母平均の区間推定と標準化（1）

岩村 英之

無限に数字のつまった壺を考えよう．この壺の中の数字 X が平均 μ，標準偏差 σ の正規分布に従うならば，次のような操作を施してつくられる新しい変数 Z は，平均 0，標準偏差 1 の正規分布（すなわち標準正規分布）に従う．

$$Z = \frac{X - \mu}{\sigma}$$

これは，いわば全員の背番号を「一定の規則に従って」付け替えただけ．

$$\text{新しい背番号} = \frac{\text{もとの背番号} - 5}{2}$$

（たとえば $\mu = 5$, $\sigma = 2$ とすれば）．したがって，最初の背番号の下で 10 番から 20 番までの人が全体の 50 パーセントを占めていたとすれば，新たな背番号の下では $(10-5)/2 = 2.5$ 番から $(20-5)/2 = 7.5$ 番までの人が全体の 50 パーセントを占めていることになる．当たり前だが重要な点なので，確認しておきたい．

今，私たちが

- この壺から試しにひとつだけ取り出したところ，X_1 という数字が出た．
- さらに，この X_1 は，もしこの壺の平均（つまり母平均）が μ_1 であるならば 95 パーセントの確率で出てくるはずの値（＝複数あるので区間になる）に入っているとする（図 1 上半分）．

すでに見たとおり，新しい背番号 $\frac{X-\mu_1}{\sigma}$ は**標準**正規分布に従う．
そして，新しい背番号の下では（つまり標準正規分布では）-1.96 から 1.96 までの人が 95 パーセントを占めていることがわかっている．
よって，X_1 が最初の背番号で 95 パーセント区間に入っているならば，彼の新しい背番号 $\frac{X_1-\mu_1}{\sigma}$ は新しい背番号の下での（つまり標準正規分布における）95 パーセント区間，すなわち-1.96 から 1.96 の区間に入っているはず（図 1 下半分）．
$\Leftrightarrow$ このことを逆から見れば，「$\frac{X_1-\mu_1}{\sigma}$ が-1.96 から 1.96 の範囲に入っていれば，X_1 は μ_1 の下での 95 パーセント区間に入っている」．

もう少し大きい母平均 μ_2（これは μ_1 より十分に大きいとする）について考えてみる．

- X_1 は，母平均が μ_2 であるならば 95 パーセントの確率で出てくるはずの区間には入っていないとする（図 2 上半分）．

もとの背番号で 95 パーセントに入っていないならば，新しい背番号でも $\frac{X_1-\mu_2}{\sigma}$ は-1.96 から 1.96 の範囲に入っていないはず（図 2 下半分）．
$\Leftrightarrow$「$\frac{X_1-\mu_2}{\sigma}$ が-1.96 から 1.96 の範囲に入っていなければ，X_1 は μ_2 の下での 95 パーセント区間に入っていない.」

以上をまとめると，次のようになる．

平均はわからないが標準偏差が σ とわかっている正規母集団から偶然取り出した値が，X_1 であったとする．

ケース A　ある μ について $\frac{X_1-\mu}{\sigma}$ が-1.96 から 1.96 の範囲に入っている
$\Rightarrow$　X_1 はその μ のもとで十分起こり得る値
$\Rightarrow$　その μ を「あり得ない」と断じる積極的理由はない

ケース B　ある μ について $\frac{X_1-\mu}{\sigma}$ が-1.96 から 1.96 の範囲に入っていない
$\Rightarrow$　X_1 はその μ のもとで起こる可能性が低い値
$\Rightarrow$　その μ が本物である可能性はかなり低い

私たちは，偶然取り出した X_1 が「ケース A」の条件を満たすような μ を探しているわけだが，そのような μ はひとつだけではない．実際,「ある μ について $\frac{X_1-\mu}{\sigma}$ が-1.96 から 1.96 の範囲に入っている」という条件を数式で言い換えれば,

$$-1.96 \leq \frac{X_1-\mu}{\sigma} \leq 1.96$$

となるが，この式を満たす μ は明らかにひとつではない．この連立不等式を μ について解けば,

$$X_1 - 1.96\sigma \leq \mu \leq X_1 + 1.96\sigma$$

となる．これが，X_1 だけを証拠に特定することのできる,「母平均のあり得る区間」である．95 パーセント信頼区間と呼ぶ.

なぜ標準化するのか．いちいち標準化などせずに，もとの背番号のままで,「もし母平均が ◯◯ だったら何番から何番までが 95 パーセントを占めるか．偶然取り出した X_1 はその範囲に入っているか」と考えればよいのではないだろうか．標準化した世界で考えることのメリットはなんだろうか？

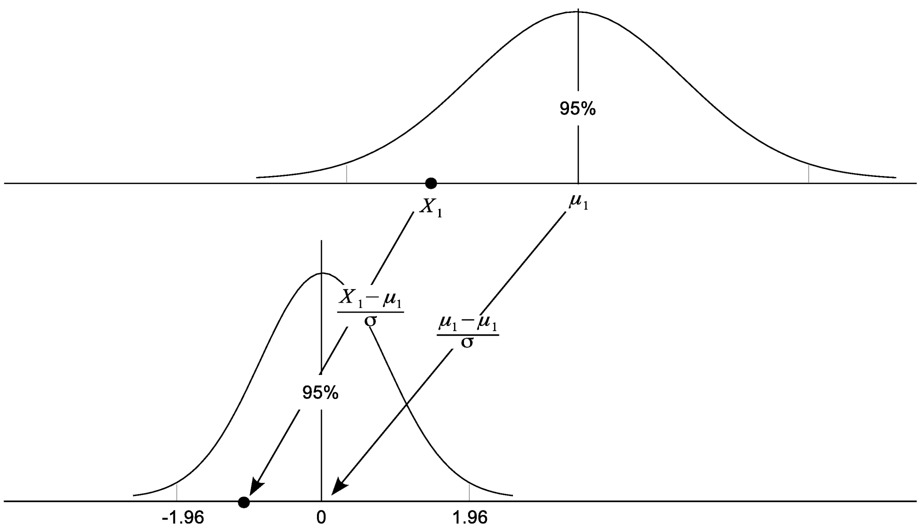

図 1: $\mu = \mu_1$

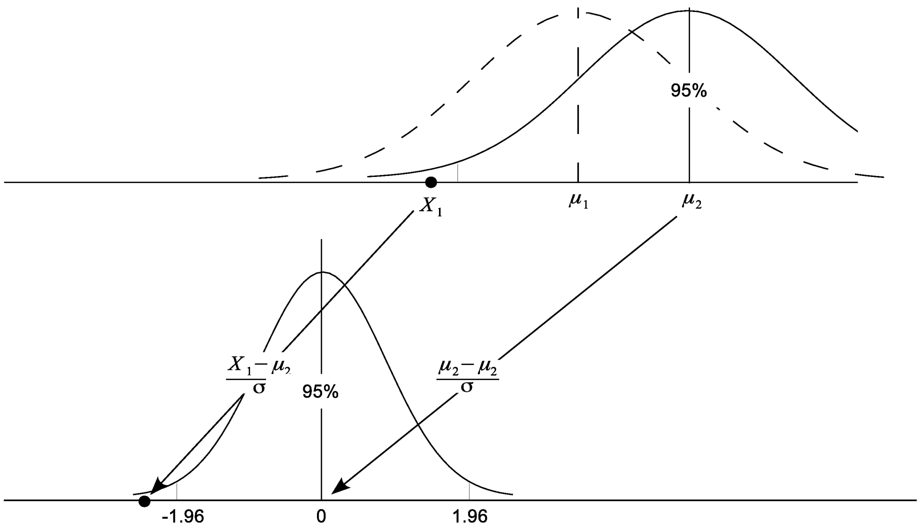

図 2: $\mu = \mu_2$

今，分散が σ^2 である母集団から 2 回ランダムサンプリングを行い，その平均を計算することを考える．1 回目に出る値を x_1，2 回目に出る値を x_2 とする．したがって，2 回の平均は $\frac{x_1+x_2}{2}$ となる．今，この $\frac{x_1+x_2}{2}$ がもとの母集団に比べてどれくらいばらつくのかを考えてみる．

$$
\begin{aligned}
\frac{x_1+x_2}{2}\text{の分散} &= \frac{x_1}{2}\text{の分散} + \frac{x_2}{2}\text{の分散} \\
&= \left(\frac{1}{2}\right)^2 \times (x_1\text{の分散}) + \left(\frac{1}{2}\right)^2 \times (x_2\text{の分散}) \\
&= \frac{1}{4} \times [(x_1\text{の分散}) + (x_2\text{の分散})] \\
&= \frac{1}{4} \times \left[\sigma^2 + \sigma^2\right] \\
&= \frac{1}{2}\sigma^2
\end{aligned}
$$

分散の平方根が標準偏差である．したがって，

$$
\frac{x_1+x_2}{2}\text{の標準偏差} = \frac{\sigma}{\sqrt{2}}
$$