統計（医療統計）
前期・第3回　確率変数と確率分布（1）

授業担当：徳永伸一

東京医科歯科大学教養部　数学講座

前回（第2回）の授業の概要：

- 第１回（教科書第９章「順列・組合せと確率」ほぼ全部）の復習
- 教科書第１０章「記述統計」

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］<u>ベイズの定理 Bayes' Theorem</u>

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.

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

[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

$A_1 = A$, $A_2 = \bar{A}$ として

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

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] 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。

<u>条件：</u>

P（B | A）= 0.99

P（B | A^C）= 0.07

P（A）= 0.01

P（A^C）= 1 － 0.01 = 0.99

[復習]例題(p.75)の解答と考察

P(A | B)

$= (P(A)P(B \mid A)) / (P(A)P(B \mid A) + P(A^C)P(B \mid A^C))$

$= (0.01 \times 0.99) / (0.01 \times 0.99 + 0.99 \times 0.07)$

= <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。

[復習]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）

［復習］Ⅰ. 統計データの種類 &Ⅱ. 度数分布

Ⅰ．統計データの種類

◎ 定性的データ

◎ 定量的データ

- 離散的discreteデータ
- 連続的continuousデータ

★「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

◎ 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

◎ ヒストグラム

［復習］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

[復習] Ⅲ. データの特性値（2-3）

1-代表値

［１］平均mean

データ $x_1,x_2,\ldots, x_n$ に対し,

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

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

$\bar{x}$ ：＝ （1/n）$\sum x'_k f_k$

と計算（一種の近似計算）。

［２］メディアンmedianmean＝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）

[復習] Ⅲ. データの特性値（4-5）

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 σ^2 :＝{ Σ(x_kー $\bar{x}$)^2 } / n

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

σ^2 :＝ (1/n) Σ (x'_k − $\bar{x}$)$^2 f_k$

標準偏差＝「σ^2の正の平方根」、すなわち

σ :＝√(σ^2)

[復習] 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,

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

★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）

[復習] 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。

（復習ここまで）

第11章 確率変数と確率分布

はじめに

確率変数は，確率・統計の学習において
もっとも基本的かつ重要な概念
であるが、きちんと理解するのは意外と難しい.
（一度わかってしまえば簡単だが）

ということを頭に留めておきましょう.

第11章 確率変数と確率分布

Ⅰ. 確率変数と確率分布の定義

Ⅱ. 確率変数の特性値
- 期待値（平均），分散など

Ⅲ. 確率変数の独立性

Ⅳ. 代表的な確率分布
- 2項分布，正規分布など

Ⅴ. 中心極限定理と正規近似

Ⅵ. 標本分布

Ⅰ．確率変数と確率分布の定義（1）

1-確率変数の定義

［定義］ 標本空間Ω上の実数値関数

（各根元事象に実数を対応させたもの）を確率変数 random variable という.

- とり得る値が離散的→離散型確率変数
- とり得る値が連続的→連続型確率変数

Ⅰ．確率変数と確率分布の定義（2）

教科書p.83例1

Ω： サイコロを振ったときの，目の出方で定まる事象全体の集合.

⊙ 「サイコロを振って１の目が出る」は 事象.

⊙ 「サイコロを振ってｉの目が出る」 という事象ω_i に整数 i を対応させる関数をX（＝$X(\omega_i)$）とおくと，Xは（離散型）確率変数 となる.

⊙ 確率変数Xに対し，

- 「X＝１」「X≦４」
- 「Xは偶数」

などは事象.

Ⅰ．確率変数と確率分布の定義（3）

2-離散型確率変数の確率分布

［定義］離散型確率変数Xのとる値xと， Xがその値をとる確率P(X=x)との対応関係を（Xの）確率分布という.

教科書p.84例３

X：サイコロを1回振ったときの目の値.

Xの確率分布（離散型）：

★関数 f(x)=P(X=x) を「Xの確率分布」とよんで差し支えない。

Ⅰ．確率変数と確率分布の定義（4）

離散型確率変数の性質：

離散型確率変数Xの取り得る値を$x_1,x_2,\cdots$とする.

f(x) = P（X=x）とおくと，f は確率の性質（公理）より

$$f(x_k) \geqq 0 \quad (k=1,2,\cdots) \quad かつ \Sigma f(x_k)=1$$

を満たすことがただちに導ける.

次に連続型確率変数へ

Ⅰ．確率変数と確率分布の定義（5）

3-連続型確率変数の確率分布

教科書p.83例2：

「ある短大の1年生から無作為に選んだ1名の身長」をXcmとすると，Xは連続型確率変数.

（とり得る値が連続的になっただけ）

では、

Xが連続型確率変数のとき，離散型の場合と同様に

「確率変数Xのとる値xと，確率P(X=x)との対応関係」

（もしくは関数 f(x)=P(X=x) そのもの）

を（連続型）確率分布と呼んで良いだろうか？

Ⅰ．確率変数と確率分布の定義（6）

そもそも

「連続型確率変数Xと確率との対応関係」

とは？

［注意］Xが連続型確率変数のとき，

（特殊な例を除き）ほとんどすべての値ｘに対して

P(X=x)＝0である！

つまり

Ⅰ．確率変数と確率分布の定義（7）

連続型確率分布は
f(x)=P(X=x)のような関数で表すことはできない.

そこでこれに代わるものとして確率密度関数を導入.
［定義］

$f(x) \geqq 0$， $\int_{-\infty \leqq x \leqq \infty} f(x)dx = 1$であり，

$$P(a \leqq X \leqq b) = \int_{a \leqq x \leqq b} f(x)dx$$

であるような関数 f を，連続型確率変数Xの
確率密度関数という.
★すなわち連続型確率分布は，確率密度関数により表される.

連続型確率分布の例

教科書p.85例4〈一様分布〉

a,bを定数とするとき，密度関数

f(x)＝P(X=x)＝1/（b−a）（a≦x≦b）
f(x)＝P(X=x)＝0　　　（x＜aまたはx＞b）

であらわされる確率分布を一様分布という.

- このときXは一様確率変数または**一様乱数**
- EXCEL課題で用いるRAND関数の値はa=0,b=1とした一様乱数.

Ⅰ．確率変数と確率分布の定義（8）

［注意］

F(x)＝P(X≦x)

をXの累積分布関数という．

- ◎ 図11-1(b), 11-2(b)でイメージをつかんでください．
- ◎「累積」を省略して分布関数と呼ばれることも多く，紛らわしいので気をつけましょう．
- ◎ Excelの関数「BINOMDIST」で4つ目の引数を「TRUE」にした場合がこれに相当（→Excel実習の際に確認を）

Ⅱ．確率変数の特性値（1）

<u>1-期待値と分散・標準偏差の定義</u>

確率変数Xの平均(＝期待値expectation)E(X)

を次式で定義：

$E(X):=\sum x_k P(X=x_k)$　（Xが離散型）

$E(X):=\int x f(x)dx$　（Xが連続型）

(ただしf(x)はXの確率密度関数)

Xの値を繰り返し取り出したとき，それらの平均値は回数を増やすほどE(X)に近づくと考えられる

Ⅱ．確率変数の特性値（2）

μ ＝E(X)とするとき，
確率変数の分散varianceV(X)を

$$V(X):=E((X-\mu)^2)$$

で定義．すなわち，

- $V(X)=\Sigma(x_i-\mu)^2 P(X=x_i)$（Xが離散型）
- $V(X)=\int(x-\mu)^2 f(x)\,dx$（Xが連続型）

分散V(X)は，Xのばらつき，変動の指標となる．
$V(X)=\sigma^2$と表すことも多い．

◉ Xの標準偏差$\sigma=\sigma(X):=\sqrt{\sigma^2}$

Ⅱ．確率変数の特性値（3）

<u>期待値（平均）Eの性質：</u>

Xを確率変数，
a，b を定数(constant)とするとき，

$E(X+b) = E(X)+b$
$E(aX) = aE(X)$

が成り立つ．
以上合わせて

$E(aX+b) = aE(X)+b$

より一般には，定数a,bと関数 f,g に対して

$E(af(X)+bg(X)) = aE(f(X)) + bE(g(X))$

(教科書には載っていません)

Ⅱ．確率変数の特性値（4）

分散の性質：（Xは確率変数，a，b は定数）

$$V(X+b) = E\left((X+b-E(X+b))^2\right) = E\left((X+b-E(X)-b)^2\right) = E\left((X-E(X))^2\right) = V(X)$$

$$V(aX) = E\left((aX-E(aX))^2\right) = E\left((aX-aE(X))^2\right) = E\left(a^2(X-E(X))^2\right) = a^2V(X)$$

以上合わせて　$V(aX+b) = a^2V(X)$

Ⅱ．確率変数の特性値（5）

★以下は有名な公式ですが，教科書には載っていません．

分散の公式：（$\mu = E(X)$とする）

$$V(X) = E(X^2) - E(X)^2$$

［証明］

$$V(X) = E\left((X-\mu)^2\right) = E\left((X^2-2X\mu+\mu^2)\right) = E(X^2) - 2\mu E(X) + \mu^2 \quad \cdots(*) = E(X^2) - E(X)^2$$

注意：（＊）で公式

$$E(af(X)+bg(X)) = aE(f(X)) + bE(g(X))$$

を使ってます．

Ⅱ．確率変数の特性値（6）

<u>教科書p.87例5</u>

X：サイコロを1回振ったときの目の値　とする.

Xの確率分布（離散型）：

E（X）＝ ΣkP（X=k）＝（1+2+・・・+6）/6 ＝7/2＝ 3.5

V（X）＝ Σ（k－3.5）^{2}P（X=k）

　＝（（1－3.5）^{2}+（2－3.5）^{2} +・・・+（6 －3.5）^{2} ）/6

　＝ 35/12 ＝ 2.916666・・・

教科書p.87問題4

Z：サイコロを2回振ったときの目の和の値　とする.
このときZの確率分布（離散型）は：

E（Z）＝ ΣkP（Z=k）

　＝ 2・1/36 + 3・2/36 + ・・・+12/36

　＝ 7 ＝ 2×3.5

V（Z）＝ Σ（k－7）^{2}P（Z=k）

　＝ ・・・ ＝ 35/6 ＝ 2×35/12

期待値の加法性（その1）

実は・・・
任意の確率変数X，Yに対し

$$E(X+Y) = E(X)+E(Y)$$

が成り立っている！（期待値の加法性）

先の例2だと，サイコロを2回振ったとき
　X：1回目に出る目の値，Y ：2回目に出る目の値
とすれば，

$$E(X)=E(Y) = 3.5$$

となり，Z＝X＋Yなので

$$E(Z) = 3.5+3.5 = 7$$

期待値の加法性（その2）

Z_n：サイコロをn回振ったときの目の和
とすれば，

$$E(Z_n)=3.5n$$

も成り立つ．

さらに一般に，
　任意の定数$a_1,a_2,\cdots,a_n$と
　任意の確率変数$X_1,X_2,\cdots,X_n$に対し

$$E(\Sigma a_k X_k) = \Sigma a_k E(X_k)$$

が成り立つ（期待値の線形性）．

ところで，分散については？

分散の加法性と確率変数の独立性

先のサイコロを2回振る例では，分散についても

$$V(Z) = 2 \times 35/12$$

が成り立っていた．
実は

Z_n：サイコロをn回振ったときの目の和

とすれば，

$$V(Z_n) = n \times (35/12)$$

も成り立っている．

しかし，「分散の加法性」

$$V(X+Y) = V(X) + V(Y)$$

は（「期待値の加法性」と違って）いつでも成り立つわけではない！

成り立つための（十分）条件：

→ **確率変数の独立性**

（詳しい説明は次回）

第11章 確率変数と確率分布

Ⅰ．確率変数と確率分布の定義

Ⅱ．確率変数の特性値

- 期待値（平均），分散など

＊＊＊今日はこの辺まで＊＊＊

Ⅲ．確率変数の独立性

Ⅳ．代表的な確率分布

- 2項分布，正規分布など

Ⅴ．中心極限定理と正規近似

Ⅵ．標本分布