ヒストグラム、散布図

1　集団の位置尺度

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。

2　バラツキの尺度

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。

問 2.1
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) $2.3, 2.5, 2.4, 2.1, 2.7$

問 2.2
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。

問 2.3
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。