確率・統計

講義ノート 2020

第1章　基礎概念

1.0.1　確率の基本

同じ状態で繰り返し行なうことができて、その結果が偶然に支配される実験や観察を**試行**という。試行の結果起こる事柄を**事象**という。事象を，$A, B, C, ...$ などで表す。事象 A に対し,「A が起こらない」という事象を A の**余事象**といい，A^c で表す。

また，試行においていくつかの事象のどれが起こることも同程度に期待できるとき，これらの事象は同等に確からしいという。

例 1 白玉 *3* 個、赤玉 *2* 個の袋から球を *1* 個取り出す試行を行う。出る球の組み合わせが事象である。

また、*2* 個の硬貨を同時に投げる試行を行う。*2* 個の硬貨の表裏の組み合わせが事象である。

古典的な確率の定義（ラプラス）

試行において起こりうる場合の数が N 個あって，それらは同等に確からしいとする。起こりうる N 個の場合のうち，ある事象 E が起こる場合の数が r 個あるとき，E の起こる確率 $P(E)$ を $\frac{r}{N}$ で定義する。

例 2 *1* 枚の *10* 円玉を投げる試行において，表が出るという事象を A とする。古典的な確率の定義により

$$P(A) = \frac{1}{2}$$

である。

例 3　１つのさいころを投げる試行において，３または５の目が出るという事象を A とする。古典的な確率の定義により

$$P(A) = \frac{2}{6} = \frac{1}{3}$$

である。

例 4 赤球 *4* 個，白球 *2* 個の入った箱から，*2* 球を同時に取り出し色を見る試行を行う。

（順列）全体で $6 \times 5 = 30$ 通りの場合があり，それらは同等に確からしい。2球とも赤であるという事象を A，赤球と白球1球づつであるという事象を B とする。

A の起こる場合は $4 \times 3 = 12$ 通り，B の起こる場合は $2 \times (4 \times 2) = 16$ 通りあるから，$P(A) = \frac{12}{30} = \frac{2}{5}, P(B) = \frac{16}{30} = \frac{8}{15}$。

*(*組み合わせ*)* 各球は色を除いて互いに区別がつかないとして

$$P(A) = \frac{{}_4C_2}{{}_6C_2}, \quad P(B) = \frac{{}_4C_1 \times {}_2C_1}{{}_6C_2}$$

としてもよい。ただし，ここで ${}_nC_k$ は n 個から k 個とりだす組合わせの数を表す。

組み合わせの公式は区別するものの種類が3種類以上に増えても成立する（**同じものを含む順列**）。n 個のもののうちで，p 個が同じ，q 個が同じ，r 個が同じであるとき，n 個の並べ方の総数は $\frac{n!}{p!q!r!}$ である。もっと一般に、n 個のものが d 種類のグループ（各 $p_1, p_2, ..., p_d$ 個）に分かれているとき、それらの並べ方の総数は

$$\frac{n!}{p_1!p_2!\cdots p_d!}$$

である。

問 1 赤球2個，白球1個の入った箱から次のようにして *2* 個の球を取り出すとき，赤球 *2* 個が取り出される確率を次の各々の場合に求めよ。

（1）1個取り出してもとに戻し，さらに1個取り出す。

（2）1個取り出して，もとに戻さずにまた1個取り出す。

（3）同時に2個の球を取り出す。

問 *2* *(1)*A, A, A, B, C の *5* 文字を一列に並べる並べ方は何通りあるか。

(2) サイコロを *10* 回投げる。*1* の目が *1* 回、*3* の目が *4* 回、*6* の目が *2* 回、残り *3* 回が他の目が出る確率はいくらか。

「同じものを繰り返し取ってよいという約束のもとで」できる順列を**重複順列**という。(「同じものを繰り返し取ってよいという約束」は、通常「重複を許して」という言葉で表現される。) 異なる n 個のものから重複を許して r 個取ってできる順列の総数は n^r である。n^r を ${}_n\Pi_r$ と書くこともある。

また、n 個の異なるものから重複を許して r 個を選んだときにできる組み合わせ(**重複組み合わせ**)の総数は、 ${}_nH_r = {}_{n+r-1}C_r$ である。(*4* 章 *4.0.6* 負の二項分布を参照)

問 *3* *(1)* 1、2、3、4、5、6、7、8、9 の *9* 個の数字から *1* つづつ数字を選び *4* 桁の整数をつくる。千のくらい、百のくらい、十の くらい、一のくらいをそれぞれ a, b, c, d とする。$a < b < c < d$ をみたす整数はいくつあるか。

(2) 1、2、3、4、5、6、7、8、9 の *9* 個の数字から重複を許して *4* 個を選んで *4* 桁の整数をつくる。$a \leq b \leq c \leq d$ をみたす整数はいくつあるか。ただし、*4* つの数字の中に使われない数字があってもよい。

問 *4* りんご，なし，もも　の *3* つのフルーツから，重複を許して *5* 個選ぶ。選び方は何通りあるか求めよ。ただし，一つも選ばれないものがあってもよいとする。

なお、いくつかのものの中から選ぶ場合に、区別のできないものの場合でも、番号などを用いて区別できると考えることができる。また、起こりうる場合が無限個 *($N = +\infty$)* ある場合，それらが同等に確からしいとするとどの場合の確率も 0 である *($\lim_{N\to\infty}(r/N) = 0$)*。これでは使えないので，この定義を採用する場合には起こりうる場合は有限個ということを暗黙のうちに仮定しておく。

このように，古典的な確率の定義では,「N 個の場合の確からしさの同等性」を根拠にして確率を決めている。

A, B を事象とする。「A または B が起こる」という事象を A と B の**和事象**といい，$A \cup B$ とかく。また,「A と B がともに起こる」という事象を A と B の**積事象**といい，$A \cap B$ とかく。A が起こるとき必ず B が起こるならば，$A \subset B$ とかく。このとき，A は B の**部分事象**という。

ある事柄に該当する事象が存在しないとき，その事象を**空事象**といい $\emptyset$ で表す。空事象の余事象を**全事象**といい，Ω で表す。

このとき，次の性質が成り立つ。

（ア）

$$P(A) = 1 - P(A^c)$$

（イ）

$$P(\emptyset) = 0, \quad \text{よって} \quad P(\Omega) = 1$$

（ウ）

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

とくに，$A \cap B = \emptyset$ ならば $P(A \cup B) = P(A) + P(B)$

この場合をとくに**加法定理**という。

（エ）$A \subset B$ ならば $P(A) \leqq P(B)$。

（オ）$P(A) = 1 - P(A^c)$。

なお、ドモルガンの法則より $(A \cap B)^c = A^c \cup B^c$ であるから、

$$P(A^c \cup B^c) = 1 - P(A \cap B)$$

である。

問 5 A, B を事象とする。$A = (B \cap A) \cup (B^c \cap A)$，$(B \cap A) \cap (B^c \cap A) = \emptyset$ であることに注意して，下の問いに答えよ。

（1） $P(A) = 0.3, P(B \cap A) = 0.2$ であるとき，$P(B^c \cap A)$ をもとめよ。

（2） $P(A) = 0.3, P(B \cap A) = 0.2, P(A^c \cap B) = 0.4$ であるとき，$P(A \cup B)$ をもとめよ。

問 6 *(1)* $P(A) \geq \frac{1}{3}, P(B) \geq \frac{1}{4}, P(A \cup B) \leqq \frac{1}{2}$ のとき，$P(A \cap B) \geq \frac{1}{12}$ を示せ。

(2) A, B, C が一般の事象のとき

$$P(A \cap B \cap C) \geq 1 - P(A^c) - P(B^c) - P(C^c)$$

をしめせ。*(*数式で示すこと。*)*

2つの事象 A, B に対し、$A \cap B = \emptyset$ がなりたつとき，A と B は**排反**であるという。また3つの事象 A, B, C が排反とは、$A \cap B \cap C = \emptyset$、かつ、どの2つをとっても排反となることである。このとき

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

である。

例 5 2つのさいころを同時に投げる試行において
$A =$「2つの目の数の和が5である」
という事象と
$B =$「2つの目の数がともに偶数である」
という事象は排反である。

確率の性質 (オ) より、ある事象の確率がわかる（計算できる）ということは、その余事象の確率がわかる（計算できる）ということと同じである。これから「確率が計算できる」事象の族を考えることができる。

標本空間 Ω は一般の抽象集合とする。標本空間 Ω の部分集合の族 $\mathcal{F}$ で次の性質を満たすものを**完全加法族**または σ-**加法族**という。そして、$\mathcal{F}$ の各元を**事象**と呼ぶ。

- $\Omega \in \mathcal{F}$ (Ω は事象である。これを**全事象**という).
- $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ ($\mathcal{F}$ は補集合をとる操作に関して閉じている)
- $A_n \in \mathcal{F}(n = 1, 2, ...) \Rightarrow \cup_{n=1}^{\infty} A_n \in \mathcal{F}$ ($\mathcal{F}$ は可算和をとる操作に関して閉じている).

σ とは「可算的な」という意味である。可算演算に関しても、つぎのド・モルガンの法則が成立する。

$$(\cup_{n=1}^{\infty} A_n)^c = \cap_{n=1}^{\infty} A_n^c, (\cap_{n=1}^{\infty} A_n)^c = \cup_{n=1}^{\infty} A_n^c$$

なお、可算無限個の事象列 $\{A_n\}$ にたいして、その上極限事象、下極限事象をそれぞれ、

$$\limsup_{n\to\infty} A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n, \quad \liminf_{n\to\infty} A_n = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n$$

によって定義する。これらはまた事象である。上極限事象は $\{A_n\}$ **のうち無限個が起こる**という事象を表し、下極限事象は、$\{A_n\}$ **のうちある番号から先のすべてが起こる**という事象を表す。

同じ条件のもとで同じ試行を何回か行い、各回の試行が独立であるとき、このような試行を**反復試行**という。

1個のさいころを n 回続けて投げて、1の目がちょうど r 回出る確率は

$$_nC_r p^r (1-p)^{n-r}$$

である。ただし、$p = 1/6$。

例 6 n 人のグループのうちでちょうど r 人が誕生日を特定の日（例えば *12*月 *31* 日）にもつ確率は，反復事象の確率により

$$p_r = {}_nC_r(\frac{1}{365})^r(\frac{364}{365})^{n-r}$$

である。ただし，*1*年を *365* 日とした。

例 7 相撲で *10* 番勝負（*10* 組出場）があるとし、j 番目の取り組み $(j = 1, ..., 10)$ で東方が勝つ *(*西方が負ける*)* ことを $\omega_j = o$ とかき、東方が負ける *(*西方が勝つ*)* ことを $\omega_j = \times$ とかく。

このとき $(\omega_1, ..., \omega_{10})$ の出方には $(\times, \times, ..., \times)$ から $(o, o, ..., o)$ まですべての出方がある。相撲中継を見ていると、場合によって、勝敗の出方が $(o, o, o, o, o, \times, \times, \times, \times, \times)$ と並ぶ時もあれば $(o, \times, o, \times, o, \times, o, \times, o, \times)$ と並ぶ時もあるが、これらはいずれも上で述べたすべての出方の一つである。

じっさい、$\#\{\omega_j = o\} = 5, \#\{\omega_j = \times\} = 5$ となる確率は（東方、西方の力士の実力が同等であり各回の取り組みが独立におこなわれたととすると）

$${}_{10}C_5\frac{1}{2^{10}}$$

であり、2^{10} 個の中の現れ方の中で $\#\{\omega_j = o\} = 5, \#\{\omega_j = \times\} = 5$ なるものに着目する理由は（先験的には）見当たらない。増して、不思議な力により、o,$\times$ が交互に並ぶ方が o,$\times$ が *5* 個ずつ並ぶ方より出やすい、と信じる理由はない。

1.0.2　確率の基本 2

では,「同等に確からしい」とは何か。つまり，扱っている対象の何と何を同一視し，何と何を区別すればよいか。そして，いかにして「どれが起こることも同程度に期待できる」と考えうるのか。

問 7 次の確率計算は実験にあうか。

（1）明日の天気は，雨が降るか（A），降らないか（A^c）のどちらかである。よって，$P(A) = 1/2$。

（2）水野は矢島を好きか（A），好きでないか（A^c）のどちらかである。よって，$P(A) = 1/2$。

じつは，それは先験的には知りえない。「同等に確からしい」事象は，実験に合うように設定するものなのである。

例 8 N 個の箱に r 個の玉を入れる問題

＜仮定 I＞　1つの玉がどの箱に入るかは同等に確からしく、各々$1/N$である。

次のように書いても同じ：

＜仮定 I'＞　いろいろな玉の入れ方は全部で N^r 通りあるから、これらの書く場合を同等に確からしいと仮定して、各々$1/N^r$ とする。

これは重複順列 ${}_N\Pi_r = N^r$ 個の場合をすべて同等とみなす立場。

この仮定は，箱を空間の小領域、玉を気体の分子と見たときに,「マックスウエル-ボルツマンの統計」とよばれる。しかし、この統計（確率の決め方）では実験にあわなかった（黒体輻射の実験を説明できない)。

＜仮定 II＞　玉は区別がつかない。

区別がつくのは、どの箱に何個づつ玉が入っているか、という様相のみである。したがって、重複組み合わせ（異なる N 種類のものから重複を許して r 個とる組み合わせ）${}_NH_r = {}_{N+r-1}C_r$ 個の場合をすべて同等とみなす立場。言い換えると、

＜仮定 II'＞　玉の盛り分け方は全部で ${}_NH_r$ 通りあり、それらをすべて同等とする。

問 8　例えば、$N = 3, r = 2$ ならば、${}_3H_2 = 6$。これら *6* 通りをすべて書き出せ。

この仮定は,「ボーズ-アインシュタインの統計」とよばれる（光子 (ボーズ粒子）を取り扱うときに用いられる)。この場合実験に当う。

さらに、次の仮定を考える。

＜仮定 III＞　1つの箱に玉は1つしか入らない

とする。この場合には重複を許さない組み合わせ ${}_NC_r$ 通りの場合がある。言い換えると、

＜仮定 III'＞　玉の盛り分け方は全部で ${}_NC_r$ 通りあり、それらをすべて同等とする。

この仮定は、電子や陽子（フェルミ粒子）を取り扱うときに用いられ,「フェルミ-ディラクの統計」とよばれる（この場合実験に当てはまる)。(物理では＜仮定 III＞は,「パウリの排他原理」(異なる粒子は同時に同一状態を取ることはない）に対応する。)

問 9　$N = 4, r = 3$ とする。仮定 I – III のもとで同等に確からしい場合を、各々すべて書き出せ。

上の仮定 II, III を現実離れしたものと思ってはいけない。

ある種の色盲の人は赤と緑の区別がつかない。青，赤，緑の 3 種類のランプ（発光ダイオード）を暗闇で等頻度で発光させるとする。この色盲の人には，赤のランプと緑のランプは区別がつかず，事象 $A =$「青が発光する」，事象 $B =$「赤（$=$ 緑）が発光する」，の 2 つの事象に対して，$P(A \cup B) = 1, P(A) = 1/3, P(B) = 2/3$ と感じるであろう。

また，モンシロチョウは紫外線が見えるが，ヒトには紫外線が見えない。紫外線の発光によってある事象 A の発生が伝えられたとき，モンシロチョウには $P(A) > 0$ だとしても，ヒトには $P(A) = 0$ でなければ正しい確率と思われない（ヒトの目から見た実験に合わない）。

このように，何をもって「同等に確からしい」とするかは先見的には決まらない。確率は，光や熱のように物理量として実在するものではなく，我々の脳の中で「実験に合うように」設定されるものなのである。

問　題

1 2つのさいころを同時に投げる試行を T とし，出た目の数の和を X とする。試行 T を 10 回反復して行ったとき，$X = 7$ となるのが 9 回である確率をもとめよ。

2 A さん, B さん, C さんの 3 人がじゃんけんをして勝者を 1 人選ぶ。3 人あいこならばじゃんけんを繰り返し，2 人勝ちならば勝った 2 人で決戦をするものとする。次の確率をもとめよ。

（1）A さんが 1 回目で優勝する確率

（2）A さんが 2 回目で優勝する確率

（3）A さんが 3 回目で優勝する確率

（4）3 回目が終わっても勝者が決まらない確率

3 n を 2 以上の整数とする。中の見えない袋に $2n$ 個の玉が入っていて，そのうち 3 個が赤で残りが白とする。A さんと B さんが，A さんから始めて交互に 1 個ずつ玉を取り出し，先に赤の玉を取り出したほうが勝ちとする。ただし，取り出した玉は袋に戻さないとする。B さんが勝つ確率をもとめよ。

4 「幸福な家庭はみな同じように似ているが，不幸な家庭は不幸なさまもそれぞれ違うものだ」（「アンナ・カレーニナ」トルストイ）という命題について，確率論の観点から解説せよ。

5 n を 4 以上の整数とし、$1, 2, ..., n$ の数字がそれぞれ 1 つづつ書かれた n 枚のカードが箱に入っている。この箱から 4 枚のカードを同時に取り出

し、それらを無作為に横 1 列に並べる。このときカードの数字が小さい順に並んでいる確率を求めよ。

第2章　条件つき確率

2.0.3　条件つき確率

第1節で述べたように，確率 P は次の2つの性質を満たす。

(1)　$P(\emptyset) = 0, \quad P(\Omega) = 1, \quad 0 \leq P(A) \leq 1$

(2) $A \cap B = \emptyset$ ならば,

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

近代的な（大学以上での）確率の定義では，上の2つを満たす P を**確率**とよぶ。この場合，試行の結果起こりうる各々の場合は等確率と仮定しなくてもよいし[1]，確率の値は有理数でなくてもよい。有理数でない確率の例は3章末で示す。(3章末の [発展] を参照。)

なお、無限個の事象を扱う場合には、次の2つの条件を確率の定義とする：

σ-加法族 $\mathcal{F}$ の各元 A に対して、ある実数 $P(A)$ で $0 \leq P(A) \leq 1$ をみたすものが定まってつぎの性質を満たすとき、$P(A)$ を事象 A の**確率**という。

$$(1)\ P(\Omega) = 1(\text{全事象の確率は } 1).$$

$$(2)\ A_n \in \mathcal{F}, A_n \cap A_m = \emptyset (n \neq m) \Rightarrow P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)\ (\text{確率の完全加法性の公理}).$$

A, B を事象とする。A を前提として B の起こる確率のことを、条件 A のもとでの B の**条件つき確率**といい、$P_A(B)$ または $P(B|A)$ で表す。

---

[1]実際，花びらを使った恋占いでは，試行の結果には「愛してる」と「愛してない」の2つの場合があり，花びらの数の奇数・偶数に「愛してる」と「愛してない」を対応させている。偶数と奇数は交互に並んでいるから，これはこれら2つの事象がほぼ同等に確からしいことを（暗黙のうちに）仮定していることになる。この仮定を常に用いるのは非現実的であろう。

条件つき確率の数学的定義

（1）$P(A) > 0$ の場合

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

である。

（2）$P(A) = 0$ の場合

$$P(B|A) = 0$$

である[2] 。

例 9 $A =$「ある夜月にカサがかかる」, $B =$「その翌日雨が降る」，とする。ただし，ここでカサとは月の周囲にできる光の輪のことである *(*暈（ハロ）*)*。

このとき,

$P(A \cap B) =$ 夜月にカサがかかり，かつ，翌日雨が降る確率

であるから,

$P(A \cap B) =$ （夜月にカサがかかる確率）$\times$ （前夜月にカサがかかったとき、翌日雨が降る確率）

$= P(A).P(B|A)$。

よって，$P(A) > 0$ のとき，$P(B|A) = \frac{P(A \cap B)}{P(A)}$ である。

これより導かれる

$$P(A \cap B) = P(A).P(B|A)$$

を、確率の**乗法定理**という。

なお、3 つの事象 A, B, C について，上の定義で A の代りに $B \cap C$ をとり，B の代りに A をとれば,

$$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$$

であることがわかる。

例題 1 ある町には *2*つの宝くじ売り場があり，当たりくじの出る確率は各々*0.3, 0.1*である。ある日，花子はこのうちどちらか一方の売り場を選

[2] 0 の代わりに $[0, 1]$ 内の他の値を採用する場合もある。

んでそこで宝くじを買ったところ，はずれであった。それを聞いた太郎は，他方の売り場に宝くじを買いに行った。

花子がはずれたとき，太郎が宝くじに当たる条件つき確率を求めよ。また，花子がはずれたとき，太郎が花子と同一の売り場に宝くじを買いに行った場合，太郎が宝くじに当たる条件つき確率を求めよ。

解 1 花子がはずれるという事象を A，太郎が当たるという事象を B とする。$P(A\cap B)=P(A)P(B|A)$ であり，$P(A)=(1/2)\times 0.7+(1/2)\times 0.9=0.8$。はじめの場合，$P(A\cap B)=(1/2)\times 0.7\times 0.1+(1/2)\times 0.9\times 0.3=0.17$。よって，$P(B|A)=\frac{0.17}{0.8}=\frac{17}{80}$。後の場合，$P(A\cap B)=(1/2)\times 0.7\times 0.3+(1/2)\times 0.9\times 0.1=0.15$。よって，$P(B|A)=\frac{0.15}{0.8}=\frac{15}{80}$。つまり，前者のほうが太郎にとって有利な戦略である。

なお，花子の当たりはずれの情報がはいらない場合には，太郎が宝くじに当たる確率は上の 2 つの値の中間値 $\frac{16}{80}=\frac{1}{5}$ $(\frac{15}{80}<\frac{16}{80}=\frac{1}{5}<\frac{17}{80})$ になる。なぜならば、

$$P(A\cap B)=\frac{1}{2}\times 0.17+\frac{1}{2}\times 0.15=\frac{1}{2}\times 0.32.$$

よって、

$$P(B|A)=\frac{1}{0.8}\times\frac{1}{2}\times 0.32=\frac{16}{80}=\frac{1}{5}=\frac{1}{2}\times 0.3+\frac{1}{2}\times 0.1=P(B).$$

このように、一般には A(先に起こった事象) は B(後に起こった事象) に影響する。これを**因果律**という。

例題 2 旅館「道後館」には３つの部屋があり，中に各々女性２人，男性２人，男女各１人が入っている。１つの部屋をノックしたところ，女性の声で「誰か来たわよ。あなた出てちょうだい」と聞こえた。男性が出てくる確率はいくらか。

解 2 男性２人の部屋を A，女性２人の部屋を B，男女各１人の部屋を C とする。さらに，3 人の女性をそれぞれ O_1,O_2,O_3 で表す。女性が返事をするという事象を F とし，C の部屋の女性が返事をするという事象を G とする。以下では，例えば，O_1 が B の部屋にいる状態を (B,O_1) とかく。このとき

$$F=\{(B,O_1),(B,O_2),(C,O_3)\}$$

$$G=\{(C,O_3)\}$$

どの部屋をノックし，部屋の中の 2 人のうちどちらが返事をするかは同等に確からしいから，$P(F) = 1/2$，$P(G) = 1/6$。これより求める条件つき確率 $P(G|F)$ は

$$P(G|F) = \frac{P(F \cap G)}{P(F)} = \frac{P(G)}{P(F)} = \frac{1/6}{1/2} = \frac{1}{3}$$

となる。

なお、「(あることが) 起こらなかった」も事象である。

原因の確率（ベイズ (Bayes') の定理）
(簡単な場合)

例題 3 ある病気に感染しているかどうかを判定する検査がある。*100* 人に *1* 人が、この病気に感染している。感染者はこの検査によって *99*% の確率で陽性だと判定できる。また、感染していない人が陽性だと判定されてしまう確率が *2*% ある。ある人がこの検査を受けて陽性だと判断されたとき、この人がこの病気に感染している確率を求めよ。

解 3 事象 A が「検査結果が陽性である」であり、事象 B が「感染している」とする。

$$P(A \cap B) = P(B)P(A|B) = \frac{1}{100} \times \frac{99}{100} = \frac{99}{10000}$$

よって

$$P(A) = P(A \cap B) + P(A \cap B^c) = \frac{99}{10000} + \frac{99}{100} \times \frac{2}{100} = \frac{297}{10000}$$

これより

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{99}{10000}}{\frac{297}{10000}} = \frac{99}{297} = \frac{1}{3}$$

(つまり $P(B^c|A) = \frac{2}{3}$ *であった。)*

このような検査をふるい分け検査という。10000 人の検査をすると仮定する。この中で 100 人は病気にかかっており、9900 人は病気にかかっていない。病気にかかっている 100 人のうち陽性と出る人は

$$10^4 \times P(B|A)P(A) = 10^4 \times \frac{1}{3}.\frac{297}{10000} = 99$$

人であり、陰性と出る人は

$$10^4 \times P(B|A^c)P(A^c) = 10^4 \times \frac{P(B \cap A^c)}{P(A^c)} P(A^c) = 10^4 \times P(B)P(A^c|B) = 10^4 \times \frac{1}{100} . \frac{1}{100} = 1$$

人である。一方、病気にかかっていない 9900 人のうち陽性と出る人は

$$10^4 \times P(B^c \cap A) = 10^4 \times P(B^c|A)P(A) = 10^4 \times \frac{2}{3} . \frac{297}{10000} = 198$$

人であり、陰性と出る人は 9702 人である。

このように 10000 人の被験者の内で病気にかかっていそうな人は $99 + 198 = 297$ に絞られたことになる。ふるい分け検査の目的は被験者の数を膨大な数からより少ない数にふるいにかけて減らすことである。

問 10 １枚の硬貨を３回投げ，表が出た回数を X とする。次にさいころを X 回投げる。そうして，１または２の目が出た回数を Y とする。ただし，$X = 0$ の場合には，$Y = 0$ ときめる。$Y = 0$ という条件のもとで，$X = 2$ である条件つき確率をもとめよ。

問 11 ある病気 B の患者３００人のうち２００人は喫煙者である。また全人口のうち喫煙者の割合は３０％である。喫煙者は非喫煙者に比べてどのくらい病気 B にかかりやすいか。

(一般の場合)

互いに素かつ網羅的 ($i.e.\ B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$) な事象 $B_1, ..., B_n$ が起こった時に A という事象が起こる確率を考える。$P(B_j)$ と $P(A|B_j)$ がわかっているとする。このときに、条件つき確率 $P(B_j|A)$ を計算したい。

$$P(A \cap B_j) = P(A|B_j)P(B_j) = P(B_j|A)P(A)$$

だから

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}$$

となる。B_j は互いに素であるから $P(A) = P(A \cap B_1) + \cdots P(A \cap B_n) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)$ であるから

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)}$$

と書ける。これを **ベイズの公式** という。

問 n に関する帰納法により、これを証明せよ。

2.0.4　事象の独立の定義

2つの事象 A, B が

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

を満たすとき、A と B は**独立**であるという。

つまり，A と B は独立であるとは，$P(A) > 0$ のとき

$$P(B|A) = P(B)$$

が成り立つことである。

２つの事象は独立でないとき**従属**という。

問 事象 A が自分自身と独立になるとき $P(A) = 0$ or 1 であることを示せ。

例題 4 *10* 本のうち *3* 本が当たるくじがある。*A, B, C* の３人がこの順にこのくじを引くとき，それぞれの人が当たる確率をもとめよ。また,「A が当たる」という事象と「B が当たる」という事象は独立か。

解 4 当たる確率は３人とも $\frac{3}{10}$ である。なぜならば，A については明らか。B の当たる確率は

$$\frac{3}{10} \times \frac{2}{9} + \frac{7}{10} \times \frac{3}{9} = \frac{27}{90} = \frac{3}{10}$$

同様に考えて，C の当たる確率は

$$\{3 \times 2 \times 1 + 3 \times 7 \times 2 + 7 \times 3 \times 2 + 7 \times 6 \times 3\} \times \frac{1}{10 \times 9 \times 8} = \frac{216}{720} = \frac{3}{10}$$

また

$$P(「A \text{ が当たる}」 \cap 「B \text{ が当たる}」) = \frac{3}{10} \times \frac{2}{9} = \frac{1}{15}$$

$$P(「A \text{ が当たる}」) \times P(「B \text{ が当たる}」) = \frac{3}{10} \times \frac{3}{10} = \frac{9}{100}$$

であるから，２つの事象は独立でない。

事象の独立性は事象間の意味的なつながりによって定義されるのではなく，確率 P によって定義されることに注意せよ。複雑な事象を考える場合，事象間の意味的なつながりにより独立性を判断するのは困難な場合が多い（下の例題参照)。

例題 5 ある家族が n 人の子を持つ場合，子供には男女それぞれの場合がある。例えば，$n=2$ ならば，子供の可能性には $\{(男,女),(女,男),(男,男),(女,女)\}$ の 4 通りあり，それらは同等に確からしいと考えられる。n を固定し，A を「n 人の子供は男女両児からなる」という事象，B を「n 人の子供はたかだか一人の女児しか含まない」という事象とする。A と B は独立か。

解 5 （1）$n=2$ の場合。起こりうる場合は上の 4 つであるから，$P(A)=\frac{2}{4}=\frac{1}{2}, P(B)=\frac{3}{4}, P(A\cap B)=\frac{2}{4}=\frac{1}{2}$。よって，$A$ と B は独立でない。

（2）$n=3$ の場合。起こりうる場合は

$$\{(男,女,男),(男,女,女),(女,男,男),(女,男,女),(男,男,男),$$

$$(男,男,女),(女,女,男),(女,女,女)\}$$

の 8 通り。したがって，$P(A)=\frac{6}{8}=\frac{3}{4}, P(B)=\frac{4}{8}=\frac{1}{2}, P(A\cap B)=\frac{3}{8}$。よって，$A$ と B は独立である。

問 12 （独立性を直感で判断してはいけない。）

区別のできる 2 つのさいころを同時に投げる試行を行う。
はじめのさいころの目が４であると言う事象を A、
2 つめのさいころの目が２であると言う事象を B、
2 つのさいころの目の和が３であると言う事象を C、
2 つのさいころの目の和が９であると言う事象を D、
2 つのさいころの目の和が７であると言う事象を E とおく。
（1）A と B は独立か。
（2）A と C は独立か。
（3）A と D は独立か。
（4）A と E は独立か。

3 つ以上の事象の独立性

(a) ３つの事象 A,B,C が独立とは、どの２つの事象も独立であり、かつ、

$$P(A'\cap B'\cap C')=P(A')P(B')P(C')$$

となるときを言う。ただし、$A' = A$ or A^c, $B' = B$ or B^c。

(b) n 個の事象 $A_1, ..., A_n$ が独立であるとは、これらの任意の k 個 $(k \leq n)$ $A_{j_1}, ..., A_{j_k}$ に対して

$$P(A_{j_1} \cap A_{j_2} \cap ... \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k}) \qquad (*)$$

が成り立つことである。

例 n 個の事象の独立性では、上の $(*)$ の $k = n$ のときのみの成立では不十分である。たとえば、$\Omega = \{1,2,3,4,5,6\}, A_1 = \{1,2,3,4\}, A_2 = A_3 = \{4,5,6\}$ とすると、

$$P(A_1 \cap A_2 \cap A_3) = P(\{4\}) = \frac{1}{6},$$

$$P(A_1)P(A_2)P(A_3) = \frac{4}{6} \cdot \frac{3}{6} \cdot \frac{3}{6} = \frac{1}{6}$$

となるので、$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ は成り立つ。

一方、$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(\{4\}) = \frac{1}{6} \neq P(A_1)P(A_2) = P(A_1)P(A_3) = \frac{4}{6} \cdot \frac{3}{6} = \frac{2}{6} = \frac{1}{3}$, $P(A_2 \cap A_3) = \frac{3}{6} = \frac{1}{2} \neq P(A_2)P(A_3) = \frac{3}{6} \cdot \frac{3}{6} = \frac{1}{4}$ となって、A_1, A_2, A_3 のどの 2 つも独立にならない。

問　題

1 ある事件で，本当のことを言う確率が 80 %である証人 X_1, X_2, X_3 がいる。いま 3 人とも「Y が犯人だ」と証言した。本当に Y が犯人である確率はいくらか。ただし，Y は上の 3 人とは別人である。

2 鳥インフルエンザ検査の正確さが 98% だとする。つまり、鳥インフルエンザにかかっている鶏がこの検査を受けた場合に陽性と出る確率が 98% であり、鳥インフルエンザにかかっていない鶏がこの検査を受けた場合 98% の確率で陰性と出る。さらに、実際に鳥インフルエンザにかかっている鶏の割合は 0.5% だとする。

ある鶏がこの検査を受けたところ、結果は陽性であった。この鶏が鳥インフルエンザにかかっている確率はいくらか。

3 ある製品を製造する工場 A, B があり，A 工場の製品には 5 %，B 工場の製品には 3 %の不良品が含まれている。A 工場の製品と B 工場の製品を 2：3 の比で混ぜた中から 1 個を取り出すとき

（1）それが不良品である確率を求めよ。
（2）不良品であったとき，それがA工場の製品である確率を求めよ。

4 各問が3つの選択肢からなる小問が10問ある。この選択肢のうち1つが正解で他の2つは不正解である。受験者は理解している問いには正しい選択肢を選び，理解していないものについてはでたらめに（すなわち各選択肢を等確率で）選んで解答する。

ある受験者がちょうど7問正解したとき,「この受験者はどの問にもでたらめに回答した」という判断が正しい確率を次のようにしてもとめよう。

（1）A を「受験者がちょうど7問正解した」という事象とし，$B_i, i = 0, ..., 10$ を「この受験者がちょうど i 問正解を知っていた」という事象とする。$P(A|B_i)$ をもとめよ。ただし，$P(A|B_8) = P(A|B_9) = P(A|B_{10}) = 0$ とする。

（2）$P(B_0) = \cdots = P(B_7)$ という仮定のもとで，問題の確率を

$$P(B_0|A) = \frac{P(B_0 \cap A)}{P(A)} = \frac{P(B_0).P(A|B_0)}{\sum_{i=0}^{7} P(B_i).P(A|B_i)}$$

により計算せよ。

5 E君は一目ぼれの彼女に熱烈なメイルを出したが、ついに返事は来なかった。ただし出した先は私信のメールもチェックするので悪名高い女子寮で、検閲に引っかかって彼女のもとに渡らない確率が0.3, 彼女がそれを見ても一笑に付してメイルを削除する確率が0.5, 見て好意を抱いてくれても羞恥心から返事を書かない確率が0.7である。

1週間たっても返事はこなかった。E君にはどれくらいの確率で望みが残されているか。

第3章　平均と分散

3.0.5　確率分布

試行の結果によってその値が定まる変数を**確率変数**という。確率変数 X のとる値が $x_1, x_2, ...$ であるとき、$X = x_i$ となる確率 $P(X = x_i)$ を p_i と表わす。確率変数と確率の組を確率分布という。

確率変数 X がとびとびの値のみをとるとき、X は**離散確率変数**とよび、その分布を**離散確率分布**という。

確率変数 X が特定の値を正の確率でとることがないとき，すなわち，X の値が区間全体に亘り、すべての $x \in \mathbf{R}$ について $P(X = x) = 0$ となるとき、、X は**連続確率変数**とよび、その分布を**連続確率分布**という。

期待値、分散の定義

(1) 離散分布の場合

確率変数 X のとる値が $\{x_1, x_2, ..., x_n\}$ であるとき、

$$m = \sum_{i=1}^{n} x_i P(X = x_i),$$

$$v = \sum_{i=1}^{n} (x_i - m)^2 P(X = x_i)$$

を、各々X の**分布の平均、分散**という。誤解が生じないときには，たんに X の**平均，分散**という。m を $E[X]$、v を $V(X)$ と書くこともある。また，平均のことを**期待値**ともいう。X のとる値 x_i とその確率 p_i の組 $((x_i, p_i); i = 1, ..., n)$ を**確率分布**という。

これを表にすると次のようになる：

X	x_1	x_2	$\cdots$	x_n	計
P	p_1	p_2	$\cdots$	p_n	1

（2）連続分布の場合

この場合、$\rho(x) \geq 0, \int_{-\infty}^{\infty} \rho(x)dx = 1$ を満たす関数 $\rho(x)$ が存在して[1],

$$P(a \leq X \leq b) = \int_a^b \rho(x)dx$$

と表される。 $\rho(x)$ を **(確率) 密度関数**という。

$$\rho(x) = \frac{dP}{dx}(-\infty < X \leq x)$$

と書いてもよい。

$$m = \int_{-\infty}^{\infty} x\rho(x)dx,$$
$$v = \int_{-\infty}^{\infty} (x-m)^2\rho(x)dx$$

を，各々X の **(分布の) 平均，分散**という。m を $E[X]$，v を $V(X)$ と書くこともある。また，平均のことを**期待値**ともいう。密度関数 $\rho(x)$ と X のとる値の範囲 (たとえば (a,b)) の組 $(\rho(x); x \in (a,b))$ を**確率分布**という。

なお，離散分布と連続分布が組み合わさった分布を**混合分布**という。

平均は X のとる値の代表値である。また，分散は X の値の平均からのずれの期待値であり，X の分布のばらつきを表す。なお，離散分布，連続分布いずれの場合でも，

$$\sigma(X) = \sqrt{V(X)}$$

を，X の（分布の）**標準偏差**という。また

$$F(x) = P(X \leq x),\ x \in \mathbf{R}$$

を X の**分布関数**という。

期待値の意味

期待値は公平な賭けを行う場合にその公平さの基準となる値である。ある賭けで、もし賭けが当たりならば 10000 円もらえ、はずれならば 0 円もらえるとする。この賭けの参加料を a 円とすると、$a = E[X]$（得点の期待値）のときこの賭けは双方（参加側、受け側）にとって公平である。

[1]実数全体で定義された関数 $f(x)$ に対して $\int_{-\infty}^{\infty} f(x)dx$ は $f(x)$ の広義積分を表す。

このとき、$a < E[X]$ ならば参加側にとって（平均的に）有利であり、$a > E[X]$ ならば参加側にとって（平均的に）不利である。同じことは双方が商売（売り方、買い方）をするときにも成り立つ。つまり、$a = E[X]$（円）で売買が成立する。

実際の例として、次のような賭けを考える。

参加料 100 円を支払えば、6 面サイコロ 1 個を 1 回投げることができる。サイコロの目に応じて、次の金額を貰える。1：20 円, 2：50 円, 3：100 円, 4：100 円, 5：150 円, 6：150 円

このとき、もらえる金額の期待値を求めると、

$$E = 20 \times \frac{1}{6} + 50 \times \frac{1}{6} + 100 \times \frac{1}{6} + 100 \times \frac{1}{6} + 150 \times \frac{1}{6} + 150 \times \frac{1}{6} = 95$$

(円) である。

得られる金額の期待値 95 円が参加費 100 円を下回ることから、この賭けは参加者が得をする可能性もあるものの平均的には損をするということが分かる。特に回数を増やすほど、試行ごとに 5 円の損をした状態に限りなく近づく。

更に実際の例として、ジャンボ宝くじの例をとると、参加費（一口の値段）300 円であり、得られる金額の期待値は約 140 円である。

離散分布の場合

例題 6 さいころを 2 回投げて，出る目の数の和を X とする。X の分布の分散 $V(X)$ を求めよ。

解 6 X のとりうる値は，*2, 3, ..., 12* である。起こりうる *36* 通りのうち，X の値が，n になる場合は
$n-1$ 通り, ($n = 2, 3, ..., 6$ のとき)
$13-n$ 通り, ($n = 7, ..., 12$ のとき)
あるから *(*次頁表参照*)*,

$$E[X] = \sum_{n=2}^{6} n\frac{n-1}{36} + \sum_{n=7}^{12} n\frac{13-n}{36} = \frac{252}{36} = 7$$

よって,

$$V(X) = \sum_{n=2}^{6} (n-7)^2\frac{n-1}{36} + \sum_{n=7}^{12} (n-7)^2\frac{13-n}{36} = \frac{35}{6}$$

例題 7　2つの特製さいころがあり，*1*つ目のさいころの各面には *1, 3, 4, 5, 6, 8* という目が書かれており，*2*つ目のさいころの各面には *1, 2, 2, 3, 3, 4* という目が書かれている。この2つの特製さいころを同時に投げて出る目の数の和をYとする。Yの分布の分散$V(Y)$を求めよ。ただし，これら特製さいころの各面の出方は同等に確からしいとする。

表 3.1: 例題6

X/Y	1 2 3	4 5 6
1	2 3 4	5 6 7
2	3 4 5	6 7 8
3	4 5 6	7 8 9
4	5 6 7	8 9 10
5	6 7 8	9 10 11
6	7 8 9	10 11 12

表 3.2: 例題7

X/Y	1 3 4	5 6 8
1	2 4 5	6 7 9
2	3 5 6	7 8 10
2	3 5 6	7 8 10
3	4 6 7	8 9 11
3	4 6 7	8 9 11
4	5 7 8	9 10 12

解 7 Yのとりうる値は，*2, 3, ..., 12*である。起こりうる*36*通りのうち，Yの値がnになる場合は，
$n-1$通り，($n = 2, 3, ..., 6$のとき)
$13-n$通り，($n = 7, ..., 12$のとき)
ある。したがって，Yの分布はXの分布と同じであり，

$$E[Y] = \frac{252}{36} = 7,$$

$$V(Y) = \frac{35}{6}$$

問 13 例題６，例題７における起こりうる場合を，各々すべて書き出せ。

次の例のように、期待値は人の行動の適否を決める手がかりをあたえる。

例題 8 さいころを１回または２回投げ，最後に出た目の数を得点とするゲームを考える。１回投げて出た目を見た上で，２回目を投げるか否かを決めるのであるが，どのように決めるのが有利であるか。

また，３回投げることも許されるとしたら，２回目，３回目を投げるか否かの決定はどのようにするのが有利か。

ただし、新たにさいころを投げると直近の得点は無効になる。

解 8 有利であるかどうかは得点の期待値を計算して比較することにより判断する。

まず，最大２回まで投げることができる場合。２回目に出る目の数の期待値は $\frac{7}{2} = 3.5$ であるから，１回目に出た目の数が３以下ならば２回目を投げるほうが有利であり，そうでなければ２回目を投げないほうが有利である。

次に，最大３回まで投げることができる場合。最大２回投げて，２回目まで投げるルールを上のようにして行動すると，その場合の得点の期待値は

$$\frac{3}{6} \times 3.5 + \frac{1}{6}(4 + 5 + 6) = 4.25$$

である。よって，もし１回目の目の数が４以下ならば，２回目以降を続行するほうが有利である。もし１回目の目の数が５以上ならば，２回目を投げないほうが有利である。

１回目の目の数が４以下の場合，２回目、３回目をどうするかは上の場合と同様に考えて，もし２回目の目の数が３以下なら３回目を投げ，そうでなければ３回目を投げないほうが有利である。

問 さいころを投げて出た目の数だけ得点がもらえるゲームがある。ただし、出た目が気に入らなければ１回だけ投げなおすことができる。このゲームでもらえる得点の期待値が最大になるように振舞ったとき、その期待値を求めよ。

生保（生命保険）への応用

例 10 x 歳で加入，n 年契約，死亡保険金 C 円の生命保険（掛け捨て）を考え、その価格（保険料）を求める。被保険者の余命を X で表す。

年利率を i で表す。つまり，A 円を *1* 年間預けると $(1+i)A$ 円になる。A 円を j 年間預けると $(1+i)^j A$ 円になる。$v = \frac{1}{1+i}$ を**現価率**という。j 年後の A' 円の現在（契約時）価値は $v^j A'$ 円である。

この契約についての保険会社側の支払い金額の現在価値を Z とする。$Z = Cv^j,\ (j-1 \le X < j),\ j = 1, ..., n$ である。この保険の保険料を B 円とすると，$B = (Z$ の期待値$)$ であるから

$$B = E[Z] = C\sum_{j=1}^{n} v^j P(j-1 \le X < j)$$

となる。したがって，X の分布がわかれば B を求めることができる。

なお、$B = (Z$ の期待値$)$ なることを**収支相等の原則**という。これが生命保険商品の値決めの基本である。

連続分布の場合

Ω を定義域とし、$\mathbf{R}$ を値域とする関数 $X(\omega)$ が任意の実数 a, b にたいして

$$X^{-1}((a,b]) \text{ が可測（=測れる）}$$

を満たすとき（どんな $a < b$ に対しても, 事象 $\{a < X \le b\}$ が測れるとき）、X を**確率変数**という。上の左辺の集合は関数 X による $(a,b] \subset \mathbf{R}$ の逆像を表し、それはまた $\{X \in (a,b]\}$ など略記される。

離散確率空間の場合には、確率変数は Ω の上の任意の関数であると定義すればよかったが、事象の集合 $\mathcal{F}$ が Ω の部分集合全体に必ずしも等しくない場合には、その定義に上のような制限を設ける必要があるのである。

標本空間 (X の像空間) が $\mathbf{R}$ に等しいとき、部分集合族としては $\mathbf{R}$ の部分集合の全体を考えるのではなく、ボレル集合と呼ばれる特別な部分集合の全体（これを**ボレル集合族**といい $\mathcal{B}$ で表す）を採用するのが標準的である。$\mathcal{B}$ はつぎの条件で特徴づけられる。

(1) $[a,b)$ のタイプの区間はボレル集合である。すなわち $[a,b) \subset \mathcal{B}$.

(2) $\mathcal{B}$ は補集合と可算和をとる操作に関して閉じている。

(3) $\mathcal{B}$ は上の (1),(2) の性質をもつ集合族のうち最小のものである。

$\mathbf{R}$ の部分集合でボレル集合でないものが存在することが知られている。しかし、常識的に考えられるほとんどの集合はボレル集合である。たとえば、あらゆるタイプの１次元区間はボレル集合であり、１点からなる集合やその加算和である有理数の集合、その補集合である無理数の集合などはみなボレル集合である。

確率変数の分布や結合分布という概念を導入することができる。X を確率変数とすると、任意のボレル集合 B に対してその X の逆像の確率 $P(X^{-1}(B))$ が定義可能となる。それを $\mu(B)$ とおく。μ は $(\mathbf{R}, \mathcal{B})$ 上の確率である。これを確率変数 X の**確率分布**といい、μ_X で表す。

任意の $x \in \mathbf{R}$ に対して

$$F_X(x) = P(\{X(\omega) \leq x\})$$

を、確率変数 X の**分布関数**という。$F_X(x)$ は x に関して単調増加な関数になる。

$F_X(x)$ が、ある非負積分可能関数 $\varphi(x)$ を用いて

$$F_X(x) = \int_{-\infty}^{x} \varphi(t)dt$$

と表現されるとき、$\varphi(x)$ を**確率密度関数**という。

例 11 確率変数 X のとる値 x の範囲が $0 \leq x \leq 2$ で，その密度関数 $\rho(x)$ が

$$\rho(x) = \frac{1}{2} \quad (0 \leq x \leq 2)$$

で与えられるとき,

$$P(0 \leq X \leq 2) = 1, \quad P(0 \leq X \leq \frac{3}{2}) = \int_0^{3/2} \frac{1}{2}dx = \frac{3}{4}$$

$$E[X] = \int_0^2 x\frac{1}{2}dx = 1, \quad V(X) = \int_0^2 (x-1)^2\frac{1}{2}dx = 1/3$$

問 14 確率変数 X のとる値 x の範囲が $1 \leq x \leq 3$ で，その密度関数 $\rho(x)$ が

$$\rho(x) = -\frac{3}{4}x^2 + 3x - \frac{9}{4} \quad (1 \leq x \leq 3)$$

で与えられるとき，X の平均と分散をもとめよ。

問 15 確率密度関数 $\rho(x)$ がつぎで与えられる確率分布にしたがう確率変数の期待値（平均），分散，標準偏差をもとめよ。

（1）$\rho(x) = \frac{1}{\beta-\alpha}$ $(\alpha \leq x \leq \beta)$

（2）$\rho(x) = \frac{3}{4}(1-x^2)$ $(-1 \leq x \leq 1)$

平均と分散の性質

X, Y を確率変数とする。X, Y の分布が離散であっても連続であっても次の性質が成り立つ。

$$E[X+Y] = E[X] + E[Y]$$

$$E[cX] = cE[X], \quad c \text{ は定数}$$

これらより，定数 a, b に対して，$E[aX+bY] = aE[X]+bE[Y]$, $E[aX+b] = aE[X]+b$ であることがわかる。また

$$V(aX+b) = a^2V(X), \quad a, b \text{ は定数}$$

がなりたつ。

証明（離散の場合）

簡単のため $X = x_1, x_2, Y = y_1, y_2, y_3$ のときのみ考える。

X/Y	y_1 y_2 y_3	計
x_1	r_{11} r_{12} r_{13}	p_1
x_2	r_{21} r_{22} r_{23}	p_2
計	q_1 q_2 q_3	1

このとき

$$E[X+Y] = (x_1+y_1)r_{11}+(x_1+y_2)r_{12}+(x_1+y_3)r_{13}+(x_2+y_1)r_{21}+(x_2+y_2)r_{22}+(x_2+y_3)r_{23}$$

$$= x_1(r_{11}+r_{12}+r_{13})+x_2(r_{21}+r_{22}+r_{23})+y_1(r_{11}+r_{21})+y_2(r_{12}+r_{22})+y_3(r_{13}+r_{23})$$

$$= x_1p_1 + x_2p_2 + y_1q_1 + y_2q_2 + y_3q_3 = E[X] + E[Y]$$

となる。証明終

問　同じ設定で $V(aX+b)=a^2V(X),\ \ a,b$ は定数 を示せ。

X,Y を離散分布をもつ確率変数とする。X のとる任意の値 a と Y のとる任意の値 b について

$$P(X=a,Y=b)=P(X=a)P(Y=b)$$

が成り立つとき，X **と** Y **は（互いに）独立**であるという。

X,Y を連続分布をもつ確率変数とする。X のとる任意の値 $a_1<b_1$ と Y のとる任意の値 $a_2<b_2$ について

$$P(a_1<X<b_1,a_2<Y<b_2)=P(a_1<X<b_1)P(a_2<Y<b_2)$$

が成り立つとき，X **と** Y **は（互いに）独立**であるという。

X と Y が（互いに）独立であるとき次の性質が成り立つ。

$$E[XY]=E[X]E[Y]$$

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

証明（離散の場合；前ページと同じ設定とする）

X/Y	$y_1\ y_2\ y_3$	計
x_1	$p_1q_1\ p_1q_2\ p_1q_3$	p_1
x_2	$p_2q_1\ p_2q_2\ p_2q_3$	p_2
計	$q_1\ q_2\ q_3$	1

このとき

$$E[XY]=(x_1y_1)p_1q_1+(x_1y_2)p_1q_2+(x_1y_3)p_1q_3+(x_2y_1)p_2q_1+(x_2y_2)p_2q_2+(x_2y_3)p_2q_3$$

$$=x_1p_1(y_1q_1+y_2q_2+y_3q_3)+x_2p_2(y_1q_1+y_2q_2+y_3q_3)$$

$$=(x_1p_1+x_2p_2)(y_1q_1+y_2q_2+y_3q_3)=E[X]E[Y]$$

となる。証明終

問同じ設定で $V(X+Y)=V(X)+V(Y)$ を示せ。

例 12 例題7において，1つ目のさいころの目の数をY_1，2つ目のさいころの目の数をY_2とすると，Y_1とY_2は独立であり，$Y = Y_1 + Y_2$。また，

$$E[Y_1] = (1+3+4+5+6+8)/6 = \frac{27}{6} = \frac{9}{2},$$

$$V(Y_1) = ((1-9/2)^2+(3-9/2)^2+(4-9/2)^2+(5-9/2)^2+(6-9/2)^2+(8-9/2)^2)\times\frac{1}{6} = \frac{59}{12}$$

$$E[Y_2] = (1+2+2+3+3+4)/6 = \frac{15}{6} = \frac{5}{2},$$

$$V(Y_2) = ((1-5/2)^2+(2-5/2)^2+(2-5/2)^2+(3-5/2)^2+(3-5/2)^2+(4-5/2)^2)\times\frac{1}{6} = \frac{11}{12}$$

であるから，

$$E[Y_1] + E[Y_2] = 7 = E[Y], \quad V(Y_1) + V(Y_2) = 35/6 = V(Y)$$

が成り立っている。

確率の公理的構成

近代的な確率の定義（コルモゴロフ）

事象の集合の上で定義され，実数値をとる関数$P(A)$で次の2つの条件を満たすものを**確率**という。

(1) 任意の事象Aについて$0 \le P(A) \le 1$かつ　$P(\emptyset) = 0,\ P(\Omega) = 1$

(2) $A \cap B = \emptyset$ならば，

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

近代的な確率は事象の “面積” のようなものである。各事象は無限集合でもよいし，事象の確率は等確率とは限らず，（有理数でも無理数でも）$[0,1]$中の様々な値を取り得る。

上の（1），（2）から次の性質を導くことができる。

（ア）

$$P(A) = 1 - P(A^c)$$

（イ）

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

（ウ）$A \subset B$ならば，$P(A) \le P(B)$，かつ，$P(B \setminus A) = P(B) - P(A)$。

（エ）A, B, C のどの２つをとっても共通部分が空事象ならば

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

$A_1, A_2, ..., A_n$ のどの２つをとっても共通部分が空事象ならば

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$$

（オ）全事象 Ω がいくつかの部分事象 $A_1, ..., A_n$ の互いに重なり合わない和事象となっているとき

$$P(B) = P(B \cap A_1) + ... + P(B \cap A_n)$$

問 16 次をしめせ。

（１）$A_1, ..., A_n$ を事象とするとき

$$P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

（２）A, B, C を事象とするとき

$$P(A \cap B \cap C) \geq 1 - P(A^c) - P(B^c) - P(C^c)$$

さらに，独立性の定義と $P(A \cap B^c) = P(A) - P(A \cap B)$ を使うと，次の４つの性質は同値であることがわかる。

（１）A と B は独立である

（２）A と B^c は独立である

（３）A^c と B は独立である

（４）A^c と B^c は独立である

問 17 上の４つが同値であることを証明せよ。

問　題

1 原点 O を出発して，数直線上を動く点 P がある。さいころを投げて，５か６の目が出れば P は＋ 1 だけ移動し, それ以外の目が出れば P は－ 1 だけ移動する。この試行を 10 回繰り返した後の，点 P の座標を X とする。

（１）X の期待値（平均）と分散を計算せよ。

（2）X^2 の期待値（平均）を計算せよ。

2 10 枚のカードがあり、その各々に 0，2，6 のいずれかの数を記入する。これら 10 枚のカードから 1 枚を選び，そのカードに記された数を X とするとき，その平均が 3 で分散が 6 以下になるようにしたい。数 0，2，6 の記されたカードの枚数をそれぞれいくらにすればよいか。

3 X は連続分布に従う確率変数で，その密度関数が

$$\rho(x) = C\frac{1}{(1+|x-1|)^5}$$

で与えられるとする。ただし，C は正定数である。

（1）定数 C を計算せよ。

（2）X の期待値と分散を計算せよ。

（3）$|X| \leq 1$ となる確率を計算せよ。

4 集団の中から無作為に 13 人を選ぶとき，日曜日生まれの人の数を X，土曜日生まれの人の数を Y とする。ただし，どの曜日に生まれる確率も $\frac{1}{7}$ とする。

（1）$X = k, Y = m$ となる確率 $P(X = k, Y = m)$ を k, m の式として表せ。ただし，$k \geq 0, m \geq 0, k + m \leq 13$ とする。

（2）$P(X = k), P(Y = m)$ をもとめよ。X, Y は独立か。

5 X は離散型または連続型確率分布にしたがう確率変数とし，X の平均 $E[X]$，分散 $V(X)$ ともに有限とする。次をしめせ。

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

6 さいころを 2 回投げ、1 回目の目を X, 2 回目の目を Y とする。$U = X + Y, V = X - Y$ とおく。U と V は独立か。

総　合　問　題

問題 1　n を 2 以上の整数とする。1 つのさいころを n 回続けて投げ，同じ目が初めて 2 回続けて出るまで投げた回数を X とする。ただし，n 回までに続けて同じ目が出なかったときには，$X = n + 1$ とする。X の期待値（平均）を求めよ。

問題 2　1 枚の硬貨を投げて，表が出れば A さんに 1 点を与え，裏が出れば B さんに 1 点を与える。硬貨を n 回投げるとき，A さんの総得点を X，B さんの総得点を Y とする。2 人とも持ち点 0 から始めるとして

（1）X の確率分布と期待値を求めよ。

（2）$X-Y$ の確率分布と期待値を求めよ。

（3）n 回投げて $X=i$ であったとき，1 回目に表が出ていた条件つき確率を，$i=1,2,...,n$ についてもとめよ。

問題3　1 枚の硬貨を 6 回投げ，各回ごとに表が出たら次の規則にしたがって点を与え，裏が出たらその回は 0 点として，6 回の合計点を X とする。

1 回目 $\cdots$ 3 点

2, 3 回目 $\cdots$ 2 点

4, 5, 6 回目 $\cdots$ 1 点

（1）X の期待値（平均）と分散を求めよ。

（2）$P(X=k)$ を $k=2,3,7,8$ について求めよ。

問題4　3 つの箱 A，B，C があり，箱 A には 4 個の赤球と 2 個の白球が入っている。箱 B には 1 から 8 までの数字を 1 つづつ書いた札が計 8 枚入れてあり，箱 C には 4 から 11 までの数字を 1 つづつ書いた札が計 8 枚入れてある。

まず箱 A から球を 1 個取り出して，もしそれが赤球ならば箱 B から札を 1 枚取り出し，もしそれが白球ならば箱 C から札を 1 枚取り出す。このようにして取り出される札の数を X とする。

（1）$3 \leq X \leq 5$ となる確率をもとめよ。

（2）X の期待値（平均）と分散をもとめよ。

問題5　つぼ A には数字 2, 4, 6, 8 がひとつずつ書かれた札が計 4 枚，つぼ B には数字 1, 3, 5, 7 がひとつずつ書かれた札が計 4 枚入っている。まず，つぼ A から無作為に 1 枚取り出しその札に書かれた数を X とする。

もし $X=8$ ならば，つぼ B から無作為に 1 枚取り出しその札に書かれた数を Y とする。

もし $X \neq 8$ ならば，つぼ A の残りの 3 枚から無作為に 1 枚取り出しその札に書かれた数を Y とする。

$Z=X+Y$ とする。

（1）　$E[X]$ および $V(X)$ をもとめよ。

（2）$E[Y]$ および $V(Y)$ をもとめよ。

（3）X と Y は独立か。

（4）$E[Z]$ および $V(Z)$ をもとめよ。

第4章　いろいろな分布

離散分布，連続分布各々について，代表的な分布の性質を調べる。

4.0.6　離散分布

結果が2種類の実現の可能性しかなく (それらを成功および失敗と呼ぶことが多い)，それらの起こる確率 p と q $(p+q=1)$ がどの回も同じである独立試行を**ベルヌーイ試行**という。$X_1, X_2, ..., X_n$ を n 個の確率変数とし，それらは同一の確率分布にしたがうとする。

(1) **一様分布**　X のとる値は $\{1, 2, ..., n\}$ であり，$P(X=k)=\frac{1}{n}, k=1, ..., n$ であるとき，X は（離散型の）一様分布にしたがうという。

このとき

$$E[X]=\sum_{k=1}^{n} k\frac{1}{n}=\frac{n+1}{2}$$

$$E[X^2]=\sum_{k=1}^{n} k^2\frac{1}{n}=\frac{(n+1)(2n+1)}{6}$$

よって

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

$$=2(n+1)(2n+1)/12-3(n+1)^2/12=(1/12)(4n^2-3n^2+6n-6n+2-3)=\frac{n^2-1}{12}$$

である。

(2) **二項分布**　X のとる値は $\{0, 1, 2, ..., n\}$ であり，

$$P(X=k)={}_nC_k p^k(1-p)^{n-k}, k=0, 1, ..., n$$

であるとき，X は二項分布にしたがうという。ただし，$0 \leq p \leq 1$ である。

このとき

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

である。

証明

以下で $q = 1 - p$ とおく。

$$E[X] = \sum_{k=0}^{n} k_n C_k p^k q^{n-k} = \sum_{k=1}^{n} k_n C_k p^k q^{n-k}$$

ここで，$k_n C_k = n_{n-1} C_{k-1}$ であるから (左辺 $= k\frac{n!}{k!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} =$ 右辺),

$$E[X] = \sum_{k=1}^{n} n_{n-1} C_{k-1} p^k q^{n-k} = n \sum_{l=0}^{n-1} {}_{n-1}C_l \; p^{l+1} q^{n-l-1}$$

$$= np \sum_{l=0}^{n-1} {}_{n-1}C_l \; p^l q^{(n-1)-l} = np(p+q)^{n-1} = np$$

（ここで $l = k - 1$）

同様に計算すると

$$E[X(X-1)] = n(n-1)p^2$$

であることがわかるので,

$$V(X) = E[X^2] - (E[X])^2 = E[X(X-1)] + E[X] - (E[X])^2$$

$$= n(n-1)p^2 + np - (np)^2 = np(1-p) = npq$$

であることがわかる。

例題 9　1 枚の硬貨を 5 回投げるとき，表の出る回数から裏の出る回数を引いた数 X の期待値および分散をもとめよ。

解 9 表の出る回数を Y とすると，Y は二項分布に従うから

$$E[Y] = 5 \times \frac{1}{2} = \frac{5}{2},\; V(Y) = 5 \times \frac{1}{2} \times \frac{1}{2} = \frac{5}{4}$$

$X = Y - (5 - Y) = 2Y - 5$ であるから

$$E[X] = 2E[Y] - 5 = 2 \times \frac{5}{2} - 5 = 0$$

$$V(X) = 4V(Y) = 2^2 \times \frac{5}{4} = 5$$

である。

問 18 ○，×で答える 6 つの問題が与えられている。この解答をするのに考えないででたらめに○，×をつけるとき，そのうちの正解数を X とする。X の期待値，分散および標準偏差をもとめよ。

問 19 日本人の血液型は，*10* 人に 3 人の割合で O 型である。*5* 人の日本人を任意に選んだとき，そのうちの O 型の人数を X とする。X の期待値，標準偏差をもとめよ。

(3) **幾何分布**　1 枚の硬貨を投げつづけ，初めて表の出るまでの投げる回数の確率分布を考える。

X のとる値は $\{1, 2, ...\}$ であり，$P(X = k) = p(1-p)^{k-1}, k = 1, ...$ であるとき，X はパラメータ p の幾何分布にしたがうという。ただし，$p > 0$ である。上の例では $p = \frac{1}{2}$ の場合に当たる。実際，事象 $X = k$ は $A \cap B$ に等しい。ここで

$$A = \text{はじめの } (k-1) \text{ 回裏が出る}$$

$$B = k \text{ 回目に表が出る}$$

である。A, B は独立であり、よって $P(X = k) = P(A \cap B) = (1-p)^{k-1}p$ となる。

このとき

$$E[X] = \frac{1}{p},\ V(X) = \frac{1-p}{p^2}$$

である。

証明　$q = 1 - p$ とおく。

$$E[X] = \sum_{k=1}^{\infty} kP(X = k) = p\sum_{k=0}^{\infty} kq^{k-1}$$

$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ であるから，$\sum_{k=0}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$ (両辺を q で微分する)。したがって

$$E[X] = p\frac{1}{p^2} = \frac{1}{p}$$

一方

$$E[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)pq^{k-1}$$

$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$ の両辺を微分して

$$\sum_{k=2}^{\infty} k(k-1)q^{k-2} = \frac{2}{(1-q)^3}$$

よって，$E[X(X-1)] = pq\frac{2}{(1-q)^3} = 2\frac{q}{p^2}$。これより

$$V(X) = E[X(X-1)] + E[X] - (E[X])^2 = 2\frac{q}{p^2} + \frac{1}{p} - \frac{1}{p^2}$$

$$= \frac{2q+p-1}{p^2} = \frac{1-2p+p}{p^2} = \frac{1-p}{p^2}$$

となる。

(4) **負の二項分布**独立なベルヌーイ試行を行ったときに、r 回の「成功」をする前に失敗した試行回数の分布を負の 2 項分布という。その分布は

$$P(X=k) = {}_{k+r-1}C_k p^k (1-p)^r, k = 0, 1, ...,$$

である。別の言い方をする (主語を言い換えると) と、 r 回の成功を x 回目の試行で達成する確率は

$$P(X'=x) = {}_{x-1}C_{r-1} p^r (1-p)^{x-r}, x = 0, 1, ...,$$

である。

$r = 1$ の場合が幾何分布である。上側の表現から、$X_1, ..., X_r$ が独立で幾何分布にしたがうとき、$X = X_1 + ... + X_r$ は負の二項分布にしたがう。したがって

$$E[X] = E[X_1] + ... + E[X_r] = r\frac{1}{p} = \frac{r}{p}$$

である。同様にして

$$V[X] = V[X_1] + ... + V[X_r] = r\frac{1-p}{p^2} = \frac{r(1-p)}{p^2}$$

である。

(5) **ポアソン分布**

X のとる値は $\{0, 1, 2, ...\}$ であり，$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k = 0, 1, ...$ であるとき，X はパラメータ λ のポアソン分布にしたがうという。ただし，$\lambda > 0$ である。

このとき

$$E[X] = \lambda, V(X) = \lambda$$

である。

証明

$$E[X] = \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}$$

$$= \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

一方,

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}$$

$$= \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2$$

したがって,

$$V(X) = E[X^2] - (E[X])^2 = E[X(X-1)] + E[X] - (E[X])^2$$

$$= \lambda^2 + \lambda - \lambda^2 = \lambda$$

となる。

例 13 次のような変数はポアソン分布に従うことが知られている。

（1）単位時間にガソリンスタンドにくる自動車の数。
（2）単位時間に交換機が受ける電話の呼び出し数。
（3）一連の製造ラインで発生する不良品の個数。
（4）ある地域の火災件数。
（5）小さなスーパーマーケットのレジでの待ち人数。
（6）*1*ミリリットルの希釈された水試料中に含まれる特定の細菌の数
（7）*1*ページの文章を入力するとき、綴りを間違える回数
（8） *1*日に受け取る電子メールの件数

このように，**小さな確率で非恒常的に起こる**現象の記述にはポアソン分布が適している。

例題 10 $\frac{1}{100}$ の確率であたるくじを *100* 回試したとき，実際 r 回当たる確率をもとめよ。

解 10 *100*回のくじ引きをして当たる回数の平均は

$$\sum_{i=1}^{100} \frac{1}{100} = 1$$

そこで*100*回のくじ引きを*1*単位としてそこで平均*1*回当たると考える。実際r回当たる確率は

$$P_1(r) = \frac{1}{r!}1^r e^{-1} = \frac{e^{-1}}{r!}$$

したがって

$$P_1(0) = 0.368,\ P_1(1) = 0.368,\ P_1(2) = 0.184,\ P_1(3) = 0.061, ...$$

である。

*100*回のくじ引きをして平均１回当たるといっても「１回も当たらない」確率が約*37%*あることに注目すべきである。つまり，祭礼の餅播きなどで，参加者の約３倍の餅を用意しなければ，餅がほぼ確実に全員に行き渡るようにはならない。($(1/3)^3 = 0.037...$。)

なお，X_1, X_2が独立で，X_1がパラメータλ_1のポアソン分布にしたがい，X_2がパラメータλ_2のポアソン分布にしたがうとき，$Y = X_1 + X_2$はパラメータ$\lambda = \lambda_1 + \lambda_2$のポアソン分布にしたがう。(ポアソン分布の**再生性**)

証明。

$$P(X_1 + X_2 = n) = \sum_{k=0}^{n} P(X_1 = n-k)P(X_2 = k)$$

$$= \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^{n-k}}{(n-k)!}e^{-\lambda_2}\frac{\lambda_1^k}{k!} = e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^{n-k}}{(n-k)!}\frac{\lambda_2^k}{k!}$$

$$= e^{-(\lambda_1+\lambda_2)}\frac{1}{n!}\sum_{k=0}^{n} {}_nC_k\lambda_1^{n-k}\lambda_2^k = e^{-(\lambda_1+\lambda_2)}\frac{1}{n!}(\lambda_1+\lambda_2)^n$$

ただし、最後の等式には二項定理を使った。　証明終わり。

例題 11 正岡救急病院は道後町，秋山町を管轄区域とし４台のベッドがある。道後町，秋山町から搬送される救急患者数は，それぞれ$\lambda_1 = 2, \lambda_2 = 1$のポアソン分布にしたがっている。病院収容患者数が４人を超える確率をもとめよ。

解 11 $Y = X_1 + X_2$ とおく。道後町と秋山町は別の町内なので、X_1 と X_2 は独立。よって Y はパラメータ $\lambda = \lambda_1 + \lambda_2$ のポアソン分布にしたがうと考えられる。これより

$$P(Y = k) = \frac{1}{k!}\lambda^k e^{-\lambda}, \quad k = 0, 1, 2, ...$$

したがってもとめる確率は

$$P(Y > 4) = 1 - \sum_{k=0}^{4} P(Y = k)$$

$$= 1 - e^{-\lambda}(1 + \lambda + \frac{1}{2}\lambda^2 + \frac{1}{6}\lambda^3 + \frac{1}{24}\lambda^4)$$

となる。$\lambda = 3$ として、$\sim 1 - 0.815 = 0.185$。これは無視できない確率である。

問 20 ある家に *1* 日にかかってくる電話の回数はポアソン分布に従い，その平均は８（回）である。この家に電話が *1* 日に *10* 回以上かかってくる確率と，*1* 日に高々４回しかかかってこない確率をもとめよ。

問 21 あるレンタカーの営業所、には *3* 台の車があり *1* 日単位で貸し出す。レンタカーの需要は *1* 日平均 *2* 台でポアソン分布に従う。*1* 台貸し出すと *7000* 円の収入がある一方、営業所全体の経費として *1* 日あたり *8000* 円かかる。その営業所全体での *1* 日の利益額の期待値を求めよ。

4.0.7 連続分布

(1) **一様分布** 確率密度関数 $\rho(x)$ がつぎで与えられる確率分布を一様分布という。

$$\rho(x) = \frac{1}{b-a} \ (a \leq x \leq b)$$

この分布では区間 $[a, b]$ 上の長さの等しい（互いに重ならない）小区間に値をとる事象は，どれも同等に “確からしい” とみなされる。

例題 12 （有理数でない確率の例）

太郎と花子はあるレストラン前で待ち合わせわした。太郎は *12* 時 X 分、花子は同日の *12* 時 Y 分にそれぞれ到着する。太郎、花子ともレストラン前に到着してから、相手に会えず「待ちぼうけを喰らった」と思って店を離れるまでの時間は t_0 （$0 < t_0 < 60$）分であり、レストランを出た後、再度入店することはないものとする。なお X, Y それぞれ区間 $(0, 60)$ の一様分布に従う確率変数であり、X, Y は互いに独立とする。

このとき、太郎、花子が同時刻にレストラン前にいる確率が $\frac{1}{3}$ となるように t_0 をきめよ。また確率が $\frac{1}{\sqrt{2}}$ となる t_0 を求めよ。

解 12 太郎と花子が同時刻にレストラン前にいる確率は $P(|X - Y| < t_0)$ である。X, Y は互いに独立であり、それぞれ $(0, 60)$ の一様分布に従うことから、上記の確率は図の正方形に占める色付部分の領域の割合である。

題意から、

$$\frac{1}{3} = 1 - \frac{(60 - t_0)^2}{60^2}$$

を満たす t_0 を求めればよい。よって $t_0 = 20(3 - \sqrt{6}) \notin \mathcal{Q}$。

(2) **指数分布** 確率密度関数 $\rho(x)$ がつぎで与えられる確率分布を指数分布 $(\mathrm{Exp}(\lambda))$ という。$\rho(x) = 0\ (x < 0),\ = \lambda e^{-\lambda x}\ (x \geq 0)$ ただし，$\lambda > 0$ である。

X がこの分布にしたがうとき，$E[X] = \frac{1}{\lambda},\ V(X) = \frac{1}{\lambda^2}$ である。

証明。

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x} dx$$
$$= \frac{1}{\lambda}$$

問 これを示せ。

同様にして

$$E[X^2] = \frac{2}{\lambda^2}$$

これより

$$V(X) = E[X^2] - (E[X])^2 = \frac{1}{\lambda^2}$$

証明終わり。

問 22 ある電話局管内の電話の通話時間（分）は確率変数 X で表され，その確率密度関数 $\rho(x)$ は

$$\rho(x) = Ce^{-\frac{x}{3}}\ (0 \leq x < 180),\ = 0\ (x \geq 180)$$

である。一方，通話料は $3(n-1) \leq x < 3n$ （n は自然数）の通話時間に対して $10n$ 円である。

（1）定数 C の値をもとめよ。

（2）*1* 回の通話時間の平均をもとめよ。

（3）*1* 回の通話料の平均をもとめよ。

(3) **正規分布**

ガウス積分

$$\int_{-\infty}^{+\infty} e^{-t^2} dt = \sqrt{\pi}$$

証明　2 重積分

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(t^2+s^2)} ds dt = \pi$$

をしめせばよい。詳細は省略。

問　これを示せ。

$m \in \mathbf{R}, \sigma > 0$ に対し

$$\rho_{m,\sigma}(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-m)^2}{2\sigma^2}}$$

を確率密度関数とする連続分布を，**パラメータ m, σ^2 の正規分布**といい，その分布を $N(m, \sigma^2)$ で表す。確率変数 X の分布がパラメータ m, σ^2 の正規分布のとき，X はパラメータ m, σ^2 の正規分布にしたがうといい，$X \sim N(m, \sigma^2)$ とかく。

ここで $y = \rho_{m,\sigma}(x)$ のグラフは，$x = m$ について対称であり，$(\rho_{m,\sigma}(m-x) = \rho_{m,\sigma}(m+x))$，$x < m$ で増加，$x > m$ で減少であり，さらに $x = m - \sigma, m + \sigma$ を変曲点にもつ。また $u = \frac{t-m}{\sqrt{2}\sigma}$ と変数変換すると

$$\int_{-\infty}^{+\infty} \rho_{m,\sigma}(t) dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-u^2} \sqrt{2}\sigma du = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-u^2} du = 1$$

である。

このとき

$$E[X] = m, V(X) = \sigma^2$$

である。

証明

$$E[X] = \int_{-\infty}^{+\infty} t\rho_{m,\sigma}(t)dt$$

$u = \frac{t-m}{\sqrt{2}\sigma}$ と変数変換すると

$$R.H.S. = \int_{-\infty}^{+\infty} \frac{m + u\sigma\sqrt{2}}{\sigma\sqrt{2\pi}} e^{-u^2}\sqrt{2}\sigma du$$

$$= m\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{-u^2}du + \frac{\sigma\sqrt{2}}{\sqrt{\pi}}\int_{-\infty}^{+\infty} ue^{-u^2}du = m$$

問　これを示せ。

$$V(X) = \int_{-\infty}^{+\infty} (t-m)^2\rho_{m,\sigma}(t)dt$$

$$= \int_{-\infty}^{+\infty} \frac{2\sigma^2u^2}{\sigma\sqrt{2\pi}} e^{-u^2}\sigma\sqrt{2}du = \sigma^2\frac{2}{\sqrt{\pi}}\int_{-\infty}^{+\infty} u^2e^{-u^2}du$$

$(u^2e^{-u^2} = u.ue^{-u^2}; u \to 1, ue^{-u^2} \leftarrow -(1/2)e^{-u^2})$

$$= \sigma^2\frac{2}{\sqrt{\pi}}([-\frac{1}{2}ue^{-u^2}]_{-\infty}^{+\infty} + \frac{1}{2}\int_{-\infty}^{+\infty} e^{-u^2}du) = \sigma^2\frac{2}{\sqrt{\pi}}(0 + \frac{1}{2}\sqrt{\pi}) = \sigma^2$$

証明終わり

例 14　$X \sim N(m, \sigma^2)$ のとき，表と変数変換より

$$P(m-\sigma \le X \le m+\sigma) \sim 0.683$$

$$P(m-2\sigma \le X \le m+2\sigma) \sim 0.954$$

$$P(m-3\sigma \le X \le m+3\sigma) \sim 0.997$$

標準正規分布

とくに $m = 0, \sigma = 1$ のとき，この分布を**標準正規分布**という。つまり，$X \sim N(0, 1)$ のとき，

$$P(a \leq x \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

である。標準正規分布には $\int_x^\infty \rho_{0,1}(t)dt$ の表があり，利用することができる。**標準正規分布の分布関数**

$$F(x) = P(X \leq x)$$

について，$F(-x) = 1 - F(x)$ である。$[F(-x) = P(X \leq -x) = P(X > x) = 1 - P(X < x) = 1 - F(x), x > 0$ による。$]$ 標準正規分布の分布関数を $\Phi(x)$ で表すことが多い。

さらに，上の証明からわかるように，X がパラメータ m, σ^2 の正規分布にしたがうとき，変数変換 $Y = \frac{X-m}{\sigma}$ をおこなえば Y は標準正規分布にしたがう。標準正規分布には表があるので，これを使って X の確率をもとめることができる。

表の引き方

表中の数字は、全体の面積を 1.0 としたときの、$Z = 0$ から Z までの面積を表します。

たとえば $Z = 1.00$ の場合は「.3413」となり、斜線の部分の面積が全体の 34.13% であることがわかります。

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879

.....

1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621

.....

正規分布の例

（１）測量における誤差

誤差とは、観測値と真実の値（真値）との差のことを言う：誤差＝観測値－真値。誤差を含む測定では、測定値 は測定ごとに異なる値を取るが、この測定値 が、ある確率分布 を持つと仮定する。この確率分布の代表的なものがガウスの正規分布である。そのため正規分布を**ガウス分布**ともいう。

（２）視神経における感度特性

ヒトを含む哺乳類の網膜における視神経における分光感度特性を考察する。網膜中の視細胞は光エネルギーを電気信号に変える。

視細胞には**桿体** (rod) と**錐体** (cone) がある。このうち錐体 (cone) は明るいときに働き、色を知覚する (S-cone, M-cone, L-cone)。興奮のピーク波長は、各々、R : 700nm　G : 546.1nm　B : 435.8nm である。実際の各細胞の感度はほぼ正規分布にしたがう。

例題 13 $X \sim N(4, 3^2)$ であるとき，$P(3 < X < 6)$ および $P(X > 10)$ をもとめよ。

解 13 $Y = \frac{X-4}{3}$ とおくと $Y \sim N(0, 1)$ である。よって

$$P(3 < X < 6) = P(\frac{3-4}{3} < \frac{X-4}{3} < \frac{6-4}{3}) = P(-0.333 < Y < 0.667)$$

$$= \Phi(0.667) - \Phi(-0.333) = 0.7475 - 0.3694 = 0.3781$$

同様に

$$P(X > 10) = P(Y > \frac{10-4}{3}) = 1 - \Phi(2) = 0.0228$$

例題 14 ある高校の *3* 年男子 *500* 人の身長 X は平均 *170.9cm*、標準偏差 *5.4cm* の正規分布にしたがう。

(1) 身長 *180cm* 以上の生徒は約何人か。

(2) 高いほうから *129* 人の中に入るには何 *cm* 以上あればよいか。

解 14 *(1)* $Y = \frac{X-170.9}{5.4}$ とおくと Y は標準正規分布 *N(0,1)* にしたがう。

$$P(X \geq 180) = P(Y \geq \frac{180-170.9}{5.4}) = P(Y \geq 1.69) = 0.0455$$

身長 *180cm* 以上の生徒は全体の *0.0455* なので、$500 \times 0.0455 = 22.75$ から、約 *23* 人いる。

(2) x*cm* 以上あれば高いほうから *129* 人の中に入るとする。$\frac{129}{500} = 0.258$ だから、$P(Y \geq u) = 0.258$ となる u の値をさがす。正規分布表から $u \sim 0.65$。よって x の値は、$0.65 = \frac{x-170.9}{5.4}$ より、$x = 174.41$。よって *174.5cm* あればよい。

問 23 ある試験での成績の結果は，平均点 *64* 点，標準偏差 *14* 点であった。得点の分布は正規分布にしたがうとするとき，次の問に答えよ。

（１）得点が *36* 点から *92* 点の者が *400* 人いた。受験者の総数は約何人か。

（２）合格点を *50* 点とすると，約何人が合格することになるか。

なお，$X_1 \sim N(m_1, \sigma_1^2), X_2 \sim N(m_2, \sigma_2^2)$ であり，かつ X_1 と X_2 は独立であるとき，$Y = X_1 + X_2$ は $N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$ にしたがう。(正規分布の**再生性**) この証明には特性関数の議論が必要なため，詳細は後述する。

(4) **対数正規分布** $X \sim N(m, \sigma^2)$ であるとき，$Y = e^X$ の分布を対数正規分布という。つまり，Y の確率密度関数 $\rho(x)$ は

$$\rho(x) = 0, \quad x \leq 0$$

$$\rho(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp(-\frac{(\log x - m)^2}{2\sigma^2}), \ x > 0$$

であたえられる。このとき

$$E[Y] = e^{m+\frac{\sigma^2}{2}},$$

$$V(Y) = (e^{\sigma^2} - 1)e^{2m+\sigma^2}$$

である。

対数正規分布は，株価過程の数学的モデル化の際に用いられる。

問　題

1　50歳の夫と48歳の妻が20年後まで生存する確率は，夫が0.15，妻が0.2である。現在夫が50歳，妻が48歳である夫婦が10組いるとして，このうち20年後に夫婦のうち少なくとも一方が生存している組の数をXとする。Xが二項分布に従うとして，Xの期待値，標準偏差をもとめよ。

2 じゃがいもの山積みがあって，その不良品率は10%である。この中からでたらめに25個取り出すとき

（1）不良品がちょうど3個含まれる確率をもとめよ。

（2）良品が少なくとも20個含まれる確率をもとめよ。

（3）良品が20個以下である確率をもとめよ。

3 人口100万人のある県では，ある病気Bの発生が1年間に平均10件みられる。この県で病気Bにかかる人が1年間に2人以下である確率をもとめよ。

4 E大学で花粉症が流行し，対策に追われた当局は花粉の数を調べるため構内に10000個の観測皿を設置してその中の花粉数を調べた。その結果，10000個の観測皿のうち花粉が2個入っていたものが109個，3個入っていたものが6個あった。

（1）10000個の観測皿全体に何個の花粉が降りそそいだと考えられるか。

（2）花粉が1つも入っていない観測皿はいくつあると推測されるか。

5 次の表（次ページ）のような宝くじ（1枚200円）の賞金額の期待値を計算せよ。

等級	当選金	本数
1等	40000000	7
1等前後賞	10000000	14
1等組違い賞	200000	903
2等	10000000	5
2等組違い賞	100000	645
3等	1000000	130
4等	140000	130
5等	10000	1300
6等	1000	26000
7等	200	1300000
はずれ	0	(引き算)
合計		13000000

6 次の問に答えよ。

(1) $X \sim N(4, 2^2)$ のとき $P(X \leq 6)$ をもとめよ。

(2) $X \sim N(3, (1.5)^2)$ のとき，$P(X \leq x) = 0.4218$ となる x をもとめよ。

(3) $X \sim N(5, 2^2)$ のとき $P(2.5 \leq X \leq 6.5)$ をもとめよ。

(4) $X \sim N(6, 2^2)$ のとき，平均を中心とする区間でその上の確率が 0.9 になるようなものをもとめよ。

第5章　モーメント母関数

5.0.8　モーメント母関数

定義 5.1　確率変数 X に対して e^{tX} の期待値

$$M_X(t) = E[e^{tX}]$$

を X を**モーメント母関数**（または積率母関数）という。

X が離散型の確率変数で、X の取りうる値が $X = \alpha_1, \alpha_2, , , ,$ のとき

$$M_X(t) = \sum_{k=1}^{\infty} e^{t\alpha_k} P(X = a_k)$$

で与えられる。X が連続型の確率変数のときは、X の確率密度関数 $f_X(x)$ を用いて

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

と与えられる。

ここで、$M_X(t)$ を t で微分すると

$$M'_X(t) = E[Xe^{tX}]$$

となり、$t = 0$ を代入すると

$$E[X] = M'_X(0)$$

となる。$M'_X(t)$ をもう一度 t で微分すると

$$M''_X(t) = E[X^2 e^{tX}]$$

となり、$t = 0$ を代入すると

$$E[X^2] = M''_X(0)$$

となる。以下同様に計算できる。すなわち、モーメントはモーメントを生成する関数となっている。

5.0.9　モーメント母関数の例

(1) 二項分布 $B(n;p)$。ただし、$q=1-p$ である。

$$M_X(t)=\sum_{k=0}^{n}e^{tk}P(X=k)=\sum_{k=0}^{n}{}_nC_k(pe^t)^kq^{n-k}=(pe^t+q)^n$$

(2) ポアソン分布 $Po(\lambda)$。

$$M_X(t)=\sum_{k=0}^{\infty}e^{tk}P(X=k)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!}=e^{-\lambda}\exp\{\lambda e^t\}=\exp\{\lambda(e^t-1)\}$$

(3) 幾何分布 $Ge(p)$。ただし、$q=1-p$ である。

$$M_X(t)=\sum_{k=0}^{\infty}e^{tk}P(X=k)=\sum_{k=0}^{\infty}e^{tk}q^kp=p\sum_{k=0}^{\infty}(e^tq)^k=\frac{p}{1-qe^t}$$

(4) 正規分布 $N(\mu,\sigma^2)$

$$M_X(t)=e^{\mu t+\frac{1}{2}\sigma^2t^2}$$

(5) 指数分布 $\mathrm{Exp}(\lambda)$

$$M_X(t)=\int_0^{\infty}e^{tx}\lambda e^{-\lambda x}dx=\lambda\int_0^{\infty}e^{-(\lambda-t)x}dx=\frac{\lambda}{\lambda-t}$$

証明　(4)

$$M_X(t)=\frac{1}{\sqrt{2\pi}\sigma}\quad\int_{-\infty}^{\infty}e^{tx}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad dx$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\quad\int_{-\infty}^{\infty}e^{-\frac{1}{2\sigma^2}((x-\mu-\sigma^2t)^2-(\sigma^2t)^2+2\sigma^2t(x-\mu)-2\sigma^2tx)}dx$$

$$=e^{\mu t+\frac{1}{2}\sigma^2t^2}\frac{1}{\sqrt{2\pi}\sigma}\int e^{-\frac{z^2}{2\sigma^2}}dz\quad(z=x-\mu-\sigma^2t)$$

$$=e^{\mu t+\frac{1}{2}\sigma^2t^2}.$$

q.e.d.

モーメント母関数はもっと重要な性質をもっている。

定理 5.1 確率変数 X, Y について、$M_X(t) = M_Y(t)$ が成り立つとき、X, Y の確率分布は等しくなる。

たとえば、X のモーメント母関数を計算してみて $M_X(t) = \exp(\frac{1}{2}t^2)$ になったとする。これは正規分布のモーメント母関数であるので、X は正規分布に従うことがわかる。

定義 5.2

$$P(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = P(a_1 \leq X \leq b_1)P(a_2 \leq Y \leq b_2)$$

がつねに成立するとき確率変数 X と Y は**独立**であるという。

たとえばコインを 10 回投げた時、1 回目から 5 回目に表の出た数 X を、6 回目から 10 回目に表の出た数を Y とすると、X と Y は独立になる。

定理 5.2 (A) X と Y は独立となるとき、次の (1)-(3) が成り立つ。

$$(1)E[XY] = E[X]E[Y]$$

$$(2)M_{X+Y}(t) = M_X(t)M_Y(t)$$

$$(3)V(X+Y) = V(X) + V(Y)$$

(B) 有限個の場合も上の (1)-(3) が成り立つ。

$$(1)E[X_1X_2\cdots X_n] = E[X_1]E[X_2]\cdots E[X_n]$$

$$(2)M_{X_1+X_2+\cdots+X_n}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}$$

$$(3)V(X_1 + X_2 + \cdots + X_n) = V(X_1) + \cdots + V(X_n)$$

証明は帰納法による。

定理 5.2(A)(2) および定理 5.1 より、ポアソン分布、正規分布などの分布の再生性がわかる。

ただし、X,Y が独立で同一の型の分布（たとえば正規分布 $N(\mu_1,\sigma_1^2), N(\mu_2,\sigma_2^2)$）に従うとき、$X+Y$ も同一の型の分布（正規分布 $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ にしたがうことを**再生性**という。たとえば

$$M_X(t)M_Y(t) = e^{\mu_1 t+\frac{1}{2}\sigma_1^2t^2}e^{\mu_2 t+\frac{1}{2}\sigma_2^2t^2} = e^{(\mu_1+\mu_2)t+\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2} = M_{X+Y}(t).$$

5.0.10　独立確率変数の和・差・積・商の計算法

確率変数 X,Y は独立で、それぞれの確率密度関数が $f_X(x), f_Y(y)$ で与えられるとする。また $f(x,y) = f_X(x)f_Y(y)$ とおく。$Z_1 = X+Y, Z_2 = X-Y, Z_3 = XY, Z_4 = \frac{X}{Y}$ で与えられる確率変数 Z_1, Z_2, Z_3, Z_4 の確率密度関数を求める。

(1) $W_1 = Y$ とおき、X,Y を Z_1 と W_1 で表すと、$X = Z_1 - W_1, Y = W_1$ となる。このとき (Z_1, W_1) の同時確率密度関数 $g(z_1,w_1)$ は

$$g(z_1,w_1) = f(z_1-w_1,w_1)|\frac{\partial(x,y)}{\partial(z_1,w_1)}|$$

であたえられる。ここで

$$\frac{\partial(x,y)}{\partial(z_1,w_1)} = |\begin{pmatrix} \frac{\partial x}{\partial z_1} & \frac{\partial x}{\partial w_1} \\ \frac{\partial y}{\partial z_1} & \frac{\partial y}{\partial w_1} \end{pmatrix}| = |\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}| = 1 \tag{5.1}$$

となるので

$$g(z_1,w_1) = f_X(z_1-w_1)f_Y(w_1)$$

となる。Z_1 の周辺確率密度関数 $f_{Z_1}(z_1)$ は $g(z_1,w_1)$ を w_1 について積分することにより

$$f_{Z_1}(z_1) = \int_{-\infty}^{\infty} f_X(z_1-w_1)f_Y(w_1)dw_1$$

となる。

(2) $W_2 = Y$ とおき、X,Y を $Z_2 = X-Y$ と W_2 で表すと $X = Z_2 + W_2, Y = W_2$ となる。このとき

$$\frac{\partial(x,y)}{\partial(z_2,w_2)} = |\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}| = 1 \tag{5.2}$$

となるので (Z_2, W_2) の同時確率密度関数 $g(z_2, w_2)$ は

$$g(z_2, w_2) = f(z_2+w_2, w_2)|\frac{\partial(x,y)}{\partial(z_2, w_2)}| = f(z_2+w_2, w_2) = f_X(z_2+w_2)f_Y(w_2)$$

であたえられる。したがって Z_2 の周辺確率密度関数 $f_{Z_2}(z_2)$ は $g(z_2, w_2)$ を w_2 について積分することにより

$$f_{Z_2}(z_2) = \int_{-\infty}^{\infty} f_X(z_2 + w_2)f_Y(w_2)dw_2$$

となる。

(3) $W_3 = Y$ とおき、X, Y を $Z_3 = XY$ と W_3 で表すと

$$X = \frac{Z_3}{W_3},\ Y = W_3$$

となる。(Z_3, W_3) の同時確率密度関数 $g(z_3, w_3)$ は

$$g(z_3, w_3) = f(\frac{z_3}{w_3}, w_3)|\frac{\partial(x,y)}{\partial(z_3, w_3)}|$$

となる。$x = \frac{z_3}{w_3}, y = w_3$ であるから

$$\frac{\partial(x,y)}{\partial(z_3, w_3)} = \begin{vmatrix} \frac{1}{w_3} & -\frac{z_3}{w_3^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{w_3}$$

となるので

$$g(z_3, w_3) = f_X(\frac{z_3}{w_3})f_Y(w_3)\frac{1}{|w_3|}$$

したがって Z_3 の周辺確率密度関数 $f_{Z_3}(z_3)$ は $g(z_3, w_3)$ を w_3 について積分することにより

$$f_{Z_3}(z_3) = \int_{-\infty}^{\infty} f_X(\frac{z_3}{w_3})f_Y(w_3)\frac{1}{|w_3|}dw_3$$

となる。

(4) $W_4 = Y$ とおき、X, Y を $Z_4 = \frac{X}{Y}$ と W_4 で表すと

$$X = Z_4W_4,\ Y = W_4$$

となる。(Z_4, W_4) の同時確率密度関数 $g(z_4, w_4)$ は

$$g(z_4, w_4) = f(z_4w_4, w_4)|\frac{\partial(x,y)}{\partial(z_4, w_4)}|$$

となり、$x = z_4 w_4, y = w_4$ であるから

$$\frac{\partial(x,y)}{\partial(z_4,w_4)} = \left| \begin{array}{cc} w_4 & z_4 \\ 0 & 1 \end{array} \right| = w_4$$

となる。

$$g(z_4, w_4) = f_X(z_4 w_4) f_Y(w_4) |w_4|$$

なので $Z_4 3$ の周辺確率密度関数 $f_{Z_4}(z_4)$ は $g(z_4, w_4)$ を w_4 について積分することにより

$$f_{Z_4}(z_4) = \int_{-\infty}^{\infty} f_X(z_4 w_4) f_Y(w_4) |w_4| dw_4$$

となる。

問　題

1　(Max と Min の分布) $X_1, ..., X_n$ は独立同分布の確率変数で、指数分布 Exp(λ) に従うとする。

(1) $Z_1 = \max(X_1, ..., X_n)$ とおく。Z_1 のモーメント母関数を求めよ。

(2) $Z_2 = \min(X_1, ..., X_n)$ とおく。Z_2 のモーメント母関数を求めよ。

hint: (1) $P(Z_1 \leq x)$ を調べる。$f_{Z_1}(x) = \frac{d}{dx}P(Z_1 \leq x)$

(2) $P(Z_2 \geq x)$ を調べる。$P(Z_2 \leq x) = 1 - P(Z_2 \geq x)$

2　（1）X が $[0, 1]$ 上の一様分布にしたがうとき、$Y = aX + b$ $(a > 0)$ の確率密度関数を求めよ。

（2）X が平均 μ, 分散 σ^2 の $\mathbf{R}$ 上の正規分布にしたがうとき、$Y = aX + b$ の確率密度関数を求めよ。

hint: (1) $y = ax + b$ より $x = \frac{y-b}{a} = y^{-1}(y)$ とおく。$f_Y(y) = f_X(y^{-1})|\frac{dx}{dy}| = 1.\frac{1}{a}$

(2) $Y = aX + b$ と変換すると、平均 $\mu \to \mu + b$, 分散 $\sigma^2 \to a^2\sigma^2$。また再生性より Y も正規分布に従う。$\frac{dy}{dx} = a$。

3　X と Y が、平均 0, 分散 1 の $\mathbf{R}$ 上の正規分布にしたがう独立な確率変数であるとき、$Z = X + Y$ の確率密度関数を求めよ。

4　X と Y が、$[0, 1]$ 上の一様分布にしたがう独立な確率変数であるとき、$Z = X + Y$ の確率密度関数を求めよ。

hint: 5.0.10 (1) を適用するがまず分布関数を求める:$P(X_1 + X_2 \leq t), 0 < t < 2$。ただし $0 < t < 1$ のときと $1 < t < 2$ のときで場合分け（$x_2 = -x_1 + t$)。

第6章　正規分布の応用

6.0.11　チェビシェフの不等式

補題 X を確率変数で、平均 $E[X]$, 分散 $V(X)$ がともに有限とする。このとき任意の $\epsilon > 0$ に対し

$$P(|X - E[X]| > \epsilon) \leq \frac{V(X)}{\epsilon^2}$$

証明

$m = E[X]$ とおく。X を連続型確率変数とし、その密度関数を $f(x)$ とする。

$$V(X) = \int_{-\infty}^{+\infty} (x-m)^2 f(x)dx$$

$$\geq \int_{-\infty}^{m-\epsilon} (x-m)^2 f(x)dx + \int_{m+\epsilon}^{\infty} (x-m)^2 f(x)dx$$

$$\geq \int_{-\infty}^{m-\epsilon} \epsilon^2 f(x)dx + \int_{m+\epsilon}^{\infty} \epsilon^2 f(x)dx$$

$$= \epsilon^2 \int_{\{x;|x-m|>\epsilon\}} f(x)dx$$

$$= \epsilon^2 P(|X-m| > \epsilon).$$

両辺を ϵ^2 で割ればよい。X が離散型確率変数の場合も同様にできる。

証明終

6.0.12　大数の弱法則

定理１（大数の弱法則）－ラプラス 1814

$X_1, X_2, X_3, ...$ を独立で同分布な確率変数、

平均 $m = E[X_1] = E[X_2] = ...,$
分散 $v = V(X_1) = V(X_2) = ...$
はいづれも有限とする。このとき任意の $\epsilon > 0$ に対し

$$\lim_{n\to\infty} P(|\frac{X_1 + ... + X_n}{n} - m| > \epsilon) = 0$$

例 15 確率変数 X_i を
$X_i = 1$　さいころを投げて i 回目に 3 の目が出るとき, $X_i = 0$ それ以外
とおく ,$i = 1, 2, ...$。このとき $E[X_i] = \frac{1}{6} = m, i = 1, 2, ...$ である。$X_1 + ... + X_n$ は n 回中 3 の目が出る回数を表し、$\frac{X_1+...+X_n}{n}$ はその頻度を表す。このとき

$$\lim_{n\to\infty} P(|\frac{X_1 + ... + X_n}{n} - \frac{1}{6}| > \epsilon) = 0$$

である。

定理 1 では X_i の分布は何でもよい。E をある事象とし、もし E が起これば $X_i = 1$、そうでなければ $X_i = 0$ とすると、$m = E[X_i]$ は E の発生確率を表す。よってこの定理は、観測（頻度の測定）により事象の確率を推定することの正当性を保証している。

定理 1 の証明
$Y_n = X_1 + ... + X_n$ とおく。(X_i) の独立性から

$$V(Y_n) = V(X_1) + ... + V(X_n) = nv$$

であり、よって

$$V(\frac{Y_n}{n}) = \frac{1}{n^2}V(Y_n) = \frac{v}{n}$$

一方

$$E[\frac{Y_n}{n}] = \frac{1}{n}E[Y_n] = \frac{1}{n} \cdot nm = m$$

補題より

$$P(|\frac{Y_n}{n} - m| > \epsilon) \leq \frac{\frac{v}{n}}{\epsilon^2} = \frac{v}{n\epsilon^2}$$

これより $0 \leq \lim_{n\to\infty}$ 左辺 $\leq \lim_{n\to\infty} \frac{v}{n\epsilon^2} = 0.$
証明終

上の証明において、

$$P(|Y_n - nm| > n\epsilon) \leq \quad \frac{v}{n\epsilon^2}$$

である。そこで $\epsilon = c\sqrt{\frac{v}{n}}$ とおくと $(c > 0)$、

$$\text{左辺} = P(Y_n - nm| > c\sqrt{vn})$$

$$= P(|\frac{1}{\sqrt{vn}}(X_1 + ... + X_n - nm)| > c) \leq \frac{1}{c^2}$$

ということになる。この評価は n によらないので、

$$\frac{1}{\sqrt{vn}}(X_1 + ... + X_n - nm) = \frac{1}{\sqrt{\frac{v}{n}}}(\frac{X_1 + ... + X_n}{n} - m)$$

は、n が大きいとき、n によらない分布に近づくことが予想される。実際、次が成り立つ。

6.0.13　中心極限定理

定理 2（中心極限定理）
$X_1, X_2, X_3, ...$ を独立で同分布な確率変数、
平均 $m = E[X_1] = E[X_2] = ...,$
分散 $v = V(X_1) = V(X_2) = ...$
はいづれも有限とする。さらに

$$E[|X - m|^3] < \infty$$

を仮定する。このとき、任意の $a < b$ について、

$$\lim_{n\to\infty} P(a < \frac{\frac{X_1+...+X_n}{n} - m}{\sqrt{\frac{v}{n}}} < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

となる。

これを、「$\frac{1}{\sqrt{nv}}(X_1 + ... + X_n - nm)$ が標準正規分布に法則収束する」という。

X_i の分布が何であっても、収束先が正規分布になるのがこの定理の強力なところである。

2 項分布に対する中心極限定理－ドモアブル・ラプラス

中心極限定理とはやや異なるが、次のことが知られている：*2*項分布 $B_{n,p}$ にしたがう確率変数 X は、n が大きいとき、近似的に正規分布 $N(np, npq)$ にしたがう。ただし、$q = 1 - p$。したがって、

$$T = \frac{X - np}{\sqrt{npq}}$$

とおくと、T の分布は、n が大きいとき、近似的に標準分布 $N(0, 1)$ にしたがう。*($X = Z_1 + ... + Z_n, Z_i = 0$ or 1; where $V(Z_i) = pq$)*

この証明は比較的易しい。

例題 15

解 15

6.0.14　中心極限定理の証明

弱収束

グリベンコの定理

定理の証明

問　題

問 24 *(1)* あるくじは *1000* 本発行され、このうち *2* 本が当たりである。このくじを何本か買って当たる確率を $\frac{1}{2}$ 以上にするためには、少なくとも何本買わなければならないか。

(2) このくじは毎週発行される。太郎君は毎週このくじを *1* 枚買う。少なくとも *1* 回当たる確率を $\frac{1}{2}$ 以上にするためには、少なくとも何本買わなければならないか。

問 25 次を示せ。

$$\int_0^x e^{-\frac{t^2}{2}} dt \geq \sqrt{\frac{\pi}{2}}(1 - \frac{1}{x^2})$$

問 26 $(X_i)_{i=1,2,...}$ を平均 $\lambda = 1$ のポアソン分布をもつ独立同分布の確率変数列とし、$S_n = X_1 + X_2 + ... + X_n$ とする。

(1) S_n の分布は何か。

(2) $P(S_n \leq n) = e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!}$ を示せ。

(3) 中心極限定理を使って

$$\lim_{n\to\infty} e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!} = \frac{1}{2}$$

を示せ。

問 27 ある町には *20000* 人の住人がいて、このうちある病気の免疫を持っていない人の割合は *0.4* である。この病気になる人の割合を全体の *1* 割以下にする確率が *0.9* をこえるためには、ワクチンを少なくとも何個用意しなければならないか。

第7章　記述統計

7.0.15　記述統計の イントロダクション

例:一般（ランキング）
(i) 食物・味付けの分布　(display-2-b-00.pdf 後半)
(ii) 棒グラフ (display-1.pdf)

ある集団を**母集団**とよぶ。ここからいくつかのデータをとるとき、これを**標本**とよぶ。母集団から抽出された標本はさまざまな値をとる。そこで、この標本１つ１つを**変数**とよぶ。変数が離散的な数値（ばらばらの値）をとるとき、**離散型変数**という。変数が連続的な値をとるとき、**連続型変数**という。

例 16 世帯の子供の数、*1*日の交通事故の件数、窓口を訪れる客の数は離散型変数である。長さ、重さ、時間、温度は連続型変数である。

なお、データが数値を取らない場合もある。
例：血液型の分布 (display-2-b-00.pdf 前半)

7.0.16　分類と度数分布

あることを調査しようとするとき、資料を集めて分類することからはじめることが多い。さらに、得られたデータが数値で表されている場合には、度数分布を見るとわかりやすい。

例 17 *143*人の通学時間を調べた。$x_1, ..., x_{143}$ がデータであり、x_i は i さんの通学時間を表す。
(1) $\max_{1\leq i\leq n} x_i$ を**最大値**、$\min_{1\leq i\leq n} x_i$ を**最小値**とよび、$(\max_{1\leq i\leq n} x_i) - (\min_{1\leq i\leq n} x_i)$ を**範囲**とよぶ。

(2) データをいくつかのグループにわけるとき、各グループを**階級**という。階級の範囲を階級の幅という。また、各階級の真ん中の値を**階級値**という。

階級の数は通常 *10* 個程度にし、階級の幅はデータの測定単位 *(*精度*)* の整数倍になるようにとる。

(3) 各階級に入るデータの個数を**度数**という。度数がわかると**度数分布表**がつくれる。

度数分布表を図にしたものが**ヒストグラム**である。ヒストグラムによりデータの分布の形が直観的に把握できる。また、ヒストグラムを折れ線で結んでえられるものを**度数折れ線**という。

これらの処理は表計算ソフト（エクセルなど）によって行うことができる。またこれらを連続変数の場合に行ったものが確率密度分布である。

累積度数分布表、累積度数分布曲線も同様に作れる。

例:新生児の体重の累積度数分布表 (display-2-b-00.pdf 前半)

例：薬累積度数分布曲 i 線は薬の効力の記述や毒性の評価などにも利用される。投与量を対数曲線上にとるか、あるいは投与量の対数値を算術目盛上にとり、投与量を少しずつ増加したとき投与量に対する反応率は一般に S 字型の累積度数分布曲線を示す。このような曲線を用量–反応曲練という (図 3・5)

7.0.17　代表値

ある集団の特徴をひとつの数値で代表するとき、これを統計量（または代表値）という。例えば、平均値、中央値、最頻値などがこれにあたる。$x_1, ..., x_n$ をデータとする。

（1）$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ を**平均値**という。たんに**平均**ということもある。

例 18 $n = 5,\ x_1 = 10, x_2 = 2, x_3 = 5, x_4 = 7, x_5 = 4$ のとき、$\bar{x} = \frac{1}{5}(10 + 2 + 5 + 7 + 4) = \frac{28}{5} = 5.6$。

（2）x_i を大きさの順に並べた時、その中央に位置する値を**中央値（メディアン）**といい、$\tilde{x}$ とかく。これは頻度が五分五分の位置にある x_i の値といってもよい。

標本数 n が奇数のときは、ちょうど $\frac{n+1}{2}$ 番目の値が中央値になる。n が偶数のときは、$\frac{n}{2}$ 番目と $\frac{n}{2} + 1$ 番目の 2 つの値の平均値が中央値になる。

例 19 $n = 5, x_1 = 4500, x_2 = 13300, x_3 = 9500, x_4 = 18600, x_5 = 7200$ のとき、$\tilde{x} = 9500$。

例：新生児の体重

（3）もっとも頻繁に現れる（度数の大きい）値を**最頻値（モード）**という。(à la mode.) 多くの場合、最頻値は度数分布表に整理したうえで考える。

また、集団の分布の特徴を表す数値を散布度という。範囲、**四分位範囲**、**標本分散**、**標準偏差**などがこれにあたる。

ここで、四分位範囲とは次のようにして決まる：中央値を境にしてデータの個数が等しくなるように２つの部分に分ける。

２つに分けるたうち、最小値を含む方のデータの中央値を求める。これを**第１四分位数**という。同様に、最大値を含む方のデータの中央値を求める。これを**第３四分位数**という。

第３四分位数から 第１四分位数までの範囲（または、(３四分位数) - (第１四分位数)）を**四分位範囲**という。四分位範囲の半分の大きさを四分位偏差という。

なお、**箱ひげ図**という統計的グラフを使って、範囲、四分位範囲、平均値、中央値を同時に簡潔に表すことができる。また、箱ひげ図を使うと、平均値や中央値が等しくてもデータの分布の様子が大きく異なることがあることが、よくわかる。箱ひげ図を使ってデータのばらつきを視覚的にとらえることもできる。

この他、医療関係では幹葉図も用いられる。幹業図は、生データを幹と楽に対応する数に分けて分布を作成する方法でヒストグラムと違って生データを再現できる。

例 20 東京都における、各月の、日ごとの平均気温。

（4）$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ を**標本分散**という。$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ を**不偏標本分散**という。$\sigma = \sqrt{s^2}$ を**標準偏差**という。

例 21 *(1)* の例では $\bar{x} = 5.6$,

$$s^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - 5.6)^2 = \frac{37.2}{5} = 7.44$$

$$s_{n-1}^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 5.6)^2 = 9.3$$

（5）$\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$ を**平均偏差**という。

問 28 次のデータはある高校の運動選手 *6* 人の身長である（単位 cm）。このとき標準偏差を小数第 *2* 位まで求めよ*:*

174　169　175　171　168　181

7.0.18　平均値と標準偏差の計算

性質

（1）$u_i=x_i-a, i=1,...,n$ とすると、$\bar{u}=\bar{x}-a, s^2(u)=s^2(x), s^2_{n-1}(u)=s^2_{n-1}(x)$

（2）$v_i=bx_i, i=1,...,n$ とすると、$\bar{v}=b\bar{x}, s^2(v)=b^2s^2(x), s^2_{n-1}(v)=b^2s^2_{n-1}(x)$

例題 16 あるアルバイトの日給は、実労働時間を x 時間とすると、$1000+800x$（円）で計算される。標本調査によると、そのアルバイトをしている人の *1* 日の実労働時間の平均は *7.5*、標準偏差は *1* である。日給の平均と標準偏差はいくらか。

解 16 $u=1000+800x$ とおくと、$\bar{u}=1000+800\bar{x}=1000+800\times 7.5=7000$。 $s^2(u)=(800)^2s^2(x)=(800)^2.1^2$。よって、$s(u)=800$。

度数分布表からの平均値と標準偏差の計算

次の表はある高校の篭球部の部員 10 人でフリースローを 1 回づつ行った結果である。

点数 x	度数 f	xf	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2f$
0	2	0	-2	4	8
1	3	3	-1	1	3
2	1	2	0	0	0
3	2	6	1	1	2
4	1	4	2	4	4
5	1	5	3	9	9
計	10	20		19	26

この表から点数の標準偏差を計算してみる。

xf の値は上のようになるから、点数の平均値 $\bar{x}$ は $\bar{x} = 20/10 = 2$ 点。さらに $(x-\bar{x})^2 f$ の値は上のようになるから、点数の標本分散は $s^2 = 26/10 = 2.6$。よって標準偏差は $s = \sqrt{2.6} \sim 1.61$ 点となる。

7.0.19　2 次元データのまとめ方

１つの対象について２つの質的変数を調べる。
例：喫煙の習慣の有無と肺がんの有無
薬物の服用の有無と症状
支持政党と内閣の支持・不支持
...
2 つの変数からなる大きさ M の標本を考えたとき、身長 (cm) と体重 (kg)，薬剤投与 (投与群，非投与群) と最低血圧値、薬剤投与 (新薬群，偽薬群) と効果 (効来あり、効果なし) のようなそれぞれのタイプの 2 つの変数の組の関連を図示するにはどのようにしたら良いだろうか。

座標平面 XY を考えて、各個体の X 変数 Y 変数を座標平面上の点 (X, Y) として表すと座標平面上に n 個の点が図示される。この図を**散布図**という (図 3・9) 関連性がある場合は、X の値の変化に応じて Y の値が変化することから、関連性の傾向を示すことができる。
例：一般　(display-1.pdf)

例 22 語学とコンピュータ科目の成績

例 23 ある商品の *CM*（*X)* とその商品の購入（*Y)*

例 24 乗用車による重大事故について、ドライバーの生死（*X)* とシートベルト着用の有無（*Y)*

7.0.20　特論　保健婦（保健師）の統計

日本では保健師を保健師助産師看護師法（以下、保助看法と記す）において、「厚生労働大臣の免許を受けて、保健師の名称を用いて、保健指導に従事することを業とする者」と定めており、大学や保健師養成校に

て所定の教育を受けた後、保健師国家試験に合格して得られる国家資格（免許）である。

保健師は、主に都道府県・市町村などの保健所、保健センター等で保健行政に従事する行政保健師と企業の産業保健スタッフとして勤務する産業保健師、学校等で学生と教職員の心身の健康保持に努める学校保健師（養護教諭）の3つに大別される。

保健師助産師看護師法は、保健師・助産師及び看護師の資質を向上し、もって医療及び公衆衛生の普及向上を図ることを目的とする日本の法律である（同法1条）。

用語

1.1 有病率

有病とは、ある時点で疾病にかかっている状態のことである。

$$\text{有病率} = \frac{\text{ある時点における有疾病者の人数}}{\text{ある時点における観察対象者の人数}}$$

と定義する。

1.2 罹患率

罹患とは、**一定期間に**新たにある疾病を発症することである。

罹患率　をどう定義するか。

ある地域の人間集団は開集団であり、一定期間の人口には増減がある。そこで

$$\text{罹患率} = \frac{\text{一定期間に新たに疾病を発症した人の人数}}{\text{1人1人の観察期間を全部足したもの}}$$

と定義する。

要因の分析（ベイズの定理）

2.1 暴露

疫学で言う**暴露**とは、疾病の発症にかかわっているかもしれない要因のことである。「性別」、「年齢」、「生活習慣」いずれも暴露（要因）である。健康に良い影響を与えるもの、健康に悪い影響を与えるもの、のいずれにも使う。

例:喫煙の習慣。

2.2 発症率

$$\text{発症率} = \frac{\text{疾病を発症した人の人数}}{\text{(一定期間に) ある要因に暴露された人数}}$$

例：食堂で発生した食中毒について、その原因を推定する。

食べた・食べない	発症	未発症	発症	未発症	計
パン	96	200	4	20	320
牛乳	95	150	5	70	320
焼き魚	50	100	50	120	320
合計	241	450	59	210	960

(1) 食べた人の発症率

$$\text{パンの発症率} = \frac{\text{パンを食べて食中毒を発症した人の人数}}{\text{パンを食べた人数}} = \frac{96}{296} = 0.32$$

$$\text{牛乳の発症率} = \frac{\text{牛乳を飲んで食中毒を発症した人の人数}}{\text{牛乳を飲んだ人数}} = \frac{95}{245} = 0.39$$

$$\text{焼き魚の発症率} = \frac{\text{焼き魚を食べて食中毒を発症した人の人数}}{\text{焼き魚を食べた人数}} = \frac{50}{150} = 0.33$$

(2) 食べてない人の発症率

$$\text{パンの発症率} = \frac{\text{パンを食べてなくて食中毒を発症した人の人数}}{\text{パンを食べてない人数}} = \frac{4}{24} = 0.17$$

$$\text{牛乳の発症率} = \frac{\text{牛乳を飲んでなくて食中毒を発症した人の人数}}{\text{牛乳を飲んでない人数}} = \frac{5}{75} = 0.07$$

$$\text{焼き魚の発症率} = \frac{\text{焼き魚を食べてなくて食中毒を発症した人の人数}}{\text{焼き魚を食べてない人数}} = \frac{50}{170} = 0.29$$

ここで**オッズ比**（[食べた人の発症率]/[食べてない人の発症率]）を見ると
パン 1.88
牛乳 5.57
焼き魚 1.14
となり、牛乳が一番危険度（リスク比）が高いことがわかる。

2.3 データの誤差

症例対照研究

症例群：研究の対象となる疾病を発症した人たち
対照群：発症してない人たち
このような人たちを集めてきて、両群の過去の曝露状況を調べる
疾病を発症しているかどうかという違い以外は、両群が同じような人により構成されていることが重要である。

3.1 紅茶は花粉症を予防するか？(上表参照)

紅茶を飲む・飲まない	花粉症あり	花粉症なし	計
飲む	8	16	24
飲まない	32	24	56
合計	40	40	80

3.1.2 ノイズ（バイアス）
ノイズには、**偶然誤差**と**系統誤差**がある。ここで、系統誤差とは一定の方向にデータが偏ってしまうことをいう。対象者の集め方、データの集め方によってはこのような偏りが生じる。系統誤差を**バイアス**ともいう。
この例におけるバイアス
対照群の人たちに、「紅茶は花粉症の発症を予防するかもしれない」と説明したうえで、「あなたは紅茶を飲んでいましたか」と聞くとする。その人は「自分がいま花粉症出ないのは紅茶を飲んでいたせいかもしれない」と考えて、「習慣的に紅茶を飲んでいた」という人が出てくる可能性がある。また現在花粉症の人は、「紅茶を飲んでいなかったから花粉症に

なった」と考えて、紅茶を飲んでいた記憶を思い出させにくくなる可能性がある。このようなバイアスを**思い出しバイアス**という。

3.2 オッズ比

オッズとは、　ある事象が起きた割合と起きなかった割合の比のことを言い、オッズ比とは、オッズを 2 つの群で比較した比である。

例：紅茶と花粉症

症例群　$(8/40)$　/ $(32/40) = 1/4$

対照群　$(16/40)$ / $(24/40) = 16/24 = 2/3$

オッズ比 $(1/4)$ / $(2/3) = 3/8$

この結果から

「花粉症の有無は紅茶を飲む習慣に 3/8 倍影響する」

といえる。

行について同じことをすれば、

症例群　$(8/24)$　/ $(16/24) = 1/2$

対照群　$(32/56)$ / $(24/56) = 32/24 = 4/3$

オッズ比 $(1/2)$ / $(4/3) = 3/8$

この結果から

「紅茶を飲む習慣の有無は花粉症に 3/8 倍影響する」

といえる。

オッズ比の意味

オッズ比 $= 1$ 曝露因子と疾病発症に関係はない

オッズ比 > 1 曝露因子は疾病発症の危険因子である

オッズ比 < 1 曝露因子は疾病発症の軽減因子である

7.0.21　相関関係

例 25 円盤投げと砲丸投げの得点分布

例 26 醤油消費とソース消費

例 27 都道府県別の若年未婚率と生涯未婚率

このような図を**散布図**（または相関図) という。

散布図を見て、一方の変数の値が大きくなるとき他方の変数の値が直線的に大きくなる傾向があるとき、２つの変数の間には**正の相関がある**という。また、一方の変数の値が大きくなるとき他方の変数の値が直線的に小さくなる傾向があるとき、２つの変数の間には**負の相関がある**という。相関関係が見られないとき、**無相関**という。

例 28 *1* 人当たりの国内家計支出と世帯でのピアノの普及率

なお、データ全体からみて、分布から極端に離れているデータを**外れ値**という。

２つの変数 (x,y) の**相関係数** r_{xy} を次のように定義する。データ数を n とする。

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, s_x = \sqrt{s_x^2}$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, s_y = \sqrt{s_y^2}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

とおくとき、

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

を、**相関係数**という。なお、s_{xy} を x と y の**共分散**という。

$-1 \leq r \leq 1$ である。また、x と y に正の相関があるとき $r > 0$ であり、x と y に負の相関があるとき $r < 0$ である。

なお、x と y に正の相関があるからといって、両者の間に**因果関係**があるとは限らない。

例 29 人がある健康食品を買う回数と、風邪をひく回数に正の相関がみられたとする。これから「その健康食品を食べると、風邪をひきやすくなる」と判断してよいか？

例 30 ある学部で、クラスの中でメガネをかけている学生の割合と、学力テストのクラス平均の結果に正の相関がみられたとする。これから「メガネをかければ頭がよくなる」と判断してよいか？

納豆、サカナサカナサカナ。。。
また、場合によっては外れ値を取り除いて相関係数を計算しなおしてみるとよい結果を得られることがある。

例 31 肥満比率と心臓病死亡者数 $(r = 0.3658, 0.3423)$

7.0.22 回帰直線

２つの変量の間に強い正の相関があるとき、データは散布図の上である直線にそってばらついていると考えられる。そこで変量 x と y の間に

$$y = bx + a$$

という１次関係を想定する。この直線を**回帰直線**という。
回帰直線を求めるために最小２乗法を用いる。すなわち、データ $(x_i, y_i), i = 1, ..., n$ に対し、変量

$$S(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

を考え、$S(a, b)$ を最小にする (a, b) を求める。
そのため、a, b を α, β と書き、$S(\alpha, \beta)$ を α, β について偏微分して極値を求める。つまり、a, b は連立１次方程式

$$a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

の解である。

問　題

問 29 次の数値は健常成人 *50* 人の血糖値データ *(g/dl)* である。このデータをもとに度数分布表を作れ。また、箱ひげ図を作成せよ。

108, 94, 87, 88, 91, 81, 86, 87, 87, 87
79, 92, 87, 97, 84, 86, 96, 89, 96, 93
92, 89, 79, 94, 79, 70, 82, 89, 93, 89
88, 114, 93, 102, 92, 82, 109, 98, 93, 107
96, 120, 100, 103, 87, 87, 82, 107, 74, 85

問 30 売り上げ

150 円、*150* 円、*155* 円、*155* 円、*155* 円、*160* 円、*160* 円、*160* 円、*160* 円、*165* 円

の平均値、中央値、分散をもとめよ。

問 31 次の数値はある病院における *10* 人の患者の血圧データ *(mmHg)* である。このデータの箱ひげ図を作成せよ。

118, 148, 128, 141, 139, 120, 125, 123, 134, 144

問 32 次の数値はある都市における *10* 日間の最高気温 *(° C)* である。これから平均値、分散、最頻値、中央値を求めよ。

26, 27, 27, 28, 28, 28, 28, 27, 26, 25

問 33 次の表（次ページ）は A 組と B 組における英語の試験の結果である。

（1）各組の平均値 m_A, m_B および標準偏差 s_A, s_B を求めよ。

（2）各組について、$m_A - s_A$ と $m_A + s_A$ の間、および、$m_B - s_B$ と $m_B + s_B$ の間の点数をとる人数をそれぞれ求め、さらにこれらの全体に対する割合を言え。

組	点数	度数
A	0	0
A	10	1
A	20	1
A	30	4
A	40	3
A	50	4
A	60	6
A	70	8
A	80	9
A	90	3
A	100	1
B	0	0
B	10	0
B	20	2
B	30	6
B	40	2
B	50	7
B	60	8
B	70	4
B	80	7
B	90	1
B	100	3

問 34 組 A,Bで試験を行ったところ下のようになった。このとき生徒全体の平均値と標準偏差を求めよ。

組	人数	平均	標準偏差
A	45	58	20
B	55	60	22

問 35 A 組は p 名からなり平均点は a 点、B 組は q 名からなり平均点は b 点であった。両組をあわせての平均を式で表せ。

問 36 次の表（次ページ）は世界各地の年平均気温と人口 1 万人あたりの死亡率である。気温と死亡率との相関係数をもとめよ。

地名	気温	死亡率
ストックホルム	*7*	*11*
ベルリン	*9*	*13*
東京	*14*	*13*
長崎	*16*	*16*
カイロ	*21*	*24*
マドラス	*28*	*36*

問 37 変量 x と y を 5 人で測定した結果が次の表のようになった。

番号	x	y
1	*1*	*5*
2	*2*	*7*
3	*3*	*1*
4	*4*	*3*
5	*5*	*9*

(1) x と y の平均と標準偏差を各々求めよ。
(2) x と y の相関係数を求めよ。
(3) x と y の回帰直線を求めよ。

第8章　推定

8.0.23　推定の一般論

ある集団を母集団とよぶ。ここから n 個のデータ $X_1, ..., X_n$ をとる。これを**標本**とよぶ。標本から全体の特徴をしようとすることを**標本調査**という。また、母集団から標本をぬき出すことを**抽出**といい、母集団に含まれる要素の個数を**標本の大きさ**という。

母集団から偏りのない標本を抽出するためには、乱数さい（とよばれる特別なさいころ）やコンピュータに発生させた乱数などを用いる。このような標本の抜き出し方を**無作為抽出**という。なお、母集団から標本を抽出するとき、抽出のたびに要素を元に戻し、あらためて次を抽出する方法を**復元抽出**という。一方、元に戻さないで続けて抽出する方法を**非復元抽出**という。

$X_1, ..., X_n$ はどのような値を取るか事前にはわからないので、これを独立同分布な確率変数とみて、$E[X_i] = m, i = 1, 2, ..., n$、$V(X_i) = v, i = 1, 2, ..., n$ とする。ただし、m の値は未知。m を**母平均**、v を**母分散**とよぶ。

例 32 母集団の総数を N、標本数を n と表す。
(1) 家計調査 $N =$ 約 *2700* 万、$n =$*8000*
(2) 労働力調査　$N =$ 約 *1.2* 億、$n =$*40000*
(3) 賃金構造基本調査　$N =$ 約 *3000* 万、$n =$*140* 万
*(*厚労省の賃金統計の集計に不正が発覚し、数値が削除された公表資料毎月勤労統計調査 *(2009/1))*

問題はデータ $X_1, ..., X_n$ から母平均 m（または母分散 v）を推定することである。なお、母分散の正の平方根を**母標準偏差**という。

8.0.24　点推定

問題はデータ $X_1, ..., X_n$ から母平均 m を推定することである。

$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$

を**標本平均**という。

ここで、標本平均を使って母平均を推定する。つまり、データの値が $X_i = x_i, i = 1, ..., n$ のとき、m の推定値を

$$\bar{X}|_{X_1=x_1,...,X_n=x_n} = \frac{x_1 + ... + x_n}{n}$$

とする。

ここで、

$$E[\bar{X}] = \frac{1}{n}E[X_1 + ...X_n] = \frac{1}{n}nm = m$$

である。また、$X_1, ..., X_n$ は独立であるから

$$V(\bar{X}) = \frac{1}{n^2}(V(X_1) + ... + V(X_n)) = \frac{n}{n^2}v = \frac{1}{n}v$$

である。

このようにデータの関数としてある値を推定する方法を**点推定**という。

点推定の例

47 都道府県にある映画館の合計スクリーン数の点推定を行ってみます。47 都道府県全てのデータを調べるのは面倒なので、無作為に 10 都道府県のデータを抽出しました。

これら 10 都道府県にある映画館の合計スクリーン数の平均は「93.8」でした。この「93.8」を 47 都道府県（＝母集団）の平均値（つまり母平均）と見なしてしまおう、とするのが母平均の点推定です。

点推定は簡便であるが次のような推論のリスクがある。

例 33 *10*円玉を n 回投げて表の出る回数を数える試行をおこなう。$X_i = 1$（表が出た場合）、$X_i = 0$（裏が出た場合）とし、$X = X_1 + ... + X_n$ とおく。

場合 1　$n = 10$ で *5* 回表が出た場合

$\bar{x} = \bar{X}|_{X_1=x_1,...,X_n=x_n} = \frac{5}{10} = \frac{1}{2}$ である。

場合 2　$n = 10000$ で *4998* 回表が出た場合

$\bar{x} = \bar{X}|_{X_1=x_1,...,X_n=x_n} = \frac{4998}{10000} \sim \frac{1}{2}$ である。

２つの場合の推定値はともにおよそ $\frac{1}{2}$ であるが、推定の精度が高いのは場合２のほうであろう。点推定ではこの違いを結果に反映させることができない（誤差評価がない）。

点推定の例（I） - 不偏推定量

ある値 θ の推定値としてデータのある関数 $T(X_1, ..., X_n)$ の実現値をとる。上の例では $\theta = m$ であり、$T(x_1, ..., x_m)$ が θ の推定値である。

（１）$\lim_{n\to\infty} T(X_1, ..., X_n) = \theta$ となるものを**一致推定量**という。

（２）$E[T(X_1, ..., X_n)] = \theta$ となるものを**不偏推定量**という。

一致推定量については、左辺の n をいくつにすれば θ との誤差がいくつ以下になるか、という判断基準が一般にはない。(定量評価ではない。)

不偏推定量については、左辺の平均が θ になると言っているだけで、各々の $T(X_1, ..., X_n)$ の値は θ に近い保証はない。

例 34 上の例では $T(X_1, ..., X_n) = \frac{X_1+...+X_n}{n} = \bar{X}$ ととると、

$$E[T(X_1, ..., X_n) = \frac{1}{n}(E[X_1] + ... + E[X_n])$$

$$= \frac{1}{n}(m + ... + m) = m$$

であるから、$\bar{X}$ は m の不偏推定量である。

一方、v の不偏推定量は

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

である。これを**不偏標本分散**という。

証明

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

とおく。$E[S^2] = \frac{n-1}{n}v$ を示す。そうすれば

$$E[\hat{S}^2] = E[\frac{n}{n-1}S^2] = v$$

であることがわかる。

$$S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2 = \frac{1}{n}\sum_i (X_i - m + m - \bar{X})^2$$

$$= \frac{1}{n}\sum_i (X_i - m)^2 + \frac{2}{n}(m - \bar{X})\sum_{i=1}^{n}(X_i - m) + (m - \bar{X})^2$$

ここで第 2 項 $= 2(m - \bar{X})(\frac{1}{n}\sum_i X_i - \frac{n}{n}m)$

$$= 2(m - \bar{X})(\bar{X} - m) = -2(m - \bar{X})^2$$

これより

$$S^2 = \frac{1}{n}\sum_i (X_i - m)^2 - (m - \bar{X})^2$$

したがって

$$E[S^2] = E[\frac{1}{n}\sum_i (X_i - m)^2] - E[(m - \bar{X})^2]$$

$$= \frac{1}{n} \cdot n \cdot E[(X_i - m)^2] - V(\bar{X})$$

$$= v - \frac{v}{n} = \frac{n-1}{n}v$$

ここで $V(\bar{X}) = \frac{1}{n}v$ であることを使った。

例題 (点推定の例)：

無作為に抽出された 10 都道府県の合計スクリーン数のデータから全国のスクリーン数の不偏分散を求めよ。

1 兵庫 126 2 大阪 224 3 奈良 34 4 岩手 25 5 千葉 199 6 茨城 89 7 福岡 178 8 山梨 14 9 滋賀 38 10 鳥取 11

cf. 愛媛　59

47 都道府県にある映画館の合計スクリーン数の点推定を行ってみる。47 都道府県全てのデータを調べるのは面倒なので、無作為に 10 都道府県のデータを抽出した。

これら 10 都道府県にある映画館の合計スクリーン数の平均は「oo.o」でした。この「oo.o」を 47 都道府県（＝母集団）の平均値（つまり母平均）と見なしてしまおう、とするのが母平均の点推定である。

練習問題

(1) ある中学校の 1 年生 300 人からランダムに選んだ 10 人の英語のテストの点数は次のとおりである。この結果から、学年全体の平均点と不偏分散を求めよ。

データ：80　75　40　100　95　55　80　85　70　65

(2) ある中学校の 1 年生 300 人からランダムに選んだ 10 人の 50m 走のタイムは次のとおりである。この結果から、学年全体の平均タイムと不偏標準偏差を求めよ。

データ：7.0　6.2　8.3　10.0　9.1　6.8　7.4　8.5　9.2　7.9

点推定の例（II）　-　最尤法

未知のパラメータ θ（$\theta = m, v$ など）を含む n このデータ $x_1, ..., x_n$ が与えられたとする。さらに各々の x_i が出現する確率は $p(x_i; \theta)$ とする（既知）。このとき

$$L(\theta) = p(x_1; \theta) \cdots p(x_n; \theta)$$

を**尤度関数**（ゆうどかんすう）という。

基本的な考え方：「現実の標本は θ の値に関して最大確率となるものが出現した」と考える。そこで $L(\theta)$ を最大にするような θ の値 $(= \hat{\theta})$ を θ の推定値とする。このような $\hat{\theta}$ を θ の**最尤推定量**（さいゆうすいていりょう）という。$L'(\theta) = 0$ を**尤度方程式**という。

例題 17 インチキな硬貨を *5* 回投げたら、

$$H, H, H, H, T$$

とでた。この硬貨を *1* 回投げて H がでる確率 θ を推定せよ。

解 17

$$p(H; \theta) = \theta, p(T; \theta) = 1 - \theta$$

であるから、$L(\theta) = \theta^4(1 - \theta)$。これを最大にする θ を $\hat{\theta}$ とおくと、

$$L'(\theta) = 4\theta^3(1 - \theta) + \theta^4(-1)$$

$$= \theta^3(4 - 4\theta - \theta) = \theta^3(4 - 5\theta)$$

よって

$$L'(\theta) = 0 \quad \Longleftrightarrow \quad \theta = \frac{4}{5}$$

よって $\hat{\theta} = \frac{4}{5}$。

例題 18 袋の中に白球ばかりが何個か入っている。いま *50* 個取り出し、それを赤球に替えて元に戻す。よくかき混ぜた後に *50* 個取り出したところ、そのうち赤球は *5* 個であった。この袋に初めにあった白球の数を推定せよ。

解 18 初めにあった白球の数を N とする。$\theta = N$ であり、尤度関数は

$$L(N) = \frac{{}_{50}C_5 \cdot {}_{N-50}C_{50-5}}{{}_{N}C_{50}}$$

となる。$L(N)$ を最大にする N を求めるため、$L(N)$ と $L(N-1)$ の比をとると

$$\frac{L(N)}{L(N-1)} = \frac{{}_{50}C_5 \cdot {}_{N-50}C_{50-5} \cdot {}_{N-1}C_{50}}{{}_{50}C_5 \cdot {}_{N-50-1}C_{50-5} \cdot {}_{N}C_{50}} = \frac{(N-50)(N-50)}{N(N-50-50+5)}$$

となる。

これより

- $N \leq 50 \cdot 50/5$ のとき $L(N)$ は増加
- $N = 50 \cdot 50/5$ のとき $L(N)$ は最大
- $N \geq 50 \cdot 50/5$ のとき $L(N)$ は減少

であることがわかる。よって $\hat{N} = 50 \cdot 50/5 = 500$ ないし $500 - 1 = 499$ が最尤推定量である。

練習問題

(A)

(1) データ $X_1, ..., X_n$ が独立に正規分布 $N(m, \sigma^2)$ から得られているとする.

(a) m の値は $m = 0$ で既知のとき σ^2 を推定する．分散の範囲が $1 \leq \sigma^2 \leq 2$ であることが分かっているとき，σ^2 の最尤推定量 を求めよ.

(b) σ^2 の値は既知として m を推定する．期待値 m の範囲が $0 \leq m \leq 1$ であることが分かっているとき，m の最尤推定量 を求めよ.

(2) データ $X_1, ..., X_n$ が独立に指数分布 $\mathrm{Exp}(\lambda)$ にしたがっているとする．このときパラメータ λ の最尤推定量 を求めよ.

(B) 次の各場合データ $X_1, ..., X_n$ は独立で [oo] 分布にしたがうとする。各場合について尤度関数と最尤推定量をもとめよ。

(I)[二項分布]

(1) 10 人中 3 人に発生したとき。

(2) 100 人中 20 人に発生したとき。

(3) 30 人中 5 人に発生したとき。

(4) 20 人中 13 人に発生したとき。

(5) 400 人中 7 人に発生したとき。

(6) n 人中 10 人に発生したとき

(II)[指数分布]

(1) 1 人のデータ $x_1 = 10$ が得られたとき。

(2) 2 人のデータ $x_1 = 20; x_2 = 22$ が得られたとき。

(3) 3 人のデータ $x_1 = 20; x_2 = 30; x_3 = 40$ が得られたとき。

(4) 4 人のデータ $x_1 = 0; x_2 = 1; x_3 = 2; x_4 = 5$ が得られたとき。

(5) 5 人のデータ $x_1 = 3; x_2 = 20; x_3 = 50; x_4 = 18; x_5 = 9$ が得られたとき。

(III)[正規分布（分散既知）]

(1) 1 人のデータ $x_1 = 100$ が得られたとき ($\sigma = 5$)。

(2) 2 人のデータ $x_1 = 70; x_2 = 90$ が得られたとき ($\sigma = 13$)。

(3) 3 人のデータ $x_1 = 80; x_2 = 130; x_3 = 150$ が得られたとき ($\sigma = 14$)。

(4) 4 人のデータ $x_1 = 100; x_2 = 120; x_3 = 200; x_4 = 120$ が得られたとき ($\sigma = 10$)。

(5) 5 人のデータ $x_1 = 90; x_2 = 60; x_3 = 70; x_4 = 120; x_5 = 110$ が得られたとき ($\sigma = 15$)。

8.0.25　区間推定

1.1　母平均 m の推定

1．1．1　母分散が既知の場合

仮定：母分散 v の値は既知

$X_1, ...X_n$(データ) は独立･同分布の確率変数とする。$E[X_i] = m, V(X_i) = v, i = 1, ..., n$。

$\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$ とおくと、$E[\bar{X}] = m, V(\bar{X}) = \frac{1}{n}v$ であった。これより、$\frac{\bar{X}-m}{\sqrt{\frac{v}{n}}}$ は平均 0、分散 1 である。中心極限定理より、$n \to \infty$ のとき、$\frac{\bar{X}-m}{\sqrt{\frac{v}{n}}}$ の分布は標準正規分布に収束する。つまり、任意の $a, b, a < b,$ に対し

$$\lim_{n\to\infty} P(a < \frac{\bar{X}-m}{\sqrt{\frac{v}{n}}} < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

これより、n が大きいとき、確率 $P(a < \frac{\bar{X}-m}{\sqrt{\frac{v}{n}}} < b)$ は、

$$I(a,b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

に近い。

これを使って、データ $X_1 = x_1, ..., X_n = x_n$ が与えられたとき、確率変数 $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$ が $a < \frac{\bar{X}-m}{\sqrt{\frac{v}{n}}} < b$ を満たすという事象の確率を以下のように推定できる。

例 35 *1* から *999* までの間の奇数の乱数を *20* 個発生させて、数値の標本平均 $\bar{X}$ を計算する。この場合、母集団は *1* から *999* までの奇数であるから母平均 500、母分散 289^2 である *(***問**これを示せ*)*。よって

$$E[\bar{X}] = 500, \quad V(\bar{X}) = \frac{289^2}{20} \sim 65^2$$

である。$n = 500$ は十分大きいので、$a = -1.96, b = 1.96$ ととると、

$$P(-1.96 < \frac{\bar{X} - 500}{\sqrt{\frac{289^2}{20}}} < 1.96) \sim I(-1.96, 1.96) = 0.95$$

となる。

$I(a,b)$ で、a と b は近いほうが m の推定値としてシャープであるが、そういう確率は小さい。一方、$a \to -\infty, b \to \infty$ とすれば確率を上げることができるが、そうすると $a < \frac{\bar{X}-m}{\sqrt{\frac{v}{n}}} < b$ という主張は意味がない。

そこで妥協策として、$a = -1.96, b = 1.96$ ととると、

$$I(a,b) \sim 0.95$$

である。そこで次のように考える。

$$(1)\ -1.96 < \frac{\bar{X}-m}{\sqrt{\frac{v}{n}}} < 1.96 \ \text{となる確率はほぼ}\,0.95\,\text{である}$$

(2) したがって

$$m - 1.96\sqrt{\frac{v}{n}} < \bar{X} < m + 1.96\sqrt{\frac{v}{n}}$$

となる確率はほぼ 0.95 である。

ここで得られたデータ $X_1 = x_1, ..., X_n = x_n$ は、この確率 0.95 の事象が出現した結果だと思おう。（本当にそうかどうかはわからない。そこは**ギャンブル**である。ここではデータ $X_1 = x_1, ..., X_n = x_n$ は確率 0.95 の方のデータだと信じる（賭ける）ことにする。）

（3）こうすると

$$(*) \qquad m - 1.96\sqrt{\frac{v}{n}} < \bar{x} < m + 1.96\sqrt{\frac{v}{n}}$$

つまり

$$(*') \qquad \bar{x} - 1.96\sqrt{\frac{v}{n}} < m < \bar{x} + 1.96\sqrt{\frac{v}{n}}$$

が得られる。このことを、統計では「**信頼係数 95% で母平均 m の信頼区間は $[\bar{x} - 1.96\sqrt{\frac{v}{n}}, \bar{x} + 1.96\sqrt{\frac{v}{n}}]$ である**」という。なお信頼係数のことを**信頼度**ともいう。これは、粗く言えば「その区間推定が正しい確率」である。

やや詳しく言えばつぎのようになる。$\bar{X}$ は変数であるから、$(*')$ の区間は $\bar{X}$ の実現値 $\bar{x}$ の値によって変わってくる。しかし 95% の確率でこの区間は m を含む、つまり $(*')$ が成り立つ、といえる。

推論のリスクをもっと減らしたい場合には、a, b を 0 からもっと離してとればよい。たとえば、$a = -2.58, b = 2.58$ ととれば、$I(a, b) = 0.99$ である。よって「信頼係数 99% で母平均 m の信頼区間は $[\bar{x} - 2.58\sqrt{\frac{v}{n}}, \bar{x} + 2.58\sqrt{\frac{v}{n}}]$ である」ということになる。この場合、区間の長さはより長くなる。しかし、あまり長いと意味がない。

なお、(*), (*') における信頼区間の長さは $\sqrt{\frac{1}{n}}$ に比例する。つまり同じ信頼係数で信頼区間の長さを $\frac{1}{10}$ にするには、データを 100 倍取らないといけない。

例題 19 松山市のある水田に植えられた稲の穂 *100* 本について、１穂あたりの米粒数を数えたところ標本平均 *71.7* 粒であった。*1* 穂あたりの粒数の分散は $(19.1)^2$ であることがわかっている。この水田の稲の１穂あたりの粒数を信頼係数 *95*% で推定せよ。また信頼係数 *99*% の場合はどうなるか。

解 19 $\bar{x} = 71.7, v = (19.1)^2, n = 100$ である。信頼係数 *95*% なので、信頼区間は

$$[\bar{x} - 1.96\sqrt{\frac{v}{n}}, \bar{x} - 1.96\sqrt{\frac{v}{n}}] = [71.7 - 1.96\sqrt{\frac{(19.1)^2}{100}}, 71.7 + 1.96\sqrt{\frac{(19.1)^2}{100}}]$$

$$= [71.7 - 1.96\frac{19.1}{10}, 71.7 + 1.96\frac{19.1}{10}] = [71.7 - 1.96 \times 1.91, 71.7 + 1.96 \times 1.91]$$

$$= [67.96, 75.44]$$

である。

同様にして、信頼係数 *99*% の場合には、

$$[\bar{x} - 2.58\sqrt{\frac{v}{n}}, \bar{x} + 2.58\sqrt{\frac{v}{n}}] = [71.7 - 2.58 \times 1.91, 71.7 + 2.58 \times 1.91]$$

$$= [66.62, 76.78]$$

である。

練習問題

(1) 日本人男性 100 人をランダムに選んで身長を測定したところ、平均値は 172cm であった。日本人男性の平均身長の 95 %信頼区間を求めよ。ただし、日本人男性の身長の母分散は 5.5^2 であるとし、日本人男性の身長は正規分布に従うものとする。

(2) 日本人男性 100 人をランダムに選んで体重を測定したところ、平均値は 67kg であった。日本人男性の平均体重の 90 %信頼区間を求めよ。ただし、日本人男性の体重の母分散は 9.0^2 であるとし、日本人男性の体重は正規分布に従うものとする。

(3) 日本人女性 100 人をランダムに選んで体重を測定したところ、平均値は 49kg であった。日本人女性の平均体重の 99 %信頼区間を求めよ。ただし、日本人女性の体重の母分散は 6.0^2 であるとし、日本人女性の体重は正規分布に従うものとする。

問 38 過去の資料から 18 歳男子の身長の標準偏差は $5.8cm$ であることが知られている。18 歳男子の身長の平均値を信頼係数 95% で推定するために何人かを抽出したい。信頼区間の長さを $2cm$ 以下にするためには何人以上を調査する必要があるか。

1．1．2　母分散 v が未知の場合

上で述べたように $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ とおくと、$E[S^2] = \frac{n-1}{n}v$ である。したがって、$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ にたいし、$E[\hat{S}^2] = v$。この $\hat{S}^2$ は、n が大きいとき、ほぼ標準正規分布に従うことが導ける。

そこで v を、$\hat{S}^2$ の実現値

$$\hat{s}^2 = \hat{S}^2|_{X_1=x_1,...,X_n=x_n}$$

で代用し、1．1．1 の母平均の推定を行うことにする。

より正確には次のようになる：

（仮定）：$X_1, ..., X_n$ が $N(m, v)$ にしたがう

とき、

$$T = \frac{\bar{X} - m}{\sqrt{\frac{\hat{S}^2}{n}}}$$

は、自由度 $n-1$ の t-分布 (これを $t(n-1)$ とかく) にしたがう。ここで、「自由度 $n-1$ の t-分布」とは、密度関数が

$$f(x) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})}(1 + \frac{x^2}{n-1})^{-n/2}$$

で与えられる確率分布である。

$X_1, ..., X_n$ が $N(m, v)$ にしたがうとき、$\bar{X}$ は $N(m, v)$ にしたがい、$\frac{\bar{X}-m}{\sqrt{v/n}}$ の分布は $N(0, 1)$ にしたがう。一方、$X_1, ..., X_n$ が $N(m, v)$ に従わなくても、n が大きい時、中心極限定理から、$\frac{\bar{X}-m}{\sqrt{v/n}}$ の分布はほぼ $N(0, 1)$ にしたがう。そこで、はじめから上の（仮定）が満たされるとして差支えない。

一方、n が限りなく大きくなるとき、t-分布 $(t(n-1))$ は限りなく $N(0, 1)$ に近づくことが知られている。これにより、n が大きい時、v を $\hat{S}^2$ の実現値 $\hat{s}^2$ で代用し、$N(0, 1)$ 使った 1．1．1 の母平均の推定を行えることが正当化される。

n が大きくない時には、t-分布表を使って推定を行う方が正確である。(t-分布表は少し詳しい統計学の本の巻末などに載っている。)

例題 20 ある電池の 9 個の標本について調べたところ、

$$85, 91, 83, 87, 86, 88, 90, 84, 89 \text{（時間）}$$

という結果を得た。母平均の信頼係数 95% における信頼区間をもとめよ。

解 20 標本平均 $\bar{x} = \frac{1}{9}(85 + ... + 89) = 87$、不偏標本分散

$$\hat{s}^2 = \frac{1}{9-1}\{(85-87)^2+...+(89-87)^2\} = \frac{1}{8}(2^2+4^2+4^2+0^2+1^2+1^2+3^2+3^2+2^2)$$

$$= \frac{1}{8}65 = 8.125$$

信頼係数 95% における信頼区間は、

$$[\bar{x} - 2.30\sqrt{\frac{\hat{s}^2}{n}}, \bar{x} + 2.30\sqrt{\frac{\hat{s}^2}{n}}]$$

$$= [87 - 2.30\frac{8.125}{3}, 87 + 2.30\frac{8.125}{3}]$$

$$= [87 - 2.30 \times 0.95, 87 + 2.30 \times 0.95]$$

$$= [84.81, 89.18]$$

である。

練習問題

(1) ある 30 人のクラスからランダムに 5 人選んだときの化学のテストの結果は次のとおりであった。このとき、クラス全体の平均点の 95% 信頼区間を求めよ。ただし、化学のテストの点数は正規分布に従うとする。

データ：80　95　60　70　100

(2)

A 高校の 1 年生からランダムに 6 人選んだときの世界史のテスト結果は次のとおりであった。

平均点：80 点、不偏分散：300

また、別の B 高校の 1 年生からランダムに 8 人選んだときの世界史のテスト結果は次のとおりであった。

平均点：70 点、不偏分散：250

このとき、A 高校と B 高校の世界史のテストの平均点の差の 95

1．1．3　比の推定

ある性質について、母集団を構成する各要素が、それをもつかもたないかのどちらかであるとき、その性質をもつ要素の割合を**母比率**という。母比率を p で表す。

標本のうちある性質をみたすものの比率を $\bar{p}$ とする。標本数 n が十分大きいとき、p の信頼係数 95% における信頼区間は

$$[\bar{p} - 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}]$$

となり、信頼係数 99% における信頼区間は

$$[\bar{p} - 2.58\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + 2.58\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}]$$

となる。

実際、X_i を $X_i = 1$(標本がその性質をもつとき)、$X_i = 0$(標本がその性質をもたないとき) とし、$\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$ とすると、

$$E[X_i] = 1 \times p + 0 \times (1-p) = p,$$

$$V(X_i) = (1-p)^2 \times p + (0-p)^2 \times (1-p) = p(1-p) = pq$$

より、

$$E[\bar{X}] = \frac{1}{n} n E[X_i] = p,$$

$$V(\bar{X}) = \frac{1}{n} V(X_i) = \frac{1}{n} pq = \frac{1}{n} p(1-p)$$

$\bar{X}$ の実現値 $\bar{x}$ を $\bar{p}$ とかく。あとは１．１．１と同様である。

例題 21 ある市の全世帯から *400* 世帯を無作為抽出して、ある意見に対する賛否を調べたら、*273* 世帯が賛成であった。全世帯における賛成の比率 p を信頼係数 *95*% で推定せよ。

解 21 X を、賛成のとき１、そうでないとき０をとる確率変数とする。

$$P(X=1) = \frac{273}{400},\ P(X=0) = \frac{127}{400}$$

である。

したがって

$$E[X] = 1 \times \frac{273}{400} + 0 \times \frac{127}{400} = \frac{273}{400},\quad V(X) = \frac{273}{400} \times \frac{127}{400} = \frac{273 \cdot 127}{400^2}$$

したがって、標本平均 $\bar{X}$ の平均、分散は

$$E[\bar{X}] = \frac{273}{400} = 0.683,\ V(\bar{X}) = \frac{1}{400}\frac{273 \cdot 127}{400^2} = (0.466)^2$$

となる。

よって、母比率に対する 95% の信頼区間は

$$[\bar{p} - 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}]$$

において、$n = 400$、$\frac{\bar{p}(1-\bar{p})}{n} = (0.466)^2$ を代入すると、$[0.637, 0.729]$ になる。

練習問題

(1) A 高校の 1 年生からランダムに 6 人選んだときの世界史のテスト結果は次のとおりであった。平均点：80 点、不偏分散：300 また、別の B 高校の 1 年生からランダムに 8 人選んだときの世界史のテスト結果は次のとおりであった。平均点：70 点、不偏分散：250 このとき、A 高校と B 高校の世界史のテストの平均点の差の 95% 信頼区間を求めよ。

(2) サイコロを 400 回投げたところ、6 の目が 80 回出た。このサイコロで 6 の目が出る母比率の 95 %信頼区間を求めよ。

(3) ある選挙で立候補者 T 氏が当選するかどうかをいち早く知るために、出口調査を行うことになった。母比率の 95% 信頼区間の幅を 5 %以内にしたい場合、何人以上を対象に調査を行う必要があるか。ただし、事前調査により T 氏の得票率は 60% 程度であることが分かっているものとする。

(4) 日本人女性 400 人をランダムに選んで身長を測定したところ、平均値は 158cm であった。平均身長の 95 %信頼区間を求めよ。 ただし、日本人女性の身長の母分散は 5.3^2 であるとし、日本人女性の身長は正規分布に従うものとする。

１．２ 母分散の推定

これには χ^2-分布（カイ 2 乗分布）を用いる。

1.2.1　密度関数 $f(x)$ が

$$f_n(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{1}{2}(n-2)} e^{-\frac{x}{2}}, \quad x > 0$$

で与えられる分布を「自由度 n の χ^2-分布」という。ただし、$n = 1, 2, ...$ はパラメータ。$\Gamma(s)$ はガンマ関数 $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$。

命題 1 $N(0,1)$ にしたがう、たがいに独立な確率変数 $X_1, ..., X_n$ について、

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

は、自由度 n の χ^2-分布にしたがう。

証明

$$X_i \sim N(m, \sigma^2) \ => \ \frac{X_i - m}{\sigma} \sim N(0,1)$$

Then Y follows χ^2 distribution. 証明終

これより

命題 2 $N(m, \sigma^2)$ にしたがう、たがいに独立な確率変数 $X_1, ..., X_n$ について、

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - m)^2$$

は、自由度 n の χ^2-分布にしたがう。

証明

$$Y = \sum_{i=1}^{n} (\frac{X_i - m}{\sigma})^2$$

とかく。ここで、各 $\frac{X_i - m}{\sigma}$ は $N(0,1)$ にしたがうから、命題 1 より、Y は自由度 n の χ^2-分布にしたがう。　証明終

命題 3 $N(m, \sigma^2)$ にしたがう母集団から n 個の標本 $X_1, ..., X_n$ をとり、

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

とおくと、Y は、自由度 $n-1$ の χ^2-分布にしたがう。ただし、$\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$。

例題 22 *1* 羽の鶏が産む卵の重さは正規分布に従うと考えられる。ある鶏が産んだ *10* 個の卵の重さはつぎのようであった。（単位グラム）

68.1　70.4　71.5　67.6　70.2

74.5　68.6　70.3　71.2　69.6

このデータから、もとの正規分布 $N(m, \sigma^2)$ の母分散 σ^2 を推定せよ。信頼係数は *90*% とする。

解 22 標本平均 $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 70.2$, 標本分散 $s^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = 3.532$ である。命題 3 より、$Y = \frac{1}{s^2}\sum_{i=1}^{10}(X_i - \bar{X})^2$ は自由度 $n-1=9$ の χ^2 分布にしたがう。

今のデータより

$$y = Y|_{X_i=x_i} = \frac{10 \times 3.532}{\sigma^2} = \frac{35.32}{\sigma^2}$$

自由度 *9* の χ^2 分布における下側 *5*% 点、上側 *5*% 点は、各々 *3.33, 16.92* である。したがって信頼係数 *90*% で、$3.33 < \frac{35.32}{\sigma^2} < 16.92$。つまり、

$$2.09 = \frac{35.32}{16.92} < \sigma^2 < \frac{35.32}{3.33} = 10.6$$

である。

これでは信頼区間があまりに広いが、$n = 10$ ではこれが限界である。幅をもっと狭めるにはデータ数 n を増やさなければならない。

例題 23 いままで n 日間無事故のドライバーと、*1* 度だけ事故を起こしたドライバーの *1* 日あたりの事故率はどれくらいちがうか。

解 23 $A =$ いままで n 日間無事故のドライバーの集合、$B =$ いままで *1* 度だけ事故を起こしたドライバーの集合、とし、A のドライバーの事故率を p_0、B のドライバーの事故率を p_1 とする。確率変数 $(X_i), i = 1, ..., n$ を

$$X_i : A \cup B \to \{0, 1\},$$

$$X_i(\omega) = 1 \quad \omega \text{ さんが } i \text{ 日目に事故を起こしたとき}$$

$$X_i(\omega) = 0 \quad \omega \text{ さんが } i \text{ 日目に事故を起こさなかったとき}$$

とする。$\omega \in A \cup B$。

このとき事故回数を表す確率変数

$$S = X_1 + ... + X_n$$

は二項分布にしたがい、その平均、分散は

$$E[S1_A] = np_0, \quad V(S1_A) = np_0(1-p_0)$$

$$E[S1_B] = np_1, \quad V(S1_B) = np_1(1-p_1)$$

であたえられる。

n がおおきいとして、二項分布を正規分布で近似する。そして、*1* 日あたりの事故率の母分散は、

$$V(\frac{S1_A}{n}) = \frac{p_0(1-p_0)}{n} \quad \omega \in A \text{ の場合}$$

$$V(\frac{S1_B}{n}) = \frac{p_1(1-p_1)}{n} \quad \omega \in B \text{ の場合}$$

として母平均の推定をおこなう。

標本平均のデータは $\bar{x}_0 = \frac{0}{n} = 0, \bar{x}_1 = \frac{1}{n} = 0$ である。

これより 95% の信頼区間（片側）は

(1) $$p_0 < 0 + 1.645\sqrt{\frac{p_0(1-p_0)}{n}}$$

(2) $$p_1 < \frac{1}{n} + 1.645\sqrt{\frac{p_1(1-p_1)}{n}}$$

となる。

（１）より、

$$p_0^2 < (1.645)^2 \frac{p_0 - p_0^2}{n}$$

これより

$$(n + (1.645)^2)p_0 < (1.645)^2 \quad \Longleftrightarrow \quad p_0 < \frac{(1.645)^2}{n + (1.645)^2}$$

$n = 100$ とすると、

(3) $$p_0 < 0.0263$$

である。

一方、(*2*) の場合、

$$(p_1 - \frac{1}{n})^2 < (1.645)^2 \frac{p_1(1-p_1)}{n}$$

よって

$$(n + (1.645)^2)p_1^2 - (2 + (1.645)^2)p_1 + \frac{1}{n} < 0$$

$n = 100$ とおくと $p_1 < x$、ここで

$$x = \frac{(2+\alpha) - \sqrt{(2+\alpha)^2 - 4(100+\alpha)/100}}{2(100+\alpha)}$$

ただし、$\alpha = (1.645)^2$。これより

$$p_1 < 0.0435 \tag{4}$$

である。

100 に *1* 回という差にかかわらず、(*3*) と (*4*) をみると、事故率は意外に差が大きい (約 *2* 倍)。

問 39 ある工場の製品 *400* 個について検査したところ、不良品が *8* 個あった。これを無作為標本としてこの工場の全製品における不良率を信頼係数 *95*% で推定せよ。

練習問題

（1）あるメーカーが生産している自動車 A の 10 台の燃費を計測したところ、その不偏分散は $3.29km^2$ であった。この自動車 A の燃費の母分散の 95 %信頼区間を求めよ。

（2）ランダムに選んだ男性 5 人の身長を測定したところ、次のようなデータが得られた。男性の身長の母分散の 95% 信頼区間を求めよ。

No. 身長 [cm]
1 175.8
2 171.9
3 172.7
4 170.3
5 180.2

（3）男性 9 人をランダムに選び、40 秒間での腕立て伏せの回数を記録した。男性の腕立て伏せの回数の母分散の 90 %信頼区間を求めよ。

No. 記録
1 24
2 28
3 22
4 31
5 28
6 25
7 27
8 26
9 25

問　題

問 40 森の中に n 匹のピカチュウがいるがその正確な数は不明である。まず r 匹のピカチュウを同時に捕獲し、印をつけて森の中に放す。しばらくしてから、s 匹のピカチュウを再び捕獲する。ただし、r, s は予め定めた定数である。

いま s 匹中 x 匹に印がついていたとする。n を推定せよ。

問 41 あるテレビ番組の視聴率を調べるため、任意抽出により *200* 世帯を選んで調査したところ *56* 世帯が視聴していることがわかった。視聴率 p を *95%*の信頼係数で推定せよ。

問 42 ある県で実施された模擬試験の平均点を、信頼係数 *99%*、誤差 *2* 点以内で推定したい。少なくとも何人以上の受験者の得点を任意抽出しなければならないか。ただし、従来の経験により得点の標準偏差は *20* 点としてよい。

問 43 数学の試験を受けた生徒の中から *100* 人を任意に抽出して調査したところ、平均点は *65.2* 点であった。母標準偏差を *14.0* 点として、母平均の信頼区間を信頼係数 *95%*で求めよ。

問 44 ある県において、*18* 歳人口の身長の標準偏差は *7.5cm* である。この県の *18* 歳人口の身長の平均を誤差 *0.5cm* 以内で推定したい。*95%*の信頼係数で推定するとして、何人の標本を任意抽出しなければならないか。

問 45 ある町の駅の乗降客 *400* 人を任意抽出して調べたところ、*196* 人がその町の住人であった。乗客中その町の住人の比率を信頼係数 *99%*で推定せよ。

問 46 消費者団体が写真フィルム 1 パックあたりの値段をスーパーで調べたところ、次のような結果を得た（単位　円）:

800, 800, 660, 780, 990, 820, 750

（1）この写真フィルム 1 パックあたりの値段を確率変数 X とみなして、平均 $E(X)$ および標準偏差 $\sigma(X)$ を求めよ。

（2）上の標本により、この写真フィルム 1 パックあたりの値段の信頼区間を、信頼係数 *90%*および *99%*で各々求めよ。

第9章　検定

9.0.26　検定の一般論

検定とは統計的データをもとにある結論（判断）を導くことである。
例：
（1）治験：薬が効くか効かないか
（2）鑑定：親子鑑定、本人確認、DNA 鑑定、
（3）犯罪捜査：犯人立証
このように適用範囲はたいへん「現実」的である。また、使う道具は数学であるが、対象となる諸事象について論理的思考が求められる。
検定では、主張したい命題を否定した命題を作る。これを**帰無仮説**という。(これに対し、元々主張したい命題を**対立仮説**という。) この仮説のもとにデータを眺めたとき、データが実現する割合が「きわめて小さい」場合に、この仮説を否定する。このようにして主張したい命題を統計的に証明するのである。言わば統計的「背理法」(の応用) である。
具体的には次のような手順による。
(1) ある仮説 (H) をたてる
(2) H に基づいて推定をおこなう
(3) データを推定結果と比較する
(4) もしデータが、推定結果のうち確率のきわめて小さい部分に属しているならば、H は誤りであると結論する
このように、帰無仮説は否定されることに意味がある。まるで「流し雛」のようにはかない運命を担っている。

例題 24 さいころを *98* 回投げたとき、結果が

目	*1*	*2*	*3*	*4*	*5*	*6*
回数	*24*	*14*	*17*	*13*	*13*	*17*

であった。このさいころは「正しい」と言えるか。

解 24 見かけ上 1 の目が顕著に多いので、「正しくない」と言いたい。そこで背理法を使って

仮説 *(H):* このさいころは正しい

とする。この H が帰無仮説である。

この H のもとで、さいころを *98* 回投げたとき、*1* の目の出る回数 X は、平均 $m = 98 \times \frac{1}{6} = \frac{49}{3}$、分散 $v = 98 \times \frac{1}{6} \times \frac{5}{6} = \frac{245}{18}$ の 2 項分布にしたがう。

$n = 98$ は十分大きいので、中心極限定理から、$\sigma = \sqrt{v}$ として

$$T = \frac{X - m}{\sigma}$$

は $N(0,1)$ にしたがうと考えられる。

データより

$$t = \frac{24 - \frac{49}{3}}{\sqrt{\frac{245}{18}}} \sim 2.07807$$

であり、$T \sim N(0,1)$ のとき

$$P(T \geq t) = \int_{2.078}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \sim 0.0192$$

である。つまり、はじめのデータのような目の出方の事象が起こる確率は、H のもとで、*2*% 未満である。

.... ここでどう考えるか。

「H のもとであるデータを得るような確率が α 以下であり、かつ実際にそのようなデータを得たならば、H は正しくない」

と考える。このような α を**有意水準**（または**危険率**）という。検定を行う際には、このような α をあらかじめ決めておく。通常 $\alpha = 0.05\ (5\%), \alpha = 0.01\ (1\%)$ などをとることが多い。

帰無仮説 H を正しくないと判定することを

H を棄却する

という。

そうでないとき（H を正しくないと判定しないとき）

H を採択する

という。しかしこれは「H が正しい」と検定したわけではない。つまり、「何とも言えない」ということである。

いづれにしても、確率 α 以下の事象が起こらない保証はない。その意味で、検定は確率が大きい方にかけるギャンブルであるといえる。

したがって判断を誤ることもある。

(I) H が正しいのに棄却する誤りを**第１種の誤り**という。

(II) H が誤りなのに採択する誤りを**第２種の誤り**という。
どちらになるかは、H が正しい命題かどうかで決まる。

詳しく言うと次のようになる。
ある罪を犯した容疑者の裁判を例に取ってみます。
帰無仮説 H を「この容疑者は罪を犯した」とします。
このとき、本当はこの容疑者は罪を犯したのに無罪という判決を受ける確率が第 1 種の過誤の確率です。
一方、本当はこの容疑者は無罪なのに有罪という判決を受ける確率が第 2 種の過誤の確率です。

真実	本当は有罪	本当は無罪
裁判 有罪	正しい	**第 2 種の過誤**
裁判 無罪	**第 1 種の過誤**	正しい

容疑者の裁判の例では、第 1 種の過誤を犯す確率を下げるために容疑者を全て有罪にしてしまうと、無罪の人まで有罪になってしまうために第 2 種の過誤を犯す確率が上がってしまいます。逆に第 2 種の過誤を犯す確率を下げるために容疑者を全て無罪にしてしまうと、本来有罪の人まで無罪になってしまうために第 1 種の過誤を犯す確率が上がってしまいます。したがって、第 1 種の過誤を犯す確率と第 2 種の過誤を犯す過誤を犯す確率のバランスをとることが重要です。

有意水準に相当する確率の事象が生じるパラメータの範囲を**棄却域**という。検定法としては次の２つがある。
(1) 「ある値以下」または「ある値以上」を棄却域とする検定法を**片側検定**という。
(2) 平均を挟んで左右対称の領域を非棄却域とする検定法を**両側検定**という。
上の例題で片側検定を行い $\alpha = 0.05$ とすると H は成り立たない。$\alpha = 0.01$ とすると H は成り立たつ。ただし、データを見てから $\alpha = 0.01$ を選んではいけない。

例題
薬 A に含まれるある成分 B についての分析を行います。B の含有量を調べるため、生産された薬 A の中からランダムに 25 粒を抜き取り、成分 B の量を測定しました。その結果平均が $\bar{x} = 98\text{mg}$、不偏分散が $s^2 = 1$ で

した。

この問題では帰無仮説 H を「薬 A 中の成分 B の含有量は 100mg である」としたときに、3 通りの対立仮説 H_1

が考えられます。

(a) 薬 A 中の成分 B の含有量は 100mg ではない

(b) 薬 A 中の成分 B の含有量は 100mg より多い

(c) 薬 A 中の成分 B の含有量は 100mg より少ない

(a) は成分 B の含有量が 100mg かどうかを調べるための検定です。

(b) は成分 B の含有量が 100mg より多いかどうかを調べるための検定です。この場合、成分 B の含有量が 100mg より少ないかどうかについては考慮しません。

(c) は成分 B の含有量が 100mg より少ないかどうかを調べるための検定です。この場合、成分 B の含有量が 100mg より多いかどうかについては考慮しません。

(a) のような検定方法を「両側検定」、(b) と (c) のような検定方法を「片側検定」といいます。

正規分布を使った検定

例題 25 *(***両側検定：母分散既知の場合***)* あるコーヒー豆袋詰め機械が袋に詰めるコーヒー豆の重さは、平均 *100* グラム、標準偏差 *5* グラムの正規分布にしたがうように設定されている。機械が正しく調整されているか確かめるため、*9* 袋の無作為標本をとってコーヒー豆の重さを量ったら、平均は *102.4* グラムであった。この機械は正しく調整されているか、有意水準 *5*% で検定せよ。

解 25 正しく調整されていないとして、その状態の母平均を m とする。コーヒー豆の重さを X とすると、$X \sim N(100, 5^2)$ である。もし正しく調整されていないならば *100* グラムより多く、または少なく袋詰めしてしまうから *(H)* 帰無仮説 *:* $m = 100$ とする。

$n = 9, \sigma = 5, \bar{x} = 102.4$ から、$T = \frac{\bar{X}-m}{\sqrt{\frac{\sigma^2}{n}}}$ の実現値 $t = \frac{102.4-100}{\frac{5}{3}} = 1.44$ となる。有意水準 *5*% であるから $u(\frac{0.05}{2}) = 1.96$。*1.44* はこれより小さいから Hは棄却できない。よって「正しく調整されていないとはいえない」。

例題 26 *(***両側検定：母分散未知の場合***)* ある野球球団がシーズンに入っ

てからの、*1*試合あたりの入場者数は以下のとおりである：

1475　1420　1433　1452　1411　1466　1432　1453　1414

この球団は今季 *1* 試合あたりの平均入場者数を *1455* 人と予測して経営計画をたてた。上の結果から、平均入場者数の予測違いはないという仮説を、有意水準 *5%* で検定せよ。

解 26 入場者数が *1455* 人を上回り過ぎても下回り過ぎても予測違いである。よって

*H(*帰無仮説*)*：予測違いでない　*(*平均入場者数 $m = 1455$ である）

とする。$n = 9$ であり、標本平均 $\bar{x} = \frac{1475+...+1414}{9} = 1442.9$ 不偏標本分散

$$\hat{s}^2 = \frac{1}{9-1}(1475 - \bar{x})^2 + ... + (1414 - \bar{x})^2 = 416.1$$

となる。

ここで有意水準 *5%* だから自由度 *8* の *t*-分布における確率 $\frac{0.05}{2}$ のパーセント点 $t_8(\frac{0.05}{2}) = 2.306$、したがって棄却域は $(-\infty, -2.306) \cup (2.306, \infty)$ である。データにおける

$$t = \frac{\bar{x} - m}{\sqrt{\frac{\hat{s}^2}{n}}}$$

の実現値 $t = -1.78$ はこの棄却域に入らない。よって H を棄却できない。つまり「予測違いであるとはいえない」。

2 標本 t 検定

2 つの独立した母集団があり、それぞれの母集団から抽出した標本の平均に差があるかどうかを検定することを「2 標本 t 検定」といいます。例えば、ある学校で行ったテストの点数が 1 組と 2 組とで差があるかどうかの検定や、被験者に対してある薬を投与する前後で血圧がどう変化したかの検定に使います。

対応がない場合の 2 標本 t 検定の方法

異なる対象から抽出された 2 つの標本は「対応のないデータ（対応なし）」です。 例えば、1 組と 2 組の生徒は異なるので、それぞれのクラスから抽出された 2 つの標本は「対応のないデータ」となります。

対応がない場合の 2 標本 t 検定では、2 つの標本に対応がないことを加味した検定統計量を用いる必要があります。母分散が分からない場合、1 群目の標本平均を $\bar{x}_1$、母平均を m_1、サンプルサイズを n_1、2 群目の標本

平均を $\bar{x}_2$、母平均を m_2、サンプルサイズを n_2 としたときに、次の式から算出される統計量 t を使います。検定で用いるのは自由度 $n_1 + n_2 - 2$ の t 分布です:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

ただし s^2 は 2 群の標本を「プールした不偏分散」である。プールした分散とは、2 つの標本の不偏分散を 1 つにまとめたもので、1 群目の不偏分散を s_1^2、2 群目の不偏分散を s_2^2 とした場合、下の式から求めることができます:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

なお、帰無仮説を「母平均が等しい」という仮説が正しいとすると、$m_1 = m_2$ になります。したがって

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

となる。

ポアソン分布を使った検定

例 36 松山市では *1* 日平均 *1* 人の交通事故による死者がでる。*1* 週間の死亡者数を信頼係数 *95%* で推定せよ。また、*1* 週間に *14* 人以上の死者がでたら異常な事態といえるか。

解 27 *1* 日平均 *1* 人だから *1* 週間で平均 *7* 人である。$\lambda = 7$ となるポアソン分布を考える。

（前半）$N =$*(1* 週間の交通死亡者数*)* とすると

$$P(N = k) = \frac{1}{k!}\lambda^k e^{-\lambda} = \frac{1}{k!}7^k e^{-7}$$

これを $k = 0, 1, ..., 14$ について書いてみると

k	$\frac{1}{k!}7^k e^{-7}$
0	9.119×10^{-4}
1	6.383×10^{-3}
2	*0.0223*
3	*0.05213*
4	*0.0912*
5	*0.1277*
6	*0.1490*
7	*0.1490*
8	*0.1304*
9	*0.1014*
10	*0.071*
11	*0.045*
12	*0.026*
13	*0.014*
14	7.094×10^{-3}

これより

$$1 - P(3 \leq N \leq 12) = 0.02963 + 0.027... \sim 0.056...$$

となって、約 5.6% である。したがって、$P(N = 13) = 0.014$ であることを考慮すると、信頼係数 95% による信頼区間としては

$$3 \leq N \leq 13$$

とすればよい。

（後半）「異常である」と主張しようとして、帰無仮説 H :「異常でない」とする。

$$P(N \leq 13) = 0.987..., \ P(N \geq 14) \sim 0.013$$

であるから、有意水準 1% で片側検定すると、H は棄却できない。よって「異常であるとはいえない」が結論である。

しかし、もし *1* 週間に *15* 人以上事故死したとすると

$$P(N \leq 14) = 0.99428...$$

であるから、H は棄却される。この場合には結論は「異常である」となる。

2 標本 t 検定

2 つの独立した母集団があり、それぞれの母集団から抽出した標本の平均に差があるかどうかを検定することを「2 標本 t 検定」といいます。例えば、ある学校で行ったテストの点数が 1 組と 2 組とで差があるかどうかの検定や、被験者に対してある薬を投与する前後で血圧がどう変化したかの検定に使います。

(1) 対応がない場合の 2 標本 t 検定の方法

異なる対象から抽出された 2 つの標本は「対応のないデータ（対応なし）」です。 例えば、1 組と 2 組の生徒は異なるので、それぞれのクラスから抽出された 2 つの標本は「対応のないデータ」となります。

対応がない場合の 2 標本 t 検定では、2 つの標本に対応がないことを加味した検定統計量を用いる必要があります。母分散が分からない場合、1 群目の標本平均を $\bar{x}_1$、母平均を m_1、サンプルサイズを n_1、2 群目の標本平均を $\bar{x}_2$、母平均を m_2、サンプルサイズを n_2 としたときに、次の式から算出される統計量 t を使います。検定で用いるのは自由度 $n_1 + n_2 - 2$ の t 分布です:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

ただし s^2 は 2 群の標本を「プールした不偏分散」である。プールした分散とは、2 つの標本の不偏分散を 1 つにまとめたもので、1 群目の不偏分散を s_1^2、2 群目の不偏分散を s_2^2 とした場合、下の式から求めることができます:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

なお、帰無仮説を「母平均が等しい」という仮説が正しいとすると、$m_1 = m_2$ になります。したがって

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

となる。($(1/n_1 + 1/n_2)^{-1}$ は n_1 と n_2 の調和平均：$1/R = 1/R_1 + 1/R_2$)

(2) 対応のある 2 標本 t 検定の方法

例題：

血圧を下げる薬のテストを行います。被験者 5 人に対して薬の投与前と投与後の血圧を測定したところ、次の表のような結果が得られました。

この結果から、薬の投与によって血圧は下がったと言えるでしょうか。被検者 No. 投与前の血圧 投与後の血圧

1 180 150

2 130 135

3 165 145

4 155 150

5 140 140

対応がある場合の 2 標本の t 検定では 2 群の差が 0 かどうかについての検定を行います。この例題では、投薬前後での血圧の差が 0 かどうかを検定します。したがって、まず薬の投与前後での血圧の差とその平均値を算出します。

被検者 No. 投与前の血圧 投与後の血圧 差（投与前-投与後）

1 180 150 30

2 130 135 -5

3 165 145 20

4 155 150 5

5 140 140 0

平均 154 144 10

(i) 仮説を立てる

帰無仮説は「投薬前後の血圧は等しい＝投薬によって血圧は下がらなかった」とします。したがって、対立仮説は「投薬によって血圧に差があった＝投薬によって血圧は下がった」となります。

(ii) 有意水準を設定する $\alpha = 0.05$ とします。

(iii) 適切な検定統計量を決める

この実験では母分散が分からないので、不偏分散 s^2 を用いる t 統計量を使います。統計量 t は次の式から求められます。

$$t = \frac{\bar{d} - m}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{d} - m}{\frac{s}{\sqrt{n}}}$$

$\bar{d}$ は投薬による血圧の差の平均、m は差の母平均、n はサンプルサイズを表します。

(iv) 棄却ルールを決める

この検定で使用する分布は自由度「5-1=4」の「t 分布」です。この例題では血圧が下がったかどうかのみを考えればよいので、片側検定を行います。統計数値表から $t_{0.05}(4)$ の値を読み取ると「2.132」となっています。

(v) 検定統計量を元に結論を出す

投薬前後での血圧の差が 0 かどうかを検定するため、$m = 0$ となります。また、薬の投与前後での血圧の差の不偏分散

$$s^2 = \frac{(30-10)^2 + ... + (0-10)^2}{5-1}$$

は 212.5 になります。

この値を統計量 t の式に代入すると次のようになります。

$$t = \frac{10-0}{\sqrt{212.5/5}} \sim 1.53$$

1.53 は棄却域に入っていないことから、「有意水準 5% において、帰無仮説は棄却されない」という結果になります。つまり、「投薬によって血圧が下がったとは言えない」と結論づけられます。

問　題

問 47 ある種のメダカの黒色個体と白色個体を交配させたところ、黒色個体ばかりを得た。この第 *2* 世代の黒色個体同士を交配させたところ、黒色個体 *162* 尾、白色個体 *163* 尾を得た。もしこのメダカの体色の染色体がメンデルの法則に従うならば、第 *3* 世代の体色の分離比は *3:1* となるはずである。上の実験結果がメンデルの法則に矛盾するかしないかを、危険率 *5%* で検定せよ。

問 48 ある病気には A、B ふたつの治療薬があり、B 型の薬の有効率（服用して効き目のある確率）は *0.6* である。A 型の薬を *200* 人の患者に与えたところ *134* 人の患者に効き目があった。A 型の薬は B 型の薬より優れているといえるか。有意水準 *5%* で検定せよ。

問 49 ある牛乳業者が発売している *1l* 入りと表示された牛乳パックから、無作為に *100* 個を抽出して内容量を計ったところ、平均値は *0.98l*、標準偏差は *0.08l* であった。この会社の牛乳パックの内容量は表示どおりでないとみなしてよいか。危険率 *5%* で検定せよ。また、危険率 *1%* で検定するとどうなるか。

問 50 A、Bふたつのさいころがある。どちらも 100回投げたところ、1の目が出た回数は Aでは 26回、Bでは 10回であった。この 2つのさいころ A、Bはそれぞれ正しく作られていないといえるか。危険率 5%で検定せよ。また、危険率 1%で検定するとどうなるか。

問 51 ガソリン添加剤によるエコカー（環境に配慮した自動車）の燃費効率の改善を調べるため、ある型のエコカー 10台を集め添加剤なしで走ったところ、平均で $1l$あたり $30.3km$走った。さらにその型の他のエコカー 10台を集め添加剤を入れて走ったところ、平均で $1l$あたり $31.5km$走った。この添加剤は燃費効率に効果があると判断してよいか、5%の危険率で検定せよ。ただし、この型のエコカーの $1l$あたりの走行距離の分散は 5.8である。

問 52 2014年サッカー日本代表メンバー 23人のうち、血液型が B型である選手は 2人いた。日本人全体の血液型の分布は以下のとおりである：

血液型	O	A	B	AB
割合 $(\%)$	30.5	38.2	21.9	9.4

B型の選手の数は日本人全体の血液型の分布からみてかたよっているといえるか、危険率 5%で検定せよ。

問 53 次の表は、1つ $25.5\ kg$の $($はずの$)$強力粉玉 20個をサンプリングし、重量を測定した結果をまとめたものである。このデータを用いて、強力粉の重量は $25.5\ kg$ではないと言えるかどうか検定せよ。なお、有意水準は 0.05 とする。

項目	測定結果
サンプルサイズ	20
平均	25.29
不偏分散	$2.23\ (=1.49^2)$

問 54 あるパンメーカーでは、人気の商品であるメロンパンを 2つの工場で製造している。2つの工場で製造されているメロンパンの重量 (g)を調べた結果、A工場の 10個については平均 93、不偏分散 $13.7(=3.7^2)$であった。また、B工場の 8個については平均 87、不偏分散 $15.2(=3.9^2)$であった。この 2工場の間でメロンパンの重量 (g)に差があると言えるかどうか検定せよ。なお、有意水準は 0.05とする。

問 55 300人に開発中のチョコレート菓子を試食してもらい、おいしいか否かの 2択で回答してもらった。その結果、160人が「おいしい」と回答した。このお菓子を「おいしい」と判断する母比率が 0.5であるかどうかを有意水準 0.05で検定せよ。

問 56 あるドラマの視聴率が 30% だったという記事が、怪しい週刊誌に掲載された。この記事に信憑性があるかどうか確認するため、ランダムに選んだ 250 人にアンケートを取ったところ 55 人がそのドラマを見ていたことが分かった。この結果から、週刊誌の記事を信用できるか有意水準 0.05 で検定せよ。