データの整理

統計調査をおこなって記述されたデータの結果は、大まかに、質的属性と量的属性の２種に分けることができる。質的属性 とは、世帯主の職業とか有権者の支持政党など、調査結果が数値ではなく、いくつかの項目にいずれか属するとか属さないという形で与えられる。質的属性の分類項目を カテゴリー (category) とよぶ。また属性をカテゴリーに分類するとき、予めなんらの方法により、数量化しておいたほうが統計解析を施し易いことが多い。　しかし複雑な多面的現象や微妙な心理的特性、知能や好み具合を数量化することは非常に難しいことである。これに対して、統計調査の観測結果が数値により記録される場合、そのデータを量的データという。量的データには、$0, 1, 2, 3, \cdots$ など自然数や整数を結果としてとり得る計数データ (count data) と、小数（あるいは分数）の値をとり得ると考えられる計量データ (measure data) に分類される。長さ、重さ、温度など数直線上の目盛り、実数を測定するものも計量データである。後の確率的取り扱いと対応させるためには、前者を 離散型データ (discrete data)、そして後者を 連続型データ (continuous data) と呼んだほうが都合がよい。いずれの２つも、これらは計量データであり、理論上、測定を精密にすることにより、実数軸上の任意の値をとり得るが、測定値として記録する場合、離散化が施される。野球選手の打率は、小数第４位で丸めて、0.263, 0.320 などとなり、また身長の測定値は mm の単位で、体重は kg で測定し、$100g$ 単位で丸めるのが普通に行われている。測定値が離散的になっていても、本来の変量が連続的に変化すると考えられるならば、それは連続的変量である。連続的変量の測定値は離散化されて記録されるが、元々の離散型データとは異なる状況にある。統計解析をおこなうために，調査，観測をした結果，得られた個体の特性の表す数値を変量という。

連続型変量 数値が小数で表される，連続的な値をとる

離散型変量 整数や自然数など，とびとびの値をとる

統計データ	
質的データ	カテゴリー（項目）で分類
量的データ	計数データ　離散型変量の観測値
	計量データ　連続型変量の観測値

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。

数値によるデータのまとめ

集団の中心位置を表す尺度

平均 もっとも多く使われる算術平均であるが，重心を意味する。数学的に取り扱いやすいし，理論的な側面もきわめて重要であるが，万能というわけではなく，裾の影響に大きく左右される欠点をもつ。$\overline{X} = \frac{1}{n}\{X_1 + X_2 + \cdots + X_n\} = \sum_i X_i/n$ データの平均値 (average, mean value) とは、データの値の総和を個数で割ったものをいい、$\overline{X}$ (エックスバーとよむ) で表す。n 個のデータの総和は k 個に分けた階級値と度数をつかって表すと，$X_1 + X_2 + \cdots + X_n$ の計算の代わりに $x_1 f_1 + x_2 f_2 + \cdots + x_k f_k$ で総和を求める。

中位数（中央値） 小さいものから大きいものへと，大きさの順に並べて，ちょうど真ん中に位置する値で，小さい方も大きい方もそれぞれ５０％ずつとなる値。比較的頑健性をもち，常識にもかなっている。ただし，データがたくさんであると，大きさの順に並び替えることに労力を要する。

モード　度数分布表で求めた度数の最大となっている値をいう。ヒストグラム（度数多角形）がひとつ山の形を単峰型といい，モードが中心傾向を表す。一つ山で対称なときにはこれら３つともほぼ同じ値になるが，集団の山形に歪み（Ｌ字型は正の歪み，逆Ｌ字型は，負の歪みをもつという）があれば，これらは異なった値となる。

バラツキを表すための尺度

分散 各データと平均からのずれ（差）を平方 (square) した値の平均で，非負の値である。ゼロとなるのは，全てのデータが同じ値，すなわちまったく変動がない場合である。

$$s^2 = s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2$$

標準偏差 分散の単位はもとのデータの２乗値となるが，この平方根をとったもので，平均と同じ単位となる。Standard Deviation からＳＤと約されることが多い。

$$s = \sqrt{s^2} = \sqrt{s_X^2}$$

四分位数 中位数（メディアン）がちょうど１／２となる点であったと同様に，４つの２５％ずつに分けた値で，第１四分位数（小さい方から２５％の点），第２（中位数），第３四分位数（７５％の点）といい，第３から第２を引いた値を４分位範囲という。

パーセンタイル（百分位数）；非常に大きな集団で，これを１００個のパーセントに分けたものをいう。上位５％点，下位５％点がとくに集団の特質や特徴を表すために用いられる。

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。

多変量データの解析

１個体が複数の変量で表される（データがベクトルで与えられる）場合で、２変量のデータ（平面上のデータ）であれば、

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

平面状にプロットした図を散布図という。また３変量のデータ（空間上のデータ）であれば、つぎのような形がデータである。

$$(x_1, y_1, z_1), (x_2, y_2, z_2) \cdots, (x_n, y_n, z_n)$$

このようなデータを整理するには、いろいろな方法があるが、2 変量データについて、つぎのような概念が重要である。

共分散 2 変量の間の全体的なばらつき具合を考える。一つの変数の増加について、他方も増加であれば、正の値、逆の場合には、負の値をとる。増減がはっきりしないときには、ゼロに近い値となる。つぎの相関係数と対応する。

$$s_{XY} = \frac{1}{n} \sum_n \{(X_i - \overline{X})(Y_i - \overline{Y})\}$$

相関係数 共分散を各変量の標準偏差で基準化したもの。

$$\rho = \frac{s_{XY}}{s_X s_Y} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{1}{n} \sum_n \left(\frac{X_i - \overline{X}}{s_X} \right) \left(\frac{Y_i - \overline{Y}}{s_Y} \right)$$

$-1 \leq r \leq 1$ が成り立ち、絶対値が 1 に近いときには直線傾向が強く、絶対値がゼロに近いときには、直線傾向がみられないことを意味する。

曲線の当てはめ、最小 2 乗法 散布図からある程度の直線的な関係があるとき、これらの 2 変量データに直線をつぎの基準で当てはめる。

$$\sum_i (Y_i - a - bX_i)^2$$

を最小にするような a, b を求める。

回帰直線 この方法を最小 2 乗法といい、求めた直線を Y の X への回帰直線という。最小 2 乗法で定めた直線の係数を回帰係数という。

$$b = \frac{s_{XY}}{s_X^2} = r\frac{s_Y}{s_X}, \quad a = \overline{Y} - r\frac{s_Y}{s_X}\overline{X}$$

「練習問題」

1 5 個の数値データ $3, 5, 4, 1, 7$ の平均と分散を計算し、これをもちいて、つぎの場合の数値データの平均と分散を計算しなさい。

(a) $13, 15, 14, 11, 17$

(b) $2.3, 2.5, 2.4, 2.1, 2.7$

2 4 個の数の平均は 5、分散は 2 で、もうひとつの集団の 6 個の数値の平均は 7、分散は 3 であるという。この 2 つを合わせた 10 個に対する平均と分散を計算しなさい。

3 与えられたデータ $X_1, \cdots, X_n$ について、これを変換して平均は５０、分散は１００,（標準偏差は１０）にするにはどうしたら、よいか。これは「ｔスコア（偏差値)」とよばれます。

4 相関係数は線形変換をしても同じ値になることを示しなさい。つまり $(X_1, Y_1), \cdots, (X_n, Y_n)$ と $(aX_1 + b, cY_1 + d), \cdots, (aX_n + b, cY_n + d)$（ただし $ac > 0$）とは同じ相関係数をもつ。

5 与えられた 2 変量データ X, Y が順位データであるとき、すなわちそれぞれが $1, 2, 3, \cdots, n$ までを順位づけられていて、並び替えたものである。ただし同順位はないとする。このときには相関係数が

$$r = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}, \quad (d_i = X_i - Y_i)$$

で与えられることを示しなさい。これをスピアマンの順位相関係数という。