Technical Sheet
No. 21-23

データ解析入門 1 <要約統計量と相関係数>

キーワード：要約統計量、相関係数、アンスコムの数値例

はじめに

近年、様々な分野において人工知能や機械学習などの導入が進んでいます。機械学習などの解析技術の理解や実行には、統計学の基礎知識が求められます。本稿ではデータ解析に不可欠な基本的な統計量と相関係数について紹介します。

データの特徴を把握する方法

ある学校の生徒の身体測定結果を確認したいという状況を考えます。測定項目や生徒数が少ない場合は、数値の羅列からデータの傾向を把握することができるかもしれませんが、それでも骨の折れる作業になってしまいます。また、生徒数や測定項目が多くなると数値データのままでは、もはや手に負えなくなることは容易に想像できます。

そこで、視覚的にデータを理解できるようにグラフを用いることが重要になります。データの可視化には様々な種類のグラフが用いられています。

図 1 にヒストグラムと呼ばれるグラフを例示します。ヒストグラムでは、データをいくつかの階級に分け、階級ごとにデータの度数を表示します。したがって、ヒストグラムにより身長の大まかな分布をひと目で理解することが可能になります。ただし、階級の分け方によってグラフの印象が変わることがある点には注意が必要です。

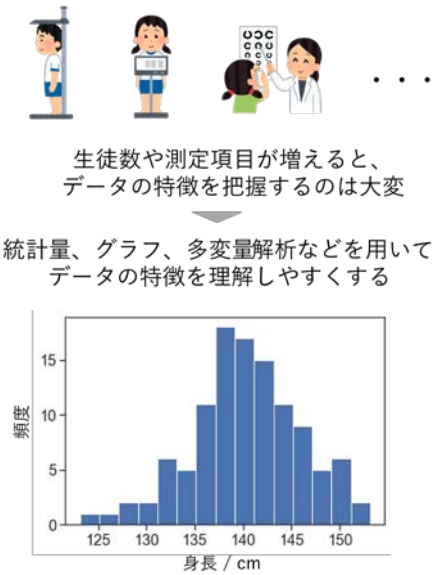

図 1 ヒストグラムによるデータの可視化

また、データの分布を要約する要約統計量を用いることもできます。表 1 に主な要約統計量を示します。たとえば、標準偏差はデータのばらつきを表します。そのほかの要約統計量や各統計量の定義式については Wikipedia[1)]などをご参照ください。

なお、標準偏差は測定値の単位と同じになります。そのため、身長の単位として、メートル（m）とセンチメートル（cm）のいずれを採用するかによって標準偏差は 100 倍変わります。逆に、この性質を利用することで、データのダイナミックレンジをそろえることができます。実際、標準偏差はスケーリングの目的でよく用いられる重要な統計量です。

表 1 代表的な要約統計量

平均値	データの総和をデータ数で割った値
中央値	全データを小さい順に並べたときの中央の値（第二四分位点）
最頻値	頻度が最大である値
分散	データのばらつきを表す指標
標準偏差	データのばらつきを表す指標（分散の正の平方根）
四分位点	小さい順に並べたデータを4等分したとき、その境界となる値

相関関係について

ここまでは測定項目ごとのグラフ作成や統計量の算出について考えてきました。次に 2 つの変数間の相関関係を考えてみます。先ほどの身体測定の例に戻りますが、身長が高ければ、体重は重くなる傾向にあると考えられ、両者の間に相関関係が予想されます。このような 2 つの変数間の相関を定量的に表す指標として相関係数があります。相関係数にはいくつか種類がありますが、一般的に相関係数というと、ピアソンの積率相関係数を指します（式(1)）。

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad (1)$$

地方独立行政法人
大阪産業技術研究所 本部・和泉センター
https://orist.jp/
〒594-1157 和泉市あゆみ野２丁目７番１号
Phone: 0725-51-2525（総合受付）

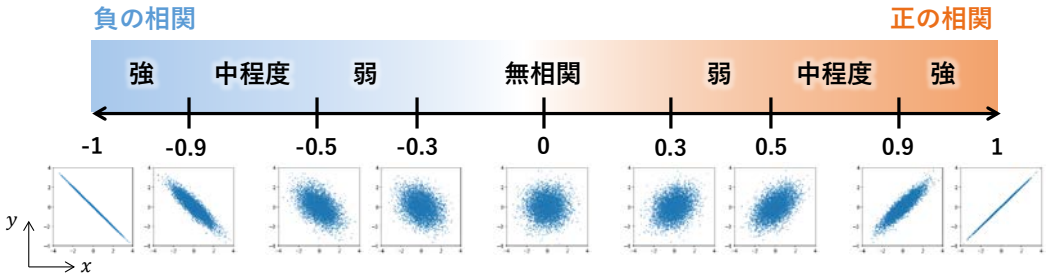

図 2　相関係数と 2 変数データ

ここで、r は変数xと変数yの相関係数、$\bar{x}$と$\bar{y}$は各変数の平均値、添え字 i はサンプルインデックス、N は全データ数を表します。

相関係数rは、$-1 \leq r \leq 1$ の値をとりますが、相関係数がどのくらいの値になると、相関が強いと判断して良いかについては絶対的な決まりはありません。また、適用分野や目的によっても解釈は異なります。そのため、あくまで目安でしかありませんが、相関係数と凡その相関の強さについて、5000 個の 2 変数データの散布図とともに示します（図 2）。要約統計量と同様に、相関係数を確認することはデータ解析を実行する上で重要です。

統計量によるデータ要約の注意点

表 1 に示した統計量はデータ解析において頻繁に用いられますが、要約統計量だけでデータの特徴を把握しようとすることは危険です。

図 3 にアンスコムの数値例と呼ばれる 4 つの 2 変数データを示します。各データの分布は明らかに異なります。一方、最小二乗法から得られる回帰直線も図示しましたが、いずれの回帰直線も $y = 3.00 + 0.500x$ となり、ほとんど同じになります。さらに、これらのデータから算出される要約統計量も、全く同じ、あるいはほとんど同じ値となります（表 2）。つまり、要約統計量ばかりに目をやっていては、データの特徴を見落としかねないということです。

以上、データの可視化の重要性について説明しましたが、実際のデータ解析では扱う変数が多すぎて、ひとつずつグラフを図示することが難しい場合があります。このような場合、データを逐一図示することはあきらめ、統計量を基にデータ処理を実行することになるかもしれませんが、統計量によるデータ要約の危険性について留意した上でデータ解析を進めていくことが推奨されます。

おわりに

本稿では、データセットの単変数あるいは 2 つの変数に着目して要約統計量などについて説明しました。一方、実際の解析対象は多変量データとなることが多く、複数の変数を同時に考慮する必要が出てきます。多変量データの取り扱いは単変量データよりも複雑になりますが、幸いなことにすでに数多くの有用な解析手法が確立しています。次稿では、主成分分析と呼ばれる手法による多変量データの可視化について紹介します。

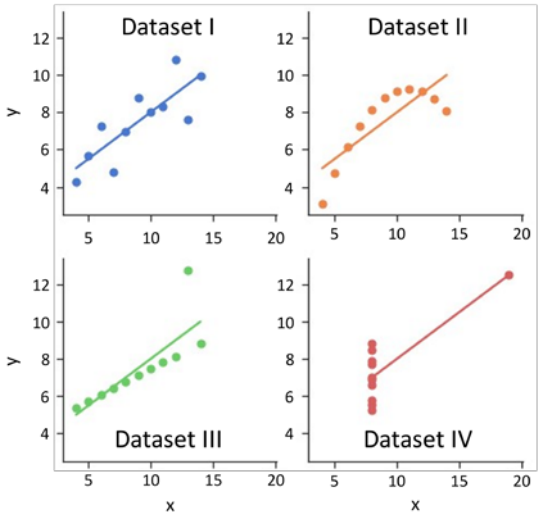

図 3　アンスコムの数値例

表 2　アンスコムの数値例から得られる各指標

Dataset	x		y		相関係数
	平均値	分散*	平均値	分散*	
I	9.00	11.0	7.50	4.13	0.816
II	9.00	11.0	7.50	4.13	0.816
III	9.00	11.0	7.50	4.12	0.816
IV	9.00	11.0	7.50	4.12	0.817

* 不偏分散[2)]

参考文献

1) Wikipedia,
http://ja.wikipedia.org/w/index.php?curid=236662
(accessed on November 10th, 2021)
2) Wikipedia,
http://ja.wikipedia.org/w/index.php?curid=8429
(accessed on November 10th, 2021)

発行日　2021 年 12 月 20 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@tri-osaka.jp