No.14 推定

NAKADA Masayuki

Kobe University Secondary School

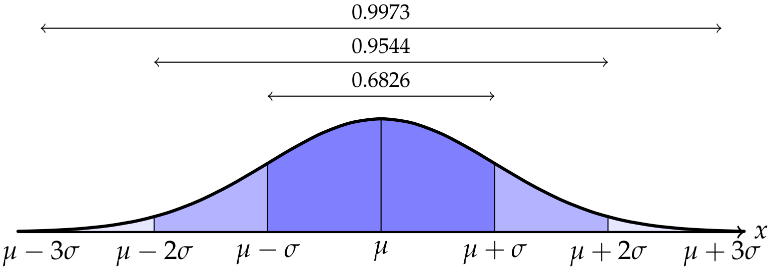
November 8, 2021

正規分布とデータの範囲

n シグマ範囲

正規分布 $N(\mu, \sigma^2)$ で，平均 μ から標準偏差 n 個分離れた範囲 $[\mu - n\sigma, \mu + n\sigma]$ のことを，**n シグマ範囲**という．

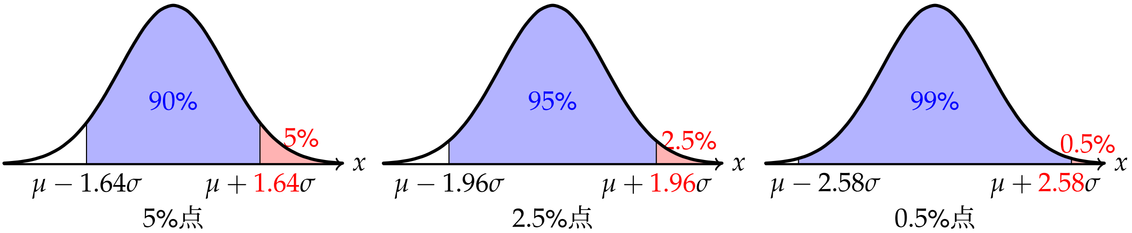
統計学では，観測された事象がどのくらい珍しいことであるかを見積もることが多い．その際，データが平均から標準偏差何個分離れているかを目安にするのである．



※標準化 $Z = \frac{X-\mu}{\sigma}$ により，n シグマ範囲 $\mu - n\sigma \leqq x \leqq \mu + n\sigma$ は $-n \leqq z \leqq n$ に対応する．

パーセント点

正規分布 $N(\mu,\sigma^2)$ において，次のような，割合とそこから逆算した点の位置関係もよく言及される．確率密度関数の中央からの左右対称範囲の面積が $1-2\alpha$ であるとき，x の座標を **100α%点**という．



パーセント点は，推定や検定などでよく利用される．確率変数を標準化することで，パーセント点は上記の x 座標における σ の係数と一致する．

推定

母平均や母分散のように，母集団の分布を特徴づける定数のことを**母数（パラメタ）**という．
実際の統計調査などでは，母集団を直接全数調査することは，困難であることが多い．標本から母数を推測することを，**推定**という．
推定には，大きく次の 2 種類がある．

- **点推定** 母数をある 1 つの値で推定する方法．誤差を伴うため，誤差評価が必要．
- **区間推定** 母数をある確率で含む区間によって推定する方法．

今回は，区間推定について取り扱う．

区間推定

母平均の区間推定: 問題意識

母平均 μ が未知であるが，正規分布に従うことが分かっている母集団 U に対して，標本調査から母平均 μ を求めたい．

考え方のフレーム:

母平均（未知） μ $\leftarrow$ **標本平均** $\bar{x}$ $\leftarrow$ **標本** $[x_1, x_2, \ldots, x_n]$

今回は区間推定を行うため，ある確率で母平均を含む区間を求める．区間が母平均を含む確率を，95%として設定することにする．

正規分布 $N(\mu, \sigma^2)$（μ: 未知，σ^2: 既知）に従う母集団 U から大きさ n の標本 $[x_1, x_2, \ldots, x_n]$ を無作為に抽出した状況を考える．
一般に，大きさ n の標本の標本平均を与える確率変数 $\overline{X}$ は，$N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$ の正規分布に従う（No.13）．
標準化した確率変数 $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ を考え，$z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ とおく．このとき，標本平均に対応する標準化量 z は，95% の確率で区間 $[-1.96, 1.96]$ に含まれる．すなわち，

$$\boxed{95\% \text{ の確率で } -1.96 \leqq \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqq 1.96 \text{ は真}}$$

※母分散 σ^2 が未知のときは，代わりに（不偏）標本分散 $\hat{S}^2$ を用いる．この場合，$\frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}$ は正規分布ではなく，**自由度 $n-1$ のティー分布**に従う．よって，数値がややずれるのであるが，分布を置き換えさえすれば，基本的な考え方は同様である．ティー分布については，後日改めて取り扱う．

$$
\begin{aligned}
& 95\% \text{ の確率で} \quad -1.96 \leqq \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqq 1.96 \quad \text{は真} \\
\Leftrightarrow \quad & 95\% \text{ の確率で} \quad -1.96 \cdot \frac{\sigma}{\sqrt{n}} \leqq \overline{x}-\mu \leqq 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \text{は真} \\
\Leftrightarrow \quad & 95\% \text{ の確率で} \quad -\overline{x}-1.96 \cdot \frac{\sigma}{\sqrt{n}} \leqq -\mu \leqq -\overline{x}+1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \text{は真} \\
\Leftrightarrow \quad & 95\% \text{ の確率で} \quad \overline{x}-1.96 \cdot \frac{\sigma}{\sqrt{n}} \leqq \mu \leqq \overline{x}+1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \text{は真}
\end{aligned}
$$

よって，標本から求めた区間 $\left[\overline{x}-1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x}+1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$ は，95% の確率で母平均 μ を含んでいると分かる．この区間を **95%信頼区間**という．信頼区間の上端，下端をそれぞれ上側信頼限界，下側信頼限界といい，設定した確率 95%を**信頼度**という．

推定してみよう

犬派 or 猫派問題の実験