

- 1 (1) 5 個の標本データ 3, 5, 4, 1, 7 の平均と分散を計算しなさい。
 (2) これの結果を用いて、つぎの各々データについて、その平均と分散を求めよ。

(2a) 23, 25, 24, 21, 27 (2b) 2.3, 2.5, 2.4, 2.1, 2.7

(解) (1) の各データを $x_i, i = 1, 2, \dots, n$ とおく。 $n = 5, \sum x_i = 20, \sum x_i^2 = 100$ が求められ、これを用いる。
 (1) は平均 $\bar{X} = \frac{1}{n} \sum x_i = \frac{20}{5} = 4$ であり、また分散は $s_X^2 = \frac{1}{n} \sum (x_i - \bar{X})^2 = \frac{\sum x_i^2}{n} - \bar{X}^2 = \frac{100}{5} - 4^2 = 4$ となる。
 (2) (2a) と (2b) の各データをそれぞれ $y_i, z_i, i = 1, 2, \dots, 5$ とおく。 $y_i = x_i + 20$ だから、 Y の平均 $\bar{Y}$ と分散 s_Y^2 は $\bar{Y} = \bar{X} + 20$ と $s_Y^2 = s_X^2$ より $\bar{Y} = 4, s_Y^2 = 4$ を得る。また Z の平均 $\bar{Z}$ と分散 s_Z^2 は $z_i = \frac{x_i + 20}{10}$ だから、 $\bar{Z} = \frac{\bar{X} + 20}{10} = 2.4$ と $s_Z^2 = \frac{s_X^2}{10^2} = 0.04$ として求められる。

コメント 基本的統計量の定義と、データを変換したときに、平均や分散がどう変わるかの関係式を理解しているかどうかを問うた。単純にそれぞれのグループごとに直接計算するのは問題の意味が通じていない。また分散の計算について、各データから平均を引いて、その 2 乗和を求めることは数を桁数が大きくならない効果があるが、各データの 2 乗和から求めるほうが計算の手間は少ないことを理解する。

- 2 確率変数 X は 2 項分布にしたがい、その平均は 4 で分散は 2 であるという。このとき、つぎを求めよ。

(1) $P(X = 3)$ (2) $P(\sqrt{2} \leq X \leq 3)$ (3) $E(X^2 - 4)$

(解) 2 項分布のパラメータを n, p とすると、平均 μ , 分散 σ^2 は $\mu = np, \sigma^2 = np(1-p)$ である。仮定から、 $\mu = 4, \sigma^2 = 2$ としているから、パラメータの値が $n = 8, p = \frac{1}{2}$ と計算できる。したがってこの 2 項分布にしたがう確率変数を X とすれば、その確率が $P(X = k) = \binom{8}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{8-k} = \frac{1}{256} \binom{8}{k}, k = 0, 1, 2, \dots, 8$ である。したがって (1) は $P(X = 3) = \frac{1}{256} \binom{8}{3} = \frac{7}{32}$ となる。(2) では 2 項分布における X の値は整数値をとることに注意する。区間 $[\sqrt{2}, 3]$ 内の整数は 2 と 3 だけだから、 $P(\sqrt{2} \leq X \leq 3) = P(1.414 \dots \leq X \leq 3) = P(X = 2 \text{ or } X = 3) = P(X = 2) + P(X = 3) = \frac{1}{256} \binom{8}{2} + \frac{1}{256} \binom{8}{3} = \frac{21}{64}$ が得られる。(3) は 2 次モーメント EX^2 を平均と分散で表せればよいから、 $E(X^2 - 4) = EX^2 - 4 = \mu^2 + \sigma^2 - 4 = 4^2 + 2 - 4 = 14$ である。

コメント 2 項分布は離散型分布の典型であって整数値しか取らないことに注意する。もしパラメータの n が大きいならば、正規分布で近似する方法で確率を計算する。 n が少しでも大きいと、階乗とか、2 項係数の計算はとてもたいへんである。不等式が与えられていれば、その満たす区間を求め、その中の整数値を拾い出すことは容易であろう。

- 3 確率変数 X は $N(3, 1)$ にしたがう。このときつぎの確率を求めよ。

(1) $P(2.5 < X < 3)$ (2) $P(0.5 < |X - 3|)$

(解) 正規分布に関する確率計算は、正規分布表をもちいる。仮定から $X \sim N(3, 1)$ であるから、 $Z = \frac{X - 3}{1} = X - 3$ は標準正規分布 $N(0, 1)$ にしたがう確率変数である。またその分布関数を Φ とおくと (1) は $P(2.5 < X < 3) = P(-0.5 < X - 3 < 0) = P(-0.5 < Z < 0) = P(0 < Z < 0.5) = \Phi(0.5) - 0.5 = 0.6915 - 0.5 = 0.1915$ ここで標準正規分布は原点で対称となることをもちいている。(2) については、絶対値が表れているから、これを不等式に直す必要がある。 $P(0.5 < |X - 3|) = P(0.5 < |Z|) = 1 - P(|Z| < 0.5) = 1 - P(-0.5 < Z < 0.5) = 1 - 2P(0 < Z < 0.5) = 1 - 2 \times 0.1915 = 0.6170$ と計算される。

コメント 最も基本的な正規分布の確率を計算する統計の問題である。一般の正規分布から、その平均と分散を用いて、標準正規分布へと変換する形も覚えておかねばならない。問いの (2) では不等式をいれて多少難があるかも知れないが、絶対値は誤差として用いられる基本量であるから、よく理解しておいてほしい。

4 「中心極限定理」について、約 150 字程度以内で説明を述べよ。数式等をもちいてよい。

(解) 平均 μ 、分散 σ^2 をもつ独立同一分布確率変数列 $X_1, X_2, \dots$ について考える。あるいは平均 μ 、分散 σ^2 をもつ母集団分布からランダムサンプリング(無作為抽出)した標本データの列を $X_1, X_2, \dots$ として考えてもよい。このとき n 個の確率変数あるいは n 個の標本データに対して、その標本平均 $\bar{X}_n = \frac{1}{n} \sum_i X_i$ は n が十分に大きければ、正規分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ に近づく。すなわち標準化した $\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}$ は標準正規分布 $N(0, 1)$ で近似される。つまり任意の a, b で $n \rightarrow \infty$ とすると $P\left(a < \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} < b\right) \rightarrow \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$ が成り立つ。

コメント CLT (central limit theorem) とか表すこともあるが、統計、確率の理論ではもっとも重要な定理のひとつである。いわゆるよく用いられる(標本)平均が、母集団での分布パラメータとして平均と分散とのかかわり、つまり標準化することの意味もこの定理で表れている。この標準化の形も覚えていたほうがよい。また正規分布の概形(数学でいう増減表)もどう対称性で、増加と減少(微係数のプラス、マイナス)とか、接線の傾きの変化(2次の微分係数)するかも覚えておいたほうが確率の計算に必要なかも知れない。

5 ある試験に対して $n = 40$ 人の得点結果 $x_1, x_2, \dots, x_n$ は、

$$\bar{X} = \frac{\sum_i x_i}{n} = 58.2, \quad u^2 = \frac{\sum_i (x_i - \bar{X})^2}{n-1} = 10.3^2$$

であった。得点が正規分布 $N(\mu, \sigma^2)$ に従うとして、平均 μ の 90% 信頼区間をもとめよ。

(解) 正規母集団の平均に関する区間推定である。分散が未知であるから、統計量としては、標本不偏分散 u^2 をもちいる。したがって標準化した $\frac{\bar{X} - \mu}{\sqrt{u^2/n}}$ は t 分布にしたがうことに注意すると、90% 信頼区間は

$$-t_{\alpha/2}(\nu) < \frac{\bar{X} - \mu}{\sqrt{u^2/n}} < +t_{\alpha/2}(\nu)$$

つまり

$$\bar{X} - t_{\alpha/2}(\nu) \sqrt{u^2/n} < \mu < \bar{X} + t_{\alpha/2}(\nu) \sqrt{u^2/n}$$

これに与えられた数値を代入すればよい。 t 分布表から $n = 40, n-1 = 39$ であるがほぼ等しいと考えて $t_{0.05}(40) = 1.684$ から、 $58.2 - 1.684 \sqrt{10.3^2/40} < \mu < 58.2 + 1.684 \sqrt{10.3^2/40}$ ゆえに $55.46 < \mu < 60.94$ が答えである。

コメント このような信頼区間を求めることが標本データ $x_1, x_2, \dots, x_n$ から計算される和と 2 乗和 $\sum_i x_i, \sum_i x_i^2$ から $\bar{X}$ と u^2 によってなされること。また分散が既知かあるいは未知かによって、正規分布、 t 分布によるかを覚えていればよい。

6 仮説検定における「第 1 種の過誤」と「第 2 種の過誤」について、合わせて約 200 字程度以内で説明を述べなさい。さらに仮説、棄却域、検定統計量に適切な数式をもちい、これらの過誤を犯す確率を表しなさい。

(解) 仮説検定のなかで母集団分布の未知パラメータ(母数) θ を推定するとき、与えられたある値 θ_0 と θ_1 に関して、帰無仮説を $\theta = \theta_0$ 、対立仮説を $\theta = \theta_1$ とし、棄却域を R 、また検定統計量を T とおく。第 1 種の過誤とは、帰無仮説が正しい(真)にもかかわらず、それを棄却することで、一方第 2 種の過誤とは、帰無仮説が正しくない(偽)とき、これを採択する誤りをいう。ここで棄却あるいは採択(棄却しない)という判断の検定統計量の値がどういう表れであったか、すなわち、棄却域のなかに含まれるか含まれないかによって定める。棄却とは $T \in R$ のときでは、採択はそうでないとき、 $T \notin R$ とする。これがネイマン・ピアソン流の単純対立仮説の場合の考え方である。数式でこの確率を表すときには、前提条件を条件つき確率とすればよい。の確率とは、帰無仮説が真である($\theta = \theta_0$)とき、帰無仮説を棄却する($T \in R$)ことの条件つき確率であり、したがって $P(\text{第 1 種の過誤を犯す}) = P(T \in R | \theta = \theta_0)$ となる。同様に $P(\text{第 2 種の過誤を犯す}) = P(T \notin R | \theta = \theta_1)$ である。