＜教育講演（第34回年次学術集会より）＞

臨床検査データを合成・縮約するための主成分分析 第1回　基本統計量と正規分布

山西 八郎

Synthesis and Reduction of the Coefficient of Variation by Principal Component Analysis

Part I. Basic Statistics and Normal Distribution

Hachiro Yamanishi

Summary　The normal distribution is a probability density function determined by the standard deviation (SD) and mean, with the area under the curve in the mean ±2SD range accounting for 95.5% of the total area. Variance refers to the degree of variation in the measured data, and variance that considers degrees of freedom is called unbiased variance. Since the unit of variance is the square of the measured data, the standard deviation is defined by its square root to restore the unit to its original dimension. On the other hand, it is meaningless to compare the SDs of two groups with different mean values, and the coefficient of variation (CV) should be used as an indicator when comparing the degree of variability. If there is no overlap in the 95% confidence intervals of the CV being compared, the difference can be considered significant at the 5% level of significance.

Ⅰ．はじめに

2024年3月に大阪で開催された「第34回 生物試料分析科学会年次学術集会」の教育講演において、「臨床検査データを合成・縮約する」というテーマで、主成分分析について述べる機会をいただきました。その後、講演内容について「生物試料分析」への投稿依頼を受けましたが、1回の総説では言葉足らずとなることが予想されたため、3回シリーズとして掲載していただくこととなりました。第1回と2回は、主成分分析にたどり着くまでのWarming-upととらえていただければ幸いです。そこでまず第1回目は、すべての統計手法の基礎となる「正規分布と基本統計量」について解説します。

Ⅱ．正規分布

正規分布とは、μ：平均、σ：標準偏差、e：ネイピア数、π：円周率としたとき、次式で定義される平均を中心とした左右対称、釣り鐘型の分布をいう。なお、以下の解説では文献[2-4]

天理大学 医療学部
〒632-0018　奈良県天理市別所町80-1 天理大学別所キャンパス

Tenri University Department of Clinical Laboratory Science, 80-1 Bessho-cho, Tenri, Nara 632-0018, Japan

連絡先：山西 八郎
Tel: +81-743-63-7811
E-mail: yamaha@sta.tenri-u.ac.jp

を参考とした。

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

正規分布（normal distribution）は発見者である数学者ガウス（1777~1855）にちなんで、ガウス分布（Gaussian distribution）ともよばれる。Fig. 1に健常日本人男性におけるHDL-コレステロール（HDL-C）の分布を示す。データベースとしては「アジア地区共有基準範囲設定国際プロジェクト2009」を使用した。若干、右に裾を引いているが、おおむね正規分布とみなすことができる。ここで、健常な状態であっても、HDL-Cは30 ～ 80 mg/dLの範囲でゆらいでいる。つまり個人間差である。では、この分布を代表するHDL-C値を一つ選ぶとするならば、その第一候補となるのは平均という統計量である。統計量とは、観測（測定）されたデータの集合を要約した、あるいは代表する値の総称である。正規分布は平均と標準偏差の2つのパラメータ

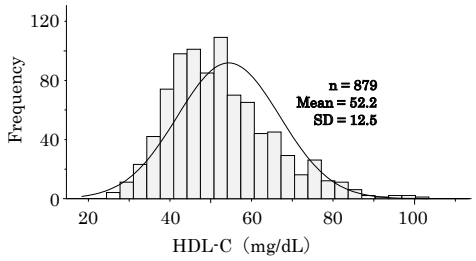

Fig. 1　Distribution of HDL-C in healthy Japanese men

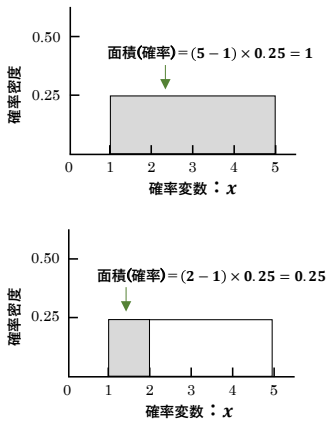

Fig. 2　Probability density vs. probability

（変数）のみで決定される。

ここで、Fig. 1に示したヒストグラムの縦軸は、各階級に属する標本数：度数であるが、無作為に選んだ標本がある階級に属する確率は、（階級度数）/（全標本数）として算出される。すなわち、階級度数は確率と同じ意味を有している。一方、各階級の中央値（階級値）を線で結ぶと折れ線となるが、階級数を無限大にすると、折れ線は曲線となる。これが正規理論曲線であるが、この曲線の縦軸を確率で定義することはできない。それは、曲線上には無限大数の実数が存在しているため、ある特定の実数が選ばれる確率は1/∞＝0である。そこで、正規理論曲線の縦軸は確率密度という別の概念で定義される。

確率密度を確率変数xが1~5の範囲の一様分布で考えると、そこからランダムに1つ取り出した実数が、① 1 ～ 5の範囲にある確率は1である。また、その実数が、② 1~2の範囲にある確率は0.25である（Fig. 2）。同様に、③ 1~3の範囲にある確率は0.5である。つまり、上述の確率をそれぞれの四角形の面積に置き換えると①の横軸の長さは“5－1＝4”、②では“2－1＝1”、③では“3－1＝2”であるので、面積と確率を一致させるためには四角形の高さは0.25となる。これが確率密度に相当する。したがって、確率密度を速度、確率変数xの範囲を走行時間に例えると、四角形の面積は走行距離を意味することになる。一様分布では分布の範囲は四角形となるが、正規分布のように曲線で定義される場合は、確率変数xの範囲における曲線下面積が確率に相当する。

Ⅲ．分散（variance）

Fig. 3にデータベースよりランダムに選んだ10人のHDL-C値を示す。10人の平均は52.8 mg/dLである。観測値と平均の差を偏差といい、観測値のバラつきの目安、つまりこの例では個人間差の指標となる。平均よりも低い観測データの偏差は負の値となるが、その絶対値の大きさでバラつきの程度を比較すると、標本No7が一番大きく、No10が最小であることがわかる。では、10個のデータ全体のバラつきの程度を評価するために、10個の偏差の平均を求めると、

$$\{2.2+(-5.8)+7.2+11.2+(-9.8)+(-4.8)+12.2+(-11.8)+(-0.8)+0.2\}\div 10=0\div 10=0$$

と、分子が0となるので平均も0と計算される。実はいかなるデータの集まりであれ、その偏差の総和は数学的に0となる。これは以下のように証明できる。データをx_1~x_n、総平均を$\bar{x}$として偏差の総和を求めると、

$$(x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})$$

$$=\left(x_1+x_2+\cdots+x_n\right)-n\times\bar{x}$$

$$=\left(x_1+x_2+\cdots+x_n\right)-\frac{n\times\left(x_1+x_2+\cdots+x_n\right)}{n}=0$$

そこで、各偏差を2乗した値の平均を考える。すると、すべての観測データが同値でない限り、その総和が0になることはない。また、偏差を2乗することにより平均からの距離（差）が増幅、あるいは圧縮されることになるが、2乗しても平均からの距離を反映していることに変わりはない。実際に計算すると偏差の2乗和（偏差平方和）は623.6であり、その平均は62.36となる。つまり、10人のデータ全体の平均からの距離、言い換えるとバラつきの程度は62.36であり、これを分散といい、標本データ全体の平均からのゆらぎの程度を表している。

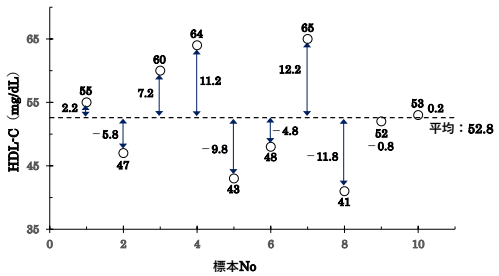

Fig. 3　Mean and deviation of the sample

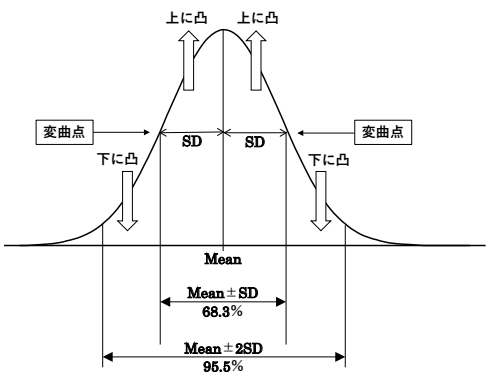

Fig. 4　Normal distribution and standard deviation

Ⅳ．自由度（degree of freedom：df）

このように、観測データのバラつきや検査データの個人間差を分散として評価することができるが、ここで問題となるのが自由度の概念である（高校数学でも分散・標準偏差は教授されるが自由度の概念は含まれていない）。Fig. 3を例とすると、偏差の総和は0となるので、任意の9個の偏差がわかれば残りの偏差は必然的に決まる。例えば、No7の偏差（12.2）が不明であったとしても、次式によりその値を求めることができる。

$$0-\{2.2+(-5.8)+7.2+11.2+(-9.8)+(-4.8)+(-11.8)+(-0.8)+0.2\}=12.2$$

同時に偏差平方和も任意の9個の偏差がわかれば自動的に決定される。「分散は偏差の2乗値の平均」と表現したが、ならば、偏差平方和に実質的に貢献しているn数は10ではなく10－1＝9と考える必要があり、この「n－1」が自由度の概念である。したがって、自由度を考慮した分散は次式で定義される。

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad — \qquad (2)$$

自由度を考慮しない分散を「標本分散」、自由度を考慮した分散を「不偏分散」といい、n数が例えば100であるとき、上式の分母が100であっても99であっても分散の値は大きく変化しないが、本項での分散はすべて不偏分散で統一する。

Ⅴ．標準偏差（Standard Deviation：SD）

分散は偏差の2乗値より算出されるため、その単位は観測データの単位の2乗となる。Fig. 3での分散は69.29（*mg/dL*）^{2}であり、観測データと加法（減法）計算をすることはできない。そこで、単位を元の次

元に戻すために、その平方根で定義されるのが標準偏差（SD）である。Fig. 3でのSDは8.32 mg/dLである。ここで最も重要となるのが正規分布におけるSDの意味である。

結論から述べると、正規分布において平均±1SDの範囲の曲線下面積は全体の面積の68.3％、平均±2SDの範囲の曲線下面積は全面積の95.5％を占める（Fig. 4）。ちなみに、全面積の95％を占める範囲は、平均±1.96SDである。また、分布曲線を左側からたどると、下に凸＞上に凸＞上に凸＞下に凸であり、上に凸、下に凸でもない変曲点が左右に1点ずつ存在する。正規分布では平均から左右の変曲点までの距離が1SDに相当する。見た目では正規分布と区別のつかないt 分布では、上述のようなSDと曲線下面積、変曲点との関係は成立しない。

Ⅵ. 変動係数（Coefficient of Variation: CV）

Fig. 5に異なる2群の観測データを示す。A群の平均は4.0、SDは1.83で、B群の観測値はA群のデータを10倍したものである。したがって、B群の平均は40.0、SDは18. 3となる。視覚的にはB群のデータの方が大きくバラついているように見えるが、A群で平均4.0から1単位バラつくことは、平均が10倍であるB群において平均40.0から10単位バラつくことと数学的なバラつきの程度は同じであると考えることができる。つまり、平均が異なる群間でSDの大きさを比べることに実質的な意味はない。また、平均が近似していても、単位の異なる群間でSDを比較することにも意味はない。バラつきの程度を比較する場合は、次式で定義される変動係数（CV値）を指標とする。

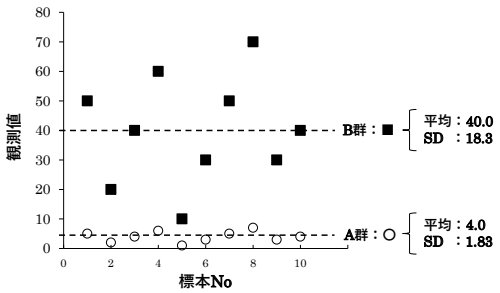

Fig.5　Relationship between sample data variability and mean and standard deviation

$$\text{変動係数} = \frac{\text{標準偏差}}{\text{平均}} \quad — \quad (3)$$

CV値は平均1単位あたりのSDの大きさを意味しており、平均の異なる群間のバラつきの程度を比較することができる。また、平均とSDは同単位であるので分母・分子で単位がキャンセルされてCV値は無単位となり、単位の異なる群間でもその大きさを比べることができる。(3) 式から算出される値に100を乗じて％表示する場合もある。Fig. 5のA群とB群のCVはともに0.458（45.8％）である。

一方､CV値の単なる大小の比較だけではなく、群間のCV値に有意差があるかどうかを検定するためには、CV値の95％信頼区間を推定する必要があるが、極めて複雑な計算が必要となるために[5-6]、少なくとも臨床検査の分野でCV値の有意差検定について議論されることはほとんどない。この点について稲田は簡便な信頼区間の直接概算法を提案している[7]。CV値を算出ための標本数は20以上、標本は正規分布にしたがうことなどの条件があるが、標本数をnとすると、CV値の95％信頼区間は次式で定義される。

$$\left[\frac{CV}{1+\sqrt{\frac{2}{n}}} \quad , \quad \frac{CV}{1-\sqrt{\frac{2}{n}}}\right] \quad — \quad (4)$$

2群間の95％信頼区間に重なりがなければ、有意水準5％でCV値に有意差を認めると判断できる。

Ⅶ. 終わりに

正規分布と基本統計量としての分散、標準偏差について解説した。特に正規分布と標準偏差の関係が重要で、パラメトリックな統計的仮説検定（有意差検定）の原理を理解するうえにおいても重要であることを付け加える。

本論文内容に関連する著者の利益相反：なし

文献

1） 山西八郎：臨床検査データを合成・縮約する．生物試料分析，47:4, 2024.
2） 市原 清志：バイオサイエンスの統計学, 262-284，南江堂，東京（1990）
3） P.G.ホーエル：初等統計学，75-104，培風館，東京（1979）
4） 山西八郎：正規分布と基本統計量，寺子屋統計教室，2-13，情報機構，東京（2022）
5） MacKay AT：Distribution of the coefficient of variation and the extended “t” distribution. J Royal Stat. Soc, 95:695-698, 1932.
6） Vangel MG: Confidence interval for a normal coefficient of variation. The American Statistician, 50: 21-26, 1996.
7） 稲田政則：変動係数の95％信頼区間の推定法．臨床化学，43：601-607，2018.