MA10_確率・統計1

確率とは二種類ある

・**頻度確率**
実験によって確かめられた確率。
頻度確率は**客観確率**ともいう。
例:「10本のうち一本だけ当 たりのクジを引いて当選する
確率を調べたところ10%で あった」という事実

※仮に実験の試行回数を増やしていくと、真の平均的な値に近づいてゆく(**大数の法則**)

・**ベイズ確率**
信念の度合いを数字で表したもの。
ベイズ確率は**主観確率**ともいう。
必ず信念の度合いを最大１００%とする。
信念の度合いを定めたのであれば、信念の度合い同士の演算が可能になる。

例：「あなたは40%の確率でインフルエンザです」
私がインフルエンザになる確率は４０%というのは、
ちゃんと実験したわけではないので、頻度確率になり得ない。
(⇒私が100人いるわけではない。私１００人いて、こういう症状だったら100にいたうちの40%なんて調べようがないので検証できない)
インフルエンザになりうる様々なファクターにより信念の度合いが高まる。
お医者さんの診断は、主観確率に過ぎない。

独立な事象の同時確率

お互いの発生には因果関係のない事象X=xと事象Y=yが、同時に 発生する確率

$$P(X = x, Y = y) = P(X = x)P(Y = y) = P(Y = y, X = x)$$

※順序は問わない。順序が逆でも同じ結果になる。

条件確率

ある事象X=x(雨が降る確率)が与えられた下で, Y=y(交通事故に遭う確率)となる確率のことを指す。
そして、XとYの積で表現することができる。
例:雨が降っている条件下で交通事故に遭う確率

$$P(Y = y \mid X = x) = \frac{P(Y=y,X=x)}{P(X=x)}$$

雨が降っている条件下で交通事故に遭う確率に対して雨が降る確率を掛け合わせることで、同時確率を算出することができるという関係性が重要

余談であるが、Xとxの違いについて
Xはx達全部を表す。
xはXの中で１つを表す。

MA11_確率・統計2

ベイズ則

ベイズ則は、同時確率と条件確率を用いることにより、導くことができる。
主観確率と客観確率のどちらでも使用することができる。

同時確率
$P(X = x, Y = y) = P(Y = y, X = x)$

条件付き確率により下記の式に変形できる
$P(X = x, Y = y) = P(X = x|Y = y)P(X = y)$ ・・・①
$P(Y = y, X = x) = P(Y = y|X = x)P(X = x)$ ・・・②

①と②を同時確率の式に代入するとベイズ則の式を導出できる

$P(X = x|Y = y)P(X = y) = P(Y = y|X = x)P(X = x)$
$P(X = x|Y = y) = \frac{P(Y=y|X=x)P(X=x)}{P(X=y)}$

MA12_確率・統計3

確率変数

事象と結びつけられた数値
事象そのものを表すといっても良い

確率分布

各事象の確率が同じことがほぼなくて、確率は事象ごとによってバラバラ。
様々な事象に対する確率を表にまとめると確率がどういう分布をしているかがわかる。
事象の発生する確率の分布のことを**確率分布**という。(ある確率変数を取る分布が確率分布)
離散値であると表で表すことができる。

確率分布の特徴を調べるには？

確率分布を見ただけでは、特徴がわからない。
特徴を見出す例として**平均**がある。(分散、共分散については後述)

MA13_確率・統計4

期待値(平均)

その分布における、確率変数の「平均の値」or「ありえそうな値」
もう少し言うと確率的に出てくる値の平均値を考えるときには、期待値と呼ばれる。
期待値はある確率変数の変数と確率変数で表され、下記の式になる。

事象: X
確率変数: $f(X = x_{k=n})$
確率: $P(X = x_{k=n})$

$E(f) = \Sigma_{k=1}^{n} P(X = x_{k=1})f(X = x_{k=1})$

連続値の場合は、積分すること。

MA14_確率・統計5

期待値（平均）だけだとわからないことが多い。
確率分布がものすごくバラバラな場合、期待値（平均値）通りにならない。
どれだけ、期待値通りになるかが**分散**と言う考え。

分散

データのちりばり具合。期待値はデータからどれだけずれているか平均化したもの。
なんで、平均化するか？
サンプル数が少ない場合はいいかもしれない。
サンプル数が多いとズレが大きいかもしれない。（例えば５００万人もいればズレそう）

分散の単位は、元の単位の二乗で表す。
絶対値の考えもあったが場合分けが必要になるため、二乗という考えを使っている。

分散 : Var $f(x)$ ＝ $E(f_{(X=x)} - E(f))^2$

1つのデータのバラツキに関しては、分散という考えを用いた。
複数のデータのバラツキを考える場合に用いるのが、**共分散**。

共分散

2つのデータのバラツキとは何かというと、2つのデータ系列の違いを示すことと同義。
⇨2つのデータ(X、Y)の散らばり方の違いをみる。傾向の違いを知る。

正の値⇨似た傾向。似てるかも。関係性があるかも
負の値⇨逆の傾向。似てないかも。無関係という意味でない。一方が増え、一方が減る関係。
ゼロ　⇨関係性が乏しい。
必ずわかるわけではないが、二つのデータの関係を知るために使用する。

共分散 : $Cov(f, g)$ ＝ $E(f_{(X=x)} - E(f))E(g_{(Y=y)} - E(g))$

MA15_確率・統計6-1

確率分布によってできることが変わってくる。
分布によって中身が変わる。一方が増えて、一方がへるなど。(確率分布がちぐはぐだったり)
分散・共分散を出すのもいいけど、本当に出せるかわからない。難しいかもしれない
数値計算できるかわからない。ただ単に数値計算でしか出せないのか?
だから、事前に確率分布してるかを知りたい。

様々ある確率分布で代表的な二つの分布を紹介する。

ベルヌーイ分布
確率変数が 0 と 1の２値の場合に使う(現象を二つに分けた場合など)
頑張ればベルヌーイ分布の組み合わせで全て表すことができる

・P(X=0)=1－μ
・P(X=1))=μ
と表すと下記の式がベルヌーイ分布の式である。ただし、
$P_{(X=x)} = \mu^x(1-\mu)^{(1-x)}$

マルチヌーイ分布
→ 確率変数が3つ以上
「カテゴリカル分布」という呼ばれるのが一般的。
サイコロを転がすイメージ

$$P(X=n) = \prod_{k=1}^{n} \lambda_k^{[x=k]}$$

λの[x=k]は条件式

MA16_確率・統計6-2

二項分布
さっきまでは、一回やった時の話だった。N回やった場合の話に入る。
N回やった場合、それが二項分布。μに相当する部分がλで表現されている。
ベルヌーイ分布の多試行版であり、式は下記である。

$$P(x \mid \lambda, n) = \frac{n!}{x!(n-x)} \lambda^x (1-\lambda)^{(n-x)}$$

組み合わせの式が入っているのは、、、
例えば、表が三回といっても順序まではいわれていない。
そういった場合、表が三回という事象が増える。
なので、順序によって事象が増えないことを考慮するために組み合わせの式が入っている。

ガウス分布(正規分布)
今までは離散的な値を考えてきた。ガウス分布は、釣鐘型の連続分布である。
中心が大きく(平均が大きく)、どんどん小さくなるグラフ。

基本的な式の形式としては、$e^{(-x^2)}$ で表現される。
また、確率が1になるように平均・分散を用いたパラメータが付与されている。

$$P(x\ ;\mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(\frac{1}{2\sigma^2}(x-\mu)^2\right)$$