確率モデル入門
確率の用語整理、確率変数、確率分布

nepia

2020/05/30

講義（セミナー前半）４０～６０分

講義（セミナー前半）４０～６０分

目標

確率がわかり、確率モデルが身近になること

講義（セミナー前半）４０～６０分

目標

確率がわかり、確率モデルが身近になること

アジェンダ

- 試行
- 標本空間と事象
- 確率変数
- 確率分布
  - 離散確率分布
  - 連続確率分布

全体の構成

- なぜ「確率」が必要なのか？
- 確率の用語整理
- 確率変数、確率分布
- 離散確率分布、連続確率分布
- Python で例を見る
  - 離散確率分布
  - 連続確率分布
  - 観測データの分布への収束
- まとめ
- ハンズオン

なぜ「確率」が必要なのか？（１）

なぜ「確率」が必要なのか？（１）

- 「記述統計」ならば「確率」は不要でした

なぜ「確率」が必要なのか？（１）

- 「記述統計」ならば「確率」は不要でした
  - ヒストグラムを書いてデータを概観する
  - 平均値、中央値を算出してデータの中心傾向を知る
  - 分散を算出してデータの散らばり具合を知る

なぜ「確率」が必要なのか？（1）

- 「記述統計」ならば「確率」は不要でした
  - ヒストグラムを書いてデータを概観する
  - 平均値、中央値を算出してデータの中心傾向を知る
  - 分散を算出してデータの散らばり具合を知る

これだけならば、確率の知識は不要！

なぜ「確率」が必要なのか？（２）

- しかし踏み込んで "背後の法則" を考えたい

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか

なぜ「確率」が必要なのか？（２）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか
    「サンプル数が多いほど信頼できる」
    とは言うけれど？

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか
    「サンプル数が多いほど信頼できる」
    とは言うけれど？
- 「統計モデリング」では

なぜ「確率」が必要なのか？（２）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか
    「サンプル数が多いほど信頼できる」
    とは言うけれど？
- 「統計モデリング」では

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか
    「サンプル数が多いほど信頼できる」
    とは言うけれど？
- 「統計モデリング」では
  - 実世界のデータから筋のよい「確率モデル」を記述したい
  - そのモデルがどれくらい本当なのかを議論したい

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか
    「サンプル数が多いほど信頼できる」
    とは言うけれど？
- 「統計モデリング」では
  - 実世界のデータから筋のよい「確率モデル」を記述したい
  - そのモデルがどれくらい本当なのかを議論したい

データがまず与えられる。

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
正規分布だとして
・平均はいくつなのか
・その平均はどれくらい信頼できるのか
「サンプル数が多いほど信頼できる」
とは言うけれど？
- 「統計モデリング」では
  - 実世界のデータから筋のよい「確率モデル」を記述したい
  - そのモデルがどれくらい本当なのかを議論したい

データがまず与えられる。
その背後にある "原則" "真理" を考える。

なぜ「確率」が必要なのか？（2）

- しかし踏み込んで "背後の法則" を考えたい
- 「統計モデル」になぜ興味があるのか
  - データそのものではなく、
    その背後にある "原則" "真理" が知りたいから
  - しかも"定量的"に知りたい
    正規分布だとして
    ・平均はいくつなのか
    ・その平均はどれくらい信頼できるのか
    「サンプル数が多いほど信頼できる」
    とは言うけれど？
- 「統計モデリング」では
  - 実世界のデータから筋のよい「確率モデル」を記述したい
  - そのモデルがどれくらい本当なのかを議論したい

データがまず与えられる。
その背後にある "原則" "真理" を考える。
→ 考える材料として確率の知識が必要！

まずは確率の用語整理から

※かいつまんで紹介

まずは確率の用語整理から

※かいつまんで紹介

> 試行 (trial)

実験や観測などを行うことです。確率モデルに基づいて論理展開を行っていくにあたっては、試行を行った結果を確率的に解釈していきます。

まずは確率の用語整理から

※かいつまんで紹介

> 試行 (trial)

実験や観測などを行うことです。確率モデルに基づいて論理展開を行っていくにあたっては、試行を行った結果を確率的に解釈していきます。

> 標本空間 (sample space)

試行の結果を要素とする集合です。

まずは確率の用語整理から

※かいつまんで紹介

> 試行 (trial)

実験や観測などを行うことです。確率モデルに基づいて論理展開を行っていくにあたっては、試行を行った結果を確率的に解釈していきます。

> 標本空間 (sample space)

試行の結果を要素とする集合です。

> 事象 (event)

標本空間の部分集合です。

確率変数とは？

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません
    それは「試行」です

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません
    それは「試行」です
  - 確率変数は「起こりうる事柄」ではありません

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません
    それは「試行」です
  - 確率変数は「起こりうる事柄」ではありません
    それは「事象」です

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません
    それは「試行」です
  - 確率変数は「起こりうる事柄」ではありません
    それは「事象」です
  - 確率変数は「取りうる値を確率で表したもの」ではありません

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は○○ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません
    それは「試行」です
  - 確率変数は「起こりうる事柄」ではありません
    それは「事象」です
  - 確率変数は「取りうる値を確率で表したもの」ではありません
    それは「確率分布」です

確率変数とは？

確率変数とは、値が確率である変数です。
ランダムに値を取るのが特徴です。

- （注意）確率変数は〇〇ではありません
  - 確率変数は「とる値の集合」ではありません
    それは「標本空間」です
  - 確率変数は「具体的な実現値」ではありません
    それは「試行の結果」です
  - 確率変数は「サイコロを振ること」でもありません
    それは「試行」です
  - 確率変数は「起こりうる事柄」ではありません
    それは「事象」です
  - 確率変数は「取りうる値を確率で表したもの」ではありません
    それは「確率分布」です

（難しい）

余談：ランダムは実はとってもむずかしい

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
すべての事象が等確率で起こるとして、
場合の数で考えられていた

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
すべての事象が等確率で起こるとして、
場合の数で考えられていた
- 確率 ＝ 求める事象/全事象
（ラプラスによる定義 古典的確率）

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
  すべての事象が等確率で起こるとして、
  場合の数で考えられていた
- 確率 ＝ 求める事象/全事象
  （ラプラスによる定義 古典的確率）
- できないことも多かった
  （すべての事象を数え上げるのがの無理なケースなど）

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
  すべての事象が等確率で起こるとして、
  場合の数で考えられていた
- 確率 ＝ 求める事象/全事象
  （ラプラスによる定義 古典的確率）
- できないことも多かった
  （すべての事象を数え上げるのがの無理なケースなど）
- 20 世紀初頭

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
  すべての事象が等確率で起こるとして、
  場合の数で考えられていた
- 確率 ＝ 求める事象/全事象
  （ラプラスによる定義 古典的確率）
- できないことも多かった
  （すべての事象を数え上げるのがの無理なケースなど）
- 20 世紀初頭
  アンドレイ・コルモゴロフというロシアの数学者が、
  確率論の公理化をみごと完成させた！
  （公理的確率論）

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
  すべての事象が等確率で起こるとして、
  場合の数で考えられていた
- 確率 ＝ 求める事象/全事象
  （ラプラスによる定義 古典的確率）
- できないことも多かった
  （すべての事象を数え上げるのがの無理なケースなど）
- 20 世紀初頭
  アンドレイ・コルモゴロフというロシアの数学者が、
  確率論の公理化をみごと完成させた！
  （公理的確率論）
- ただし、集合論・測度論・ルベーグ積分を駆使して……

余談：ランダムは実はとってもむずかしい

- 長年、確率のことについては、
  すべての事象が等確率で起こるとして、
  場合の数で考えられていた
- 確率 ＝ 求める事象/全事象
  （ラプラスによる定義 古典的確率）
- できないことも多かった
  （すべての事象を数え上げるのがの無理なケースなど）
- 20 世紀初頭
  アンドレイ・コルモゴロフというロシアの数学者が、
  確率論の公理化をみごと完成させた！
  （公理的確率論）
- ただし、集合論・測度論・ルベーグ積分を駆使して……

ランダムはむずかしい、確率変数もむずかしいので、深く考えないことがオススメです。確率変数のことは「ランダムさを含んだ変数」ぐらいにしておきます。

確率分布とは？

確率分布とは？

- 「各々の値をとる確率」を表す分布
  - 例： コインを投げたとき……
  - コインが表を取る確率： 50%
  - コインが裏を取る確率： 50%

確率分布とは？

- 「各々の値をとる確率」を表す分布
  - 例： コインを投げたとき……
  - コインが表を取る確率： 50%
  - コインが裏を取る確率： 50%
- ヒストグラムをサンプル数で割ったものと「似ています」
  - じつのところ「そうではない」のですが
  - 無限にデータ数があるなら一致していきます

離散確率分布、連続確率分布

離散確率分布、連続確率分布

- 離散確率分布とは？

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布
- 連続確率分布とは？

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布
- 連続確率分布とは？
  - 確率変数 X が連続値をとる場合の確率分布です

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布
- 連続確率分布とは？
  - 確率変数 X が連続値をとる場合の確率分布です
  - 例）花弁の長さ、16 歳男子の身長

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布
- 連続確率分布とは？
  - 確率変数 X が連続値をとる場合の確率分布です
  - 例）花弁の長さ、16 歳男子の身長
  - 正規分布、ポアソン分布、連続一様分布

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布
- 連続確率分布とは？
  - 確率変数 X が連続値をとる場合の確率分布です
  - 例）花弁の長さ、16 歳男子の身長
  - 正規分布、ポアソン分布、連続一様分布

離散ならとりうる値ごとに確率が出せますが、
連続ではそうはいきません。

離散確率分布、連続確率分布

- 離散確率分布とは？
  - 確率変数 X が離散値をとる場合の確率分布です
  - 例）コインの表裏、サイコロの出目
  - 離散一様分布、二項分布
- 連続確率分布とは？
  - 確率変数 X が連続値をとる場合の確率分布です
  - 例）花弁の長さ、16 歳男子の身長
  - 正規分布、ポアソン分布、連続一様分布

離散ならとりうる値ごとに確率が出せますが、
連続ではそうはいきません。
「身長が 170cm ジャスト」とはいえないからです。

コードで色々見てみましょう

コードで色々見てみましょう

- Python で動かしてみましょう
- 実際に動いている様子で理解のヒントになるかも

離散確率分布の例：サイコロ

連続確率分布の例：正規分布

サンプルを増やすと真の分布に漸近する

講義のまとめ

講義のまとめ

- データが与えられたら、背後の法則を考えたい

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要
- しかし、確率は基礎からしてむずかしく混乱しやすい
  一発で理解しようとするのは大変……

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要
- しかし、確率は基礎からしてむずかしく混乱しやすい
  一発で理解しようとするのは大変……

ではどうすれば？

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要
- しかし、確率は基礎からしてむずかしく混乱しやすい
  一発で理解しようとするのは大変……

ではどうすれば？

- ゆっくり丁寧にやることで理解する

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要
- しかし、確率は基礎からしてむずかしく混乱しやすい
  一発で理解しようとするのは大変……

ではどうすれば？

- ゆっくり丁寧にやることで理解する
- Python で動かすことで理解する

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要
- しかし、確率は基礎からしてむずかしく混乱しやすい
  一発で理解しようとするのは大変……

ではどうすれば？

- ゆっくり丁寧にやることで理解する
- Python で動かすことで理解する
  - 紙とペンでやるよりも
  - Python のほうがカンタンです

講義のまとめ

- データが与えられたら、背後の法則を考えたい
- 法則を考えるために「確率モデル」が利用される
- 考えるには確率への理解がどうしても必要
- しかし、確率は基礎からしてむずかしく混乱しやすい
  一発で理解しようとするのは大変……

ではどうすれば？

- ゆっくり丁寧にやることで理解する
- Python で動かすことで理解する
  - 紙とペンでやるよりも
  - Python のほうがカンタンです

講義は以上です。少し休憩をしたのち、ハンズオンに入ります。

ハンズオン

ハンズオン

- sklearn.dataset を読み込む

ハンズオン

- sklearn.dataset を読み込む
- 適当なヒストグラムを書く

ハンズオン

- sklearn.dataset を読み込む
- 適当なヒストグラムを書く
- 何の数学的分布に近いか見てみる

ハンズオン

- sklearn.dataset を読み込む
- 適当なヒストグラムを書く
- 何の数学的分布に近いか見てみる

その他、素朴な疑問について
確率に限らず拾っていく時間とします