確率論・統計学

1　導入, 記述統計

1.1　統計学の分類

- 記述統計 ... 観察対象となるデータを集め, そのデータを整理し, そのデータの特徴を記述する
  Ex: 国勢調査、試験の成績
- 推測統計 ... 収集した一部のデータから全体の性質や傾向を推測する
  Ex: 世論調査、破壊検査など

1.2　記述統計

1.2.1　データの尺度

データにも色々ある。

1. 質的データ (カテゴリデータ)
   （a）名義尺度 ... 性別, 婚姻の状態など
   （b）順序尺度 ... 階級など
2. 量的データ (数値データ)
   （a）間隔尺度 ... 摂氏温度、時刻など
   （b）比尺度 ... 絶対温度、経過時間など

名義尺度：他と区別するだけ
順序尺度：間隔に意味はないが順序に意味がある
間隔尺度：目盛が等間隔である
比尺度：原点があり、間隔や比に意味がある

1.3　データの記述

1.3.1　代表値

異なるデータを、直接比較するのは (人間の能力的にも計算機的にも) むずかしい。そこで、データの性質を表す「代表的」な値を考える。特にデータの「中心っぽい」値を代表値と呼んで、いろいろ用いられるが、ひろく使われるものは以下のようなものがある。

定義 1. 算術平均

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdot + x_n}{n}$$

定義 2. 幾何平均

$$x_G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

定義 3. 調和平均

$$x_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdot + \frac{1}{x_n}}$$

平均のお気持ち

普通の場合

年収を全部平均で置き換えると、総和が不変

比率の平均を取る場合

上昇率を全て算術平均で置き換える。この時、上昇率が $100\%, 150\%, 200\%$ だった場合、算術平均は 150% となる。そしてこれで全部置き換えると、

$$\begin{aligned} xtimes1.0 \times 1.5 \times 2.0 &= 3x \\ x \times 1.5 \times 1.5 \times 1.5 &= 3.375x \end{aligned}$$

一方、幾何平均は $\sqrt[3]{1.0 \times 1.5 \times 2.0}$
であるから、全部置き換えると当然

$$x \times \sqrt[3]{1.0 \times 1.5 \times 2.0} \times \sqrt[3]{1.0 \times 1.5 \times 2.0} \times \sqrt[3]{1.0 \times 1.5 \times 2.0} = 3x$$

となり、性質が保たれて嬉しい。したがって、掛け合わされていく数値の平均を取りたい時は幾何平均を使うのが良い。

定義 4. 中央値

データを照準に並べたとき、中央に位置する値。ただし、データの個数が偶数の場合は中央に位置する 2 つの値の平均をとる。

中央値と平均値では、中央値の方が外れ値に対して頑健。例えば世帯年収では、@abap34 のように年収 5000 兆円の人間がいるだけで日本人の平均年収は五千万円を超えてしまう。一方中央値では、@abap34 が入ったとしてもほとんど変化がない。

1.3.2 データの可視化

データをいい感じに可視化するために色々と考えられている。

1.3.3 箱ひげ図

内容についてはスライドを参考のこと。ここでは、第一四分位と第三四分位の求め方を確認しておく。

まず、データが奇数個の時は、中央を除いて左右に分割し、それぞれの中央値を第一四分位、第三四分位とする。次に、データが偶数個のときは、単にデータを左右に分割してそれぞれの中央値を採用すれば良い。

また、ひげの長さの 1.5 倍以上四分位範囲から離れたデータは外れ値としてプロットされる方式もある。が、面倒なのでテストで聞かれたら書かなくても良さそうではある。(小テストでは OK だったので)

プロットを人間にさせる方が間違っているので、楽をするのが吉

1.3.4 ヒストグラム

同じく内容はスライドを参照。