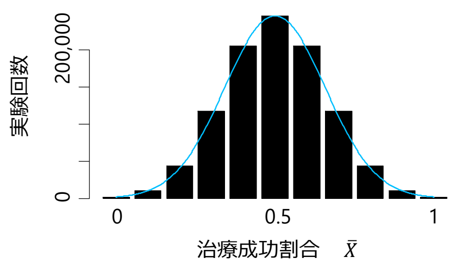
38



38

$n = 10$の場合

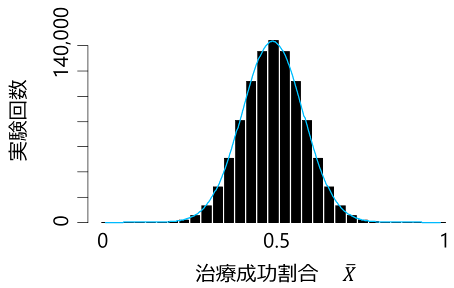
39



39

$n = 30$の場合

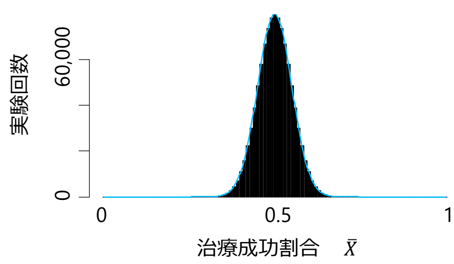
40



40

$n = 100$の場合

41



41

標本分布の導出

42

- 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

42

大数の法則

- 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

$X_1, X_2, \ldots, X_n$を互いに独立に、平均μ、分散σ^2の分布に従う確率変数とするとき、任意に小さい整数εとδに対してある整数mが存在し、$n \geq m$であれば次式が成り立つ
$Pr(|\bar{X} - \mu| < \varepsilon) > 1 - \delta$

43

中心極限定理

- 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

が標準正規分布に従う

標準誤差という
Standard Error

44

大数の法則と中心極限定理を認めれば

- 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

45

標準偏差？標準誤差？

- 前スライドの$\sqrt{\sigma^2/n}$は標準誤差とよぶ
  - Standard Error, SE
- 統計量のバラツキをしめす指標
  - 実験・思考の仮想的繰り返しによるバラツキ
  - 統計量の標準偏差といえる
    - 用語としては必ず「標準誤差」を使う

46

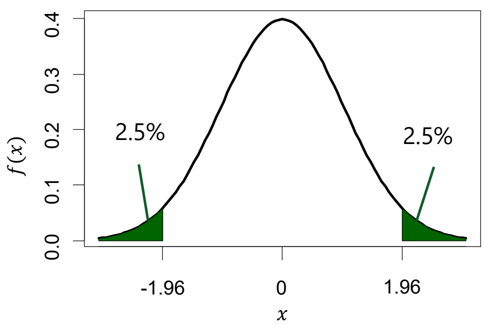
例：脚気論争

- 龍驤艦378名中169名が脚気に罹患
  - 169/378=44.7%の罹患割合
- この罹患割合の信頼度は？
  - 信頼区間 confidence interval/limit
- 大数の法則と中心極限定理より、

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

47

標準正規分布



48

割合の95%信頼区間

- 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、

$$0.447 \pm 1.96 \cdot \sqrt{\frac{0.447 \times 0.553}{378}} \approx (0.397, 0.497)$$

49

95%CI：39.7% - 49.7%

- 米食において、真の脚気の罹患割合は
  - 45%だ！　　そうかもしれない
  - 40%だ！　　そうかもしれない
  - 30%だ！　　それは違うのでは
  - 50%だ！　　それは違うのでは
- 仮説に対して、信頼区間をみれば、間違っている仮説を指摘できる

50

同じ44.7%でも

- 人数によって信頼区間は異なる
  - 得られる情報量の違い
- 同じ約44～45%でも・・
  - 4/9　　$(0.120, 0.769)$
  - 13/29　　$(0.267, 0.629)$
  - 169/378　　$(0.397, 0.497)$
  - 447/1000　　$(0.416, 0.478)$

51

龍驤と筑波で比べてみよう

- 分割表で要約

食事	脚気の発生		合計
	あり	なし	
洋食（筑波）	14（4.2%）	319	333
米食（龍驤）	169 (44.7%)	209	378

- 信頼区間をそれぞれ計算
  - 洋食（筑波）：　4.2% (2.0% - 6.4%)
  - 米食（龍驤）：　44.7% (39.7% - 49.7%)

52

信頼区間を使えば

- 例えば「洋食も米食も脚気罹患割合は20%で同じ」と仮説をおく
  - 洋食の95%CI：2.0% - 6.4%から、そんなに高くないことがいえる
  - 米食の95%CI：39.7% - 49.7%から、そんなに低くないことがいえる
- これらをまとめて、「洋食は米食より脚気発生割合が低い」

結構もったいないことをしている
どれくらい低いか言えていない

53

（軍艦じゃなくて）群間比較

- 群間差とその信頼区間を計算してみよう
- 洋食での脚気罹患割合p_1、人数をn
- 米食での脚気罹患割合p_2、人数をm
- 全体での脚気罹患割合をp
- 以下の式のμを解く

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

54

脚気論争の例で群間比較

55

食事	脚気の発生		合計
	あり	なし	
洋食（筑波）	14 (4.2%)	319	333
米食（龍驤）	169 (44.7%)	209	378

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

$$\left(\frac{14}{333}-\frac{169}{378}\right) \pm 1.96\sqrt{\frac{183}{711}\cdot\frac{528}{711}\left(\frac{1}{333}+\frac{1}{378}\right)}$$

- (-0.341, -0.469)

55

群間差の95%信頼区間から

56

- (-0.341, -0.469)
  - 洋食にすると脚気罹患割合を40%下げる！
    - そうかもしれない
  - 洋食にすると脚気罹患割合を30%下げる！
    - それは違うだろう、もっと下げるよ
  - 洋食にすると脚気罹患割合を50%下げる！
    - それはいいすぎ、そこまで下げない

56

まとめ

57

- データの要約
  - 連続量データは中央値と範囲、四分位範囲
- バラツキの分解
  - 真値、バイアス、誤差
- 大数の法則と中心極限定理
  - nを増やせば、平均値は正規分布に従う
- 信頼区間の計算
  - 結果を幅をもたせて示し、程度を議論する

57

統計数値表（標準正規分布）

58

上側確率	両側確率	%点
0.001	0.002	3.090232
0.005	0.010	2.575829
0.010	0.020	2.326348
0.025	0.050	1.959961
0.050	0.100	1.644854
0.100	0.200	1.281552
0.200	0.400	0.841621

58

次の授業のために

59

- 実際に計算をして頂きます
- 5人の誕生日（日付だけ）を
  みつけておいてください

59