総合演習「現代社会における数学」[1]

1 確率分布

1.1 確率空間

$(\Omega, \mathcal{F}, P)$ が確率空間であるとする[2]、即ち

Ω: 標本空間 (起こりうる場合全体)

$\mathcal{F}$: 事象の集まり (事象とは試行の結果として起こる事柄)

P: 確率, 事象 $A \in \mathcal{F}$ に対して $P(A)$ で A が起こる確率を表す。

とする。確率 P に対して以下が成立する:

1. $0 \leq P(A) \leq 1$ $(\forall A \in \mathcal{F})$, 特に $P(\Omega) = 1$.
2. $A, B \in \mathcal{F}$ が $A \cap B = \emptyset$(A と B は排反であるという) のとき、$P(A \cup B) = P(A) + P(B)$.

このとき、$\Omega \cup \emptyset = \Omega$ であるから 2 より $P(\Omega) + P(\emptyset) = P(\Omega)$ となり、$P(\emptyset) = 0$ が従う。また、$A \cup A^{\mathrm{c}} = \Omega, A \cap A^{\mathrm{c}} = \emptyset$ より、再び 2 から

3. $P(A^{\mathrm{c}}) = 1 - P(A)$ $(\forall A \in \mathcal{F})$.

が成立する。この授業では用いないが、通常の確率統計学では 2 より強い条件

2'. $A_n \in \mathcal{F}$ $(n \in \mathbb{N})$ が $A_n \cap A_m = \emptyset$ $(n \neq m)$ のとき、$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

を仮定することを補足しておく。(この条件を σ-加法性という[3]。) この性質は、例えば確率変数の期待値などを厳密に定義する際に必要となる。

例題 1.1 サイコロを n 回投げたとき、*1* の目が偶数回出る事象 B_n の起こる確率 $P(B_n)$ を求めよ。

解: $p_n = P(B_n)$ として p_n についての漸化式を立てる。$p_1 = \frac{5}{6}$ は明らかであろう。A_n で n 回目に投げたとき 1 の目が出る事象とすると

$$p_{n+1} = P(B_n \cap A_{n+1}^{\mathrm{c}}) + P(B_n^{\mathrm{c}} \cap A_{n+1}) = \frac{5}{6} p_n + \frac{1}{6}(1 - p_n) = \frac{2}{3} p_n + \frac{1}{6}.$$

この漸化式を解いて $p_n = \frac{1}{2} + \frac{1}{3}(\frac{2}{3})^{n-1}$ を得る。 □

問題 1.1 事象 $A, B, C \in \mathcal{F}$ に対し次を示せ。

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

---

[1] このノートは次の URL からダウンロードできます。 http://www.math.u-ryukyu.ac.jp/~sugiura/

[2] この授業では、高校数学の数学 B, 数学 C における確率統計について、大学数学の立場から概観することを目的とします。記号や用語が高校の教科書と通常 (アクチュアリー試験等で) 用いられるものと異なる場合、後者を用いるようにします。

参考文献 高等学校 数学 B, 数学 C (数研出版, 平成 7 年版)、高等学校 確率・統計 (旺文社、昭和 58 年版)

[3] 数学的な厳密さを求めるなら、$\mathcal{F}$ に対して (i) $\Omega \in \mathcal{F}$, (ii) $A \in \mathcal{F} \implies A^{\mathrm{c}} \in \mathcal{F}$, および (iii) $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ を満たすことを要求すべきである。この三つの性質をみたすとき、集合族 $\mathcal{F}$ は σ-加法族をなすという。

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

次に、事象の独立と従属について考える。

定義 1.1 *(1)* 事象 A, B について $P(A\cap B) = P(A)P(B)$ なるとき、A と B は**独立**であるという。また、*2*つの事象が独立でないとき**従属**であるという。
(2) 事象 A, B に対して $P(B) > 0$ となるとき、$P(A|B) = \frac{P(A\cap B)}{P(B)}$ によって定義される確率 $P(A|B)$ を、事象 B が与えられたときの事象 A の**条件付確率**という。*(*通常 $P(B) > 0$ のときのみ定義される。*)*

例えば、白玉 7 個と赤玉 3 個が入っている袋から、玉を 1 個ずつ 2 回取り出すとき、1 回目に白玉が出る事象 A と 2 回目に白玉が出る事象 B は独立だろうか? 答は (a) 1 回目の玉を袋に戻してから 2 回目を取り出す場合、(b) 1 回目の玉を袋に戻さないで 2 回目を取り出す場合 で異なる。(a) では独立、(b) では従属となることは明らかであろう。では、次はどうか。

問題 1.2 *1*組 *52*枚のトランプから *1*枚抜き出すとき、ハートが出るという事象 A と絵札が出るという事象 B は独立であるか。また、スペードが出るという事象を C とすると、A と C は独立か調べよ。

次に条件付確率についての例題を扱う。

例題 1.2 ある病気の判定薬があり、病気の人に対して *90%*の確率で陽性反応 *(*病気であると判定*)* を示し、健康者に対して *5%*の確率で陽性反応を示す。この病気の罹病率は *1%*であるとする。ある人がこの判定薬で陽性反応がでたとき、この人が本当にその病気に罹っている確率を求めよ。

解: 事象 A をその人がその病気に罹っているという事象とし、事象 B を陽性反応がでた事象であるとする。病気の人に対して 90%の確率で陽性反応を示すから $P(B|A) = 0.9$. 健康者に対して 5%の確率で陽性反応から $P(B|A^c) = 0.05$. また、この病気の罹病率は 1%より $P(A) = 0.01$. 求める確率は $P(A|B)$ である。
以上の式より、$P(A\cap B) = P(B|A)P(A) = 0.009$, $P(A^c\cap B) = P(B|A^c)P(A^c) = 0.0495$ で $P(B) = P(A\cap B) + P(A^c\cap B) = 0.0585$. よって、$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{0.009}{0.0585} = 0.1538\ldots$. □

問題 1.3 ある製品を製造する *2*つの工場 *A,B*があり、A 工場の製品には *2%*, B 工場の製品には *3%*の不良品が含まれていた。これら A 工場の製品と B 工場の製品を 5 : 3 の割合で混ぜた大量の製品の中から *1*個取り出すとき、*(1)* それが不良品である確率、*(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。

問題 1.4 ある工場では機械 M_1, M_2, M_3 で前製品のそれぞれ *60%, 30%, 10%*を製造していて、これらの機械で生じる不良品の割合は *2%, 3%, 4%*である。いま、*1*個の不良品が見つかったとき、それが機械 M_1 で製造されたものである確率を求めよ。

問題 1.5 独立な事象 A, B があって、$P(A) = 2/3, P(A\cap B) = 1/2$ であるとき、次の確率を求めよ。*(1)* $P(B)$, *(2)* A, B のうちどちらか一方だけが起こる確率.

1.2 確率分布 (離散型の場合)

一般に、試行の結果によってその値が定まる変数を**確率変数** (random variable) という[4]。X の取りうる値が高々可算個の点からなる場合を**離散型**, 連続的に変化する場合**連続型**という。(もちろんそれらを混合させた混合型もあるが、ここでは扱わない。)

まず、離散型の場合を考えよう。X のとりうる値が高々可算個, $a_1, a_2, \ldots, a_n, \ldots$ であるとき、$f(a_n) = P(X = a_n)(> 0)$ $(n = 1, 2, \ldots)$ を確率関数という。このとき、X の**期待値**または平均値を $E[X]$ と書き

$$E[X] = \sum_{n=1}^{\infty} a_n P(X = a_n) \tag{1.1}$$

で定める。また、関数 $\varphi : \mathbb{R} \to \mathbb{R}$ に対して $\varphi(X)$ も確率変数であり

$$E[\varphi(X)] = \sum_{n=1}^{\infty} \varphi(a_n) P(X = a_n) \tag{1.2}$$

となる。ただし、右辺が絶対収束する時のみ定義されるものとする。特に、$m = E[X]$ とし、$\varphi(x) = (x - m)^2$ としたときの $\varphi(X)$ の期待値を X の**分散**という:

$$V(X) = E[(X - E[X])^2] = \sum_{n=1}^{\infty} (a_n - m)^2 p_n = \sum_{n=1}^{\infty} {a_n}^2 p_n - m^2 = E[X^2] - E[X]^2. \tag{1.3}$$

ここで、$p_n = P(X = a_n)$ と書いた。また、$a \in \mathbb{R}$ に対して $V(aX) = a^2 V(X)$ に注意する。

例題 1.3 *(***二項分布***)* 表が出る確率が p *(*$0 < p < 1$*)* であるような *(*歪んだ*)* コインがある。このコインを n 回投げるとき、表が出る回数を X とする。このとき、*(1)* X の確率分布を求めよ. *(2)* X の平均と分散を求めよ。

解: (1) X の取りうる値は $0, 1, \ldots, n$ で、$X = k$ となるのは表が k 回、裏が $n - k$ 回出るときでその出る順序は $\binom{n}{k}$ 通りあるから、求める確率分布は $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ である[5]。
(2) $E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1 - p)^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1 - p)^{n-k} = np.$
分散については演習問題とする。答は $V(X) = np(1 - p)$. □

問題 1.6 *(***幾何分布***)* 表が出る確率が p *(*$0 < p < 1$*)* であるような *(*歪んだ*)* コインがある。このコインを表が出るまで投げるとき、初めて表が出るまでに裏の出た回数を X とする。このとき、*(1)* X の確率分布を求めよ. *(2)* X の平均と分散を求めよ。

答のみ: (1) $P(X = k) = p(1 - p)^k, k = 0, 1, \ldots$. このとき、$X$ は幾何分布 $G(p)$ に従うという。 (2) $E[X] = \frac{1-p}{p}$, $V(X) = \frac{1-p}{p^2}$. □

---

[4]厳密には、X が確率変数であるとは Ω 上の関数で $\forall a \in \mathbb{R}$ に対して $\{\, \omega \in \Omega ;\, X(\omega) < a \,\} \in \mathcal{F}$ なるものを言う。(ルベーグ積分論での、可測の概念に他ならない。)

[5] $\binom{n}{k}$ は高校では ${}_nC_k$ と記したものである。実は、${}_nP_k$ や ${}_nC_k$ は高校の教科書以外ではまず見ることのない記号である。高校の教科書では確率分布は表と書き表しているが、ここでは確率関数を求めるにとどめた。(本当の正解は X **は二項分布** $B(n, p)$ **に従う**である。)

他に次のような重要な分布がある。

- $\lambda > 0$ とする。$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ $(k = 0, 1, \ldots)$ のとき確率変数 X は **Poisson 分布** $P(\lambda)$ に従うという。このとき、$E[X] = V(X) = \lambda$ となる。ある一定時間に起こる事故の件数などが Poisson 分布に従うと考えられている。

- $\alpha > 0, p > 0$ とする。$P(X = k) = \binom{-\alpha}{k} p^\alpha (p-1)^k$ $(k = 0, 1, \ldots)$ のとき確率変数 X は**負の二項分布** $NB(\alpha, p)$ に従うという[6]。このとき、$E[X] = \frac{\alpha(1-p)}{p}, V(X) = \frac{\alpha(1-p)}{p^2}$ となる。この分布も保険のクレームの件数を表す分布として損害保険数理にしばしば見かける分布である。

X, Y を二つの確率変数とするとき、実数 a, b に対し「$X = a$ かつ $Y = b$」という事象の起こる確率を $P(X = a, Y = b)$ と表す。X のとる値が $x_1, x_2, \ldots$, Y のとる値が $y_1, y_2, \ldots$ であるとき $f(x_k, y_l) = P(X = x_k, Y = y_l)$ $(k, l = 1, 2, \ldots)$ を同時確率関数 (または**同時分布**) という。3 つ以上の確率変数に対しても同様に定義される。

このとき、$\varphi : \mathbb{R}^2 \to \mathbb{R}$ に対して

$$E[\varphi(X, Y)] = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \varphi(x_k, y_l) P(X = x_k, Y = y_l)$$

と定義する。ただし、$E[\varphi(X, Y)]$ は右辺が絶対収束する時のみ定義されるものとする。特に、$E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$ を X と Y の共分散という。

また、$P(X = x_k) = \sum_{l=1}^{\infty} P(X = x_k, Y = y_l), P(Y = y_l) = \sum_{k=1}^{\infty} P(X = x_k, Y = y_l)$ となる。このことから、

$$\begin{aligned} E[X + Y] &= E[X] + E[Y] \\ V(X + Y) &= E[(X - m_1 + Y - m_2)^2] = V(X) + V(Y) + 2E[(X - m_1)(Y - m_2)] \end{aligned}$$

が従う。ここで、$m_1 = E[X], m_2 = E[Y]$ とした。

次に独立性の概念を導入する。確率変数 $X_1, X_2, \cdots, X_n$ が**独立**であるとは

$$P(X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n) = P(X_1 = a_1)P(X_2 = a_2)\cdots P(X_n = a_n) \tag{1.4}$$

が任意の実数の組 $a_1, a_2, \ldots, a_n$ に対して成立する時にいう。このとき、$f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ に対して

$$E[f_1(X_1)f_2(X_2)\cdots f_n(X_n)] = E[f_1(X_1)]E[f_2(X_2)]\cdots E[f_n(X_n)] \tag{1.5}$$

が成立する。これより $E[(X_1 - E[X_1])(X_2 - E[X_2])] = E[X_1 - E[X_1]]E[X_2 - E[X_2]] = 0$ であるから

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n) \tag{1.6}$$

を得る。

[6] $\alpha \in \mathbb{R}$ に対し $\binom{-\alpha}{k} = \frac{(-\alpha)(-\alpha-1)\cdots(-\alpha-k+1)}{k!}$ とする。$(x+1)^{-\alpha} = \sum_{k=0}^{\infty}\binom{-\alpha}{k}x^k$ $(|x| < 1)$ に注意。

例題 1.4 表が出る確率が p $(0 < p < 1)$ であるようなコインがある。このコインを n 回投げるとき、k 回目の試行で表が出たら $X_k = 1$, 裏が出たら $X_k = 0$ とする確率変数を考える。このとき、$S = X_1 + \cdots + X_n$ の平均と分散を求めよ。*(*S は二項分布 $B(n, p)$ に従うことも容易に証明できる。*)*

以上で高校数学 B の確率分布の部分はすべて紹介しました。高校時代の教科書を引っ張り出して問題を解いて見てください。

1.3 確率分布 (連続型の場合)

前節では離散的に変化する確率変数を考えたが、例えばバスの待ち時間 (渋滞等で時間通り来るとは限らないので確率変数と思ってもいいだろう) のように連続的に変化するものもあり得る。この節では連続的に変化する確率変数を取り扱う。

確率変数 X が連続的に変化するとは $F(x) = P(X \leq x)$ が x について連続であることである。ここでは、もう少し制限を強めて $F(x)$ が区分的に微分可能な場合を考えよう。

このとき、確率変数 X に対して関数 $f : \mathbb{R} \to \mathbb{R}$ がとれて、

1. $f(x) \geq 0$ かつ $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1$,
2. X が区間 $[a, b]$ の間にある確率 $P(a \leq X \leq b) = \displaystyle\int_a^b f(x)\,dx$,

とできる。この $f(x)$ を X の確率密度関数をという[7]。このとき X の期待値 $E[X]$, 分散 $V(X)$ を

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx, \quad V(X) = \int_{-\infty}^{\infty} (x-m)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - m^2$$

で定める。ただし、$m = E[X]$ であり、期待値は $\int_{-\infty}^{\infty} |x| f(x)\,dx < \infty$ のときのみ定義されるものとする[8]。(分散も同様。)

連続型確率分布の例

- (**一様分布**) X の取りうる値が $[a, b]$ $(a < b)$ でその確率密度関数が

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0 & x \notin [a, b] \end{cases}$$

となるとき X は一様分布 $U(a, b)$ に従うという。このとき、簡単な計算で次が従う:

$$E[X] = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{a+b}{2}, \quad V(X) = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx - E[X]^2 = \frac{(b-a)^2}{12}$$

[7] X が取りうる値が $[\alpha, \beta]$ の場合は $f(x) \equiv 0$ on $x < \alpha, x > \beta$ と考える。

[8] 一見、離散型とは定義が異なるようだが、連続型は離散型の極限と見なすと全く同じ定義となる。実際、ルベーグ積分論により離散型と連続型とは統一的な取り扱いができる。
1917 年に Kolmogorov が当時完成して間もないルベーグ積分論を用いて確率論を再定義して、確率論は飛躍的に発展したことを付け加えておく。

- (**指数分布**) $\lambda > 0$ とする。X の取りうる値が $[0, \infty)$ でその確率密度関数が

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda}, & x \geq 0, \\ 0 & x < 0 \end{cases}$$

となるとき X は指数分布 $e(\lambda)$ ($\mathrm{Ex}(1/\lambda)$ と書く場合もある) に従うという。機械などが故障するまでの時間などがこれに従うと考えられている。このとき、簡単な計算で次が従う:

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

- (**Gamma 分布**) $\alpha > 0, \beta > 0$ とする。確率密度関数が

$$f(x) = \begin{cases} \frac{\beta}{\Gamma(\alpha)} (\beta x)^{\alpha-1} e^{-\beta x}, & x > 0, \\ 0 & x \leq 0 \end{cases}$$

となるとき X は Gamma 分布 $\Gamma(\alpha, \beta)$ に従うという。指数分布 $\mathrm{Ex}(1/\lambda)$ は $\alpha = 1, \beta = 1/\lambda$ となる Gamma 分布の特別な場合と考えられる。X が Gamma 分布 $\Gamma(\alpha, \beta)$ に従うとき、$E[X] = \alpha/\beta, V(X) = \alpha/\beta^2$ となる。(計算は演習問題とする。) [9]

- (**正規分布**) $m \in \mathbb{R}, \sigma > 0$ とする。X の確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad x \in \mathbb{R},$$

となるとき、X は正規分布 $N(m, \sigma^2)$ に従うという。微積分学で習った $\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$ に注意すれば、これが確率密度関数になること、および、$E[X] = m, V(X) = \sigma^2$ はすぐわかる。しかし、この確率密度関数の不定積分は初等的には求まらない。そのため、数学 C の教科書の巻末には「正規分布表」が掲載されている。これを用いるためには次の定理を必要とする。

定理 1.1 確率変数が X が正規分布 $N(m, \sigma^2)$ に従う時、$Z = \frac{X-m}{\sigma}$ とおくと、Z は標準正規分布 $N(0, 1)$ に従う。即ち、$a < b$ に対して

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}} \int_{(a-m)/\sigma}^{(b-m)/\sigma} e^{-\frac{x^2}{2}}\, dx = P\left(\frac{a-m}{\sigma} \leq Z \leq \frac{b-m}{\sigma} \right).$$

証明: $z = \frac{x-m}{\sigma}$ と変数変換すれば

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{(a-m)/\sigma}^{(b-m)/\sigma} e^{-\frac{z^2}{2}}\, dz$$

となるから従う。 □

問題 1.7 確率変数 X が正規分布 $N(4, 25)$ に従うとき、正規分布表を用いて次を求めよ。
(1) $P(X \geq 9)$ *(2)* $P(0 \leq X \leq 9)$ *(3)* $P(X \leq 12)$

[9] 自然数 n に対して X が Gamma 分布 $\Gamma(\frac{n}{2}, 1)$ に従うとき、X は**自由度 n の χ^2-分布**に従うという。χ^2-分布は統計学において重要な分布である。(この授業ではそこまで扱わない。)

2 統計処理

2.1 標本平均の分布

母平均 m, 母標準偏差 σ の母集団がら大きさ n の無作為標本 $X_1, X_2, \ldots, X_n$ を抽出するとき、標本平均 $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ の期待値と標準偏差を考える[10]。

これを数学の言葉に直すと、$X_1, X_2, \ldots, X_n$ は平均 m, 標準偏差 σ なる同じ確率分布をもつ確率変数で、独立であるとき、$\overline{X}$ の期待値と標準偏差を求めよという意味である。ただし、$X_1, \ldots, X_n$ が独立であるとは

$$P(a_1 \le X_1 \le b_1, \cdots, a_n \le X_n \le b_n) = P(a_1 \le X_1 \le b_1) \cdots P(a_n \le X_n \le b_n) \qquad (2.1)$$

なるときにいう。((1.4) の定義は X が連続型のとき $P(X = a) = 0$ $(\forall a \in \mathbb{R})$ となり用いることが出来ない。) このとき、平均値の定義から (独立性を用いないでも)

$$E[\overline{X}] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = m \qquad (2.2)$$

であり、離散型のとき同様 (1.3) の後の注意と (1.6) が成り立つことから

$$V(\overline{X}) = \frac{1}{n^2}\{V(X_1) + V(X_2) + \cdots + V(X_n)\} = \frac{\sigma^2}{n} \qquad (2.3)$$

を得る。実は、次の定理が知られている。この (2) がこの章の中心的役割を果たす。証明は 3 年次の確率統計学 ((1) は I, (2) は II) の授業で行なう。

定理 2.1 *(1) (***大数の法則***)* 任意の $\delta > 0$ に対して次が成立する*:*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - m\right| \ge \delta\right) = 0.$$

*(2) (***中心極限定理***)* $S_n = X_1 + \cdots + X_n$ とすると、任意の $a, b \in \mathbb{R}$ *$(a < b)$* に対して

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - nm}{\sqrt{n}\sigma} \le b\right) = P(a \le Z \le b)$$

となる。ここで、Z は標準正規分布 $N(0, 1)$ に従う確率変数である。

系 2.2 *(***de Moivre-Laplace の定理***)* S_n が二項分布 $B(n, p)$ に従うとき、

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx, \qquad a, b \in \mathbb{R}\ (a < b).$$

証明 $\{X_n\}$ が独立同分布で $P(X_1 = 1) = p, P(X_1 = 0) = 1-p$ とすると、$S_n = X_1+\cdots+X_n$ は二項分布 $B(n, p)$ に従い (*cf.* 例 1.4)、また、$E[X_1] = p, V(X_1) = p(1-p)$ より主張は中心極限定理の特別な場合として得られる。(スターリングの公式 $n! \sim \sqrt{2n\pi}n^n e^{-n}$ から直接証明することもできる。ただし、かなり hard である。) □

[10] **標本調査**とは集団の要素から一部だけを抜き出して調べ、その結果から全体の状況を推測することを目的とした統計調査を意味する。本来調べたい対象の全体の集まりを**母集団**、調査のために母集団から抜き出された要素を**標本**といい、母集団から標本を抜き出すことを、標本の**抽出**という。また、母集団, 標本の要素の個数を、それぞれ母集団, 標本の**大きさ**という。

例題 2.1 全国の有権者の内閣支持率が *40%*であるとき、無作為に抽出した *2,400* 人の有権者の内閣支持率を R とする。R が *38%*以上 *42%*以下である確率を求めよ。また、抽出した有権者数が *600* 人だった場合はどうか調べよ。

解: $p = 0.4, n = 2,400$ とすると投票率 R について、$nR(= S_n)$ は二項分布 $B(n,p)$ に従う。よって、系 2.2 を用いると $Z = \frac{S_n - np}{\sqrt{np(1-p)}} = \frac{R-0.4}{0.01}$ は近似的に標準正規分布 $N(0,1)$ に従う。よって、正規分布表を用いると $P(0.38 \leq R \leq 0.42) = P(-2 \leq Z \leq 2) = 2 \cdot 0.4772 = 0.9544$ を得る。$n = 600$ の場合は $Z = \frac{S_n - np}{\sqrt{np(1-p)}} = \frac{R-0.4}{0.02}$ が近似的に標準正規分布 $N(0,1)$ に従うから、$P(0.38 \leq R \leq 0.42) = P(-1 \leq Z \leq 1) = 2 \cdot 0.3413 = 0.6826$ となる。 □

例題 2.2 大手予備校の模試で、数学の成績を *10* 点きざみの度数分布に整理して平均点を計算する。平均点の誤差が *0.1* 点以内に収まる確率が *0.95* 以上にしたい。およそ、何名以上の受験生が必要か。中心極限定理を用いて求めよ。

解: 受験生を n 人とする。各受験生の得点と階級差との差 $X_1, X_2, \ldots, X_n$ は独立で、各 X_i は一様分布 $U(-5,5)$ に従うから、$E[X_i] = 0, V(X_i) = 10^2/12$ $(i = 1, \ldots, n)$. 中心極限定理より $Y_n = S_n/\sqrt{10^2 n/12}$ は漸近的に $N(0,1)$ に従う。よって、

$$P\Big(\Big|\frac{1}{n}S_n\Big| \leq 0.1\Big) = P\Big(|Y_n| \leq \frac{0.1}{\sqrt{10^2/12n}}\Big) \geq 0.95.$$

故に、$\frac{0.1}{\sqrt{10^2/12n}} \geq 1.96$. これを解いて $n \geq 3201.33\cdots$. 以上より、3,202 名以上. □

問題 2.1 サイコロを *1000* 回投げたとき、*1* が *160* 回以上出る確率の近似値を中心極限定理を用いて求めよ。また、*1* が l 回以上出る確率を *1%*以下とするためには、l の値はどの程度大きくすればよいか。

問題 2.2 サイコロを n 回投げたとき、*1* が出る回数とその平均の差が *10* 以下である確率が *0.95* 以上であるには n はどの位であればよいか。

2.2 推定

ここでは、母平均がわからないとき、それを標本平均を用いて推定する方法について考えて見よう。

母平均 m, 母分散 σ^2 をもつ母集団から、大きさ n の無作為標本 $X_1, X_2, \ldots, X_n$ を抽出するとき、標本平均 $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ について、定理 2.1 (2) より、n が大きいとき、近似的に $\frac{\overline{X} - m}{\sigma/\sqrt{n}}$ は標準正規分布 $N(0,1)$ に従うから、

$$P\Big(\Big|\overline{X} - m\Big| \leq u(\frac{\varepsilon}{2})\frac{\sigma}{\sqrt{n}}\Big) = 1 - \varepsilon$$

がわかる。ただし、$u(\varepsilon)$ で標準正規分布 $N(0,1)$ の上側 ε 点を表す: $\int_{u(\varepsilon)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \varepsilon.$

即ち、母平均 m について

$$P\Big(\overline{X} - u(\frac{\varepsilon}{2})\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + u(\frac{\varepsilon}{2})\frac{\sigma}{\sqrt{n}}\Big) = 1 - \varepsilon$$

と書くことができる。このとき、区間

$$\Big[\,\overline{X} - u(\frac{\varepsilon}{2})\frac{\sigma}{\sqrt{n}}\,,\,\overline{X} + u(\frac{\varepsilon}{2})\frac{\sigma}{\sqrt{n}}\,\Big]$$

を、母平均 m に関する**信頼度** $1-\varepsilon$ の**信頼区間**という。

例題 2.3 大量生産されたある規格の部品から *400* 個を無作為に抽出して重さを計ったところ、平均値は 98.5g であった。全製品の平均重量を、信頼度 *95%*で推定せよ。ただし、母標準偏差は σ =3.1g であるとする。

解: 標本平均は $\overline{x} = 98.5$, 母標準偏差は $\sigma = 3.1$, 標本の大きさは $n = 400$ であるから、信頼度 95%の信頼区間は

$$\Big[98.5 - u(0.025) \times \frac{3.1}{\sqrt{400}}\,,\,98.5 + u(0.025) \times \frac{3.1}{\sqrt{400}}\,\Big]$$

ここで、$u(0.025) = 1.96$ だから、求める信頼区間は $[\,98.2\,,\,98.8\,]$. (単位は g。) □

問題 2.3 あるテープの長さをある人が測定するときの誤差は mm 単位で正規分布 $N(0, 0.2^2)$ に従う。このとき、次の問いに答えよ。ただし、$u(0.005) = 2.58$ を用いよ。
(1) *10* 回測定して、その平均値によって真の長さを推定するとき、信頼度 *99%*の信頼区間の幅はいくらになるか。
(2) n 回測定して、その平均値によって真の長さを推定するとき、信頼度 *99%*の信頼区間の幅を *0.2*mm 以下にするためには、n をいくら以上にすればよいか。

母比率 p の推定区間について考える。

母比率が p である集団から大きさ n の無作為標本 $X_1, \ldots, X_n$ を抽出するとき、その標本比率 $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ に対し $n\hat{p}$ は二項分布 $B(n,p)$ に従うから、n が大きいとき、$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ は近似的に標準正規分布に従う (de Moivre-Laplace の定理)。これより、

$$P\Big(-u(\frac{\varepsilon}{2}) \le \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \le u(\frac{\varepsilon}{2})\Big) = 1-\varepsilon$$

となるから $\hat{p} - u(\frac{\varepsilon}{2})\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + u(\frac{\varepsilon}{2})\sqrt{\frac{p(1-p)}{n}}$. ここで根号内の p を標本平均 $\hat{p}$ に置き換え、信頼係数 $1-\varepsilon$ の信頼区間 $\Big[\,\hat{p} - u(\frac{\varepsilon}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\,\hat{p} + u(\frac{\varepsilon}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,\Big]$ を得る。

例題 2.4 ある選挙区における *1* 人の候補者の支持率を信頼度 *95%*の信頼区間の幅が *2%*以下であるように推定するにはどの位の大きさの標本を抽出すればよいか。

解: 支持率の推定区間は $\big[\,\hat{p} - u(\frac{\varepsilon}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\,\hat{p} + u(\frac{\varepsilon}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,\big]$. よって、

$$(\text{信頼区間の幅}) = 2u(\frac{\varepsilon}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \frac{2u(\frac{\varepsilon}{2})}{\sqrt{n}}\sqrt{\frac{1}{4} - (\hat{p} - \frac{1}{2})^2} \le \frac{u(\frac{\varepsilon}{2})}{\sqrt{n}}$$

であるから $\frac{u(0.025)}{\sqrt{n}} \le 0.02$ を解いて $n \ge 9604$. □

問題 2.4 ある工場で、製品の中から *200* 個を取り出して調べたところ、*25* 個の不良品があったという。製品全体について、不良品の割合を信頼度 *99%*で推定せよ。ただし、$u(0.005) = 2.58$ を用いよ。

2.3 検定

標本調査によって、母集団の性質を推測するのに、仮説をたて、その仮説を確率によって判定する方法がある。これを、まず、例について説明する。

例題 2.5 *2*つのサイコロ *A, B*がある。これが正しく作られたサイコロかどうか調べるために、どちらも *50*回投げてみたら、*1*の目が出た回数が、*A*では *16*回、*B*では *6*回であった。*A* はあまりにも 1 の目が出やすく、*B* は出来にくいと思われる。*A, B* とも正しくないと判定してよいか。

解: 正しいサイコロであったとし、50 回投げたときの 1 の目が出る回数を X とすると、$E[X] = 50 \cdot \frac{1}{6}, V(X) = 50 \cdot \frac{1}{6}\frac{5}{6}$ より de Moivre-Laplace の定理により $Z = \dfrac{X - 50 \cdot \frac{1}{6}}{\sqrt{50 \cdot \frac{1}{6}\frac{5}{6}}}$ は近似的に標準正規分布に従う。これより 1 の目が 16 回以上の出る確率, 6 回以下 1 の目の出る確率はそれぞれ

$$P(X \geq 16) = P(Z \geq 2.91) = 0.00181, \quad P(X \leq 6) = P(Z \leq -0.885) = 0.188.$$

これより、A については 100 回に 1 回も起こらないが、B については 5 回に 1 回くらい起こりうることだと判断できる。よって、A のサイコロは正しくないと判断されるが、B のサイコロはこの調査だけでは正しいかどうか判断できない。[11] □

上記では、はじめにサイコロが正しいという仮説をたて、その仮説をもとに、出てきた結果の確率を求めて、その確率があまりにも小さいとき、はじめの仮設が正しくないとして棄却したのである。このような方法を**仮設の検定**という。

上記の例では「A については 100 回に 1 回も起こらない」からめったに起こらないとして仮設は棄却された。このようにめったに起こらないとはどの程度なのかあらかじめ定めておく必要がある。統計の習慣では 5%または 1%以下とすることが多い。この確率を**危険率**または**有意水準**という。

問題 2.5 甲地域で行った数学の試験において、*1600*人の成績を任意抽出して、その平均値と標準偏差を求めたところ、それぞれ *58.3*点と *12.5*点であった。乙地域で同じ問題で行った試験の結果はすでに知られていて、平均点は *59*点であった。甲地域での全体の平均値は *59*点であるとみなせるであろうか。有意水準 *5%*で検定せよ。

問題 2.6 ある市長選挙で、*A, B*二人が立候補した。いま、二人の支持率に差がないかどうか調べたいので、有権者の中から無作為に *900*人を選んで調べたところ、*500*人が *A* を支持し、*400*人が *B* を指示した。二人の支持率に差があるか有意水準 *1%*で検定せよ。

問題 2.7 あるテレビ番組の視聴率は従来 *10%*であると考えられていた。任意抽出により *1600*世帯をえらんで調査したところ、*120*世帯が視聴していることがわかった。視聴率は *10%*とみなさるか。有意水準 *1%*で検定せよ。

[11] B のサイコロはついてここからわかることは、「この調査だけでは正しいかどうか判断できない」のであって、「サイコロが正しいこと」はわかっていないことに注意する。

松本裕行・宮原孝夫「数理統計入門」学術図書より