Excel による統計分析の基本

数値的指標～平均

最も多用される数値的指標としては，**算術平均**が挙げられる。n 個のデータ $\{x_1, x_2, \cdots, x_n\}$ の和をデータ数 n で除することによって得られる。和の記号 $\sum$ を用いれば，以下のように表現される。

$$\text{算術平均} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

算術平均が使用できない場合に，**幾何平均**や**調和平均**が用いられることがある。

$$\text{幾何平均} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

記号 $\prod$ は，積和（すべてのデータを掛け合わせたもの）を表す記号である。幾何平均は，たとえば，複数年の物価の上昇率の年平均を求めたい時に使用する。1 年目が 1[%](1.01 倍)，2 年目が 6[%](1.06 倍)といった時は，算術平均 $\frac{1}{2}(1+6) = 3.5$ [%]を用いるのは誤りで，正しくは，幾何平均 $\sqrt{1.01 \cdot 1.06} = 1.03469$ から約 3.469[%]を用いなければならない[1]。

$$\text{調和平均} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}$$

調和平均は，たとえば，速度の平均を求める時に用いられる。同じ道のりを，行きは 30[km/h]，帰りは 40[km/h]で移動したとすると，算術平均 35[km/h]は誤りで，調和平均 $\frac{1}{\frac{1}{2}\left(\frac{1}{30}+\frac{1}{40}\right)} = \frac{1}{\frac{1}{2}\left(\frac{40+30}{1200}\right)} = \frac{1}{\frac{1}{2}\cdot\frac{7}{120}} = \frac{240}{7} = 34.28$ から，約 34.3[km/h]を考えなくてはならない[2]。

分布の特徴を示す代表値

算術平均が意味を持つ場合であっても，データの散らばり方（分布）を無視していると，データの正確な特徴を掴むことができない。そこで，分布の数値的要約量である平均偏差や分散（あるいはその平方根の標準偏差）と組み合わせて利用する必要がある。**平均偏差**は，

---

[1] 何故か？別の例を使って説明してみなさい。
[2] これも理由を説明してみなさい。

$$\text{平均偏差} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \text{算術平均}\right|$$

として表現される。また，**分散**は，

$$\text{分散} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \text{算術平均})^2$$

となり，**標準偏差**は，$\sqrt{\text{分散}}$ として求められる。

これら以外にも，歪度（わいど），尖度（せんど）といった指標が用いられることがある。

歪度は，分布の偏りの方向と程度を示す指標である。また，尖度は，分布の尖り具合を示す指標である。

また，データの**最小値**，**最大値**，**中央値**，**最頻値**や**四分位点**を求めることで，より詳細な分布の形状を（数値的に）把握することができる。

ヒストグラム

ヒストグラムは，上述の指標値だけでは得られない，分布の特徴を直観的にとらえやすくするためのグラフ表現である。通常は，図 1 のように，データを一定の幅に区切って作った階級に反応するデータの個数（**度数**）を棒グラフとして表現したものである。また，通常は，階級の中央の値をその階級の代表値として表示する。データが離散値の場合（名義尺度の場合）には，項目の値そのものを代表値として各階級を決める。

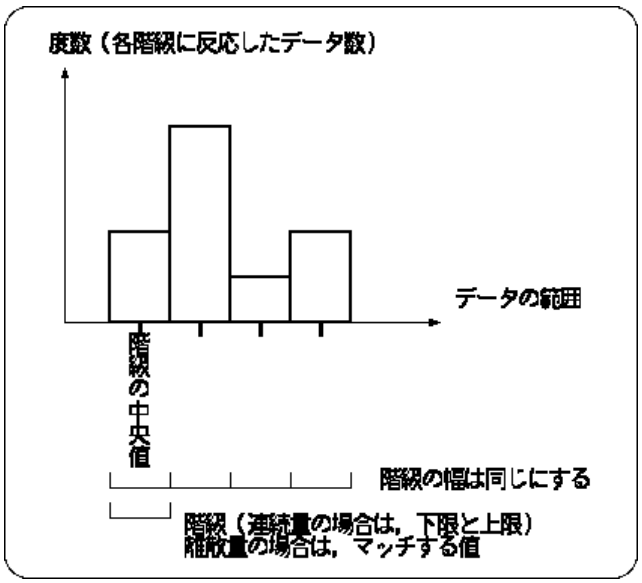

図 1：一般的なヒストグラムの様式

階級の幅やその数は分析者が任意に決めてよい。しかし，幅が狭すぎると（階級の数が

多すぎると)，歯抜けの状態になりがちで，視覚的にも適当でない（得るところの少ない）グラフとなってしまう（図 2)。

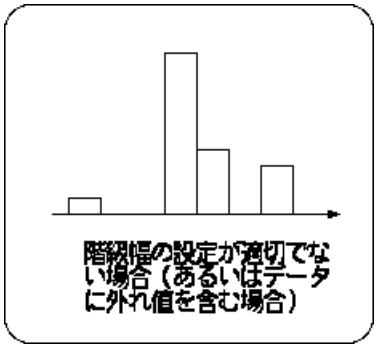

図 2：階級の設定が適切でない場合のヒストグラム

一般的には，データの数nに対して，階級の数kを，以下の**スタージェスの公式**で定めるとよい。

$$k = 1 + \log_2 n = 1 + \frac{\log_{10} n}{\log_{10} 2} = 1 + \frac{\log_{10} n}{0.30103} = 1 + 3.3219 \cdot \log_{10} n$$

これによれば，$n = 100$の場合には，

$$k = 1 + 3.3219 \cdot \log_{10} 100 = 1 + 3.3219 \cdot \log_{10} 10^2 = 1 + 3.3219 \cdot 2 = 7.6438$$

となるので，7 または 8 階級に区分するのが適当であり，$n = 1000$の場合でも，10 または 11 階級程度でよいということになる[3]。

階級数が決まれば，データの最大値と最小値の範囲をこの数で除して，階級の幅を定めればよい。なお，実際には半端な数を避け，例えば，5 刻み，100 刻みのような切りのよい値に意図的に揃えると，グラフとして見やすいものになる。

階級が決まれば，いずれの階級に属するかをすべてのデータについて検査し，その反応数（度数）をカウントしていく。この結果を棒グラフで表現するとヒストグラムが得られる。

図 3 に代表的な分布の形状を示す。同じ平均値や分散を示している場合でも，このようにヒストグラムにすることによって，偏り具合の観察や，ピークが複数あるような分布の発見により，違いを比較することが可能になる。

[3] Excel には任意の値を底とする対数を求める関数 LOG があるので，例えばデータ数が 100 の時は，適当な場所のセルで，=1+LOG(100,2) と式を入力すると上と同様の値が求まる。

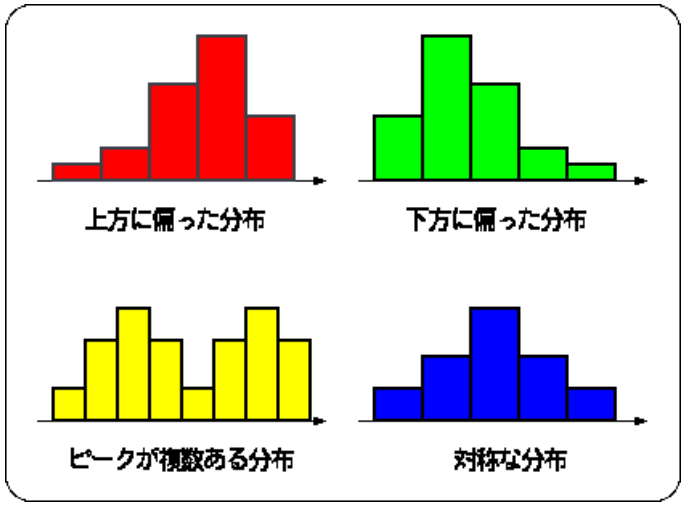

図 3：ヒストグラムで見る分布形状の代表例

演習

Web ページ上に用意したデータ（port-08ex.xls）を用いて，各種の統計量の違いを比較してみよう。

ここでは，算術平均が同じとなるようないくつかの人工的データが準備されている。各指標が，どのように異なるかについて留意しつつ，必要な部分を埋めてみよう。

ポイント

- Excel 関数による集計の仕方と，統計量の求め方
- 分布形状の違いと統計量との関係