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、

$$xtimes 1.0 \times 1.5 \times 2.0 = 3x$$
$$x \times 1.5 \times 1.5 \times 1.5 = 3.375x$$

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、

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

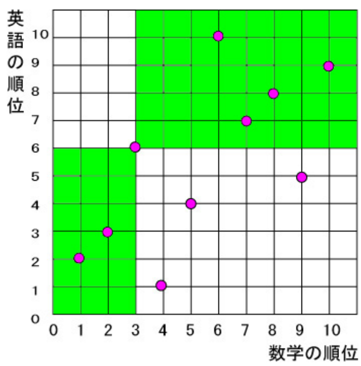
2.1.3 ケンドールの順位相関係数

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。

講義では単回帰のみを扱っていたので紹介すると、単回帰の場合はパラメータとして $a \in \mathbb{R}$ と $b \in \mathbb{R}$ を持ち、$f(x) = ax + b$ となる。(a は回帰係数、b はバイアス項とか切片などと呼ばれる。) そして a, b は解析的に求めることができて、

$$b = \frac{S_{xy}}{S_x}, \quad a = \bar{y} - b\bar{x}$$

である。(a と b について偏微分して、偏微分係数が 0 であるという必要条件から一意に定まる。)

2.1.8　決定係数

決定係数は、回帰分析の当てはまりの良さを表す指標である。

定義 12. 決定係数

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

予測誤差の二乗和を全変動の二乗和で割ったものである。定義からこれは 1 以下の実数となり、f が線形である場合は、これは相関係数の 2 乗と一致する。

3　確率、確率変数、確率分布

3.1　確率変数

世の中のいろいろな現象は (厳密には違うかもしれないが) 確率的に起こる。例えば、募金箱に手を突っ込んでお金を取り出すとき、出てくるお金は (ほぼ) 確率的に決まる。このように確率的に値が変動するものを確率変数と呼ぶ。この確率変数を少しきちんと言い直すと、以下のようになる。

定義 13. 確率変数

確率変数とは、事象から実数への写像である。

例えば、募金箱から出てくるお金を表す確率変数は、$X(100\,円を引く) = 100,\ X(10\,円を引く) = 10$ のように定義される。ではこの写像の定義域を考えると、生じる全ての事象となる。これを、標本空間といい、Ω と表す。また、起こる個々の分割できる最小の単位を標本点といい、ω と表す。当然、標本点は標本空間の元である。そして、標本点の集合 (つまり、標本空間の部分集合) を事象と呼ぶ。

例えば、コインを投げて表が出ることを 0, 裏が出ることを 1 と便宜的に書くと、標本空間は $\Omega = \{0, 1\}$ となるし、ω としては、$0 \in \Omega$ や $1 \in \Omega$ などが考えられる。そして、事象としては例えば jj

3.2　確率の公理

ここまで確率変数を定義してきたが、これをおこなってきたのは確率をよりきちんと定義するためである。

確率は、標本空間の部分集合に対して、実数を割り当てたものである。現代の確率論は、確率が満たすべき公理のみを決めて、特に「確率とは何か」ということは調べず、この公理からさまざまな事実を導いていく。(らしい)

その確率が満たすべき公理が以下である。

定義 14. コルモゴロフの公理

標本空間の冪集合から実数への写像 P が以下の条件を満たすとき、P を確率と呼ぶ。

1. $P(A) \geq 0$ $(A \subset \Omega)$
2. $P(\Omega) = 1$
3. $A_1, A_2, \ldots$ が互いに排反ならば、$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

さて、この公理を満たす写像 P を確率と呼ぶのだが、この公理だけから確率のさまざまな性質を導ける。例えば、二つの事象の確率について次のような定理が成り立つ。

3.2.1 加法定理

定理 1. 確率の加法定理

$A, B \subset \Omega$ に対して、$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

証明. $A \cup B = A \cup (B \setminus A)$ である。ここで、A と $B \setminus A$ は互いに排反であるから、$P(A \cup B) = P(A) + P(B \setminus A)$ となる。また、$B = (B \setminus A) \cup (A \cap B)$ であるから、$P(B) = P(B \setminus A) + P(A \cap B)$ となる。これらを足し合わせると、$P(A \cup B) + P(B) = P(A) + 2P(B \setminus A) + P(A \cap B)$ となる。ここで、$B \setminus A$ と $A \cap B$ は互いに排反であるから、$P(B \setminus A) + P(A \cap B) = P(B)$ となる。これを代入すると、$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ となる。 □

3.2.2 離散型確率変数と連続確率変数

確率変数はその性質で大きく二つに分けられる。確率変数の値域が高々加算である時、その確率変数を離散型確率変数といい、確率変数の値域が非加算である時、その確率変数を連続確率変数という。

3.2.3 確率分布

離散型確率変数の確率関数が満たす条件

確率変数とその確率から、確率分布が定まる。ある確率変数 X の値が x_k である確率が $f(x_k)$ となるとき、この X の値域から実数への写像 f を確率変数 X の確率分布と呼ぶ。

これを簡単に書くために、$f(x_k) = P(X = x_k)$ と書くことにする。つまり、$P(X = x_k)$ と書くことで、X が x_k となる確率を表すことにする。

また f は次のような性質をみたす。

1. $f(x_k) \geq 0$ $(k = 1, 2, \cdots)$
2. $\sum_k f(x_k) = 1$

連続型確率変数の確率関数が満たす条件

連続型確率変数に対しては、

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

なる f を確率変数 X の確率密度関数と呼ぶ。また、f は次のような性質をみたす。

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

3.2.4 累積分布関数

確率変数 X が、ある値 x 以下になる確率を表す関数を、X の累積分布関数と呼ぶ。離散型確率変数に対しては、

$$F(x) = P(X \leq x) = \sum_{x_k \leq x} f(x_k)$$

連続型確率変数に対しては、

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

として定まる F である。ここで、F は次のような性質をみたす。

1. $F(x)$ は広義単調増加
2. $F(x)$ は右連続
3. $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$

第一の性質は確率関数が非負であることから、二つ目は

$$\lim_{x \to a+} F(x) = \lim_{x \to a+} P(X \leq x) = P(X \leq a) = F(a)$$

となることからしたがう。三つ目は

$$P(X \leq -\infty) = 0$$

$$P(X \leq \infty) = P(\Omega) = 1$$

となることからしたがう。

4 積率、確率不等式

4.1 期待値

確率変数は実際に観測された値そのものではないが、(写像なので) 観測結果に対する計算を行える。例の一つが、期待値である。

4.1.1 離散型確率変数の期待値

定義 15. 離散型確率変数 X の期待値

$$E(X) = \sum_k x_k P(X = x_k)$$

例: サイコロの出る目は、それぞれ出る確率が 1/6 である離散型確率変数。その期待値は、

$$E(X) = \sum_{k=1}^{6} k\frac{1}{6} = \frac{7}{2}$$

4.1.2 連続型確率変数の期待値

定義 16. 連続型確率変数 X の期待値

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

例: 一様分布 $U(0,1)$ に従う確率変数 X の期待値この確率変数の確率密度関数は、

$$f(x) = \begin{cases} 1 & (0 \leq x \leq 1) \\ 0 & (otherwise) \end{cases}$$

(一様なので確率は一定 (定数関数)・そして積分した結果が 1 になるので、今回は $1/1 = 1$ になる) よって、

$$E(X) = \int_{-\infty}^{\infty} x1dx = \frac{1}{2}$$

4.1.3 期待値の性質

期待値は、総和・積分の線形性によって定義されていることからその線形性が保たれて、以下の性質をもつ。定数 $c \in \mathbb{R}$ と確率変数 X, Y に対して、

1. E(c) = c
2. E(X + c) = E(X) + c
3. E(cX) = cE(X)
4. E(X + Y) = E(X) + E(Y)

4.2 分散

実際の観測値に対する分散の定義は、偏差の二乗話の平均であった。これを確率変数に対しても定義する。$E(X) = \mu$ なる確率変数 X があったとき、$Y = X - \mu$ となる新たな確率変数を考える。これは確率変数が写像であることを思い出して正確に書けば、$Y(\omega) = X(\omega) - \mu$ ということである。さて。Y の期待値を考えると、期待値の線形性から、

$$E(Y) = E(X - \mu) = E(X) - E(\mu) = \mu - \mu = 0$$

ではいきなり二乗した場合、つまり、$Y = (X - \mu)^2$ とすると、

$$\begin{aligned} E(Y) &= E((X-\mu)^2) && (1) \\ &= E(X^2 - 2\mu X + \mu^2) && (2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 && (3) \\ &= E(X^2) - 2\mu^2 + \mu^2 && (4) \\ &= E(X^2) - \mu^2 && (5) \end{aligned}$$

となる。 確率変数に対してはこれが用いられる。これの計算には期待値のみを用いているので、離散型確率変数と連続型確率変数について同様の定義を用いることができる。標準偏差については、同様に分散の平方根を求めればよい。

定義 17. 確率変数 X の標準偏差

$$\sigma(X) = \sqrt{V(X)}$$

4.3 分散の性質

定数 $c \in \mathbb{R}$ と確率変数 X に対して、

1. $V(c) = 0$
2. $V(X + c) = V(X)$
3. $V(cX) = c^2 V(X)$

一つ目は、$E(X^2) = \mu^2$ から。
二つ目は、$Y = X + c$ としたとき、$E(Y) = E(X + c) = E(X) + c = \mu + c$ である。すると、
$V(Y) = E(Y^2) - (E(Y))^2$
ここで、

$$E(Y^2) = E((X + c)^2) = E(X^2 + 2cX + c^2) = E(X^2) + 2cE(X) + c^2 = E(X^2) + 2c\mu + c^2$$

$E(Y) = \mu + c$ なので、

$$(E(Y))^2 = (\mu + c)^2 = \mu^2 + 2c\mu + c^2$$

よって、

$$\begin{aligned} V(Y) &= E(Y^2) - (E(Y))^2 && (6) \\ &= E(X^2) + 2c\mu + c^2 - \left(\mu^2 + 2c\mu + c^2\right) && (7) \\ &= E(X^2) - \mu^2 && (8) \end{aligned}$$

よって、$V(Y) = V(X)$
ここで $V(c)$ は、正確には $Y(\omega) = c$ となるような Y を便宜的に書いたもの。

4.4 確率分布のかたちを表す指標

期待値や分散によって確率分布の大まかな性質がわかるが、分布の非対称性な部分はわからない。そこで、このような様子を表す指標をいくつか導入する。

4.4.1 歪度

定義 18. 確率変数 X の歪度

$$\alpha_3 = \frac{E\left((X - \mu)^3\right)}{\sigma^3}$$

これによってデータの歪み具合が評価できる。というのも、$(X-\mu)^3$ は 3 乗していることから期待値より正の部分や負の部分が大きく寄与する。したがって歪度が正の場合は、平均値よりも正の方向に裾野が広がっていることを意味し、負の場合は平均値よりも負の方向に裾野が広がっていることを意味する。

4.4.2 尖度

定義 19. 確率変数 X の尖度

$$\alpha_4 = \frac{E\left((X-\mu)^4\right)}{\sigma^4} - 3$$

この場合、4 乗していることから期待値付近より離れた値が大きく寄与する。したがって、尖度によって分布の尖り具合が評価できる。(2 乗や 6 乗ではない理由はわからないが、そもそも尖度を使っているのを見たことがないのでセーフ。) また、正規分布に対して第一項を検索すると 3 になるため、これを 0 とする補正の項として今回の定義のように第二項がつけられることがある。

4.5 モーメント

ここまで見てきたように、$E((X-\mu)^r)$ のかたちで表される数は確率分布の性質を反映しているものと考えられる。そこで、これを一般化したものをモーメントと呼ぶ。

定義 20. 確率変数 X の原点周りの r 次モーメント

$$\mu_r = E(X^r)$$

定義 21. 確率変数 X の期待値周りの r 次中心モーメント

$$\mu\prime_r = E((X-\mu)^r)$$

定義 22. 確率変数 X の標準モーメント

$$a_r = \frac{E((X-\mu)^r)}{\sigma^r}$$

ここまでの議論から分かるように、これらの値は確率分布の性質を反映しており、実は $\forall r \in \mathbb{N}$ で μ_r が一致する二つの確率分布は同じ分布である。(証明は知らず、むずかしそうなので省略。)

4.5.1 モーメント母関数

これら各 $r \in \mathbb{N}$ にたいして定義されているモーメントをいちどに表す便利な道具が、積率母関数である。

定義 23. 確率変数 X の積率母関数

$$M_X(t) = E(e^{tX})$$

つまり、新しい確率変数 $Y(\omega) = e^{tX(\omega)}$ を考えて、その期待値をとっている。したがって、その関数の具体的な計算は、離散型確率変数と連続型確率変数についてそれぞれ

$$\sum_{x\in\mathcal{X}} e^{tx} f(x)$$

$$\int_{-\infty}^{\infty} e^{tx} f(x) dx$$

となる。新しい合成写像を作ってもその確率は変わらないから、確率関数の部分は変化しないことに注意。さて、この関数が全てのモーメントを含んでいることを確認する。事実上同じ議論であるから、離散型確率変数について示す。e^{tX} をテイラー展開すると、

$$e^{tX} = \sum_{r=0}^{\infty} \frac{(tX)^r}{r!}$$

となる。これの期待値をとると、

$$E(e^{tX}) = \sum_{r=0}^{\infty} \frac{t^r}{r!} E(X^r)$$

つまり、

$$M_X(t) = 1 + \mu_1 t + \frac{\mu_2}{2!} t^2 + \frac{\mu_3}{3!} t^3 + \cdots$$

すると、

$$r \text{ 次モーメント} \leftrightarrow M_X(t) \text{ の } t \text{ についての } r \text{ 次導関数の } t = 0 \text{ の値}$$

と対応していることがわかる。

積率母関数を用いた期待値と分散の計算

積率母関数のみから、期待値と分散を求められる。まず、期待値は原点周りの一次モーメントだから直接計算すればいい。つぎに、分散は $V(X) = E(X^2) - E(X)^2$ の性質を使えば、2 回微分から二次モーメントを求め、それを用いて計算できる。

4.6 確率不等式

例えば共通テストなどは平均と分散のみが発表される。(たぶん) ここから点数が 30 点以上 40 点以下となる確率に対して評価を与えることはできるだろうか。テストの場合は得点の分布を仮定することである程度見積もれるかもしれないが、任意の確率分布についてこのような状況下でも評価を与えたくなる。この時有用なのが、種々の確率不等式である。

まず、原始的な確率不等式を紹介する。

4.6.1 マルコフの不等式

定義 24. マルコフの不等式

確率変数 X と $a > 0$ について

$$P(|X| \geq a) \leq \frac{E(|X|)}{a}$$

$a = kE(|X|)$ とすると、

$$P(|X| \geq kE(|X|)) \leq \frac{1}{k}$$

となり、期待値の値が k 倍になる確率が $\frac{1}{k}$ 以下であるという直感的な主張になる。この事実は、最も期待値の k 倍にするのが容易な状況、つまり確率変数が 0 もしくは適当な定数しか取らない場合を考えれば従うことがわかる。

4.6.2 チェビシェフの不等式

次に、期待値から一定の範囲内に存在する確率を計算する不等式についてみる。

定理 2. チェビシェフの不等式

確率変数 X について、$E(X) = \mu, V(X) = \sigma^2$ とする。この時、任意の $k > 0$ に対して、

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

証明. マルコフの不等式から示す。

$$P(|X| \leq a) = P(|X^2| \leq a)$$

より、マルコフの不等式から

$$P(|X| \geq a) \leq \frac{E(|X^2|)}{a^2}$$

となる。これが任意の確率変数 X について成立するので、とくに $Y = X - \mu, a = k\sigma(> 0)$ (σ は X の標準偏差) とすれば。

$$P(|X - \mu| \geq k\sigma) \leq \frac{E(|(X-\mu)^2|)}{k^2\sigma^2}$$

ここで、$E(|(X-\mu)^2|)$ について、これは X の分散の定義に他ならない。つまり

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

となる。 □

例題

例:

ある確率変数 X について、$E(X) = 1, V(X) = 1/3$ とする。$0 \leq X \leq 2$ となる確率を評価せよ。

解:

$P(|X - 1| \leq 1)$ を評価すれば良い。

チェビシェフの不等式から。$k > 0$ について

$$P(|X - 1| \geq k\sqrt{(}\frac{1}{3})) \leq \frac{1}{k^2}$$

$k = \sqrt{3}$ とすれば、

$$P(|X - 1| \geq 1) \leq \frac{1}{3}$$

よって、$P(|X - 1| \leq 1) = 1 - P(|X - 1| \geq 1) \underline{\geq \frac{2}{3}}$

4.6.3 コーシーシュワルツの不等式

定理 3. コーシーシュワルツの不等式

X, Y を確率変数とする。この時、

$$E(|XY|)^2 \leq E(X^2)E(Y^2)$$

これは、一般的なベクトル表記のコーシーシュワルツの不等式、

$$\boldsymbol{a} \cdot \boldsymbol{b} \leq |\boldsymbol{a}||\boldsymbol{b}|$$

の各ベクトルの成分を各確率変数の値と見て、両辺を $\frac{1}{n}^2$ で割ったものである。

4.6.4 ジェンセンの不等式

定理 4. ジェンセンの不等式

X を確率変数、f を (下) 凸関数とする。この時、

$$E(f(X)) \geq f(E(X))$$

証明. f は凸関数なので、$x = E(X)$ を通る接線 $g(x) = ax + b$ が存在し

- $x = E(X)$ の時、$f(x) = g(x)$
- $x \neq E(X)$ の時、$f(x) > g(x)$

となるから、

$$\begin{aligned} E(f(X)) &\geq E(g(X)) \\ &= E(aX + b) \\ &= aE(X) + b \\ &= g(E(X)) \\ &= f(E(X)) \end{aligned}$$

□

5 離散型確率分布

ここからは具体的に、有名な確率分布を扱う。まずは離散型確率分布を見ていく。

5.1 離散一様分布

$N \in \mathbb{N}$ 個全ての事象が等確率で起こる、つまり有限集合 $\Omega = \{x_1, x_2, \cdots, x_n\}$ の各元に対して確率が定まり、その確率関数が任意の事象 x_i に対して $P(X = k) = \frac{1}{n}, k = 1, 2, \cdots, N$ となる分布を離散一様分布という。(ちなみに、これを加算無限の濃度をもつ集合に拡張することはできない。例えば、自然数全体から一様ランダムに整数を取ってくることはできない。)

離散一様分布の性質をまとめる。

離散一様分布の性質

- $E(X) = \frac{N+1}{2}$
- $V(X) = \frac{N^2 - 1}{12}$

5.2 ベルヌーイ分布

試行結果が二種類の実験や観測を、同条件かつ独立に繰り返すことを、「ベルヌーイ試行」と呼ぶ。このベルヌーイ試行を一回行った結果を表す確率変数の分布をベルヌーイ分布という。(便宜的に、今後はひとつ目を「成功」、ふたつ目を「失敗」と呼ぶことにする。)

ベルヌーイ分布の性質

結果の一つ目を 1、二つ目を 0 として

$$P(X=1)=p, P(X=0)=1-p$$

だったとき期待値は

$$E(X)=p$$

分散は

$$V(X)=p(1-p)$$

5.3 二項分布

確率変数 X を、$X=$「n 回のベルヌーイ試行のうち、成功した回数」として定める。

この時、X の確率関数は、k 回成功する組み合わせが ${}_nC_k$ 通りあるので、

$$P(X=k)={}_nC_kp^k(1-p)^{n-k}$$

となる。これを二項分布といい、$Bi(n,p)$ と表す。 とくに $n=1$ の時はベルヌーイ分布となる。

二項分布の性質

まずは期待値を導出する。

$$
\begin{aligned}
E(x) &= \sum_{k=1}^{n} kP(x=k) \\
&= \sum_{k=1}^{n} k \cdot {}_{n}C_{k} \cdot p^{k} \cdot (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \cdot \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k)}{k!} \cdot p^{k} \cdot (1-p)^{n-k} \\
&= n \sum_{k=1}^{n} \frac{(n-1) \cdot \cdots \cdot (n-k)}{(k-1)!} \cdot p^{k} \cdot (1-p)^{n-k} \\
&= n \sum_{k=1}^{n} {}_{n-1}C_{k-1} \cdot p^{k} \cdot (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} {}_{n-1}C_{k-1} \cdot p^{k-1} (1-p)^{n-k} \\
&= np \sum_{k=0}^{n-1} {}_{n-1}C_{k} \cdot p^{k} \cdot (1-p)^{(n-1)-k} \\
&= np \sum^{n-1} \{1 + (1-p)\}^{n-1} \\
&= np
\end{aligned}
\tag{9}
$$