No.13 標本統計

NAKADA Masayuki

Kobe University Secondary School

November 2, 2021

標本調査（309のスライドより）

標本調査: **母集団**（調査の対象全体）を全数調査することが難しい場合，一部を無作為に抽出し（**標本**），全体を推測する．

- 母集団の情報:
  母平均 μ，**母標準偏差** σ
  - 直接の把握は困難
- 標本の情報:
  標本平均 $\overline{X}$，**標本標準偏差** $\sigma(X)$
  - 調査によって把握可能
  - 抽出した標本によって，値が揺れる $\leftarrow$ $\overline{X}$ や $\sigma(X)$ もまた確率変数！

把握可能な $\overline{X}$，$\sigma(X)$ から母集団の分布の情報 μ や σ を推測する．

※今回は，母集団が平均 μ，標準偏差 σ の正規分布 $N(\mu,\sigma)$ に従うものとする．

標本平均の平均と標準偏差

$X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$（$n$ 個の標本を無作為に抽出した）のとき，標本平均 $\overline{X}_n = \dfrac{1}{n}(X_1 + X_2 + \cdots + X_n)$ の平均 $E(\overline{X}_n)$ と標準偏差 $\sigma(\overline{X}_n)$ について，調べよう．

$$E(\overline{X}_n) = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \frac{n\mu}{n} = \mu,$$

$$V(\overline{X}_n) = V\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}(V(X_1) + \cdots + V(X_n)) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

($\because X_1, \dots, X_n$ は独立であると仮定した.)

上の計算は，X_i たちが同一の分布に従うならばいつでも正しい．

さらに，正規分布は再生性（後述）をもつため，$\overline{X}_n$ も正規分布に従う．

標本平均の分布

$N(\mu, \sigma^2)$ に従う大きさ n の無作為標本の標本平均は，平均 μ，標準偏差 $\frac{\sigma}{\sqrt{n}}$ の正規分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ に従う: $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

e.g. 母平均 50，母標準偏差 10 をもつ小集団から，大きさ 100 の標本を無作為抽出するとき，その標本平均 $\overline{X}$ が 52 より大きい値をとる確率を求めよう（問 25）.
$\overline{X}$ は平均 50，標準偏差 $\frac{10}{\sqrt{100}} = 1$ の正規分布に従う．その標準化 $Z = \frac{\overline{X}-50}{1} = \overline{X} - 50$ は標準正規分布に従うから

$$P(\overline{X} \geqq 52) = P(Z \geqq 2) = 0.5 - 0.4772 = 0.0228$$

により，およそ **2.3**% である． □

標本分散について

母平均が既知のとき，「標本分散」${S_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ の平均は

$$E({S_n}^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}E\left((X_i-\mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}V(X_i) = \sigma^2$$

より，$E({S_n}^2) = \sigma^2$ である．しかし，標本調査において，母平均 μ の値が既知であることは稀である．そこで，母平均 μ を標本平均 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ に置き換えた

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

を，標本分散として扱う．

標本分散について

母平均が既知のとき，「標本分散」$S_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ の平均は

$$E(S_n{}^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}E\left((X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}V(X_i) = \sigma^2$$

より，$E(S_n{}^2) = \sigma^2$ である．しかし，標本調査において，母平均 μ の値が既知であることは稀である．そこで，母平均 μ を標本平均 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ に置き換えた

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

を，標本分散として扱いたいところであるが，実はこの S では $E(S^2) = \frac{n-1}{n}\sigma^2$ と過小評価してしまう．特に，あまり大きくない標本（$n < 100$ 程度）においては影響が無視できない．

標本から母分散に近い値を取り出すためには，さらに式に修正を加えて

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

を標本分散として扱えばよい．$\hat{S}^2$ を**標本不偏分散**という．$\hat{S}^2$ は $E(\hat{S}^2) = \sigma^2$ を満たす．

※標本不偏分散が n でなく $n-1$ で割られているのは，2 乗和 $\sum_{i=1}^{n}(X_i - \overline{X})^2$ の式において，固定された $\overline{X}$ の下で $X_1, \ldots, X_n$ のうち $n-1$ 個の値が決まれば，残り 1 個の X_i の値は，条件式 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ への束縛により決まってしまう（「**自由度**が $n-1$」という）ことに由来する．

※統計量 θ（e.g. 母平均，母分散など母集団のもつパラメタ）とその標本推定量 T（e.g. 標本平均，標本分散などの確率変数）に対し，$E(T) = \theta$ が成り立つとき，その標本推定量は**不偏推定量**であるという．不偏推定量は，統計量の付近に分布することが期待されるため，標本から母集団を推測するのに都合がよい．統計的推定については，次回詳しく扱う．

自由度

n 個の変数 $x_1, x_2, \ldots, x_n$ に k 個の関係式からなる「系（system）」

$$\begin{aligned} F_1(x_1, x_2, \ldots, x_n) &= 0 \\ F_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ F_k(x_1, x_2, \ldots, x_n) &= 0 \end{aligned}$$

が与えられると，（特殊な場合を除き）自由に値を設定できる変数は $n-k$ 個である．
このとき，系の**自由度**は $n-k$ であるという．

χ^2 分布

$Z_1, Z_2, \ldots, Z_n$ が標準正規分布 $N(0,1)$ に従う独立な確率変数であるとき，確率変数

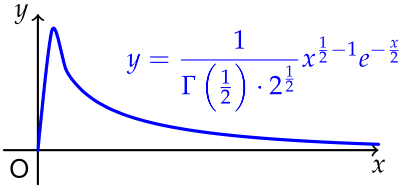
$$Y = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_n}^2$$

の従う分布のことを，**自由度 n の χ^2（カイ自乗）分布**といい，χ_n^2 で表す.

※ガンマ分布の式をよく見れば，${Z_i}^2 \sim \mathsf{G_A}\left(\frac{1}{2}, 2\right)$ であることが分かる（詳細は後述）．また，ガンマ分布は再生性をもつので，$\chi_n^2 = \mathsf{G_A}\left(\frac{n}{2}, 2\right)$ である.

ガンマ分布 $\mathsf{G_A}(\alpha, \beta)$ の確率密度関数:

$$f(x) = \frac{1}{\Gamma(\alpha) \cdot \beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$$



χ_1^2 の分布

以降は補足と発展的な話題

$Z \sim N(0,1)$ のとき $Z^2 \sim \mathrm{G_A}\left(\frac{1}{2}, 2\right)$ であることについて

置換積分の技術（数学III）を援用すると，分布関数の比較により証明できる．

$$
\begin{aligned}
P(Z^2 \leqq u) &= P(-\sqrt{u} \leqq Z \leqq \sqrt{u}) \\
&= \int_{-\sqrt{u}}^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 2\int_0^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\
&= 2\int_0^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y}{2}\right) \cdot \frac{dy}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} \int_0^u y^{-\frac{1}{2}} e^{-\frac{y}{2}}\, dy \\
&= \int_0^u \frac{1}{\Gamma\left(\frac{1}{2}\right) \cdot 2^{\frac{1}{2}}} y^{\frac{1}{2}-1} e^{-\frac{y}{2}}\, dy \qquad \left(\because \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\right)
\end{aligned}
$$

これは，ガンマ関数の分布関数を与えている．

なお，途中で $y = x^2$ として置換積分を行っている．このとき，$0 \leqq x \leqq \sqrt{u}$ において，$\frac{dy}{dx} = 2x$ により $dx = \frac{dy}{2\sqrt{y}}$ である． □

標本不偏分散の自由度について（再訪）

正規分布に従う標本 $X_1,\dots,X_n \sim N(\mu,\sigma)$ について，$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ を考える．標準化 $Y_i = \frac{X_i-\mu}{\sigma}$ を行うと，$Y_1,\dots,Y_n \sim N(0,1)$ であり，

$$
\begin{aligned}
\frac{nS^2}{\sigma^2} &= \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma} - \frac{\overline{X}-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 - 2\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}\cdot\frac{\overline{X}-\mu}{\sigma} + \sum_{i=1}^{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 \qquad (\because \quad X_1+\cdots+X_n = n\overline{X})
\end{aligned}
$$

が成り立つ．次のような変数変換 $(Y_1,\dots,Y_n) \leftrightarrow (Z_1,\dots,Z_n)$ を行う．

$$
\begin{aligned}
Z_1 &= \frac{Y_1}{\sqrt{n}} + \frac{Y_2}{\sqrt{n}} + \cdots + \frac{Y_n}{\sqrt{n}} \\
Z_2 &= \frac{Y_1}{\sqrt{1 \cdot 2}} - \frac{1 \cdot Y_2}{\sqrt{1 \cdot 2}} \\
Z_3 &= \frac{Y_1}{\sqrt{2 \cdot 3}} + \frac{Y_2}{\sqrt{2 \cdot 3}} - \frac{2 \cdot Y_3}{\sqrt{2 \cdot 3}} \\
&\vdots \\
Z_n &= \frac{Y_1}{\sqrt{(n-1)n}} + \cdots + \frac{Y_{n-1}}{\sqrt{(n-1)n}} - \frac{(n-1)Y_n}{\sqrt{(n-1)n}}
\end{aligned}
$$

この変換は，次を満たす.

$$
\sum_{i=1}^{n} {Y_i}^2 = \sum_{i=1}^{n} {Z_i}^2,\ Z_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1),\ Z_2, \ldots, Z_n \sim N(0,1)
$$

よって,

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n} Y_i{}^2 - n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i{}^2 - Z_1{}^2 = \sum_{i=2}^{n} Z_i{}^2$$

この計算結果は, $\frac{nS^2}{\sigma^2} = \sum_i (X_i - \overline{X})^2$ が自由度 $n-1$ の χ^2 分布に従うことを示している. ※変数変換 $(Y_1, \ldots, Y_n) \leftrightarrow (Z_1, \ldots, Z_n)$ はいわゆる直交変換である. 行列計算を行うと, 見通しがよくなる.

再生性

二項分布，正規分布，ガンマ分布などは，以下の性質をもつことが知られている．

$$
\begin{aligned}
&X \sim B(m,p), \quad Y \sim B(n,p) &&\Rightarrow X+Y \sim B(m+n,p)\\
&X \sim N(\mu_1,{\sigma_1}^2), \quad Y \sim N(\mu_2,{\sigma_2}^2) &&\Rightarrow \quad X+Y \sim N(\mu_1+\mu_2,{\sigma_1}^2+{\sigma_2}^2)\\
&X \sim \mathsf{G_A}(\alpha_1,\beta), \quad Y \sim \mathsf{G_A}(\alpha_2,\beta) &&\Rightarrow X+Y \sim \mathsf{G_A}(\alpha_1+\alpha_2,\beta)
\end{aligned}
$$

これらのように，同種の確率分布に従う 2 変数について，その和も同種の確率分布に従うとき，その確率分布族（一連の確率分布の集まり）は**再生性**をもつという．上記のことから，二項分布，正規分布，ガンマ分布は再生性をもつ．証明には，畳み込みなどの高度な積分技術が必要であるため，ここでは省略する．