度数分布表、データの代表値

1　度数分布表とヒストグラム

データの個数 n をデータの大きさとよぶ。以下はそれぞれ離散型データ，連続型データの度数分布表とよばれる。変量のとり得る値とその対応する調査結果，観測あるいは実験した結果の集計を表にまとめたものである。

とり得る値が k 個の種類をもつ離散型データ

変量	x_1	x_2	$\cdots$	x_k	合計
度数	f_1	f_2	$\cdots$	f_k	n

データを k 個の組に級分けした連続型データ 階級は数値の測定値から「以上」（等号を含む場合）と「未満」（含まない場合）をつかって，重なりがないよう分ける。階級値は階級の真ん中の値。

階級	$a_1^+ - a_2^-$	$a_2^+ - a_3^-$	$\cdots$	$a_k^+ - a_{k+1}^-$	合計
階級値	x_1	x_2	$\cdots$	x_k	
度数	f_1	f_2	$\cdots$	f_k	n
相対度数 (%)	f_1/n	f_2/n	$\cdots$	f_k/n	n

連続型データをグラフに表現したものが，**ヒストグラム** である。同時に２つの集団を書くと，重なってしまうので，柱の中点を順に結んでできる，**度数多角形**もよく用いられる。級の間隔や級の個数を適切に選択することは大切なことであるが，最も適切というものは判断が難しい。しかし集団の状況をさまざまな処理により，視覚により把握，判断することは，新しい発見や知見を高める上で最も重要な基本的な「探索的なデータの解析」である。コンピュータの図的表現の利用も基本的なものである。グラフによる表現には，棒グラフ，円グラフ，絵グラフ，折れ線グラフがよく用いられる。とくに棒グラフとヒストグラムは似ているが，本質的に異なることに注意する。

また複雑な状況を分かり易くするためや見栄えをよくするためにいろいろと工夫がされている。たとえば実験データについては，ボックスチャート（最大，最小，平均，４分位数を同時に表現）などがある。また単なるヒストグラムの代わりに，幹葉図，デジタルチャートというものもある。１個体について変数の組を同時に考える（多変量データ）場合には，レーダーチャートなどがよく用いられる。

2　データの標準化

与えられたデータを簡単に効率よく計算するには、データの変換をします。基本的には、データの集団を平行移動や目盛りの縮約・拡大です。これらの操作を線形変換とよびます。もとのデータ $X_1, X_2, \cdots, X_n$ から、新しいデータをつくるために、ある定数 a を加え、c 倍します。$c(X_1+a), c(X_2+a), \cdots, c(X_n+a)$ これを変数 Y とおきます。つまり $Y_i = c(X_i+a), i = 1, 2, \cdots, n$ が線形変換の形です。このとき X の平均 $\overline{X}$、分散 s_X^2 と Y の平均 $\overline{Y}$、分散 s_Y^2 とは、つぎの関係式が成立します。

$$\overline{Y} = c(\overline{X} + a), \quad s_Y^2 = c^2 s_X^2 (a \text{ にはよらない} \tag{1}$$

とくに $a = \overline{X}, c = \dfrac{1}{s_X}$ のばあいには $\overline{Y} = 0, \quad s_Y^2 = 1$ となりますから、平均ゼロ、分散は１にすることができます。