2.3　統計量と標本分布

母集団から大きさ n の標本を無作為抽出するとき、それらの変量の値を $X_1, X_2, \cdots, X_n$ で表す。ただし、母集団は極めて大きいか、そうでない場合は復元抽出をすることにする。このような $X_1, X_2, \cdots, X_n$ を大きさ n の無作為標本 (random sample) という。これは、数学的には $X_1, X_2, \cdots, X_n$ は同一の分布に従う独立な確率変数と見なせる。この $X_1, X_2, \cdots, X_n$ が従う共通の確率分布を母集団分布 (population distribution) という。特に、母集団分布が正規分布である母集団を正規母集団という。

母集団分布の平均や分散をそれぞれ母平均、母分散 (population mean, population variance) といい、これらの特性値を総称して母数 (parameter) という。

この $X_1, X_2, \cdots, X_n$ の関数を統計量 (statistic) というが、よく用いられる統計量には次のものがある。

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

$\overline{X}, S^2, U^2$ をそれぞれ標本平均、標本分散、不偏分散という。次が成立することは前回の講義で示した。

定理 2.1 $X_1, X_2, \cdots, X_n$ を母平均 μ, 母分散 σ^2 の母集団からの大きさ n の無作為標本とする。すなわち、$E[X_i] = \mu$, $V[X_i] = \sigma^2$ とする。このとき、標本平均 $\overline{X}$ と不変分散 U^2 に対して次が成立する。

$$E[\overline{X}] = \mu, \qquad V[\overline{X}] = \frac{\sigma^2}{n}, \qquad E[U^2] = \sigma^2.$$

注意 (大数の法則) n が大きくなるにつれて標本平均 $\overline{X}$ の値が母平均 μ の近くにある確率は極めて高くなる。

これはチェビシェフの不等式 $P(|X-a| \geq \delta) \leq \dfrac{1}{\delta^2}E[(X-a)^2]$, $\delta > 0$, (詳しくはより専門的な教科書を見てください) を用いて、

$$P(|\overline{X} - \mu| \geq \delta) \leq \frac{1}{\delta^2}E[(\overline{X} - \mu)^2] = \frac{1}{\delta^2}V(\overline{X}) = \frac{\sigma^2}{\delta^2 n} \to 0, \qquad n \to \infty$$

なることより、$P(|\overline{X} - \mu| < \delta) = 1 - P(|\overline{X} - \mu| \geq \delta) \to 1$, $n \to \infty$, として証明される。　　　　//

一般に、確率変数 $X_1, X_2, \cdots, X_n$ が独立で、各 X_k が正規分布 $N(\mu_k, {\sigma_k}^2)$ に従うとき、定数 $a_1, a_2, \cdots, a_n$ について、$\displaystyle\sum_{k=1}^{n} a_k X_k$ は正規分布に従うことが知られている。この性質と定理 2.1 により次の定理が成立する。

定理 2.2 (正規母集団の標本平均の分布) 正規母集団 $N(\mu, \sigma^2)$ から大きさ n の標本を無作為抽出するとき、その標本平均 $\overline{X}$ は正規分布 $N\Big(\mu, \dfrac{\sigma^2}{n}\Big)$ に従う。

例 4' 正規母集団 $N(30, 16)$ から大きさ 9 の標本を無作為抽出するとき、その標本平均 $\overline{X}$ について $\overline{X} \leq 32$ となる確率を求めよ。

解: $\overline{X}$ は $N\Big(30, \dfrac{16}{9}\Big)$ に従うから、$Z = \dfrac{\overline{X} - 30}{\sqrt{16/9}}$ は $N(0, 1)$ に従う。よって、

$$P(\overline{X} \leq 32) = P\Big(Z \leq \frac{32 - 30}{\sqrt{16/9}}\Big) = P(Z \leq 1.5) = 0.9332. \qquad //$$

一般に、$X_1, X_2, \cdots, X_n$ が正規分布に従わないときでも、n が大きいとき $\overline{X}$ は近似的に正規分布に従うことが知られている。

定理 2.3 (中心極限定理) 確率変数 $X_1, X_2, \cdots, X_n$ が独立で、平均 μ と分散 σ^2 の同一の確率分布に従うとする。n が大きいとき、$\overline{X} = \dfrac{1}{n}(X_1 + X_2 + \cdots + X_n)$ は、近似的に正規分布 $N\Big(\mu, \dfrac{\sigma^2}{n}\Big)$ に従う。

例題 1 ある工場製品の重量 (単位 g) の平均 μ, 分散 σ^2 はそれぞれ 23.4, 4.35 であることが知られている。この工場製品から 100 個の標本を無作為抽出するとき、標本平均 $\overline{X}$ が 23.0 より小さくなる確率を求めよ。

解: $\overline{X}$ は近似的に $N\Big(23.4, \dfrac{4.35}{100}\Big)$ に従うから、$Z = \dfrac{\overline{X} - 23.4}{\sqrt{4.35/100}}$ は $N(0,1)$ に従うとしてよい。よって、

$$\begin{aligned} P(\overline{X} < 23.0) &\fallingdotseq P\Big(Z < \frac{23.0 - 23.4}{\sqrt{4.35/100}}\Big) = P(Z < -1.917\cdots) \\ &\fallingdotseq P(Z < -1.92) = P(Z > 1.92) = 1 - 0.9726 = 0.0274. \qquad // \end{aligned}$$

標本比率の分布 (教科書 p.102)

ある政策について賛成か反対かを全国の有権者にたずねる場合のように、各要素が特定の性質をもつかもたないかのいずれかであるような母集団を**二項母集団**という。二項母集団において、その性質をもつとき $X = 1$, もたないとき $X = 0$ を対応させると X は確率変数となる。$X = 1$ となる要素の全体に対する割合を母比率という。

$X_1, X_2, \cdots, X_n$ を母比率 p の二項母集団から抽出した無作為標本、すなわち、$X_1, X_2, \cdots, X_n$ は独立で各 k に対して

$$P(X_k = 1) = p, \qquad P(X_k = 0) = 1 - p$$

となるとする。このとき、$S = X_1 + X_2 + \cdots + X_n$ は二項分布 $B(n, p)$ に従う。$X_1, X_2, \cdots, X_n$ で 1 となる割合を標本比率といい $\widehat{P}$ と表す:

$$\widehat{P} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{S}{n}.$$

このとき、

$$E[\widehat{P}] = \frac{1}{n}E[S] = \frac{np}{n} = p, \qquad V[\widehat{P}] = \frac{1}{n^2}V[S] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

よって、中心極限定理により、n が大きければ $Z = \dfrac{\widehat{P} - p}{\sqrt{p(1-p)/n}}$ は近似的に標準正規分布 $N(0,1)$ に従う。

問 7 ある新聞社の世論調査で、全国の有権者から大きさ 2500 の無作為標本をとり、内閣の支持率を調べた。母集団における内閣の支持率を 0.4 と仮定した場合、この調査における標本支持率 $\widehat{P}$ の平均と分散を求めよ。また、$\widehat{P} \geq 0.42$ となる確率を求めよ。

解: $E[\widehat{P}] = 0.4$, $V[\widehat{P}] = \dfrac{0.4(1-0.4)}{2500} = \dfrac{0.24}{2500}$.

よって、$Z = \dfrac{\widehat{P} - 0..4}{\sqrt{0.24/2500}}$ はは近似的に $N(0,1)$ に従うから、

$$\begin{aligned} P(\widehat{P} \geq 0.42) &\fallingdotseq P\Big(Z \geq \frac{0.42 - 0.4}{\sqrt{0.24/2500}}\Big) = P(Z \geq 2.041\cdots) \\ &\fallingdotseq P(Z \geq 2.04) = 1 - 0.9793 = 0.0207. \qquad // \end{aligned}$$

4 章 推定と検定　　1.1 点推定

母集団から無作為標本 $X_1, X_2, \cdots, X_n$ を取り出し、その母数を推定する。このとき、母数を 1 つの値で推定することを**点推定**という。

一般に、母数 θ の推定量 T について $E[T] = \theta$ が成り立つとき、推定量 T を θ の**不偏推定量**という。ここで、推定量は無作為標本 $X_1, X_2, \cdots, X_n$ の関数として定義される。また、推定量に**実現値** (実際に観測された値) を $X_1, X_2, \cdots, X_n$ の値として代入したものを推定値という。

定理 2.1 でみたように、$E[\overline{X}] = \mu$ より標本平均 $\overline{X}$ は母平均 μ の不偏推定量であり、$E[U^2] = \sigma^2$ より不偏分散 U^2 は母分散 σ^2 の不偏推定量である。では不偏推定量が複数ある場合どれが**有効**な推定量であるか。これについてはその推定量の分散が小さいものが有効であるとされる。

統計学で重要とされる推定量に**最尤推定量** (maximum likelihood estimator) がある。この講義で前提としている知識では難しすぎるので省略するが、興味のある学生は調べてほしい。