7章 標本と統計的推測

統計的推測の話の始まり.

確率部分　確率変数

確率密度関数, 分布関数に基づき,

平均と分散などの統計量,

変数変換

統計では　目の前にあるものに基づいて

もとの確率変数のことを推測する.

7.1 標本とパラメータ

用語と記号 （確率と統計では同じことを少し違う用語で表現することあり）

サイコロ投げを例に。

母集団
（標本空間）

母集団： サイコロの出目の全体 ⇐ ?
（一般的には）考える対象となる集団

母集団分布：
サイコロの目の現れ方は確率的に決まる。
母集団の確率的な挙動を表す確率分布

母集団分布

サイコロを母集団とした場合、
母集団分布は多項分布

$$\vec{X} = (X_1, X_2, \cdots, X_n)$$

n 回のサイコロ投げ

標本値 x

標本値
- X_1 : 1回のサイコロ投げの値　1, 2, 3, 4, 5, 6 の出目　$P_1 = P_2 = \cdots = P_6 = 1/6$.
- 具体的に現れる値 x

標本

x の背後に想定される確率変数 X を標本という

本書独特な捉え方.

本来の標本には、確率的な考え方は不必要。
（母集団があって、そこから取り出したものが標本）

標本から母集団の様子を、うまく、推測する
ための本書での考え方、おいおい具体的に。

無作為抽出と無作為標本

サイコロを n 回振って得られると想定される
標本 $X_1, \cdots, X_n$ があったとしよう

無作為抽出 random sampling　それぞれの標本は無作為に得られる
このように標本を作る方法

無作為標本　得られた標本

確率変数の言葉では
確率変数 $X_i, i=1:n$ が
母集団の確率的挙動を表す確率変数 X と
同一の分布に従っていて，独立

$$X_i, i=1:n \underset{i.i.d.}{\sim} X$$

パラメータ

母集団を特徴づけるパラメータ（母数）
$B(\theta)$
$N(\mu, \sigma^2)$
など

密度関数とパラメータ

必要に応じて
母集団のパラメータを記すこともある

$$f(x;\underline{\theta})$$

$$E_{\underline{\theta}}[g(X)] = \int g(x)\cdot f(x;\underline{\theta})\,dx$$

パラメータの真値

?

想定する 集団が具体的なとき，
母集団を特徴づけているθには，
何らかの真値 θ^* が存在している。

7.2 統計的推測

確率部分の流れ

$X \sim N(\mu, \sigma^2)$ としよう

X の平均と分散が μ と σ^2 であり

$Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$

$X \sim_{iid} X_{i=1:n} \to \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\bar{X} \xrightarrow[d]{} \mu$

統計の流れ

統計的推測 (statistical inference)

標本に基づいて、もとの確率変数（つまり母集団）のことを推測する

統計的推測の例

歪められたコインの表の出る確率（点推定）

- 表が出る $X = 1$, 裏が出る $X = 0$
- 表が出る確率をパラメータ θ で表わす (θ は不明)
- θ を推し量りたい
- 無作為標本 $X_{i=1:n}$ に対し $\bar{X}$ から θ を推し量る
- コインを100回振って、標本値 $x_{i=1:100}$
- 表の数が36であった. $\bar{x} = 36$, $\theta = 0.36$

7.3 標本平均と標本分散

無作為標本 $X_{i=1:n}$ に対し

標本平均 $\bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$

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

二つの統計量.

統計量 (statistic) 標本だけによって定義される量

$\bar{X}$ と S^2 の性質.

X 母集団の平均と分散 : μ と σ^2

$\bar{X}$ 標本平均 $\bar{X}$ の 平均 μ
分散 σ^2/n (⇒2.6節)

S^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$= \frac{n}{n-1}\cdot\frac{1}{n}\sum\{(X_i-\mu)-(\bar{X}-\mu)\}^2$$
$$= \frac{n}{n-1}\ \frac{1}{n}\sum\{(X_i-\mu)^2+(\bar{X}-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)\}$$
$$= \frac{n}{n-1}\ \frac{1}{n}\left\{\sum\{(X_i-\mu)^2+(\bar{X}-\mu)^2\} - \underbrace{2(\bar{X}-\mu)\sum(X_i-\mu)}_{2n(\bar{X}-\mu)^2}\right\}$$
$$= \frac{n}{n-1}\left\{\frac{1}{n}\sum(X_i-\mu)^2-(\bar{X}-\mu)^2\right\}$$

$E[S^2]$

$$E[S^2] = \frac{n}{n-1}\left\{\frac{1}{n}\sum_1^n E[(X_i-\mu)^2] - E[(\bar{X}-\mu)^2]\right\}$$

$$= \frac{n}{n-1}\frac{1}{n}\left\{\sigma^2 - \frac{\sigma^2}{n}\right\}$$

$$= \sigma^2$$

$\bar{X}$とS^2の 確率収束

大数の法則より

$X,\ \mu = E[X],\ \sigma^2 = V[X]$

$$X_{i=1:n} \sim_{i.i.d} X \Rightarrow \bar{X} = \frac{1}{n}\sum X_i \xrightarrow{P} \mu$$

算術平均

$S^2 \xrightarrow{P} \sigma^2$ の証し

$$S^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum (X_i-\mu)^2 - (\bar{X}-\mu)^2\right\}$$

$\to \mu$ 大数

$\lim_{n\to\infty}$ とすることで

少し疑問有.

$$\Rightarrow 1\left\{\underbrace{\frac{1}{n}\sum (X_i-\mu)^2}_{E[(X-\mu)^2]} - 0\right\}$$

$$\xrightarrow{P} \sigma^2$$

標本分布　統計量が内在している確率分布のこと

$$母集団分布 \sim N(\mu, \sigma^2)$$

のとき

$$標本平均\ \overline{X} \sim N(\mu, \sigma^2/n)$$

あるいは

$\overline{X}$の標本分布は$N(\mu, \sigma^2/n)$であるという

標本分散の標本分布は?

$$\begin{cases} (n-1)S^2/\sigma^2 \sim \chi^2_{n-1} \\ \overline{X} と S^2 は独立である \end{cases}$$

証し.　まず準備

$\vec{j} = (1, \cdots, 1)^t$,　$\vec{a} = \vec{j}/\sqrt{n}$　(長さが1のベクトルに)

標本平均

$$\overline{X} = \frac{1}{n}\sum_{n=1}^{n} X_i = \frac{1}{n}\vec{j}^t \cdot \vec{X} = \frac{1}{\sqrt{n}}\vec{a}^t \cdot \vec{X}$$

行列の準備　$\mathbb{A}$は直交行列 (n次)

$$\mathbb{A} = (\vec{a}, \mathbb{A}_2)$$

$$\mathbb{A}^t \cdot \mathbb{A} = \mathbb{A} \cdot \mathbb{A}^t = \mathbb{I}_n$$

$$\vec{0} = (0, \cdots, 0)^t$$

$$\mathbb{A}_2^t \cdot \vec{a} = \vec{0}$$

新しい確率変数

$$\vec{Y} = \mathbb{A}^t \cdot \vec{X}$$

$$\vec{X} \sim N_n(\mu \cdot \vec{j}, \sigma^2 \mathbb{I}_n)$$

S^2の標本分布の続き

準備

$$\mathbb{A}^t \cdot \vec{j} = \begin{pmatrix} \vec{a}^t \cdot \vec{j} \\ \mathbb{A}_2^t \cdot \vec{j} \end{pmatrix} = \begin{pmatrix} \vec{a}^t \cdot \sqrt{n}\vec{a} \\ \mathbb{A}_2^t \cdot \sqrt{n}\vec{a} \end{pmatrix} = \begin{pmatrix} \sqrt{n} \\ \vec{0} \end{pmatrix}$$

$\vec{Y}$は$\vec{X}$の線形変換なので

$$\vec{Y} = \mathbb{A}^t \cdot \vec{X} \sim N_n(\mu \mathbb{A}^t \vec{j}, \sigma^2 \mathbb{A}^t \mathbb{A})$$
$$\stackrel{d}{=} N_n(\mu \mathbb{A}^t \vec{j}, \sigma^2 \vec{I}_n)$$
$$= N_n((\sqrt{n}\mu, 0, \cdots 0)^t, \sigma^2 I_n)$$

$\stackrel{d}{=}$ は両辺の確率分布が等しいという意味,

$\vec{Y} = \mathbb{A}^t \cdot \vec{X} \sim N_n((\sqrt{n}\mu, 0, \cdots, 0)^t, \sigma^2 I_n)$ が意味するのは $\vec{Y} = (Y_1, \cdots, Y_n)^t$として

$$\left\{ \begin{array}{l} \begin{cases} Y_1 \sim N(\sqrt{n}\mu, \sigma^2) \\ Y_{i=2:n} \sim N(0, \sigma^2) \end{cases} \\ \\ Y_1, \cdots, Y_n \text{は互いに独立} \end{array} \right\}$$ が導かれた。

$\overline{X}$ と S^2 を $\vec{Y}$ で表現

$$\overline{X} = Y_1/\sqrt{n}$$

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

$\sum X_i^2 - 2X_i\overline{X} + \overline{X}^2$

$$= \sum X_i^2 - n\overline{X}^2$$

$-2\overline{X} \cdot n\overline{X}$ $\quad n\overline{X}^2$

$$= \vec{X}^t \cdot \vec{X} - n\overline{X}^2$$

$$\vec{X} = A^t \cdot \vec{Y}$$

$$(\vec{X})^t = (A^t \cdot \vec{Y})^t = A \cdot \vec{Y}^t$$

$[A_{ji}]\,[Y_i$

A_1

$$= \vec{Y}^t \cdot \vec{Y} - Y_1^2$$

$$= Y_2^2 + \cdots + Y_n^2$$

$\therefore$ $\overline{X}$ と S^2 は互いに独立

$\overline{X} = \overline{X}(Y_1)$, $\overline{S}^2 = \overline{S}^2(Y_2, \ldots, Y_n)$

$\overline{X} \sim N(\sqrt{n}\mu, \sigma^2)/\sqrt{n} = N(\mu, \sigma^2/n)$ が確認できる

$\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, $\dfrac{Y_2}{\sigma}, \ldots, \dfrac{Y_n}{\sigma} \sim N(0,1)$ なので

7.4 標準化とスチューデント化

大前提　母集団 (μ, σ^2) からの無作為標本 $X_{i=1 \sim n}$

スチューデント化　確率変数の
- 標準化　（確率の話で登場）
- スチューデント化（統計　〃　）

もしばしば

標本平均の標準化

$$Z_n = \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}}$$

標本数

標本平均のスチューデント化

$$T_n = \frac{\bar{X} - \mu}{\sqrt{S^2 / n}}$$

σ^2 に確率収束する確率変数で置き換え

母集団 ～ $N(0,1)$ の時

$Z_n \sim N(0, 1)$

$T_n \sim$ 自由度 $n-1$ の t-分布

3.2.5
χ^2分布と
t-分布

$$T_n = \frac{\bar{X}-\mu}{\sqrt{S'^2/n}} = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \cdot \frac{1}{\sqrt{S'^2/\sigma^2}}$$

$$= \frac{Z_n \sim N(0,1)}{\sqrt{\underbrace{\{(n-1)S^2/\sigma^2\}}_{\chi^2_{n-1}}/(n-1)}}$$

$$\overset{d}{=} \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

$$\overset{d}{=} t_{n-1}$$

母集団分布が不明のとき

Z_nは漸近的に$\sim N(0,1)$

中心極限定理より

T_nも漸近的に$\sim N(0,1)$

$$T_n = \frac{\bar{X}-\mu}{\sqrt{S^2/n}}$$

$$= \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}\sqrt{S^2/\sigma^2}}$$

$$\xrightarrow{d} N(0,1) \cdot 1$$