統計（医療統計）
前期・第2回　記述統計

授業担当：徳永伸一

東京医科歯科大学教養部　数学講座

再度注意しておきたいこと

- 教科書（「臨床検査学講座 数学／統計学」）について
  - 「第4刷」であることを再度確認（第3刷までのミスが訂正されている）.
  - 授業は原則として教科書の9～14章（12章は割愛）に沿って進みます.
    - 特に9～11章は徳永が執筆しているのでほぼそのまま.
    - 13・14章についても大筋は同じだが，多少方針が異なるので注意を.
- 欠席した場合は次回までに要点の確認を.
  - 「授業後３日以内（を目標）に授業スライドをｐｄｆファイルに変換してアップロードしておきます」
    と伝えましたが1回目は5月2日にアップロードしました.
    →遅れたら催促してください.
- 要点はそんなに多くない。
  - 疑問点を先延ばしするほど，確実に苦労は増えます！
  - 毎回ざっと復習してから先へ進むので確認を.

ではさっそく・・・

前回（第1回）の授業の概要:

- 統計学とは何か 〜 なぜ統計学が必要か
- 授業全体のOverview
- 教科書第９章「順列・組合せと確率」ほぼ全部
  - 確率の基礎概念 〜 ベイズの定理まで

Overview

- 確率（9章）
- 記述統計（10章）・・・・情報の要約
  - 表やグラフで表す
  - 代表値（平均など）や散布度（分散など）を求める

確率モデル（11章）

- 推測統計（13章〜）
  - 推定（点推定、区間推定）
  - 仮説検定

［復習］第9章「順列・組合せと確率」の概要

Ⅰ．順列と組合せ
Ⅱ．確率の基礎概念
Ⅲ．確率の定義と性質
Ⅳ．条件付き確率と事象の独立性
Ⅴ．ベイズの定理

- 大部分は高校数学（受験数学）の範囲です.
- とはいえ、高校数学であまり取り上げられない抽象的な概念をきちんと理解することはとても重要.
- 曖昧な理解のまま放っておくと後になって命取りとなります.

S. TOKUNAGA　5

［復習］Ⅰ．順列と組合せ

ざっと確認のみ

- 順列 ${}_nP_r = n(n-1)(n-2)\cdot\cdot\cdot(n-r+1)$
- 重複順列 ${}_n\Pi_r = n^r$
- 組合せ $\binom{n}{r} = {}_nC_r = {}_nP_r / r! = n! / (r!(n-r)!)$
  - 【注意】ほんとは$\binom{n}{r}$の形で統一したいが、パワーポイント上できれいに表示させるのはかなりやっかいなのでnCr型も併用します。
  - 11章Ⅳの「2項分布」で用いる.
  - 「2項定理」「パスカルの三角形」は知っておこう。
- 重複組合せ${}_nH_r = {}_{n+r-1}C_r$

[復習] Ⅱ. 確率の基礎概念（1）

◎ 標本空間sample space Ω
・・・確率の対象となる「結果」の全体

◎ 事象event・・・「結果」の一部. Ωの部分集合.
- 根元事象elementary event
・・・「結果」の最小単位. それ以上分割できない事象
- 全事象
・・・すべての事象を含む事象. すなわちΩと一致.
（100%の確率で起こる）
- 空事象ϕ
・・・空集合に対応する事象（起こる確率ゼロ）

[復習] Ⅱ. 確率の基礎概念（2）

◎ AとBの和事象 Union : $A \cup B$

◎ AとBの積事象 intersection : $A \cap B$

◎ A の余事象 complement : A^C または $\bar{A}$

◎ 事象Aの確率probabilityを $P(A)$ で表す.

◎ AとBが排反・・・[定義] $A \cap B = \phi$

［復習］ Ⅲ. 確率の定義と性質

公理的確率（数学的に厳密な定式化）

Ωの事象Aに実数$P(A)$が対応し，以下の3条件（＝確率の公理）を満たすとき，PをΩ上の確率という.

（1） $0 \leqq P(A) \leqq 1$

（2） $P(\Omega)=1$，$P(\phi)=0$

（3） A，Bが互いに排反事象であるとき

$$P(A \cup B)=P(A)+P(B)$$

S. TOKUNAGA 9

［復習］公理からただちに導けること：

（1） $P(A^c)=1-P(A)$

（2） $A \subset B \Rightarrow P(A) \leqq P(B)$

（3） $P(A \cup B)=P(A)+P(B)-P(A \cap B)$

（確率の加法定理）

S. TOKUNAGA 10

[復習] Ⅳ.条件付き確率と事象の独立性（要約）

◎「AのもとでのBの条件付き確率$P(B | A)$」の定義：

$$P(B | A) := P(A \cap B) / P(A)$$

◎ 確率の乗法定理：

$$P(A \cap B) = P(A) \cdot P(B | A) \quad \cdots (1)$$

◎「AとBは（互いに）独立」（定義）

$$\Leftrightarrow \text{「}P(A \cap B) = P(A) \cdot P(B)\text{」} \quad \cdots (2)$$

◎ (1)(2)より、AとBが独立のとき

$$P(B | A) = P(B), \quad P(A | B) = P(\mathrm{A})$$

[復習]あと2つ，とても大事な注意

★試験のとき「独立」と「排反」を混同する人が毎回びっくりするほど多い．再確認しておこう．

「AとBが独立」$\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$

「AとBが排反」$\Leftrightarrow A \cap B = \phi$

$\Rightarrow P(A \cap B) = 0$

★後で出てくる「確率変数の独立性」を「事象の独立性」と混同する人も多い．関連はあるが，次元の異なる概念です．

［復習］ V. ベイズの定理（1）

どんな定理？

- 「原因」（病気など）と「結果」（症状など）の関係がある程度わかっているとき，「結果」からそれがある特定の「原因」によるものである確率を求める定理.
- 確率の乗法定理の簡単な応用.
- 試験によく出る定理
  （経験的確率は90％以上）

S. TOKUNAGA 13

［復習］<u>ベイズの定理 Bayes' Theorem</u>（証明の前に再確認）

事象A_1，A_2，・・・A_r，$B \in \Omega$ について

［仮定］ ① $\bigcup_{1 \leqq k \leqq r} A_k = \Omega$ かつ
② 各A_k は互いに排反

であるとき，

［結論］ 条件付確率$P(A_1|B)$ に関して，以下の公式が成立つ.

$$\boxed{P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{\sum_{k=1}^{r} P(A_k)P(B \mid A_k)}}$$

S. TOKUNAGA 14

[復習]ベイズの定理の証明の概略（ r ＝ 3とする）

［仮定］①: $A_1 \cup A_2 \cup A_3 = \Omega$　**②:** A_1，A_2，A_3 **は互いに排反**

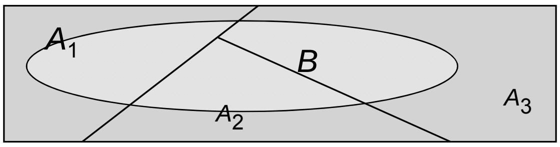
$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)}$$

$$= \frac{P(A_1 \cap B)}{P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)} \quad \leftarrow ①②より$$

$$= \frac{P(A_1)\, P(B \mid A_1)}{P(A_1)\, P(B \mid A_1) + P(A_2)\, P(B \mid A_2) + P(A_3)\, P(B \mid A_3)}$$

（証明終わり）

Ω



[復習]余談

「ピタゴラスの定理ってなんですか」と尋ねられて，

「 **AB²＋BC²＝CA²** です」と答える人は，まずいない．

しかし！

「『ベイズの定理』の内容を書きなさい」

と試験で出題すると，

$$\boxed{P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{\sum_{k=1}^{r} P(A_k)P(B \mid A_k)}}$$

だけ書く人は

少なくない．

[復習] r = 2 の場合に関する補足

r = 2 のとき，仮定の条件は

「A_2はA_1の余事象」

と言っているのと同じ。よって

$A_1 = A,\ A_2 = \overline{A}$ として

$$\boxed{P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\overline{A})P(B \mid \overline{A})}}$$

と書ける（仮定は自動的に満たされるので一般に成り立つ式となる）

（復習ここまで）.

例題（p.75）

- 事象A：「病気Xにかかっている」
- 事象B：「検診で陽性と判定される」

陽性と判定されたとき、実際にその病気にかかっている確率　P（A｜B）を求める問題。

条件：

P（B｜A）　= 0.99

P（B｜A^C）　= 0.07

P（A）　= 0.01

P（A^C）　= 1 － 0.01 = 0.99

例題(p.75)の解答と考察

P(A｜B)

=(P(A)P(B｜A))／(P(A)P(B｜A)＋P(A^C)P(B｜A^C))

= (0.01×0.99)／(0.01×0.99 + 0.99×0.07)

= <u>0.125</u> ・・・(答)

→意外と小さい？

<u>考察のポイント</u>

- 検診結果が陽性でも，実際には病気Xでない確率の方がずっと高い.
- しかし1%→12.5%だから確率は10倍以上.
- 使い方、結果の理解の仕方（患者への伝え方）が重要。

第9章の概要REPRISE：

- Ⅰ．順列と組合せ
- Ⅱ．確率の基礎概念
  - 標本空間、事象
- Ⅲ．確率の定義と性質
  - 確率の公理
- Ⅳ．条件付き確率と事象の独立性
  - 「事象の独立性」の定義
- Ⅴ．ベイズの定理
  - 仮定と結論

では10章へ.

第10章 記述統計

Ⅰ. 統計データの種類

Ⅱ. 度数分布

1. 階級と度数，度数分布表
2. 度数分布表の視覚化（ヒストグラム）

Ⅲ. データの特性値

1. 代表値（平均・メディアン・モード）
2. 散布度（分散と標準偏差、不偏分散）

Ⅰ. 統計データの種類

⦿定性的データ

・・・性別、血液型、出身都道府県etc.

⦿定量的データ

・・・数値で与えられるデータ

- 離散的discreteデータ

  個数・人数、その他とびとびの値をとるもの

- 連続的continuousデータ

  身長・体重等、実数値で与えられるもの

★「離散的」か「連続的」かで数学的な扱い方が異なる

Ⅱ．度数分布

KEYWORDS

1-階級と度数，度数分布表

- ⦿ 度数frequence，度数分布表
- ⦿ 階級class，階級値
- ⦿ スタージェスの公式
- ⦿ 相対度数，累積度数，累積相対度数

2-度数分布の視覚化

・・・ヒストグラム，折れ線グラフ 等

S. TOKUNAGA 23

Ⅲ．データの特性値（1）

代表値と散布度

⦿ 代表値：分布の中心的な位置を示す.

例：平均値mean，中央値median，最頻値mode

⦿ 散布度：分布の広がり・ばらつきの度合いを示す.

例：分散variance，標準偏差standard deviation,

四分位範囲，平均偏差

S. TOKUNAGA 24

Ⅲ．データの特性値（2）

1-代表値

［１］平均

データ $x_1,x_2,\cdots, x_n$ に対し,

平均 $\bar{x} :=(x_1+x_2+\cdots+ x_n)/ n = (1/n)\Sigma x_k$

と定義される．

度数分布表（階級数：m）が与えられているときは

階級値 $x'_1, x'_2, \cdots, x'_m$ と度数 $f_1, f_2, \cdots, f_m$ を用いて

$$\bar{x} := (1/n)\Sigma x'_k f_k$$

と計算（一種の近似計算）.

Ⅲ．データの特性値（3）

その他の代表値

［2］メディアンmedian

＝中央値（順位的に真ん中の値）

＊データが偶数個の場合は「真ん中の2つ」の平均

［3］モードmode

＝最頻値（度数が最大となる値、or階級値）

Ⅲ．データの特性値（4）

2−散布度

［1］分散variance と 標準偏差standard deviation

データ $x_1, x_2, \cdots, x_n$の平均 $\bar{x}$ に対し，

分散 $\sigma^2 := \{ \Sigma (x_k - \bar{x})^2 \} / n$

標準偏差＝「σ^2の正の平方根」、すなわち

$\sigma := \sqrt{(\sigma^2)}$

S. TOKUNAGA 27

Ⅲ．データの特性値（5）

［1］（続き）

階級値 $x'_1, x'_2, \cdots, x'_m$ と

度数 $f_1, f_2, \cdots, f_m$ を用いると

$$\sigma^2 := (1/n)\Sigma (x'_k - \bar{x})^2 f_k$$

S. TOKUNAGA 28

Ⅲ．データの特性値（6）

[2]不偏分散unbiased variance

データ $x_1, x_2, \cdots, x_n$の平均 $\bar{x}$ に対し，

不偏分散 $U^2 := \{ \Sigma (x_k - \bar{x})^2 \} / (n-1)$

★nではなく(n－1)で割る理由：不偏性（→第13章Ⅱ）
★バラツキの度合いを表す指標としては同等.
★nが十分大きいときにはnで割っても(n−1)で割っても大差ない.
（たとえばn＝10000で有効数字3桁なら無視できる）

S. TOKUNAGA 29

Ⅲ．データの特性値（7）

不偏分散についての補足

★本によっては

①「分散」を不偏分散の形で定義

②「分散」は同じだが「**標本分散**」を不偏分散の形で定義

しているケースもあり、用語の使い方が統一されていない（以前使用していた教科書でも「標本分散＝不偏分散」としていた）．

★上記①②のケースでは、標準偏差ないし標本標準偏差を不偏分散の正の平方根$U = \sqrt{U^2}$で定義。

S. TOKUNAGA 30

Ⅲ．データの特性値（8）

［3］その他の散布度

平均偏差　（$\Sigma | x_k - \bar{x} |$）／ n

というのもあるが，あまり使われない．

（散布度の指標としては自然だが数学的な分析に向かない）

他に四分位範囲など

今日（第2回）新たに学んだこと

第10章 記述統計

Ⅰ．統計データの種類

Ⅱ．度数分布

1. 階級と度数，度数分布表
2. 度数分布表の視覚化（ヒストグラム）

Ⅲ．データの特性値

1. 代表値（平均・メディアン・モード）
2. 散布度（分散と標準偏差，不偏分散）