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

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

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、@abaO34 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、@abaO34 が入ったとしてもほとんど変化がない。

1.3.2 分布の記述

代表値では、データの平均的な傾向が調べられるが、データのばらつきについては調べられない。そこで、データの分布に対して定まるデータのばらつきを表す数を導入する。

定義 5. 分散

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

定義 6. 標準偏差

$$S = \sqrt{S^2}$$

つまり、分散は平均値との差の二乗 (各データの散らばり度合い) の平均で、標準偏差はその平方根。分散は二乗の差を平均するので、単位が二乗になってしまう。そこで、標準偏差は分散の平方根をとることで、単位を元に戻している。

標準化

平均と標準偏差を使って、異なるデータの分布から得られた値にある評価を与えられる。

定義 7. 標準化

$$z = \frac{x - \bar{x}}{S}$$

これによって各データを

$X = x_1, x_2, x_3, \cdots, x_n$ から $Z = z_1, z_2, z_3, \cdots, z_n$ に変換すると。

Z は平均 0、標準偏差 1 となる。(全てのデータから x 引けば平均は x 減り、S で割れば各データの偏差は $1/S$ になり標準偏差は $1/S$ になるため)

1.3.3 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。

1.3.4 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

1.3.5 ヒストグラム

同じく内容はスライドを参照。

階級数の決め方としてスタージェスの公式

$$k = 1 + \log_2 n$$

(k: 階級数, n: 観測数)

がある。

2　相関、回帰分析

各データがある一つの数字 (あるいは記号など) ではなく、複数の数・記号などの組み合わせで表される場合がある。

例) 人の健康データを集めた際に各個人について身長、体重、血圧を集めた $\mathbb{R}^3$ の元で一つのデータとなる

このような性質を持つデータを多次元データという。(個人的には、多変量という言い方をよく聞く) この複数の変数からなるデータを解析するために複数のデータから定まる値が導入される。

2.1　相関係数

2.1.1　ピアソンの (積率) 相関係数

まず、ピアソンの (積率) 相関係数を定義する。

これは、

1. 値が 1 に近いほど、正の相関 ($= x$ が大きいほうであれば y も大きい方にある) が強い
2. 値が-1 に近いほど、負の相関 ($= x$ が大きい方であれば y は逆に小さい方にある) が強い
3. 値が 0 に近いほど、相関が弱い

という性質を持たせるように定義されている。

まず、共分散という値を導入する。

定義 8. 共分散

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

これは、x と y の偏差の積の平均である。まず、$(x_i - \bar{x})$ と $(y_i - \bar{y})$ の符号が同じならば、その積は正になる。そして、符号が異なれば、その積は負になる。

このことは、上の (書き下した) 相関係数の性質を満たすようになっている.

そして、これを使ってピアソンの (積率) 相関係数を定義する。

定義 9. ピアソンの *(*積率*)* 相関係数

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

先ほどの共分散は、二つの変数の相関関係の「方向」はあっていたが、その「強さ」は表していなかった。例えば、得点率が同じ二つのデータが、どちらも 100 点満点から 200 点満点になっただけで相関関係が 4 倍になってしまう。それでは複数のデータを比較するときに不便なので、それぞれのデータの標準偏差の積で割ることで、補正している。

どちらかというと、共分散を求める際の各偏差をそれぞれの標準偏差で割ることで、

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_x}\frac{(y_i - \bar{y})}{S_y}$$

として、各偏差の標準偏差で割って平均を引く (思い出すシリーズ: 標準化 (1.3.2)) 後に後に共分散を求める、という方がわかりやすいかもしれない。つまり、ピアソンの相関係数は、各データを標準化した後に共分散を求めているということである。

2.1.2 スピアマンの順位相関係数

実はピアソンの相関係数は、データの線形関係への近さを反映していた。

例えば、

$X = 1, 2, 3, 4, 5$ と $Y = 1, 4, 9, 16, 25$ というデータがあったとする。

これらのピアソンの相関係数を計算すると、0.9811049102515929 と高いが 1 ではない。ピッタリ 1 になる場合は、データが完全に線形関係にある場合である。

再びピアソンの相関関係の式を考える。(標準化した後に共分散を求める形)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_x}\frac{(y_i-\bar{y})}{S_y}$$

ここで、あるデータ X に対して標準化した後のデータと、$Y = aX + b$ と書けるデータの標準化したデータは、完全に一致する。(定数倍は標準偏差で、定数の加算は平均で打ち消される)

つまり、上の定義は、

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{S_x}\frac{(x_i-\bar{x})}{S_x}$$

つまり、

$$\frac{1}{n}\frac{1}{{S_x}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

ところで、

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

であるから、結局、

$$\frac{1}{n}\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 1$$

これまで見たように、ピアソンの相関係数は

「常に (x が増えるならば y も増える) が成り立つならば、最大限 1 を取る」という性質があるわけではなかった。これを成り立たせるのが、スピアマンの順位相関係数である。アイデアはごく単純で、データを順位に変換してからピアソンの相関係数を求めるだけである。

この際、順位の平均値は $\frac{n+1}{2}$ であることなど、たいていの代表値が n の関数で書ける。これらを使って、いい感じに式を変形すると、以下のようになる。

定義 10. スピアマンの順位相関係数

データを順位に変換し、その各順位を R_{x_i}, R_{y_i} とすると、

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_{x_i} - R_{y_i})^2$$

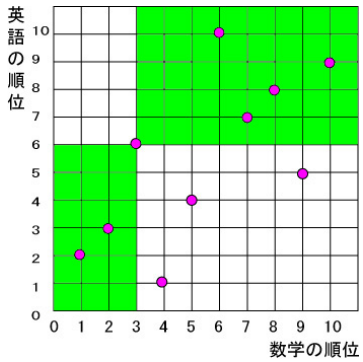
2.1.3 ケンドールの順位相関係数

他のも順位を評価する相関係数として、ケンドールの順位相関係数がある。講義資料だと、う笑という感じだが、ここでは、書き下した定義を書いておく

定義 11. ケンドールの順位相関係数

r_K = *(*各データ点について、散布図で右上か左下にある点の数*)* - *(*各データ点について、左上か右下にある点の数*)* の平均

各点以下のような感じで数えて、全ての組合わせの数 $(n(n-1)$ で割ればいい。$)$



引用元: http://www.tamagaki.com/math/Statistics610.html

2.1.4 偏相関係数

なんか扱いが小さいので飛ばす。https://manabitimes.jp/math/1400 を参照。

2.1.5 相関関係と因果関係

相関係数は、二つのデータの関係性を数値化する。この値によってデータが「数値的にどれだけに関係しているか」を表せるが、あくまでそれだけであって、相関関係があるため直ちに因果関係 (原因と結果のお関係) があるとは言えない。

相関関係はデータさえあれば計算できるが、一般に、因果関係の推定はかなりむずかしい。一方で、ビジネスや研究で気になるのは相関関係ではなく因果関係である。そのため、因果関係を推定する (因果推論) のはかなり需要があり、広く研究されている。結構面白いのでおすすめ。「効果検証入門」とかが良かった。(サンプルコードは R)

2.1.6 回帰

回帰分析とは、データ間の関係を表す構造 (モデル) を求めることである。つまり、データ y が $x_1, x_2, \ldots, x_n$ によって決まるとすると、$y = f(x_1, x_2, \ldots, x_n)$ の f に当たる部分を推定する。この際、f にどのような制約を課すかによって、この作業に色々と名前がつく。

例えば、y が $x_1, x_2, \ldots, x_n$ の線形結合で表されると仮定したとき、つまり f が (内部に固有にもつ) パラメータとして $a_1, a_2, \ldots, a_n$ を持つとすると、

$$f(x) = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

とした場合、これは線形回帰と呼ばれる。

逆にこれを仮定しない場合、例えば

$$f(x) = a x_1 + b x_2 + c x_3^2$$

のような場合、これは非線形回帰と呼ばれる。ここまで多項式で書き下されるものだけを例に出したが当然もっと複雑でも良い (例えばディープニューラルネットワークなど)

2.1.7 最小二乗法

さて、f はふつうパラメータを持ち、これが定まることで計算可能になる。例えば線形回帰においては、$a_1, a_2, \ldots, a_n$ がパラメータである。

つまり回帰は、

1. モデルに課す制約を定める
2. モデルのパラメータを求める

という二つのステップに分かれる。ここでは二つ目のステップについて述べる。単純な回帰問題で一番素直な方法は最小二乗法と呼ばれる方法である。

これは、

$$L = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}))^2$$

と定まる関数 L を最小化するようなパラメータを求める方法である。(つまり、データとして $Y = y_1, y_2, \ldots, y_n$ というものが与えられているとき、各予測値 $f(\boldsymbol{x})$ との差の二乗を考えている。) ここで注意が

必要なのは、この関数 L はここでは $a_1, a_2, \ldots, a_n$ の関数であるということである。

この関数を最小化する $a_1, a_2, \ldots, a_n$ は解析的に求めることができる。データを $X \in \mathbb{R}^{n \times m}$ とする。(つまり m 次元データが n 個ある) そして、パラメータを $\boldsymbol{a} \in \mathbb{R}^m$ とする。そして、回帰するデータを $\boldsymbol{y} \in \mathbb{R}^n$ とする。

すると、結局解きたい問題は

$$\arg\min_{\boldsymbol{a} \in \mathbb{R}^m} \|X\boldsymbol{a} - \boldsymbol{y}\|^2$$

である。