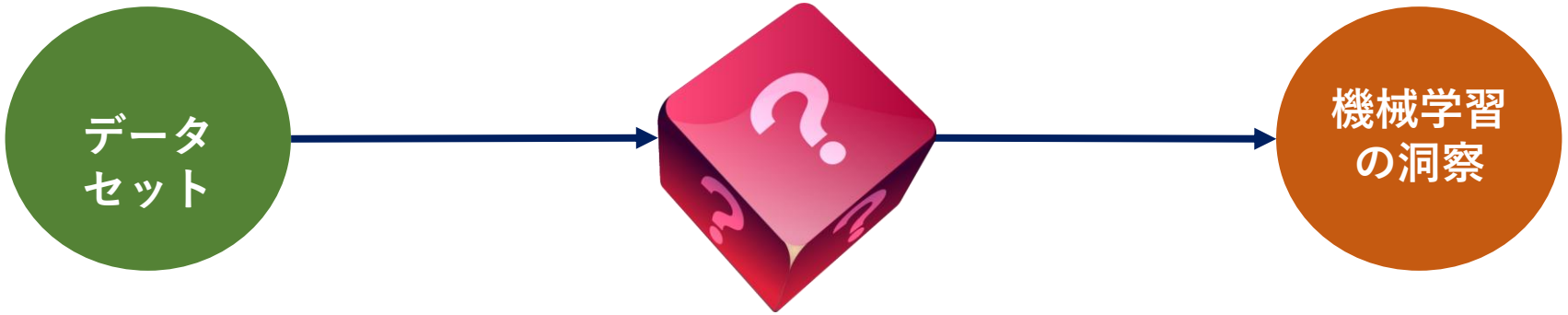
365 DataScience

統計学とデータサイエンスのための確率

確率の公式 | 標本空間 | 期待値 | 補集合

はじめに



統計学と機械学習の世界に飛び込む前に、基本的な確率のルールについて理解をしておく必要があります。 統計学において得られる多くの指標についてしっかりと理解をすることが、データサイエンスの知見を高める上では不可欠です。そして、確率についての知識を学ぶことで、統計学への橋渡しが可能となり、その結果としてデータサイエンスを学ぶ中で多くの知見を得ることができるようになるのです。

ベイズ推定 は数学の世界で広く使われており、複雑な事象を表現することに適しています。 ベイズ推定に関連した表記法を使うことによって、要素、集合、事象の関係を理解することができるようになります。そして、こうした関係を理解することが、データサイエンスにおいて知見を得る上で重要になってきます。

分布 はデータを分類する中心的な方法です。データセットがある分布に従っていると仮定するのであれば、それに基づいて確率を求めることができるようになります。そして、多くの分布には確率と起こる結果の間に美しい法則がありますので、データの特徴を把握することは非常に有用ということができます

確率とは

確率とは、何らかの事象の起こりやすさを示すものです。

そして、その公式は以下の通りです。

$$P(X)=\frac{\text{対象とする結果}}{\text{標本空間}}$$

確率の公式:

- 事象Xが起こる確率は、標本空間全てにおける、対象とする結果が起こる数で表現することができます。
- 対象とする結果というのは、私たちが起こると想定している事象のことを言います。
- 標本空間とは、起こり得る全ての結果のことを言います。

二つの事象が独立の場合:

二つの事象が独立の場合の確率は以下の式で計算することができます

P(A♥) = P(A) . P(♥)

期待値

試行 – ある単一結果

実験 – 試行を複数回行ったもの

経験的確率 – 実際行った経験に基づいて計算した確率

期待値 – 実験を行った際に想定される結果

具体例: 試行

コインを投げて結果を記録する

具体例: 実験

コインを20回投げてそれぞれの結果を記録する

経験的確率とは、20回実際にコインを投げて表が出た回数を20で割ったものである。

期待値とは、ある実験などを行った際に期待される結果のことを言う

カテゴリ変数の期待値

$$E(X) = n \times p$$

量的変数の期待値

$$E(X) = \sum_{i=1}^{n} x_i \times p_i$$

確率頻度分布

確率頻度分布とは
それぞれの事象の起こる確率をまとめたもの

なぜ頻度分布が必要なのか？
期待値が得られない場合に起こり得る確率を予想したいため

頻度とは?
頻度とは、ある標本空間において事象が発生する回数のことである

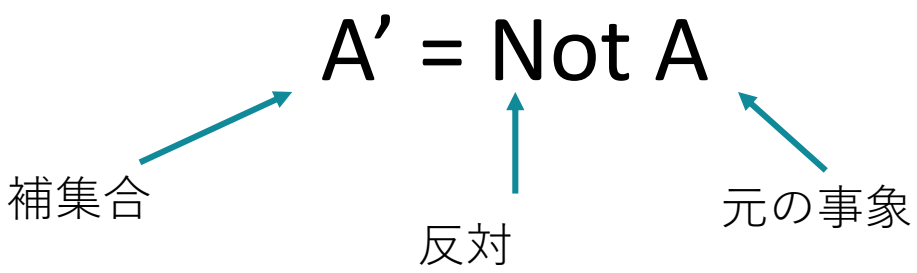
頻度分布表とは？
それぞれの結果と頻度をまとめて整理した表のことである

確率頻度分布から頻度分布表を作るには？
それぞれの頻度を標本空間の数で割ることによって求めることができる

Sum	Frequency	Probability
2	1	1/36
3	2	1/18
4	3	1/12
5	4	1/9
6	5	5/36
7	6	1/6
8	5	5/36
9	4	1/9
10	3	1/12
11	2	1/18
12	1	1/36

補集合

事象の補集合とは、対象とする事象以外のことである



補集合の特徴:

- 同時には起こらない
- 元の事象を足すと標本空間となる (A + A’ = **標本空間**)
- 足すと確率は1となる (P(A) + P(A’) = 1)
- 補集合の補集合は元の集合となる ((A’)’ = A)

具体例:

- 事象Aをスペードを引くこととするとP(A) = 0.25.
- A’ はスペードを引かないことなのでクラブかダイヤかハートを選ぶことであり、P(A’) = 1 – P(A)で P(A’) = 0.75となる