回帰分析I

6. 確率変数：期待値、分散、標準偏差

確率変数

- 回帰分析では、変数を「**確率変数**」として扱います。
- なので、回帰分析の（基本的な）理論面を理解するためには、確率変数についての理解が必要となります。
- 確率変数とは？
- 難しい言葉を使って言えば、

「確率変数は標本空間から実数への<u>関数</u>」

- 例を使って、この意味を分かりやすくしてみましょう。
- サイコロを投げて出た目によって賞金をもらうゲーム、を考えます。

- サイコロを投げて出た目を50倍した賞金がもらえるゲーム、を考えましょう。
- サイコロ投げ：標本空間は$S=\{1,2,3,4,5,6\}$
- 確率は$P(\{1\})=P(\{2\})\cdots=P(\{6\})=1/6$。
- 出た目をωで表す。賞金を$X(\omega)$と表す。
- $X(\omega)$は「サイコロの目が決まれば、貰える賞金も決まるよ」ということ。

出た目:標本空間の要素ω	1	2	3	4	5	6
賞金：確率変数の値$X(\omega)$	50	100	150	200	250	300

- この$X(\omega)$は確率変数の例。
- 確率変数は関数っていうとぴんと来ないけど、サイコロの目に応じて賞金を決定するという意味で関数。

- この確率変数X（賞金）のとりうる値とその確率の対応関係（**確率分布**）は以下の通り。

$$P(X=50)=\frac{1}{6}$$
$$P(X=100)=\frac{1}{6}$$
$$\ldots\ldots$$
$$P(X=300)=\frac{1}{6}$$

- その背後には標本空間（サイコロの目）がある。

$$P(X=50)=P(\{1\})=\frac{1}{6}$$
$$P(X=100)=P(\{2\})=\frac{1}{6}$$
$$\ldots\ldots$$
$$P(X=300)=P(\{6\})=\frac{1}{6}$$

- けれどX（賞金）が興味の対象なら、標本空間（サイコロの目）のことは特に意識せず、Xの確率分布だけを考えれば十分。

確率変数の例

- 例:コインを二枚投げる実験。標本空間は

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

で、$P(\{H,H\}) = P(\{H,T\}) = P(\{T,H\}) = P(\{T,T\}) = 1/4$

- Xを「表の枚数」として定義される確率変数とする。

一枚目・二枚目:標本空間の要素ω	*H,H*	*H,T*	*T,H*	*T,T*
表の枚数:確率変数の値$X(\omega)$	**2**	**1**	**1**	**0**

このとき確率変数の Xの分布は

$$P(X=0) = P(\{T,T\}) = 1/4$$
$$P(X=1) = P(\{H,T\}) + P(\{T,H\}) = 1/4 + 1/4 = 1/2$$
$$P(X=2) = P(\{H,H\}) = 1/4$$

離散型確率変数

- 標本空間と確率変数の対応はなんとなく理解できたと思います。
- これ以降は、<u>興味の対象が確率変数のときには、標本空間は特に意識せずに確率変数に直接フォーカスしていくことにします。</u>
  - 回帰分析では、興味の対象は確率変数。
  - なので、ほとんどの場合、おおもとの標本空間を意識しなくてOK。
- これまで見てきた確率変数の例は、起こりうる候補の値は有限個。
  - 最初の例では6個、次の例では3個
- このように起こりうる候補の値が有限個のとき、その確率変数は「**離散型**」と呼ばれます（**離散型確率変数**）。

離散型確率変数

- 離散型確率変数　X
- k個の起こりうる候補の値（実現値）$\{x_1, \cdots, x_k\}$
  - x_j $(j = 1, \cdots, k)$は一般的に書いているだけで、実際は何らかの具体的な数。
- $X = x_j$が起こる確率を$Pr(X = x_j)$と表す。
- 実現値$\{x_1, \cdots, x_k\}$とその確率$Pr(X = x_j)$の対応関係を

$$Pr(X = x_j) = f(x_j), \quad j = 1, \cdots, k$$

で表したとき、$f(x_j)$を<u>**確率質量関数**</u>（probability mass function）と言う。

- $f(x_j)$ $(j = 1, \cdots, k)$は確率なので、以下の性質を満たす。

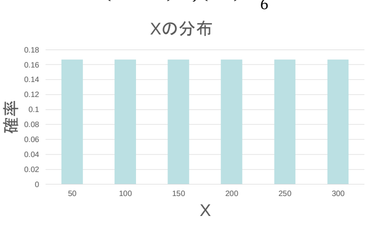
$$Pr(X = x_j) = f(x_j) \geq 0, (j = 1, \cdots, k)\text{: 確率は非負}$$

$$\sum_{j=1}^{k} Pr(X = x_j) = \sum_{j=1}^{k} f(x_j) = 1\text{:確率の和は1}$$

離散型確率変数

- 最初の例では、確率質量関数は、

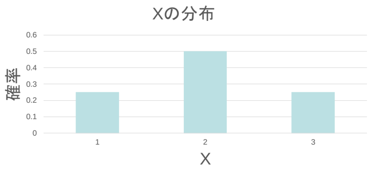
$$P(X = 50) = f(50) = \frac{1}{6}$$
$$P(X = 100) = f(100) = \frac{1}{6}$$
$$\cdots\cdots$$
$$P(X = 300) = f(300) = \frac{1}{6}$$



離散型確率変数

- 次の例では、確率質量関数は、

$$P(X=0)=f(0)=1/4$$
$$P(X=1)=f(1)=1/2$$
$$P(X=2)=f(2)=1/4$$



- ということで、確率質量関数って名前はごついけど、別に大したことはない。x_j (起こりうる値) $\underset{f}{\rightarrow}$ $f(x_j)$（それが起こる確率）ってだけのことです。

離散型確率変数の例1: ベルヌーイ確率変数

- 結果が“成功”（success） か“失敗”（failure）であるような試行を考えよう。
- 成功なら1の値、失敗なら0の値を取る確率変数Xを考える。
- 成功の確率を0.3とすると、Xの確率質量関数は

$$P(X=0)=f(0)=0.7$$
$$P(X=1)=f(1)=0.3$$

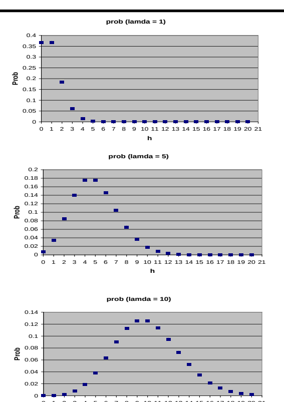
- このタイプの確率変数は**ベルヌーイ確率変数**と呼ばれる。
- 「成功」、「失敗」という言葉を使っているが、そうである必要は特にない。
- 応用例で特に多いのは、何かが「起きる」「起きない」とか、人または企業が「する」「しない」とかいうケース。
  - このタイプの従属変数をモデル化したいことはよくある。
  - そういう場合の簡単なモデル化の仕方については、本講義で扱う予定。

離散型確率変数の例2: ポアソン確率変数

- 0, 1, 2, 3, … のいずれかの値をとる確率変数Xが次の確率質量関数を持つとする:

$$p(i)=P\{X=i\}=\exp(-\lambda)\frac{\lambda^i}{i!},\ i=0,1,2,\ldots$$

- このときXはパラメーターλを持つ**ポアソン確率変数**と呼ばれる。
- 注: exp(A) = e^A てこと。e^Aの別の書き方。計量分析ではこの表記の仕方が良く使われる。
- このコースでは扱わないが、この確率分布は従属変数が<u>非負の整数のみ</u>を取る場合のモデリングに使われる。例えば、、、
  - 市で発生する一年間の殺人事件の数
  - 消費者が一年間に行く旅行の回数
  - 国がオリンピックで獲得する金メダルの数



連続型確率変数

- 離散型確率変数は、起こりうる候補の数が有限個（か可算）。
- 一方で、起こりうる数が連続した数である場合もある。
- 例えば、新車が故障するまでの時間（単位は秒）。
- 例えば、これを確率変数Xとすると、起こりうる候補の値は0より大きい実数すべて。
- これは離散型のときのように番号付けできない。
- 1秒、2秒、って考えるかもしれないけど、0.1秒、0.2秒もあるし、0.01秒、0.02秒もあるし、0.001秒、0.002秒も、、、以下略。

13

連続型確率変数

- このように実現値が連続的な数の集合で与えられる確率変数を「**連続型確率変数**」と呼びます。
  - 本講義で扱う従属変数はこのタイプの確率変数（が多い）。
- 厳密には連続型でないものも、連続型として分析することがよくある。
  - テストのスコア　ー　取りうる値は非負の整数で一点刻みが普通。70.928546…とかいうスコアはなくても連続型と見なして分析することが一般的。
  - 所得　ー　取りうる値は非負の整数で一円刻みが普通。お金を銭や厘の単位でもらうことはないけれども連続型と見なして分析することが一般的。
- ただし、市で発生する一年間の殺人事件の数、消費者が一年間に行く旅行の回数のように、取る値が0、1、2、3、4、、、（最大値はあまり大きくない正数）みたいなときは、連続型と見なすのはあまり良くないかも。
  - そういう従属変数を適切にモデル化したい場合は、このコースのレベルを超えた計量分析の学習が必要。

14

連続型確率変数

- すべての実数 $x \in (-\infty, \infty)$ について、次の性質を持つ非負の関数$f(x)$が存在するとき、確率変数Xは**連続型確率変数**という:

  実数のいかなる集合Bついて、

$$\boxed{P(X \in B) = \int_B f(x)dx}$$

- この関数$f(x)$を連続型確率変数Xの**確率密度関数**（**probability density function**）という。

15

連続型確率変数

$$P(X \in B) = \int_B f(x)dx$$

- この式の意味は、

「確率変数Xが集合Bの中の値をとる確率は、確率密度関数$f(x)$を集合B上で積分することで得られる。」

- 例えば、、、

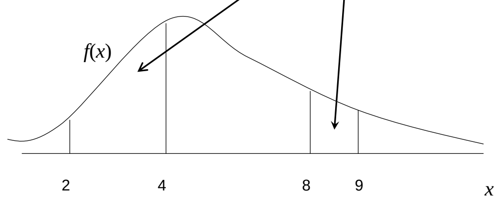
16

B = [2, 4] ∪ [8,9]

$$P(X \in B) = \int_B f(x)dx$$

$P(X \in B)$はこれらの面積の和

確率変数Xが2以上4以下または
8以上9以下の値を取る確率



連続型確率変数

- Xの確率に関するいかなることにも、確率密度関数$f(x)$を使って答えることが出来る。
- 例えば、$B = [a, b]$ なら

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

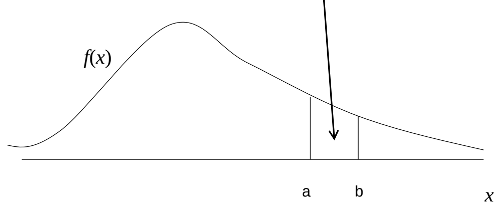
- $B = [a, \infty)$ なら

$$P(a \le X) = \int_a^\infty f(x)dx$$

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

この面積

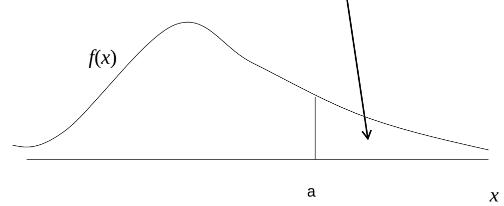
確率変数Xが
a以上b以下の値を取る確率



$$P(X \ge a) = \int_a^\infty f(x)dx$$

この面積

確率変数Xがa以上の値を取る確率



- もし$B = [a, a]$とすると

$$P\{X = a\} = \int_a^a f(x)dx = 0$$

- これは、

  連続確率変数が任意の実数をとる確率がゼロである

  ことを示します。

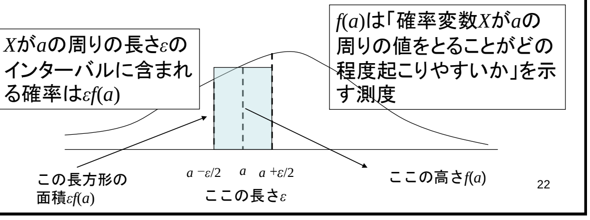
- 確率がゼロって、なんか変な感じがするけれども、連続型は実現値が無限に存在するので、個々の実現値にピンポイントで確率を与えることはできず区間の確率を考える、ぐらいの理解でいいです。

- それでは確率密度関数って？

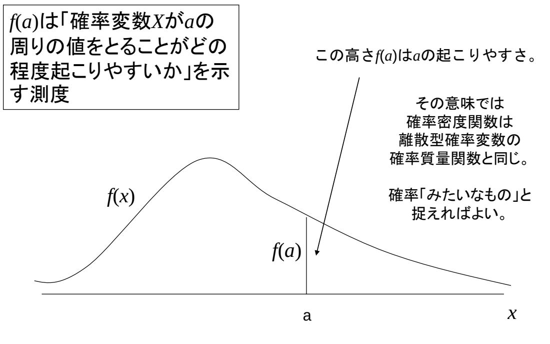
確率密度関数

- 直感的に理解するための説明は以下の通り。
- ε が小さいとき、$B = [a - \varepsilon/2, a + \varepsilon/2]$ の確率は

$$P\{a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)dx \approx \varepsilon f(a)$$



確率密度関数

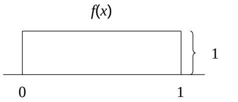


連続型確率変数の例

1. 一様確率変数

- 確率変数Xの確率密度関数が

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



  で与えられるとき、Xはインターバル$(0,1)$に一様に分布しているという。

- $0 < a < b < 1$のとき

$$P\{a \le X \le b\} = \int_a^b f(x)dx = b - a$$

2. 正規確率変数

- 確率変数Xの確率密度関数が

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\quad -\infty<x<\infty$$

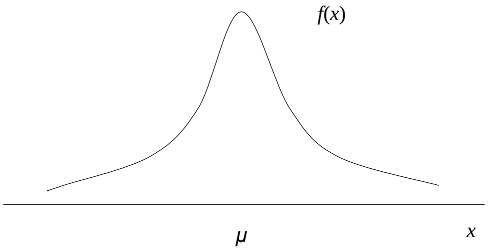
のとき、Xは平均が μ で分散がσ^2 の**正規確率変数**という。

- または、Xは平均が μ で分散が σ^2 の**正規分布**（Normal Distribution）に従うという。

- 次のように表すこともある: $X\sim N(\mu,\sigma^2)$

- 注:平均、分散は後で定義します。

- 確率密度関数関数は、μ を中心に対称なベル型（bell-shaped）のカーブ:



確率変数の期待値

- データの特徴を要約するもの:標本平均や標本分散

- 確率変数の挙動を要約するもの:平均（期待値ともいう）や分散

- これらの関係についてはいずれ詳しく説明しますが、今の時点では
  - 標本平均はデータを使って確率変数の平均を推定したもの
  - 標本分散はデータを使って確率変数の分散を推定したもの

ぐらいに思っておいてください。

- 以下では平均ではなく期待値という言葉を使います。
  - 「確率変数の平均」と「確率変数の期待値」は全く同じ意味。
  - ただ標本平均とはいいますが標本期待値とはいいません。

確率変数の期待値

- 確率変数X

- 事前にどれぐらいの値が出ると「期待」できる？
  - （ある意味での）代表的な値

- これが期待値。

- ここでは離散型の場合を考え、取りうる値の候補を $x_1, x_2, \dots, x_k$とする。

- このときXの期待値は以下のように定義される。

$$E(X)=x_1P(X=x_1)+x_2P(X=x_2)+\cdots x_kP(X=x_k)$$

- Xの期待値は、Xが取りうるそれぞれの値に重みをつけたもの（加重平均）

- 「重み」はその値が起こる確率

確率変数の期待値

- 離散型の場合を考え、取りうる値の候補を $x_1, x_2, \ldots, x_k$とする。
- 確率質量関数 $f(x_j), j = 1, \ldots, k$ を使えば、期待値は

$$\boxed{E(X) = \sum_{j=1}^{k} x_j f(x_j)}$$

と表される。

- ちなみに連続型の場合は、和のオペレーターを積分記号に変え（確率質量関数を確率密度関数に変えて）最後にdxをつけるだけ。

$$E(X) = \int x f(x) dx$$

確率変数の期待値 （離散の場合）

- 例1：離散型確率変数Xの確率質量関数が次で与えられるものだとする。

$$f(1) = f(2) = 1/2$$

- このとき

$$E[X] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = 1$$

- 例2：離散型確率変数Xの確率質量関数が次で与えられるものだとする。

$$f(1) = 1/3, \quad f(2) = 2/3$$

- このとき

$$E[X] = 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{5}{3}$$

- Q：Xをフェアーなさいころを投げたときに出る目とする。このとき$E[X]$は？
- 確率質量関数は

$$f(1) = f(2) = \cdots = f(6) = 1/6$$

よって期待値は

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

- Q：Xがパラメーターpをもつベルヌーイ確率変数だとする。このとき$E[X]$は？

$$E[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

- 1回の試行の期待成功回数は、「成功」の確率に等しい。

確率変数の一次関数の期待値

- 回帰分析の理論でよく使うルールが次のもの。
- Xは確率変数。ここで、この確率変数をa倍してbだけ足したものを考えよう（ただしa, bは定数）。

$$Y = aX + b$$

- このYも確率変数になる。
- ではYの期待値$E[Y]$は？
- Xが離散型確率変数だとすると、

$$E(aX + b) = \sum_{j=1}^{k} (a x_j + b) f(x_j) = a \underbrace{\sum_{j=1}^{k} x_j f(x_j)}_{\boxed{=E(X)}} + b \underbrace{\sum_{j=1}^{k} f(x_j)}_{\boxed{=1}}$$

期待値・超頻出ルール

もしaとbが定数なら

$$E[aX+b]=aE[X]+b$$

- これは確率変数が連続型でも成り立つ。

33

モーメント

- 確率変数Xの期待値$E[X]$は、Xの**1次のモーメント**(first moment)と呼ばれることもある。
- $E[X^n],\ n\geq 1$は「n次のモーメント」と呼ばれる。
- この「**モーメント**」という言葉、覚えておいてください。

34

分散

- 確率変数Xの**分散**は次のように定義される:

$$VAR(X)=E[(X-E[X])^2]$$

- 分散は確率変数の分布の仕方を表す指標の1つ。
  - 「**散らばりかた**」を示す1つのメジャー。
- 分散(散らばり)が小さければ、期待値の周りの実現値が起こりやすく、期待値から離れた実現値は起こりにくい。

35

分散

- 分散は次のようにも表すことができる:

$$VAR(X)=E[X^2]-(E[X])^2$$

- 宿題: 上の定理を証明しなさい。

36

分散

- 分散の特性1：定数の分散はゼロ。定数cについて、

$$VAR(c) = 0$$

- 分散の特性2:いかなる定数a、bについて

$$VAR(aX + b) = a^2 VAR(X)$$

- 例：Xは確率変数。$Y = 32 + (9/5)X$のときYの分散は

$$VAR(Y) = VAR(32 + (9/5)X) = (9/5)^2 VAR(X)$$

- 宿題:分散の特性2を証明しなさい。

期待値と標本平均、分散と標本分散

- 期待値：$E(X)$
- 標本平均：$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
- 分散：$VAR(X) = E(X - E(X))^2$
- 標本分散：$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \approx \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$（もし$N$が大きければ）
- $E(\,)$ を $\frac{1}{N}\sum_{i=1}^{N}$　で置き換えると、

期待値　→　標本平均
分散　　→　標本分散

に何となくなっているかな？、と感じてください。

標準偏差

- **標準偏差**（standard deviation）は、分散のルート。

$$SD(X) = \sqrt{VAR(X)}$$

- 標準偏差の特性1：定数の標準偏差はゼロ。すなわち定数cについて、

$$SD(c) = 0$$

- 標準偏差の特性2：いかなる定数a、bについて

$$SD(aX + b) = |a|\, SD(X)$$

標準偏差

- 標準偏差の特性2により、分散よりも標準偏差の方が便利なことあり。
- X: 確率変数
- 例えば、Xを1,000円の単位で測られた所得とする。
  - $X = 1$なら1000円、$X = 0.1$なら100円ということ。
- このとき$Y = 1{,}000X$と定義すると、Yは1円の単位で測られた所得。
- $E(X) = 20$、$SD(X) = 6$とする。このとき$E(Y) = \underline{1{,}000}E(X) = 20{,}000$、また$SD(Y) = \underline{1{,}000}SD(X) = 6{,}000$。従って、平均と標準偏差は同じ係数倍（ここでは1,000倍）になる。

標準偏差

- 分散はこの限りではない。

$$VAR(Y) = (1{,}000)^2 VAR(X)$$

- この例では、Yの分散はXの分散の1,000,000倍になる。
- これは標本分散と標本標準偏差のときにと同じパターン。