生物科学研究所 研究報告 2024 年 12 月 9 日

ロバスト z スコア：中央値と四分位数で，非正規分布，外れ値を含む標準化

井口豊*
*生物科学研究所，長野県岡谷市

DOI: https://doi.org/10.5281/zenodo.14336057

1. 標準化とは

統計学で最もよく知られた標準化あるいは基準化は，確率変数 X を平均 0，標準偏差 1 となるように変数変換することである。変換された確率変数を z で表すと以下のようになる。

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

この z は標準スコアあるいは z スコア と呼ばれることもある。一般的には，平均 μ，標準偏差 σ の正規分布に従う確率変数 X に対して標準化が行われ，z は $\mu = 0$，$\sigma = 1$ の標準正規分布 $N(0, 1)$ に従う確率変数となる。ただし，正規分布以外でも，標準化は行われる。

2. 中央値と四分位数を用いた標準化

平均と標準偏差を用いた標準化は良く知られているが，中央値と四分位数を用いた標準化は馴染みが薄い。それについて，簡単に説明しよう。なお，四捨五入の誤差があるため，以下の等式では，両辺の数値が必ずしも一致しない場合もあるので，正確を期すためには，各自で実際に計算してほしい。

まず，第 3 四分位数 Q_3（75 %点）から第 1 四分位数 Q_1（25 %点）を引いた値を四分位範囲（interquartile range, IQR）と言う。

この計算を，標準正規分布の確率変数に対応させると 1.3489 となる。すなわち z の確率分布関数を $F(z)$ とすると，次のようになる。

$$\begin{aligned} IQR &= F(0.75) - F(0.25) \\ &= 1.3489 \end{aligned}$$

これは EXCEL 関数でも，次のように求められる。

$$\mathrm{NORMSINV}(0.75) - \mathrm{NORMSINV}(0.25) = 0.6744 - (-0.6744)$$
$$= 1.3489$$

要するに，標準正規分布の平均 μ と標準偏差 σ を使って，四分位数を以下のように置き換えているのである。

$$\begin{aligned} Q_1 &= \mu - 0.6744\sigma \\ Q_3 &= \mu + 0.6744\sigma \\ IQR &= Q_3 - Q_1 \\ &= 1.3489\sigma \end{aligned}$$

ここで，$\mu = 0$，$\sigma = 1$ である。標準偏差と四分位偏差の対応関係については，井口（2024a）も参照してほしい。

そして，IQR を正規分布と関係付けるために，この 1.3489 で割る。これを正規四分位範囲（normalized IQR, $NIQR$）と言う。

$$\begin{aligned} NIQR &= \frac{IQR}{1.3489} \\ &= 0.7413IQR \end{aligned}$$

各種文献では，割った形でなく，逆数をかけた後者の形で示されることが多い。

$NIQR$ の定義は，四分位偏差 QD $(= IQR/2)$ を使って，以下のように書き直すこともできる。

$$NIQR = \frac{QD}{0.6744}$$

最後に，測定値を X_i，中央値（メディアン）を X_m とすると，冒頭で述べた確率変数の標準化に類似して，z が以下のように定義される。

$$z = \frac{X_i - X_m}{NIQR}$$

あるいは，これを四分位偏差 QD を使って，以下のようにも表せる。

$$z = \frac{X_i - X_m}{\frac{QD}{0.6744}}$$

この z を**ロバスト z スコア (robust z score)** と呼ぶ。
つまり，平均を中央値に，標準偏差を標準正規分布に対応させた四分位範囲（あるいは四分位偏差）に変換した z スコアということになる。

3. ロバスト z スコアの利点と利用分野

ロバスト z スコアを使う利点は以下の通りである。

1. 中央値は，分布の位置の代表値として，分布形に影響されない。
2. 四分位範囲あるいは四分位偏差は，ばらつきの統計量として，分布両端の外れ値に影響されない。

つまり，非正規分布のデータの標準化に，ロバスト z スコアが使えるのである。また 2 に関して， *NIQR* は外れ値を切り取ったトリム標準偏差 (trimmed standard deviation) と言える。

ロバスト z スコアは，食品化学の分野で，成分分析に使われることがある。例えば， Puwastien, P. (2002) p.997 左段にそれが書かれている。あるいは，保母ほか (2008) p . 366 式 (4) がそれである。

なお， 四分位数には複数の定義があるので，その点は注意すべきである（井口, 2024b）。

参考文献

井口 (2024a) 四分位偏差とは何か: 変動係数と長野県岡谷市「きなこ石」の話題も含めて生物科学研究所 研究報告 2024 年 12 月 9 日 . Zenodo. https://doi.org/10.5281/zenodo.14328888

井口 (2024b) 四分位数と四分位群：複数定義と用語の区別，その歴史. 生物科学研究所 研究報告 2024 年 10 月 17 日. Zenodo. https://doi.org/10.5281/zenodo.13889521

Puwastien, P. (2002) Issues in the development and use of food composition databases. Public health nutrition, 5(6A): 991-999.

保母敏行ほか (2008) 日本分析化学会における標準物質の開発. 分析化学, 57(6): 363-392.