

「確率・統計」高校と大学の狭間で

流山中央高等学校 中村 明

はじめに

昨年、千葉大での補習授業の前期「微分積分」を報告しましたので、今回はその続編、後期「確率・統計」分の概要を下記の通り報告します。
(このレポートは、その補習授業の反省を踏まえて加筆、訂正したもの)

1 第1回補習授業出席者8名の学生の内訳

- (1) 学部の内訳は園芸4名、工学(夜間)4名。
- (2) 男女各4名。
- (3) 全員推薦入試合格者。
- (4) 出身地域は、千葉県4名。他の4名は関西。

2 補習授業の目的

千葉大に推薦入試で合格した(意欲はあるが数学の知識の不足している)学生に対して高校の数学を補い、大学の「統計学」の補習授業をすることで、結果として大学の「統計学」の単位が取れるように援助すること。

3 後期分10回の授業内容 (1回の授業時間: 規定 60 分: 実際平均 70 分)

◎1回目授業

- (1) 高校時代の数学の履修状況の把握(アンケートを実施)
- (2) 中村の自己紹介(ノートパソコン持参で動画・静止画を見せての自己紹介)
- (3) 高校数学Iの順列・組合せ・確率の復習

◎2回目授業——確率変数と確率分布の考え方の重要性について

例1 コインを3回投げる試行を考える。確率変数 X : この試行において表の出る回数を X とする。

X	0	1	2	3	計
$P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

「この問題を、表の出る確率 $a = \frac{1}{2}$ 、裏の出る確率 $b = \frac{1}{2}$ 、 $a + b = 1$ として

$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ (の様に) 展開すると求めやすいかも?

この確率分布はコインを3回投げたときの表の出る状況を網羅しているので、今までの単なる一つの確率の計算とは違う価値のある表(ひょう)であることを理解する必要がありますね。」

例2 箱の中に 2, 3, 4, 5, 6 の番号のついた球が5個入っている。この中から2個の球を取り出すとき、次の各問に答えよ。

(1) 復元抽出①, 非復元抽出②, 同時抽出③の3つの場合において、標本空間の要素の数を求めよ。

(解) ① 復元抽出 2, 3, 4, 5, 6 の5通りから $5 \times 5 = 5^2 = 25$ (答)

② 非復元抽出 ${}_5P_2 = 5 \times 4 = 20$ (答)

③ 同時抽出 ${}_5C_2 = \frac{5 \times 4}{2 \times 1} = 10$ (答)

(2) 非復元抽出②の場合において、 $X = (\text{得られた2つの球の番号の差の絶対値})$ とする時、 X の確率分布 $P(X)$ を求めなさい。また、 X の平均 μ_x , 標準偏差 σ_x を求めなさい。

(解)

	2	3	4	5	6
2		1	2	3	4
3	1		1	2	3
4	2	1		1	2
5	3	2	1		1
6	4	3	2	1	

左の表より

X	1	2	3	4	計
$P(X)$	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	1

$$\begin{aligned}
 \text{平均 } \mu_x &= 1 \times \frac{4}{10} + 2 \times \frac{3}{10} + 3 \times \frac{2}{10} + 4 \times \frac{1}{10} \\
 &= \frac{1}{10}(1 \times 4 + 2 \times 3 + 3 \times 2 + 4 \times 1) \\
 &= \frac{1}{10} \times 20 \\
 &= 2 \text{ (答)}
 \end{aligned}$$

X	1	2	3	4	計
$P(X)$	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	1
$XP(X)$	$\frac{4}{10}$	$\frac{6}{10}$	$\frac{6}{10}$	$\frac{4}{10}$	2
$X^2P(X)$	$\frac{4}{10}$	$\frac{12}{10}$	$\frac{18}{10}$	$\frac{16}{10}$	5

$E(X)$

$E(X^2)$

$$\begin{aligned}
 \text{標準偏差 } \sigma_x &= \sqrt{\sigma^2} = \sqrt{E(X^2) - E(X)^2} \quad (\text{公式}) \\
 &= \sqrt{5 - 2^2} = \sqrt{1} = 1 \text{ (答)}
 \end{aligned}$$

◎3回目授業——条件付確率と二項分布

例3 今、稀にしか起こらない特別の病気を発見するのに、ある検査法が有効である。この検査法を、実際に病気に冒されている人に適用するならば、0.97の確率で陽性反応を示す。また、この検査法を健康な人に適用すると、0.05の確率で陽性反応を示す。さらに、別の軽い病気(風邪)にかかっている人に適用すると、0.10の確率で陽性反応を示す。また、多人数からなる母集団では、ここで考えた3種類の人々の割合は0.01, 0.96, 0.03であることが分かっているとする。次の各問いに答えなさい。

(1) 母集団から無作為に選ばれた1人が陽性反応を示す確率を求めなさい。

(2) このとき、その人が本当に病気に冒されている確率を求めなさい。

(解) A:病気に冒されている, B:健康, C:別の軽い病気にかかっている, D:陽性反応が出た, とする。題意より $P_A(D) = 0.97$, $P_B(D) = 0.05$, $P_C(D) = 0.10$
 $P(A) = 0.01$, $P(B) = 0.96$, $P(C) = 0.03$

$$\begin{aligned}(1) \quad P(D) &= P(A)P_A(D) + P(B)P_B(D) + P(C)P_C(D) \\ &= 0.97 \times 0.01 + 0.05 \times 0.96 + 0.10 \times 0.03 \\ &= 0.0097 + 0.0480 + 0.003 \\ &= 0.0607 \text{ (答)}\end{aligned}$$

(2) 条件付確率の定理 $P_X(Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)}$ より

$$P_D(A) = P(A|D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A)P_A(D)}{P(D)} = \frac{0.01 \times 0.97}{0.0607} = 0.1598 \text{ (答)}$$

例 4 ある店の商品の一つ「ヒューズ」の 10 % が不良品である。購入した 10 個の「ヒューズ」が 8 個以上の良品を含む確率と、購入した 10 個の「ヒューズ」が 2 個以上の不良品を含む確率を求め、両者を比較しなさい。また、電卓を用いてこの確率を小数点以下第 3 位まで求めなさい。

(解) (1) 10 個中 8 個の良品を含む確率 p_1 は

$$\begin{aligned}p_1 &= {}_{10}C_8 \left(\frac{9}{10}\right)^8 \left(\frac{1}{10}\right)^2 + {}_{10}C_9 \left(\frac{9}{10}\right)^9 \left(\frac{1}{10}\right) + \left(\frac{9}{10}\right)^{10} \\ &= \frac{9^8}{10^{10}} ({}_{10}C_8 + {}_{10}C_9 \times 9 + 9^2) = \frac{9^8}{10^{10}} ({}_{10}C_2 + 9 \times {}_{10}C_1 + 81) \\ &= \frac{9^8}{10^{10}} (45 + 90 + 81) = \frac{9^8}{10^{10}} \times 216 = 0.9298 \quad 0.930 \text{ (答)}\end{aligned}$$

(2) 10 個中 2 個の不良品を含む確率 p_2 は

$$\begin{aligned}p_2 &= {}_{10}C_8 \left(\frac{9}{10}\right)^8 \left(\frac{1}{10}\right)^2 + {}_{10}C_7 \left(\frac{9}{10}\right)^7 \left(\frac{1}{10}\right)^3 + \cdots + \left(\frac{1}{10}\right)^{10} \\ &= 1 - \left({}_{10}C_9 \left(\frac{9}{10}\right)^9 \left(\frac{1}{10}\right) + \left(\frac{9}{10}\right)^{10} \right) \\ &= 1 - \frac{9^9}{10^{10}} (10 + 9) = 1 - \frac{9^9}{10^{10}} \times 19 \\ &= 1 - 0.7360 = 0.2640 \quad 0.264 \text{ (答)}\end{aligned}$$

◎ 4 回目授業——分布の特性値の特徴について

(紙面の都合により省略)

◎ 5 回目授業——正規分布について

例 5 統計学の試験の得点分布は、平均 60 点、標準偏差 10 点の正規分布であった。50 点以上を合格とするとき、この試験で合格となる学生の割合はいくらか？
 また、上位 5 % に特優の成績を与えたい。何点以上とすべきか？

(解) $Z = \frac{X - \mu}{\sigma} \dots \textcircled{1}$ (標準化公式) に代入

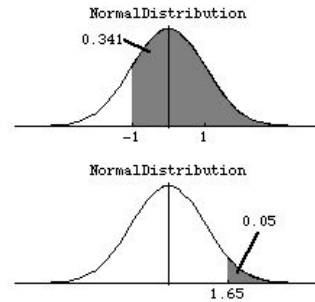
$$(1) Z = \frac{50 - 60}{10} = -1$$

$$\rho = 0.341 + \frac{1}{2} = 0.841 \text{ より, (答) } 84.1 \%$$

(2) 5%の右半分の残りの面積は 0.45 だから,
(正規分布表より)

$X = 1.65$ 等を $\textcircled{1}$ に代入

$$1.65 = \frac{X - 60}{10} \quad \therefore X - 60 = 16.5 \quad \therefore X = 76.5 \text{ 点 (答)}$$

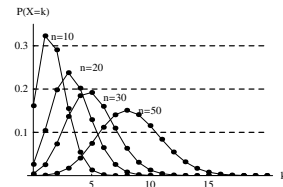


ガウスはどのようにして二項分布から正規分布を導き出したのであろうか？

我々も考えよう！まずは二項分布の数の変化とグラフの推移に着目。

二項分布 $B(n, p)$ において,

$p = \frac{1}{6}$ のときの $n = 10, 20, 30, 50$ のそれぞれに対する確率 $P(X = k)$ の値を求めてグラフにすると、次のようになる。



では、いよいよ離散型確率関数 $p_B(x) = {}_n C_x p^x q^{n-x} (p + q = 1) \dots \textcircled{1}$ から連続型正規分布を導いてみよう！

$\textcircled{1}$ を $B(n, p)$ と表し, $f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \dots \textcircled{2}$ を $N(\mu, \sigma^2)$ で表す。

基本知識 平均 $E_B(X) = np = \mu \dots \textcircled{3}$, 分散 $V_B(X) = npq \dots \textcircled{4}$ を利用する。

n を 100, 1000, 10000 等と十分に大きくしていく場合 $n \gg 0$ と表現。

$p_B(x) = {}_n C_x p^x q^{n-x}$ の自然対数をとったものを新たに $g(x)$ とおく。すなわち

$$\begin{aligned} g(x) &= \log p_B(x) = \log \left(\frac{n!}{x! (n-x)!} p^x (1-p)^{n-x} \right) \\ &= \log n! - \log x! - \log(n-x)! + x \log p + (n-x) \log(1-p) \dots \textcircled{5} \end{aligned}$$

ここで離散型関数 $h(x) = \log x! (x \gg 0)$ のとき, (乱暴にも) $\Delta x = 1$ として,

$$\begin{aligned} h'(x) &\doteq \frac{h(x) - h(x - \Delta x)}{\Delta x} = \frac{h(x) - h(x-1)}{1} \\ &= \log x! - \log(x-1)! \\ &= \log \frac{x!}{(x-1)!} = \log x \end{aligned}$$

$$\therefore (\log x!) \doteq \log x \dots \textcircled{6} (x \gg 0 \text{ のとき})$$

注意 この計算で離散型関数 $h(x) = \log x! (x \gg 0 \text{ のとき})$ について計算ソフト「カルキングJ」等で計算すれば、 $100! \approx 9.33 \times 10^{157}$ になってしまうので、上記

$$\frac{h(x) - h(x-1)}{1} = \log x! - \log(x-1)! = \log \frac{x!}{(x-1)!} = \log x$$

の意味 (平均変化率 $\Rightarrow$ 微分係数 $\Rightarrow$ 導関数) を考え、「 $x \gg 0$ のとき」の条件で

$$h'(x) \doteq \frac{h(x) - h(x-\Delta x)}{\Delta x} = \frac{h(x) - h(x-1)}{1} \text{ を中村は可とした。 (注意終)}$$

$$\left(\log(n-x)! \right)' \doteq \left(\log(n-x) \right) (n-x)' = -\log(n-x) \cdots \textcircled{7}$$

以上より⑤の $g(x)$ の x における微分係数は

$$\begin{aligned} g'(x) &= -(\log x!)' - \left(\log(n-x)! \right)' + \log p - \log(1-p) \\ &= -\log x + \log(n-x) + \log p - \log(1-p) \cdots \textcircled{8} \\ &= \log \frac{p(n-x)}{(1-p)x} \end{aligned}$$

ここで、 x を連続型変数とみなして $g'(x) = 0$ のとき $\log \frac{p(n-x)}{(1-p)x} = 0$ より $\frac{p(n-x)}{(1-p)x} = 1$

分母を払うと $np - px = x - px \quad \therefore x = np (= \mu)$

逆に、 $x = np (= \mu)$ のとき $g'(x) = 0$ となり、 $x = np$ で p_B は極値をとる。

⑧の $g'(x)$ をさらに x で微分すると

$$g''(x) = -\frac{1}{x} + \frac{-1}{n-x} = -\frac{n}{x(n-x)}$$

この x に $\mu (= np)$ を代入して

$$g''(\mu) = -\frac{n}{\mu(n-\mu)} = -\frac{n}{np(n-np)} = -\frac{1}{np(1-p)} = -\frac{1}{npq} = -\frac{1}{\sigma^2} \cdots \textcircled{9}$$

$g(x)$ を $x = \mu$ のまわりにテイラー展開すると、

$$g(x) = g(\mu) + \frac{g'(\mu)}{1!}(x-\mu) + \frac{g''(\mu)}{2!}(x-\mu)^2 + \frac{g'''(\mu)}{3!}(x-\mu)^3 + \cdots$$

$$g'(\mu) = 0, \quad g''(\mu) = -\frac{1}{\sigma^2}, \quad \frac{(x-\mu)^k}{k!} \doteq 0 \text{ (右は全て0) より}$$

$$g(x) \doteq g(\mu) - \frac{1}{2\sigma^2}(x-\mu)^2 \cdots \textcircled{10}$$

ここで、 $g(x) = \log p_B(x)$ より⑩は

$$\log p_B(x) \doteq \log p_B(\mu) + \log e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \log ce^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\log p_B(\mu) = c \text{ とおいた})$$

以上より、 $n \gg 0$ 、すなわち $x \gg 0$ のとき二項分布 $p_B(x)$ は、正規分布の密度関数

$f_N(x) = ce^{-\frac{(x-\mu)^2}{2\sigma^2}}$ に近づく。ここで、定数 c を求める。

$$\int_{-\infty}^{\infty} f_N(x) dx = 1 \text{ より } c \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 \cdots \textcircled{11}$$

ここで $z = \frac{x-\mu}{\sigma}$ とおくと、 $x: -\infty \rightarrow \infty$ のとき $z: -\infty \rightarrow \infty$ 、 $dz = \frac{1}{\sigma} dx$ より $dx = \sigma dz$ 。

$$\text{以上より⑪は } c \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \sigma dz = 1$$

$c\sigma$ 以外の定積分の値は $\sqrt{2\pi}$ より $\sqrt{2\pi}c\sigma = 1$ 。

$c = \frac{1}{\sqrt{2\pi}\sigma}$ とわかったので、②の $f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ が導かれたことになりますね。

お疲れさまでした。

付録

テイラー展開等の公式の簡単な覚え方の一つを紹介。

接線の公式 (数学 II) $f(x) - f(x_1) = f'(x)(x - x_1)$ の $f(x_1)$ を移項し

$$f(x) = f(x_1) + f'(x_1)(x - x_1) \cdots \textcircled{14}$$

を「基本の式」とみなして分子に微分係数，分母に階乗，全体には $(x - x_1)^n$ を次々に付け加えて無限級数を作る。すなわち

$$\begin{aligned} f(x) = f(x_1) + \frac{f'(x_1)}{1!}(x - x_1) + \frac{f''(x_1)}{2!}(x - x_1)^2 \\ + \frac{f^{(3)}(x_1)}{3!}(x - x_1)^3 + \cdots + \frac{f^{(n)}(x_1)}{n!}(x - x_1)^n + \cdots \quad \textcircled{15} \end{aligned}$$

⑬に $x_1 = 0$ を代入すると有名なマクローリン展開が出現する。すなわち

$$\begin{aligned} f(x) &= f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots \\ &= \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k \quad (\text{ただし } 0! = 1, f^{(0)}(0) = f(0)) \end{aligned}$$

◎ 6 回目授業——大数法則について

ある試行を 1 回行って事象 A の起こる確率 p が数学的に分からなくても，この試行を n 回行い，そのうち x 回だけ事象 A が起こったとき， $\frac{x}{n}$ を計算する。そして，この n を無限に大きくしていったとき，それは確率 p に限りなく近づく。

すなわち $\lim_{n \rightarrow \infty} \frac{x}{n} = p$ このことを大数法則と呼ぶ。

(この x は n 回行ったあとの結果であるので， x は n のある関数と考えることができる)

(解説) 高校の授業ではサイコロや硬貨の実験やコンピュータを使った模擬実験でこの大数法則を説明されるが，今回は前回学習した正規分布を使って証明する。

(証明) まず， n 回この試行を行って，事象 A の起こる回数 x は二項分布に従う。

$$p_B(x) = {}_n C_x p^x q^{n-x} (\text{ただし } p + q = 1) \cdots \textcircled{1}, \text{平均 } E_B(X) = np = \mu \cdots \textcircled{2},$$

$$\text{分散 } V_B(X) = npq \cdots \textcircled{3} \text{ とする。}$$

①の n をどんどん大きくしてゆくと正規分布

$$f_N(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (\mu = np, \sigma^2 = npq) \cdots \textcircled{4}$$

となる。ここで新たに確率変数 $\bar{x} = \frac{x}{n}$ とおき， $f_{\bar{x}}(\bar{x})$ を求める ($\lim_{n \rightarrow \infty} \bar{x} = \lim_{n \rightarrow \infty} \frac{x}{n} = p$ を示せば大数法則が証明されたことになる)。

$\bar{x} = \frac{x}{n}$ より $x = n\bar{x} \dots$ ⑤だから $x: -\infty \rightarrow \infty$ のとき $\bar{x}: -\infty \rightarrow \infty \dots$ ⑥

⑤から $\frac{dx}{d\bar{x}} = n \dots$ ⑦

⑤, ⑥, ⑦から $\int_{-\infty}^{\infty} f_{\bar{x}}(\bar{x}) d\bar{x} = \int_{-\infty}^{\infty} f_x(x) dx = \int_{-\infty}^{\infty} f_x(n\bar{x}) d\bar{x} \frac{dx}{d\bar{x}} = 1 \dots$ ⑧

⑧から $f_{\bar{x}}(\bar{x}) = n f_x(n\bar{x}) = n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(n\bar{x}-\mu)^2}{2\sigma^2}} \dots$ ⑨

⑨に④の $\mu = np, \sigma^2 = npq$ を代入すると

$$f_{\bar{x}}(\bar{x}) = n \frac{1}{\sqrt{2\pi}\sqrt{npq}} e^{-\frac{(n\bar{x}-np)^2}{2npq}} = \frac{1}{\sqrt{2\pi}\sqrt{\frac{pq}{n}}} e^{-\frac{(\bar{x}-p)^2}{2\frac{pq}{n}}} \dots$$
 ⑩

⑩より確率変数 $\bar{x}$ は平均 $\mu_{\bar{x}} = p$, 分散 $(\sigma_{\bar{x}})^2 = \frac{pq}{n}$ の正規分布 $N\left(p, \frac{pq}{n}\right)$ に従う。

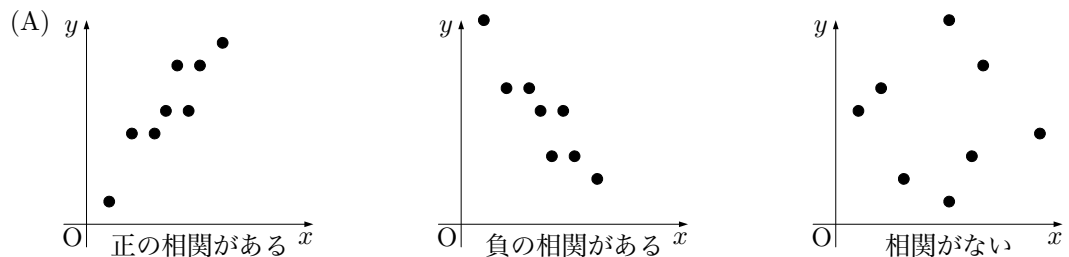
ここでさらに, $n \rightarrow \infty$ とすると $\lim_{n \rightarrow \infty} (\sigma_{\bar{x}})^2 = \lim_{n \rightarrow \infty} \frac{pq}{n} = 0 (\because p, q \text{ は定数}, n \rightarrow \infty)$

分散が限りなく 0 に近づくと, 分布に散らばりがなく, すなわち

$$\lim_{n \rightarrow \infty} \bar{x} = \lim_{n \rightarrow \infty} \frac{x}{n} = p$$

(証終)

◎ 7 回目授業——2 変数データの整理, 相関係数と最小二乗法



(B) 共分散・相関係数の定義等

2 変数データ (x_i, y_i) ($i = 1, 2, 3, \dots, n$) について

(1) x の分散 σ_x^2 , y の分散 σ_y^2

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 \dots$$
 ①

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2 \dots$$
 ②
(定義) (計算式)

$$(2) x \text{ と } y \text{ の共分散 } \sigma_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y} \dots$$
 ③

$$(3) \text{ 相関係数 } \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} (-1 \leq \rho_{xy} \leq 1) \cdots \textcircled{4}$$

- ① ρ_{xy} が 1 に近いほど正の相関が強い。
 ② ρ_{xy} が -1 に近いほど負の相関が強い。
 ③ ρ_{xy} が 0 に近いほど相関がない。

例 6 次に示すのは、10 の小さな宝石店についての広告費支出 (総経費に対する%) と純利益 (売り上げに対する%) を調べたものである。広告費 (x) と利益 (y) の相関係数を求めなさい。ただし、電卓使用可とする。

広告費 (x)	1.2	0.7	1.5	1.8	0.5	3.4	1.0	3.0	2.8	2.5
利益 (y)	2.7	2.4	2.7	3.3	1.1	5.8	2.2	4.2	4.4	3.8

$$\begin{aligned}
 (\text{略解}) \quad \bar{x} &= \frac{1}{10}(1.2 + 0.7 + 1.5 + 1.8 + 0.5 + 3.4 + 1.0 + 3.0 + 2.8 + 2.5) = 1.84 \\
 \bar{y} &= \frac{1}{10}(2.7 + 2.4 + 2.7 + 3.3 + 1.1 + 5.8 + 2.2 + 4.2 + 4.4 + 3.8) = 3.26
 \end{aligned}$$

$$\bar{x} \times \bar{y} = 1.84 \times 3.26 = 5.9984 \approx 6.00$$

次に、電卓で

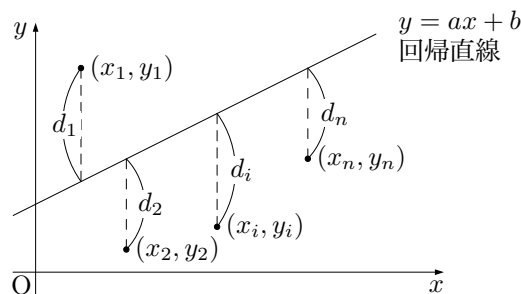
$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 9.46, \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 16.08, \quad \sum_{i=1}^{10} x_i y_i = 71.8$$

ところで

$$\begin{aligned}
 \sigma_{xy} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \times \bar{y} \\
 &= \frac{1}{10} \times 71.8 - 6.00 = 1.18
 \end{aligned}$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{1.18}{\sqrt{\frac{9.46}{10}} \sqrt{\frac{16.08}{10}}} = \frac{11.8}{3.076 \times 4.010} = \frac{11.8}{12.3} = 0.96 (\text{答})$$

最小二乗法について



(紙面の都合で計算は一部省略)

上図のように、データ各点 (x_i, y_i) が与えられたとき、回帰直線の切片 b と傾き a を求める方法の一つを紹介しよう。

$$S = \sum_{i=1}^n \left(y_i - (ax_i + b) \right)^2 = d_1^2 + d_2^2 + \cdots + d_n^2 \cdots \textcircled{5}$$

S を最小にする直線を求める。この未知数を求める方法を最小二乗法という。上の⑤式 $S(a, b)$ を a, b で偏微分して 0 とおいた方程式の解を代入すると回帰直線が得られる (何故かは各自考える)。

すなわち a, b は

$$S_b = \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - ax_i - b)^2 = \sum_{i=1}^n 2(y_i - ax_i - b)(-1) = 0 \text{ より}$$

$$\sum_{i=1}^n (y_i - ax_i - b) = 0 \text{ すなわち } a \sum_{i=1}^n x_i + b \sum_{i=1}^n 1 = \sum_{i=1}^n y_i \cdots \textcircled{6}$$

$$S_a = \frac{\partial}{\partial a} \sum_{i=1}^n (y_i - ax_i - b)^2 = \sum_{i=1}^n 2(y_i - ax_i - b)(-x_i) = 0 \text{ より}$$

$$\sum_{i=1}^n (x_i y_i - ax_i^2 - bx_i) = 0 \text{ すなわち } a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \cdots \textcircled{7}$$

⑥, ⑦からなる連立方程式 (この方程式を正規方程式という) の解である。

(ここはかなりの計算が必要ですが) その結果,

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2} = \rho_{xy} \frac{\sigma_y}{\sigma_x} \cdots \textcircled{8}, \quad b = \bar{y} - a\bar{x} \cdots \textcircled{9}$$

まとめると, n 個の 2 変数データ (x_i, y_i) がある。このとき y の x への回帰直線は次の式で求められる。

$$y = ax + b \iff \frac{y - \bar{y}}{\sigma_y} = \rho_{xy} \frac{x - \bar{x}}{\sigma_x}$$

中心極限定理 (証明は難しいので省略) (こんなすごい定理はあまり見かけない!)

互いに独立な $X_1, X_2, X_3, \dots, X_n$ が平均 μ , 分散 σ^2 の同一の確率分布 (確率密度) に従うとき $\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$ とおく。ここでさらに確率変数 Z を $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ で定義すると Z は $n \rightarrow \infty$ のとき, 標準正規分布 $N(0, 1)$ に従う。

◎ 8 回目授業——区間推定の一部

母集団から取られた標本から母集団の特徴を推測することを推定という (前半は大標本法, 後半は小標本法の解説)。

(1) 母集団平均 μ の推定。

母集団の変数 X の平均を μ , 標準偏差を σ , 標本を $X_1, X_2, \dots, X_n$, 標本平均を $\bar{X}$, 標本標準偏差 s とする。

標準偏差 σ が既知, 任意の母集団分布, 標本の大きさが $n \geq 25$ のとき,

中心極限定理: 母集団が任意の分布で σ が既知, $n \geq 25$ のときの標本平均 $\bar{X}$ は $\mu_{\bar{X}} = \mu$ (未知), $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (既知) の正規分布で近似できる。標本平均 $\bar{X}$ の標準化を行う。

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \cdots \textcircled{1}$$

一方、 Z が標準正規分布に従うことより $P(|Z| < 1.96) = 0.95 \cdots \textcircled{2}$

①を②に代入すると、0.95 の確率で以下の式が成立。

$$|Z| = \frac{|\bar{X} - \mu_{\bar{X}}|}{\sigma_{\bar{X}}} < 1.960 \implies -1.96 \sigma_{\bar{X}} < \mu_{\bar{X}} - \bar{X} < 1.96 \sigma_{\bar{X}} \cdots \textcircled{3} \text{ より}$$

$$\textcircled{a} \quad \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

①を大標本による母集団 μ の 95 % の信頼区間という。中心極限定理を用いるこの方法を「大標本法」という。 μ は 95 % の確率でこの区間に入る。

$$\textcircled{b} \quad \bar{X} - 1.96 \frac{s}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{s}{\sqrt{n}}$$

②は母集団標準偏差 σ が未知のとき、 σ の代わりに標本標準偏差 s を代用 (ただし、 $n \geq 100$)。③を大標本による母集団 μ の (近似的な) 95 % 信頼区間という¹⁾。

例 7 ある学校で 100 人の生徒の無作為標本が選ばれ、これらに知能テストが行われた。テストの実施により 100 人の生徒の点が決まり、それらを標本値として $\bar{X} = 112.0$, $s = 11.0$ が得られた。この標本値をもとにして、この学校全生徒の知能指数に対する 95 % 信頼区間を求めよ。

(解) $n = 100$, $\bar{X} = 112.0$, $s = 11.0$ である。 σ が未知より③を使うと、

$$112.0 - 1.96 \frac{11.0}{\sqrt{100}} < \mu < 112.0 + 1.96 \frac{11.0}{\sqrt{100}} \implies 109.8 < \mu < 114.2 (\text{答})$$

後半の小標本法の解説に移る。

母集団平均 μ の区間推定 (母集団の変数 X の平均を μ , 標準偏差を σ ,

標本を $X_1, X_2, \dots, X_n$, 標本標準偏差を s とする)

(母集団分布が正規分布と仮定できると任意の標本の大きさ n で使える定理。この定理を小標本法という。)

定理 小標本法, スチューデント t 分布

母集団分布が正規分布で標準偏差 σ が未知のとき、任意の大きさ n の標本をとると、

$$t = \frac{\bar{X} - \mu}{s} \sqrt{n} \text{ は 自由度 } \nu = n - 1$$

小標本法による母集団 μ の 95 % 信頼区間は、 $\bar{X} - t_0 \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_0 \frac{s}{\sqrt{n}}$

ただし、 t_0 はスチューデント t 分布表の自由度 $\nu = n - 1$, 両側確率 $P = 0.05$ (片側確率 $P = 0.025$) の値を用いる

例 8 ある製パン工場で作るパンの重さは正規分布に従うことがわかっている。検査員がパンの重さを検査するため 15 個の無作為標本をとった。これら 15 個のパンの重さの平均と標準偏差は 15.8 オンスと 0.3 オンスであった。 μ の 95 % 信頼区間を求めよ。

¹⁾もし、90 % の信頼区間を求めたければ、③式の 1.96 を 1.654 とすればよい。

(解) t_0 の値を求める。 $\nu = n - 1 = 15 - 1 = 14$, 片側確率 $P = \frac{1 - 0.95}{2} = 0.025$ より

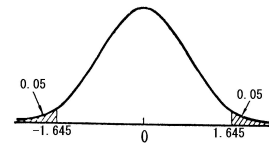
スチューデント t 分布表より $t_0 = 2.145$ 以上であるから, 上記の式より 95%信頼区間は

$$15.8 - 2.145 \frac{0.3}{\sqrt{15}} < \mu < 15.8 + 2.145 \frac{0.3}{\sqrt{15}} \Rightarrow 15.63 < \mu < 15.97 \text{ (答)}$$

◎ 9 回目授業——検定のごく一部

仮説検定においては有意水準 (危険率) $\sigma = 0.01$, $\sigma = 0.05$, $\sigma = 0.10$ で行われることが多い。通常は $\sigma = 0.05$ で行われることが多いので, 境界点 Z_0 の値は, 両側検定で 1.960, 片側検定で 1.645 を用いる。

	Z_0 (両側検定)	Z_0 (片側検定)
$\sigma = 0.01$	2.576	2.326
$\sigma = 0.05$	1.960	1.645
$\sigma = 0.10$	1.645	1.282



例 9 新生生の適性検査 (両側検定)

ある学校の過去数年間の新生生についての記録は, 適性検査の得点が平均 115, 標準偏差 20 であった。今年の新生生が今までの学生より優れているか, 劣っているかを調べよう。事前に知識がないので両側検定とする。新生生から $n = 50$ の標本平均 $\bar{X} = 118$ を得た。

(解) 未知の新生生の平均を μ とすると,

仮説 $H_0 : \mu = 115$ (新生生は今までと同じ)

対立仮説 $H_1 : \mu \neq 115$ (新生生は今までと異なる)

H_0 のもとで, 中心極限定理より, $n = 50$ の標本平均 $\bar{X}$ は, 平均 $\mu_{\bar{X}} = 115$, 標準偏差 $\sigma_{\bar{X}} = \frac{20}{\sqrt{50}} \approx 2.83$ の正規分布で近似できる (n の値が少し小さいけれども近似することにする)。

$|Z| = \left| \frac{118 - 115}{2.83} \right| \approx 1.06 < 1.960$ ゆえに有意水準 $\sigma = 0.05$ の両側検定で仮説 H_0 は採択される (対立仮説 H_1 は棄却される)。

例 10 (1) 1 枚の硬貨 K_1 を 10 回投げたところ, 表が 7 回出たとする。このとき, この実験を基準に考えて, この硬貨 K_1 の表が出る確率は $p_1 = \frac{1}{2}$ と考えてよいのか? 仮説

$H_0 : p_1 = \frac{1}{2}$ を危険率 0.10 で片側検定せよ。

(2) (1) と同様で硬貨 K_2 を 100 回投げたところ, 62 回表がでた。この硬貨 K_2 の表が出る確率を仮説 $H_0 : p_2 = \frac{1}{2}$ を危険率 0.01 で片側検定せよ。

(3) (1),(2) を出題した出題者の出題意図を予想せよ (ここは各自考える)。

(解) (1) まずこの硬貨 K_1 の表の確率を $\frac{1}{2}$ と仮定する。10 回中表の回数を X とすると、 X は二項分布 $\left(10, \frac{1}{2}\right)$ に従う。

$$\text{したがって, } m = np = 10 \times \frac{1}{2} = 5, \sigma = \sqrt{npq} = \sqrt{10 \times \frac{1}{2} \times \frac{1}{2}} = \frac{\sqrt{10}}{2}$$

$|Z| > Z_0 = 1.282$ ならば、有意水準 $\sigma = 0.10$ で棄却。ところが、

$$|Z| = \left| \frac{X - m}{\sigma} \right| = \left| \frac{7 - 5}{\frac{\sqrt{10}}{2}} \right| = \frac{2\sqrt{10}}{5} = 1.26 < Z_0 = 1.282$$

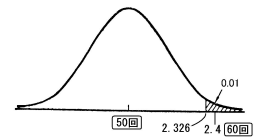
ゆえに、危険率 $\sigma = 0.10$ で仮説 $H_0 : p_1 = \frac{1}{2}$ は採択。

$$(2) m = np = 100 \times \frac{1}{2} = 50,$$

$$\sigma = \sqrt{npq} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5$$

$$|Z| = \left| \frac{X - m}{\sigma} \right| = \left| \frac{62 - 50}{5} \right| = 2.4 > Z_0 = 2.326$$

ゆえに、危険率 $\sigma = 0.01$ で仮説 $H_0 : p_2 = \frac{1}{2}$ は棄却。



◎ 10 回目授業——後期最後の授業。今までの復習と過去問題の解説

4 おわりに

- (1) 実際の 10 回の授業はノートパソコンに「Excel で学ぶ統計学入門」第 1 巻確率・統計編²⁾の付属ソフトを組み込み、数値を代入すると平均、分散、グラフ等が自動的に表示される「道具」を多用した。
- (2) 学生には毎回 10 枚以上のプリントを配布して授業を行った。このレポートはその配布したプリントの約 10 分の 1 である。以上のことより、このレポートは証明抜きで定理 (例えば、 $B(n, p)$ について $E_B(X) = np$, $\sigma^2(X) = npq$ 等) が使われているので了解していただきたい。
- (3) 紹介した 10 の例題のように試験に出る問題に学生の関心が集まり、微分・積分を多用した計算にはあまり関心を示さなかった。
- (4) 前期 (微分・積分) でも感じたことであるが、今回の後期 (確率・統計) の補習も「10 回の授業で当初の目的を果たすことはかなり難しい」が、私の率直な感想です。
- (5) 私にとって、今回の後期 (確率・統計) の補習は始めて教えることが多く、「自分の勉強のために学生に教えた」のが実情です。
- (6) 今回もかなり長くなってしまったレポートにお付き合いいただき恐縮しております。有難うございました。

今回は特に (5) の事情がありますので、ご批判・ご意見を「akirana8133@yahoo.co.jp」までいただけますと助かります。

²⁾長谷川勝也著、技術評論社