統計（医療統計）
第７回　前期の復習

授業担当：徳永伸一

東京医科歯科大学教養部　数学講座

あらためて注意しておきたいこと（前期のはじめに注意したこと＋α）

- 後期の授業は今日を含め（たった）６回。
- 成績評価は後期試験で。
  - 今年度は前期中間試験を実施しませんでした。
- 欠席した場合は次回までに要点の確認を。
  - 次回の授業までに授業スライドをｐｄｆファイルに変換してアップロードする予定。なかなかアップロードされない場合は催促してください。
- 要点を効率よく押さえましょう。
  - 「計算方法」だけでなく「統計的手法」を理解すること。
  - その前提となる概念・原理の理解も重視（当然だ！）。
  - 試験で出題のポイントとなる（＝合否の分かれ目となる）事項は授業でも強調しているつもりなのですが・・・

今日（後期第1回）の授業の概要：

- 授業全体（前期＋後期）のOverview
- 前期の内容の復習＋新たな注意

［再確認］教科書について

- 引き続き
  「臨床検査学講座 数学／統計学」（医歯薬出版）
  を使用します（13章途中から）。
- 徳永は後期の授業範囲（13章・14章）は執筆していませんが、徳永が担当した
  - 第11章 確率変数と確率分布（18ページ）
  は後期の範囲を理解する上でも重要なので随時参照してください。
- でも教科書は絶対的なものではありません。
  - 判明しているミスプリ等については授業で（すなわち授業スライドでも）指摘するので必ず確認し修正しておくこと。
  - あえて教科書と違うやり方で説明する部分もあります。

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）（13章）
  - 仮説検定（14章）

第9章の概要

- 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

［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

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

［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）

$\Leftrightarrow$「$P(A \cap B) \;=\; P(A) \cdot P(B)$」 $\cdots (2)$

- (1)(2)より、 AとBが独立のとき

$$P(B|A) = P(B), \quad P(A|B) = P(A)$$

[復習] あと2つ，とても大事な注意

★試験の答案で「独立」と「排反」を混同して用いている人が多い（毎年強調しているので減ってはきたが）。

再確認：

「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．→中間試験[5]

［復習］ベイズの定理 Bayes’ Theorem

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．

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

［復習］もういちど念押し

「定理」というものは，必ず**仮定**と**結論**から成り立っています．

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，

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

だけ（すなわち結論だけ）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

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

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公式となる）。

平成２０年度統計中間試験問題［１］

問題文：

事象Xが起こる原因として、A,B,Cという3つの事象が考えられるとする。

（1）ベイズの定理を用いて$P(A \mid X)$を求めるとき、A,B,Cが満たすべき条件を、適切な数学用語または事象の関係式を用いて答えよ。

（2）A,B,Cが（1）の条件を満たすとき、確率$P(A \mid X)$はどのように表されるか。条件付き確率を用いた式で記述せよ。

（3）（2）の式を条件付き確率の定義に基づいて証明せよ。ただし（1）の条件をどこで用いたかを明記すること。

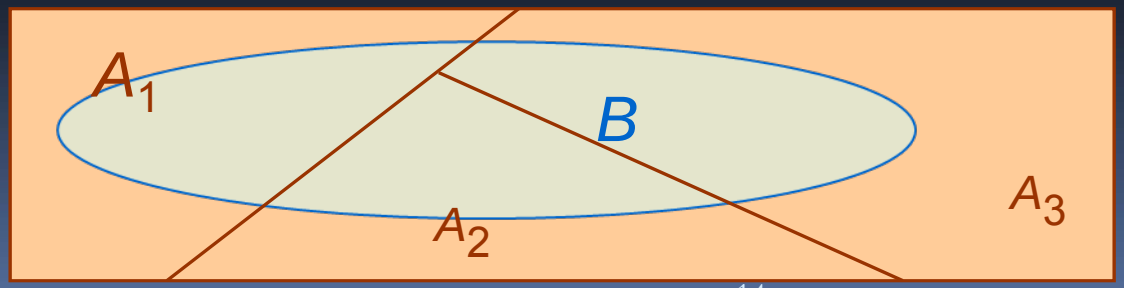
［復習］ベイズの定理の証明の概略（ $r = 3$とする）

［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）

[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] Ⅲ. データの特性値

1-代表値

[1]平均mean

[2]メディアンmedian＝中央値（順位的に真ん中の値）

[3]モードmode＝最頻値（度数が最大となる値、or階級値）

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

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

[復習] Ⅲ. データの特性値（続き）

[2]不偏分散unbiased deviation

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（前期の最後の方）

★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（補足）

【重要】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。毎回確認すべし！

★いずれのケースも「標準偏差」＝「分散の平方根」

（「分散」の定義が異なれば標準偏差も異なる！）

★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。

★特に「標準偏差」の定義はよく確認する習慣を！

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

［復習］　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

［復習］　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(ω_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．

［復習］　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）：

k	1	2	3	4	5	6
P(X=k)	1/6	1/6	1/6	1/6	1/6	1/6

★関数 f(x)=P(X=x) を「Xの確率分布」とよんで差し支えない。

［復習］　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.

［復習］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

$$f(x)=P(X=x)=1/(b-a) \quad (a \leqq x \leqq b)$$
$$f(x)=P(X=x)=0 \quad (x<a \text{ または } x>b)$$

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課題で用いたRAND関数の値はa=0,b=1とした一様乱数.

→19年度・20年度中間試験［3］

平成２０年度統計中間試験問題［３］（１）（２）

問題文：

［３］EXCELにおけるRAND関数は 0 以上 1 未満の値をランダムに取る。この値 X を確率変数と見なしたとき、以下の問いに答えよ。

（１）X が従う分布の確率密度関数 $f(x)$ はどのような関数となるか。適切に記述せよ。

（２）X の期待値 μ と分散 σ^2 を、（連続型）確率変数の期待値と分散の定義に基づいて求めよ。

［復習］　Ⅱ．確率変数の特性値

1−期待値と分散・標準偏差の定義

確率変数Xの平均（＝期待値expectation）E(X)の定義：

- $E(X) := \Sigma\, x_k\, P(X = x_k)$　　（Xが離散型）
- $E(X) := \int x\, f(x)\, dx$　　（Xが連続型）

$\mu = E(X)$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）

（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

[復習]　期待値と分散の性質まとめ１

（以下Xは確率変数，a，b 等は定数）

期待値（平均）Eの性質：

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

分散の性質：

$$V(aX+b) = a^2V(X)$$

確率変数の標準化（上の性質の応用）

$E(X)=\mu$，$V(X)=\sigma^2$ のとき

$Z=(X-\mu)/\sigma$ はXの標準化変数と呼ばれ，

$$E(Z)=0,\ V(Z)=1$$

［復習］　期待値と分散の性質まとめ２

期待値の加法性
任意の確率変数X，Yに対し
$$E(X+Y) = E(X)+E(Y)$$
さらに一般には，
任意の定数$a_1,a_2,\cdots,a_n$と
任意の確率変数定数$X_1,X_2,\cdots,X_n$に対し
$$E\left(\sum a_k X_k\right) = \sum a_k E(X_k)$$
が成り立つ（**期待値の線形性**）．

分散の加法性
確率変数X，Yが互いに独立のとき
$$V(X\pm Y) = V(X)+V(Y)$$
さらに一般には，
互いに独立な確率変数 $X_1,X_2,\cdots,X_n$ と
任意の定数 $a_1,a_2,\cdots,a_n$ 対し
$$V\left(\sum a_i X_i\right) = \sum a_i^2 V(X_i)$$
が成り立つ．

平成２０年度統計中間試験問題［２］（１）

問題文:

［2］(1)立方体の6つの面にそれぞれ1，1，1，2，3，4の目が描かれた特殊なサイコロ(ただし各面はそれぞれ確率１／６で出る)を振る試行を繰り返すとき、i回目に出た目の値を X_i で表すことにする。

$X = \sum_{i=1}^{10} X_i$ とおくと、E(X)=[(a)]、V(X)=[(b)]。

また $X_1, \cdots, X_{10}$ の平均 $\bar{X}$ の標準化変数を Z とおくと、

$Z = (\bar{X} - [(c)]) / [(d)]$。

- 基本問題です。

［復習］　Ⅲ.確率変数の独立性（1）

まず(確率変数ではなくて)事象の独立性について再確認

2つの事象A, Bの独立性は

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

と定義された.

これは

「AとBの成否が互いにもう一方の影響を受けない」

という状態を表していた.

確率変数の独立性についても

確率変数X, Yが

「互いにもう一方の影響を受けない」

という状態を数式を用いて明確に定義したい.

［復習］　Ⅲ.確率変数の独立性（3）

離散型確率変数の場合
『X，Yのとり得るすべての値x，yについて
事象「X=x」と事象「Y=y」が独立』
ならば「XとYは独立」としよう.

すなわち言い換えると:

［定義（離散型の場合）］
X，Y：離散型確率変数 のとき
『X，Yのとり得るすべての値x，yについて
P（X=x かつ Y=y）＝ P（X=x）P（Y=y）』
ならば「XとYは独立」という.

【注意】この定義を連続型確率変数の場合にそのまま当てはめることはできない.

［復習］Ⅲ．確率変数の独立性（5）

［定義（一般の場合）］
確率変数X，Yについて
『任意の実数a,b,c,dに対し
事象「a≦X≦b」と事象「c≦Y≦d」が独立』，
すなわち
『任意の実数a,b,c,dに対し
P（a≦X≦b かつ c≦Y≦d）
＝ P（ a≦X≦b ）P（c≦Y≦d ）』
ならば「XとYは独立」という．

【注意1】「任意の」という条件は本質的．
【注意2】上の定義は離散的確率変数にも適用できる．

平成２０年度統計中間試験問題［５］

問題文：

［５］サイコロを２回振ったとき、出た目の大きい方をX、平均値をYとする。

（１）XとYは離散型確率変数とみなせる。XとYが独立であるとすれば、どのようなことが成り立っていなければならないか。「確率変数の独立性」の定義に基づいて説明せよ。

（２）XとYが独立かどうかを判定せよ。

［復習］ 11章Ⅳ.代表的な確率分布（1）

1- 2項分布binomial distribution

X～ B(n,p)のとき

$$P(X=x)={}_nC_x \cdot p^x \cdot (1-p)^{(n-x)}$$

Xの期待値と分散の公式：

$$E(X)=np\ ,\ V(X)=np(1-p)$$

∵2項分布B(n,p)に従う確率変数Xは，B(1,p)に従うn個の独立な確率変数$X_1, X_2, \cdots, X_n$の和とみなすことができるので、それぞれ$nE(X_i)$, $nV(X_i)$として求められる.

2- ポアソン分布Poisson distribution

X～ $Po(\lambda)$のとき

$$f(X)=P(X=x)=e^{-\lambda}\ \lambda^x / x!$$

$$E(X)=V(X)=\lambda$$

［復習］　Ⅳ.代表的な確率分布（2）

3- 正規分布normal distribution

[1]正規分布の線形変換と標準化

Xが正規分布に従うとき，
その線形変換（Y＝aX+bの形の変換）も正規分布に従う.
したがって，X～ $N(\mu, \sigma^2)$ のとき

Xの標準化変数 $Z=(X-\mu)/\sigma$ は
標準正規分布N(0,1) に従う.

[2]正規分布の再生性

確率変数 X_1, X_2 が互いに独立で、
$X_1 \sim N(\mu_1, \sigma^2_1)$, $X_2 \sim N(\mu_2, \sigma^2_2)$ のとき，

$$X_1+X_2 \sim N(\mu_1+\mu_2,\ \sigma^2_1+\sigma^2_2)$$

平成２０年度統計中間試験問題［2］（3）（4）

問題文:

［2］

（3）X～N(0,1）のとき、P（［（g）］≦X≦1.29）＝0.203

（4）X～N(10,25）のとき、

P（15≦X≦20）＝ P（［(h)］≦Z≦ ［(i)］ ）＝ ［(j)］

ただしZはXの標準化変数とする。

［復習］　Ⅴ.中心極限定理と正規近似（1）

中心極限定理1

［仮定］$X_1, X_2, \cdots, X_n$が

（任意の！）同じ分布に従う独立な確率変数

ならば,

［結論］$n \to \infty$のとき,

和$X_1 + X_2 + \cdots + X_n$の分布は

正規分布に収束する！

[復習]　Ⅴ.中心極限定理と正規近似（2）

中心極限定理2（言い換え）

「互いに独立な確率変数$X_1, X_2, \cdots, X_n$の分布が同一で，

$$E(X_k) = \mu,\ V(X_k) = \sigma^2 \quad (k = 1,2,\cdots,n)$$

であるとき，nが十分大きければ，和ΣX_kの分布は$N(n\mu, n\sigma^2)$にで近似できる」

【注意】仮定すべき条件は独立性と同一分布性のみ．元の分布は任意．

[復習]　Ⅴ.中心極限定理と正規近似（3）

二項分布の正規近似

中心極限定理により，nが十分大きいとき，
B(n,p)はN(np, np(1−p))で近似できる.

よって標準化変数

$$Z = (X - E(X)) / \sqrt{V(X)}$$

$$= (X - np) / \sqrt{(np(1-p))}$$

は近似的にN(0,1)に従う.

∵ B(n,p)に従う確率変数は，B(1,p)に従う 独立なn個の確率変数の和と見なせるから.

［復習］　Ⅴ.中心極限定理と正規近似（4）

半整数補正

- nが大きければかなり良い近似であると思われるが，　nが小さいときはどのくらい誤差が出るのだろうか？
  - p.97問題10のケースで厳密値と正規近似の値を比較せよ.
  - そもそも離散型の確率変数を連続型で近似することに無理がある。
    - 余事象の確率を足しても1にならない？

nが小さいときに少しでも誤差を減らす方法を考えよう

- X～B(n,p)とする．整数a,bに対しP(a≦X≦b)を正規近似で求める際，P(a−0.5≦X≦b+0.5)と補正してから計算した方が誤差が減る．この補正を「不連続補正」ないし「半整数補正」といい，特にnが小さいときに効果的.
  - 区間を広げる方向に0.5ずらす（図で確認 ）.
    - すなわち確率は補正前より増える！
  - 再びp.97問題10で誤差の減少を確認.

平成２０年度統計中間試験問題［２］（５）

問題文：

［2］（4）X ～ B(9,1/3) なる X に対し、

P(X ≦ 2) の厳密値を求める式は［(k)］となる
（注：計算しなくてよい）。

またP(X≦2)を正規近似を用いて求めるとき、

半整数補正を行うと［(l)］、半整数補正なしだと［(m)］となる。

［復習］ Ⅵ.標本分布（1）

1−母集団分布と標本分布

KEYWORDS：母集団⇔標本，無作為抽出，母集団分布，統計量，標本分布

★母集団から無作為抽出した個々のデータの値を確率変数をみなして，確率分布の理論を適用することができる！

2−標本平均の分布

- 個々の標本データの値 $X_1, X_2, \cdots, X_n$ はもちろん確率変数と見なすことができる.
- 標本平均 $\bar{X}$ も1つの確率変数とみなすことができる！
  （一定の大きさの標本を繰り返し抽出し，その度に標本平均の値を計算すれば，「標本平均の分布」を観察することができる）.

よって・・・

［復習］ Ⅵ.標本分布（2）

標本平均の期待値と分散･標準偏差

$X_1, X_2, \cdots, X_n$ を平均 μ，分散 σ^2 である母集団から無作為抽出した標本とするとき，

$X_1, X_2, \cdots, X_n$ はそれぞれ，期待値 μ，分散 σ^2 の互いに独立な確率変数と見なせる．

よって標本平均 $\bar{X}$ について

$$E(\bar{X}) = \mu \times n \times (1/n) = \mu$$

$$V(\bar{X}) = \sigma^2 \times n \times (1/n)^2 = \sigma^2/n$$

（期待値･分散の加法性↑）　（↑積に関するE,Vの性質より）

$$\sigma(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{(\sigma^2/n)} = \sigma / \sqrt{n}$$

［復習］ Ⅵ.標本分布（2）

正規母集団を仮定すると･･･

定理（正規分布の性質より）

$X_1, X_2, \cdots, X_n$をN(μ, σ^2)に従う母集団から無作為抽出した標本とすると

- 和 $\Sigma X_k \sim N(n\mu, n\sigma^2)$
- 標本平均 $\bar{X} \sim N(\mu, \sigma^2/n)$

さらに $\bar{X}$ の標準化変数Zについて:

- $Z = (\bar{X} - \mu) / \sqrt{(\sigma^2/n)} \sim N(0,1)$

［復習］ Ⅵ.標本分布（3）

さらに！

$$n \times \bar{X} = X_1 + X_2 + \cdots + X_n$$

であるから，

nが十分大きければ，母集団分布が正規分布でなくても中心極限定理によって標本平均の分布を正規分布で近似できる！

注意：

- 同一分布性：同一の母集団から抽出したから
- 独立性：無作為抽出により保証される
  （非復元抽出だと厳密には独立にはならないが近似的に）
- 正規分布に従う確率変数はnで割っても正規分布．

したがって・・・

［復習］ Ⅵ.標本分布（4）

定理（中心極限定理の系）

$X_1, X_2, \cdots, X_n$を平均μ，分散σ^2である任意の母集団から無作為抽出した大きさnの標本とするとき，

標本平均 $\bar{X}$ の分布は，nが十分大きければ，
正規分布$N(\mu, \sigma^2/n)$で近似できる．

さらに $\bar{X}$ の標準化変数$Z = (\bar{X} - \mu)/\sqrt{(\sigma^2/n)}$は
標準正規分布N(0,1)で近似できる．

【注意】母集団分布が任意でよいことにあらためて注目．これにより，（十分大きい標本さえ得られれば）未知の分布を持つ母集団の母平均を推定・検定する際，正規分布が利用できる！

[復習] 標本平均の分布・まとめ（対比して再確認）

定理（正規分布の性質より）

$X_1, X_2, \cdots, X_n$を正規分布$N(\mu, \sigma^2)$に従う母集団から無作為抽出した標本とすると

$$\text{標本平均 } \bar{X} \sim N(\mu, \sigma^2/n)$$

定理（中心極限定理の系）

$X_1, X_2, \cdots, X_n$を平均μ，分散σ^2である任意の母集団から無作為抽出した標本とするとき，

標本サイズnが十分大きければ，近似的に

$$\text{標本平均 } \bar{X} \sim N(\mu, \sigma^2/n)$$

となる．

★「3−その他の重要な標本分布」は後回し（13・14章にて）．

19年度・20年度中間試験問題より

（19年度）

［3］（3）EXCELでRAND関数により10（列）×1000行の乱数表を作成し、さらに各行において10個の値の平均を求めた。この平均の値はどのような分布をとるか。理由と共にわかりやすい日本語で説明せよ。

（20年度）

［3］EXCELにおけるRAND関数は 0 以上 1 未満の値をランダムに取る。この値 X を確率変数と見なしたとき、以下の問いに答えよ。

（1）X が従う分布の確率密度関数 $f(x)$ はどのような関数となるか。適切に記述せよ。

（2）X の期待値 μ と分散 σ^2 を、（連続型）確率変数の期待値と分散の定義に基づいて求めよ。

（3）RAND関数で得た10個の値の平均を Y とする。Y はどのような確率分布で近似できると考えられるか。理由と共に述べよ。

【重要】「標本数」と「標本サイズ」

「標本」とは，母集団から取り出したデータの値の「集合」である！

- １回の標本抽出により、１つの標本が得られる。
- ［標本に含まれるデータの値の個数］
  ＝「標本サイズ（標本の大きさ）」

★EXCEL実習3aの例では10列×1000行の値を標本サイズ10、1000標本（標本数1000）と見なして扱った。

★教科書p.105「取りだした標本の数を標本の大きさという」は不適切な記述。

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

1-確率変数の定義

・・・　離散型と連続型

2-離散型確率変数の確率分布

3-連続型確率変数の確率分布

・・・　確率密度関数

Ⅱ．確率変数の特性値

1-期待値と分散・標準偏差の定義

2-確率変数の期待値と分散の性質

・・・　期待値の加法性

3-確率変数の標準化

第11章 確率変数と確率分布

Ⅲ. 確率変数の独立性

1. 確率変数の独立性の定義
2. 独立な確率変数の性質

Ⅳ. 代表的な確率分布

- 2項分布，正規分布など
- 正規分布の線形変換・標準化・再生性

Ⅴ. 中心極限定理と正規近似

1. 中心極限定理
2. 2項分布の正規近似・・・半整数補正

Ⅵ. 標本分布

- 標本平均の分布

第13章 推定

Ⅰ. 母集団と標本

Ⅱ. 点推定

- 不偏性，不偏推定量

Ⅲ. 区間推定

Ⅳ. 母平均の区間推定

1. 母分散が既知のとき ←前回ここまで
2. 母分散が未知のとき

Ⅴ. 母分散の区間推定

Ⅵ. 母比率の区間推定

[復習] Ⅰ. 母集団と標本

KEYWORDS：

- 母集団⇔標本sample，無作為標本
- 母平均，母分散
- 層別，層別抽出stratified sampling
- 乱数

[復習] Ⅱ．点推定（1）

KEYWORDS：

- 母数，母集団パラメータ
  - 母平均，母分散
- （標本）統計量，統計値（統計量の実現値）
  ・・・標本平均，標本分散，標本比率
- 推定量（母数の推定に用いる統計量），
  推定値（推定量の実現値）

点推定と区間推定

- 点推定：「無作為に選んだ標本から数値を得て，それから推定量の推定値を計算し，その推定値がイコール母数であるとする推定方法」
- 区間推定：「推定値から，ある確率である数値の区間の中にある，とする推定方法」

［復習］ Ⅱ．点推定（2） ［不偏性について］

推定量が「不偏unbiasedである（偏りがない）」とは：

- 「対応する母数より大きい（or小さい）値が得られやすい」といった傾向がない．
- その推定量を繰り返し実測し、得られた値（推定値）の平均値は，繰り返しの回数を増やすほど対応する母数に近づく．

厳密には：

- ［定義］母数θの推定量θ’に対し，

$$E(\theta') = \theta$$

のときθ’をθの不偏推定量（不偏性を持つ推定量）であるという．

[復習] Ⅱ. 点推定（3）

$X_1, X_2, \cdots, X_n$に対し

不偏分散 $U^2 := \{ \Sigma (X_k - \bar{X})^2 \} / (n-1)$
は，母分散σ^2の不偏推定量.

- すなわち，$E(U^2) = \sigma^2$ である（証明は割愛）.

ということは

- 分散$S^2 := \{ \Sigma (X_k - \bar{X})^2 \} / n$ については，
  $E(S^2) = E(((n-1)/n)U^2) = ((n-1)/n)\sigma^2$
- つまりS^2の実測値は，母分散より小さめの値をとる傾向にある（すなわち不偏でない）.

[復習] 不偏性に関する補足

- 標本平均 $\bar{X} := \Sigma X_k/n$も母平均μの不偏推定量.

  $\because E(\bar{X}) = E(\Sigma X_k/n) = \Sigma E(X_k)/n = \mu$

- たとえば n=3 のとき$X' = (X_1 + X_2 + 2X_3)/4$ としてもX'はμの不偏推定量(確認せよ).
- 不偏分散の平方根$U=\sqrt{(U^2)}$は，σの不偏推定量ではない！

  $\because E(U) = E(\sqrt{(U^2)}) \neq \sqrt{(E(U^2))} = \sigma$

平成２０年度統計中間試験問題［２］（２）

問題文：

［2］（2）X_1, X_2, X_3 を同一の母集団から無作為抽出した標本とする。このとき

$$Y = (1/4)X_1 + (1/4)X_2 + [(e)]X_3$$

は母平均の不偏推定量となる。

また母分散 $\sigma^2 = 1$ ならば、V（Y）＝［（f）］となる。

第13章 推定

Ⅰ. 母集団と標本

Ⅱ. 点推定

- 不偏性，不偏推定量 ←ここまで終わった

Ⅲ. 区間推定 ←次ここ

Ⅳ. 母平均の区間推定

1. 母分散が既知のとき
2. 母分散が未知のとき

Ⅴ. 母分散の区間推定

Ⅵ. 母比率の区間推定

［復習］ Ⅲ．区間推定

区間推定とは

- 「母数の値をズバリ推定するより，母数が（高い確率で）存在する区間を推定する」という考え方．
- 「推定値が母数にどのくらい近いか（誤差がどのぐらいあるか）」も含めて推定する．

具体的には

- 「母数θが区間Ｉに含まれる確率が○％」といった形の推定を行なう．
  - 「θの推定値θ’の誤差が確率○％でd以下」と考えても同じ。
  - $|\theta-\theta'| \leqq d \Leftrightarrow \theta \in [\theta'-d,\ \theta'+d]$（＝Ｉ）
- ↑における
  - 区間・・・（○%）信頼区間
  - 確率・・・信頼度（または信頼係数）・・・95%，99%など

[復習] Ⅳ. 母平均の区間推定

母平均μの区間推定(母分散既知の場合)

問題設定:

- 標本サイズn, 標本平均 $\bar{X}$ の実測値が与えられており, 母分散σ^2は既知とする.
- 母集団分布・・・正規分布なら好都合だが, 任意の分布でもnが大きければ, 標本平均の分布は中心極限定理により正規分布で近似可能.

以上の条件のもとで,

「μの 100γ%信頼区間」

を求める(γは信頼度＝信頼係数で, 具体的には 0.95, 0.99, 0.90など).

→続いて信頼区間の求め方, でもその前に・・・

［再確認］ Ⅵ.標本分布（2）

標本平均の期待値と分散・標準偏差

$X_1, X_2, \cdots, X_n$ を平均 μ，分散 σ^2 である母集団から無作為抽出した標本とするとき，

$X_1, X_2, \cdots, X_n$ はそれぞれ，期待値 μ，分散 σ^2 の互いに独立な確率変数と見なせる．

よって標本平均 $\bar{X}$ について

$$E(\bar{X}) = \mu \times n \times (1/n) = \mu$$

$$V(\bar{X}) = \sigma^2 \times n \times (1/n)^2 = \sigma^2/n$$

（期待値・分散の加法性↑）　（↑積に関するE,Vの性質より）

$$\sigma(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{(\sigma^2/n)} = \sigma/\sqrt{n}$$

［再確認］標本平均の分布・まとめ（対比して再確認）

定理（正規分布の性質より）

$X_1, X_2, \cdots, X_n$を正規分布$N(\mu, \sigma^2)$に従う母集団から無作為抽出した標本とすると

$$\text{標本平均 } \bar{X} \sim N(\mu, \sigma^2/n)$$

定理（中心極限定理の系）

$X_1, X_2, \cdots, X_n$を平均μ，分散σ^2である任意の母集団から無作為抽出した標本とするとき，

標本サイズnが十分大きければ，近似的に

$$\text{標本平均 } \bar{X} \sim N(\mu, \sigma^2/n)$$

となる．

★以上を踏まえて区間推定の具体的な方法へ

[復習] Ⅳ. 母平均の区間推定

母平均μの区間推定(母分散既知)の解法

- $\bar{X} \sim N(\mu, \sigma^2/n)$と近似(中心極限定理による).
  - 注意:正規母集団を仮定すれば厳密.
- $Z=(\bar{X}-\mu)/(\sigma/\sqrt{n})$ と標準化すると $Z \sim N(0, 1)$
- $P(-z(\alpha/2) \leqq Z \leqq z(\alpha/2))=\gamma$
  - ただし$\gamma=1-\alpha$ で、$z(\alpha)$ は $P(Z \geqq z(\alpha))=\alpha$ を満たす値.
  - たとえば$\gamma=0.95$ のとき$\alpha/2=0.025$, $z(\alpha/2)=z(0.025)=1.96$
  - $z(\alpha/2)$は「上側$100(\alpha/2)$%点」と呼ばれる.
- ↑を同値変形すると

$$P(\bar{X}-z(\alpha/2)\,\sigma/\sqrt{n} \leqq \mu \leqq \bar{X}+z(\alpha/2)\,\sigma/\sqrt{n})=\gamma$$

よって「μの(100×γ)%信頼区間」は

$$(\bar{X}-z(\alpha/2)\,\sigma/\sqrt{n},\ \bar{X}+z(\alpha/2)\,\sigma/\sqrt{n})$$

[復習] Ⅳ. 母平均の区間推定

教科書p.107例題1

コレステロールの平均値μの区間推定

- 母集団・・・成人男子
  （正規分布はp.106で仮定）
- 標本サイズn＝36（人）
- 標本平均$\bar{X}$＝63（mg/dl）・・・μの点推定値
- 母標準偏差σ＝12（ mg/dl ）

以上の条件のもとで,

「μの95%信頼区間を求めよ」という問題.

（信頼度95%）

［復習］Ⅳ．母平均の区間推定

例題1に関する補足

- 正規母集団の仮定がどこにも明記されていなければ，下記のいずれかの方針で考える．
  - 正規母集団に十分近い分布であると考えて，正規母集団の仮定を導入．
  - n＝36を十分大きいと見なし，標本平均の分布を正規分布で近似（中心極限定理より）．
  - 教科書では13章「Ⅰ」の末尾（p.106）において正規母集団が仮定されている．
    ★ただしカッコ内の説明はおかしいので無視してください．
    「母集団が小さければ正規母集団を仮定できない」は正しいが、
    「母集団が大きければ正規母集団を仮定できる」とは限らない。

- 公式（$\bar{X} - z(\alpha/2)\sigma/\sqrt{n}$　$\bar{X} + z(\alpha/2)\sigma/\sqrt{n}$）
  に値を代入すれば一応答えは出ます．

- 母標準偏差σの値が既知というのは，かなり虫のいい仮定．
  - 現実的にはあまりないケースと思われるが，基本的な原理と手法を理解するためにあえて導入している．

[復習] Ⅳ. 母平均の区間推定

「μの(100×γ)%信頼区間」:

$$\left(\bar{X} - z(\alpha/2)\,\sigma/\sqrt{n} \;,\; \bar{X} + z(\alpha/2)\,\sigma/\sqrt{n} \right)$$

について:

- 標本平均(の実測値)を中心とする区間である.
- $\sigma/\sqrt{n}$ は**標準誤差**と呼ばれる.

[その他の重要な考察]

- γを大きくすると・・・
  - $\alpha = 1-\gamma$ は小さくなる → $z(\alpha/2)$ は大きくなる.
    →信頼区間の幅が大きくなる.
  (外れる確率を減らすのだから、幅を大きく取る必要があるのは当然)
- nを大きくすると → 信頼区間の幅が小さくなる.
  - 情報量が増えるのだから，誤差が減るのは当然

［復習］Ⅳ．母平均の区間推定

「母分散既知の場合」のポイントをまとめると

- 母分散σ^2を用いて標本平均の分布が表せる．
- 母集団分布が正規分布なら標本平均の分布も正規分布．
- 正規母集団を仮定せずとも，標本サイズnが十分大きければ標本平均の分布は正規分布で近似でき，いずれにしても正規分布の問題に帰着できる．
- だがその（標本平均が従う）正規分布$N(\mu, \sigma^2/n)$は，母分散σ^2を用いて表されているのだから，σ^2の値がわからないと推測できない．
- 現実にはσ^2は未知のケースが多い！

→「母分散未知」のケースへ（次回）

S. TOKUNAGA

第13章 推定

Ⅰ. 母集団と標本

Ⅱ. 点推定

- 不偏性，不偏推定量

Ⅲ. 区間推定

Ⅳ. 母平均の区間推定

1. 母分散が既知のとき ←前期ここまで
2. 母分散が未知のとき ←次回ここから

Ⅴ. 母分散の区間推定

Ⅵ. 母比率の区間推定 ←（医のみ）ここを先取り