令和2年度「専修学校による地域産業中核的人材養成事業」

統計学 I

令和2年度「専修学校による地域産業中核的人材養成事業」

統計学 I

Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

目 次

AIシステム科　　シラバス　　20XX年度入学生

科目名	統計学Ⅰ					週合計駒数	1駒	作成日	
区分	必修 講義	開講時期	1年次 前期	週講義駒数 週実習等駒数	1駒 0駒	総時間数 総単位数	30時間 2単位	担当教員	

目標	概要
人工知能を学ぶ上で必要な基礎数学を習得するとともに、記述統計学の習得を目標とする。	本講義では、統計学および人工知能を学ぶ上で必須となる基礎数学を学習し、データ処理の基本知識である記述統計学について学習する。

履修前提 ※選択・エクステンションのみ記入	テキスト・参考文献
	オリジナルテキスト
評価方法	関連科目
小テスト／中間テスト／期末テスト、提出課題、授業に取り組む姿勢（出席率、授業態度）	統計学Ⅱ、データマイニング、AIプログラミングⅠ・Ⅱ、機械学習Ⅰ・Ⅱ・Ⅲ、AIシステム開発

1	学習目標 数の体系・種類を理解し、基本的な算術が出来る。	学習項目 基礎数学（1）：数の体系と種類を理解した上で、数に関する基本事項（算術演算、べき乗、階乗、約数・倍数、数の偶奇、素数、など）を学習する。
	理解度確認：練習問題、小テスト	
2	学習目標 ステートメントを数理論理学的に捉えることが出来る。論理的推論が出来る。	学習項目 基礎数学（2）：論理学の歴史と発展を踏まえた上で、数理論理学の基礎について学習する。具体的には、命題論理と一階述語論理を中心に、論理演算、量化、三段論法、逆・対偶・裏、背理法などについて学習する。
	理解度確認：練習問題、小テスト	
3	学習目標 集合とは何かを説明出来る。集合演算が出来る。	学習項目 基礎数学（3）：集合論の歴史と発展を踏まえた上で、集合論の基礎について学習する。集合のナイーブな定義から始め、集合演算、Vennダイアグラム、写像について学習する。
	理解度確認：練習問題、小テスト	
4	学習目標 各初等関数の性質を説明出来る。	学習項目 基礎数学（4）：関数とは何かを踏まえた上で、初等関数論について学習する。具体的には、各初等関数の定義・性質および初等関数の関係、多変数関数について学習する。
	理解度確認：練習問題、小テスト	
5	学習目標 微分の計算が出来る。	学習項目 基礎数学（5）：微積分を中心とした解析学について学習する。ここでは、微分の定義から始め、練習問題を解くことにより、その幾何学的意味の理解および計算方法に慣れる。
	理解度確認：練習問題、小テスト	
6	学習目標 積分の計算が出来る。	学習項目 基礎数学（6）：微積分を中心とした解析学について学習する。ここでは、積分の定義から始め、練習問題を解くことにより、その幾何学的意味の理解および計算方法に慣れる。
	理解度確認：練習問題、小テスト	
7	学習目標 データの性質を捉えることが出来る。統計データの分析プロセスおよび統計資料の活用について説明出来る。	学習項目 統計学の歴史と発展を踏まえた上で、データに関する基本事項（質的・量的、尺度、次元、時系列・クロスセクション・パネル、など）、統計データの分析プロセスと統計資料、データの活用について学習する。
	理解度確認：練習問題、小テスト	
8	学習目標 度数分布表を作成出来る。ヒストグラムを描くことが出来る。各種代表値の定義とその意味を説明出来る。	学習項目 1次元データに関する度数分布とヒストグラム、各種代表値（平均値、メディアン、モード、分散、標準偏差、など）について学習する。
	理解度確認：練習問題、小テスト	
9	学習目標 組み合わせの数および順列の数を計算出来る。確率とは何かを説明出来る。	学習項目 組み合わせの数および順列の数を踏まえた上でナイーブな確率の導入を行い、標本空間と事象について学習する。
	理解度確認：練習問題、小テスト	
10	学習目標 確率変数とは何かを説明出来る。Chebyshevの不等式の意味を説明出来る。	学習項目 確率変数について学習する。具体的には、確率分布・確率密度関数、期待値、Chebyshevの不等式について学習する。
	理解度確認：練習問題、小テスト	
11	学習目標 代表的な確率分布の説明が出来る。	学習項目 自然現象を例に挙げながら代表的な確率分布（超幾何分布、Bernoulli分布、Gauss分布、Poisson分布、一様分布など）を学習する。
	理解度確認：練習問題、小テスト	
12	学習目標 多次元の確率分布の特徴・性質について説明出来る。同時確率密度関数と周辺確率密度関数について説明出来る。	学習項目 多次元における確率分布およびその関連事項である同時確率密度関数、周辺確率密度関数、確率変数の独立性について学習する。
	理解度確認：練習問題、小テスト	

13	学習目標 大数の法則について説明出来る。	学習項目 大数の法則とその意味について学習する。併せて、大数の法則のコンピュータシミュレーションも行う。
	理解度確認: 練習問題、小テスト	
14	学習目標 中心極限定理について説明出来る。	学習項目 中心極限定理とその意味について学習する。併せて、中心極限定理のコンピュータシミュレーションも行う。
	理解度確認: 練習問題、小テスト	
15	学習目標 これまでに学習した内容を復習し、理解を確実なものにする。	学習項目 これまでの学習内容の総復習を実施する。
	理解度確認: 確認テスト	

第1回：基礎数学1

アジェンダ

- 数の体系
- 算術演算
- べき乗
- 階乗
- 約数･倍数･素数
- 数の偶奇

全15回の講義について

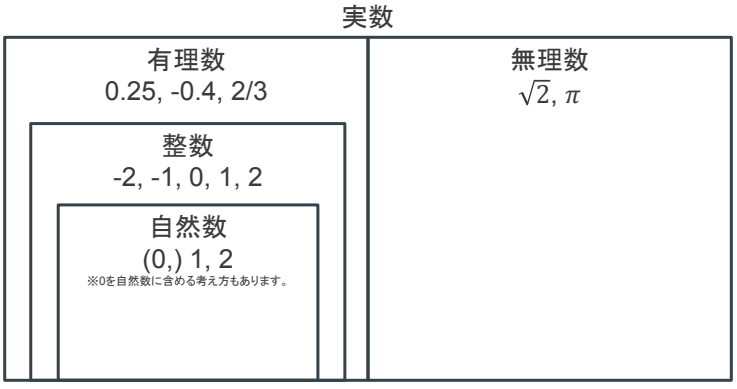
- 統計学および人工知能を学ぶ上で必須となる基礎数学を学習し、データ処理の基本知識である記述統計学について学習します。

数の体系

自然数、整数、有理数、実数には以下の包含関係があります。

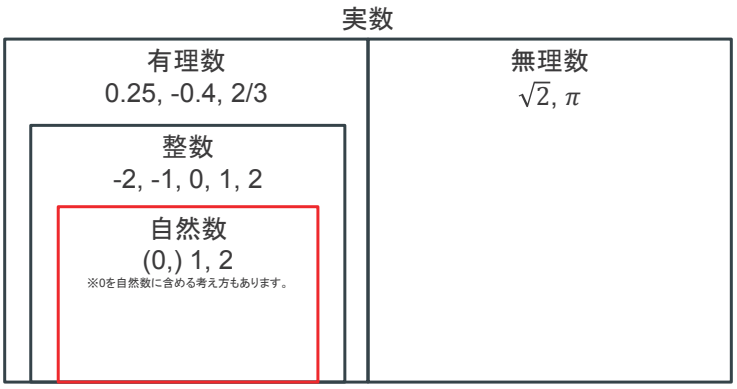
$$N \subset \mathbb{Z} \subset Q \subset R$$

- N=自然数全体の集合
- Z=整数全体の集合
- Q=有理数全体の集合
- R=実数全体の集合



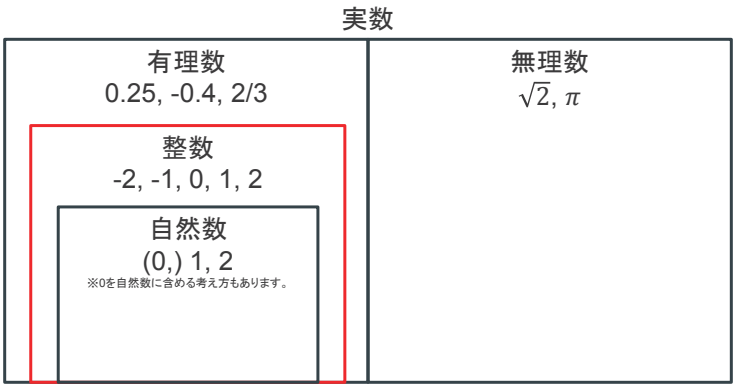
自然数とは

- 1, 2, 3, 4, 5と続く数の総称を自然数といいます。0を含めることもあります。



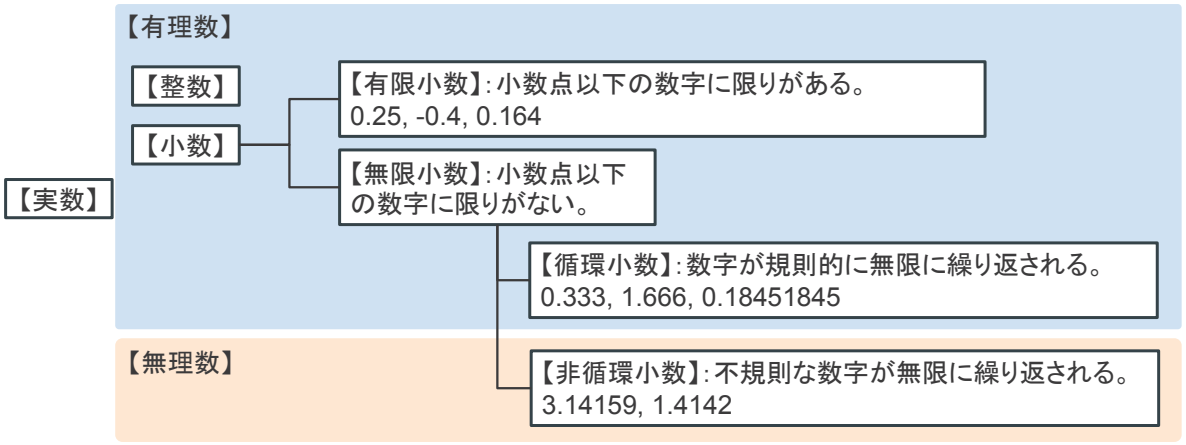
整数とは

- 0と、0に1ずつ加えて得られる自然数（1, 2, 3, 4, …）および1ずつ引いて得られる数（−1, −2, −3, −4, …）を整数といいます。



有理数/無理数とは

- 実数のうち、非循環小数を無理数といいます。
- 無理数以外の実数を有理数といいます。



算術演算

- 算術演算とは、以下の4つの基本的な計算のことです。
  - 足し算（加算）
  - 引き算（減算）
  - 掛け算（乗算）
  - 割り算（除算）

累乗とべき乗

- 累乗もべき乗も a^n（aのn乗）で表現できます。
- 「n」が自然数（正の整数）に限定される場合を累乗といいます。
- 「n」が自然数（正の整数）に限定されない場合をべき乗といいます。

べき乗の定義

- 「n」が自然数のとき、aをn回かけ合わせます。
  - $2^4 = 2 * 2 * 2 * 2$
- 「n」が0の場合は、1となります。
  - $2^0 = 1$
- 「n」が負の整数の場合は、下記のようになります。
  - $2^{-4} = 1 / (2 * 2 * 2 * 2)$

べき乗の定義

- 「n」が有理数（分数）の場合は、下記のようになります。
  - $a^{p/q} = \sqrt[q]{a^p}$
  - $2^{3/4} = \sqrt[4]{2^3}$

- 「n」がπなどの無理数の場合は、挟み込んで極限を探っていきます。
  - $2^{3.141} < 2^{\pi} < 2^{3.142}$

階乗

- 1からnまでの全ての整数をかけ合わせた値のことを「nの階乗」といい、n!と表現します。
  - 4! = 4 * 3 * 2 * 1 = 24

- 0の階乗は1と定義されます。
  - 0! = 1

約数・倍数・素数

- 約数：ある数を割ることができる整数のことです。
  - 12の約数は1, 2, 3, 4, 6, 12
- 素数：約数が1とその数しかない整数のことです。
  - 例えば2, 3, 5, 7, 11,などです。
- 公約数：複数の整数に共通する約数のことです。
  - 12の約数は1, 2, 3, 4, 6, 12
  - 18の約数は1, 2, 3, 6, 9, 18
    - 12と18の公約数は1, 2, 3, 6
- 最大公約数：最大の公約数のことです。
  - 12と18の最大公約数は6

約数・倍数・素数

- 倍数：ある整数を整数倍した整数のことです。
  - 3の倍数は3, 6, 9, 12など
- 公倍数：複数の整数に共通した倍数のことです。
  - 3の倍数は3, 6, 9, 12など
  - 4の倍数は4, 8, 12, 16など
    - 3と4の公倍数は12、24など
- 最小公倍数：最小の倍数のことです。
  - 3と4の最小公倍数は12

奇数と偶数

- 2で割り切れる自然数を偶数、割り切れない自然数を奇数といいます。
  - 偶数：2, 4, 6,,,
  - 奇数：1, 3, 5,,,
  - ※定義を「2で割り切れる整数」とした場合は、0は偶数に含まれます。

- 偶数と奇数の演算は、以下のようになります。
  - 偶数 ± 偶数 = 偶数
  - 偶数 ± 奇数 = 奇数
  - 奇数 ± 奇数 = 偶数
  - 偶数 × 偶数 = 偶数
  - 偶数 × 奇数 = 偶数
  - 奇数 × 奇数 = 奇数

演習問題

演習1：数の体系

- 自然数、整数、有理数、無理数、実数の定義について説明してください。
- 自然数、整数、有理数、無理数、実数の包含関係について説明してください。

演習2：べき乗

- 2^5を計算してください。
- 3^0を計算してください。
- 2^{-5}を計算してください。

演習3：階乗

- 5!を計算してください。
- 0!を計算してください。

演習4：約数・倍数

- 12と20の最大公約数を求めてください。
- 4と5の最小公倍数を求めてください。

第2回：基礎数学2

アジェンダ

- 数理論理学
- 命題論理
- 一階述語論理
- 背理法

数理論理学とは

- [Wikipediaより]論理学（ろんりがく、英: logic）とは、「論理」を成り立たせる論証の構成やその体系を研究する学問である。ここでいう論理とは、思考の形式及び法則である。これに加えて、思考のつながり、推理の仕方や論証のつながりを指す。論理学は、伝統的には哲学の一分野である。数学的演算の導入により、数理論理学（記号論理学）という分野ができた。

- [Wikipediaより]数理論理学（独: mathematische Logik、英: mathematical logic）は、論理学（形式論理学）の数学への応用の探求ないしは論理学の数学的な解析を主たる目的とする、数学の関連分野である。局所的には数理論理学は超数学、数学基礎論、理論計算機科学などと密接に関係している。[1]数理論理学の共通な課題としては形式体系の表現力や形式証明系の演繹の能力の研究が含まれる。

- 数理論理学とは、「推論のやり方と、その正しさ」を扱う学問です。

命題論理

命題論理とは

- [Wikipediaより]命題論理（めいだいろんり、英: propositional logic）とは、数理論理学（記号論理学）の基礎的な一部門であり[1]、命題全体を1つの記号に置き換えて単純化し、論理演算を表す記号（論理記号・論理演算子）を用いて、その命題（記号）間の結合パターンを表現・研究・把握することを目的とした分野のこと。

命題論理の例

次の1～3（命題といいます）は、全て事実だとします。

1. 釣りが好きな人は、ブリが好きである。
2. 男の人は、釣りが好きである。
3. お酒が好きな人は、イカの塩辛が好きであり、かつ男である。

このとき、次のア～オの中で確実にいえることを全て選んでください。

ア. 男の人は、イカの塩辛が好きである。
イ. お酒が好きな人は、男である。
ウ. イカの塩辛が好きな人は、ブリが好きである。
エ. イカの塩辛が好きでない人は、釣りが好きでない。
オ. ブリが好きでない人は、お酒が好きでない。

命題を論理式にする

「1. 釣りが好きな人は、ブリが好きである。」を論理式にすると、下記のようになります。

釣り→ ブリ

この場合、「釣り」を**前言**、「ブリ」を**結言**といいます。このように文を記号で表現したものを**論理式**といいます。

2つ目の命題を論理式に直すと、以下のようになります。

男→ 釣り

3つ目の命題には「かつ（and）」という表現が含まれています。この場合は「かつ」を∧の記号を使って表現します。

お酒→ イカの塩辛 ∧男

もし命題に「または（or）」という表現が含まれていた場合は、∨の記号を使います。

もし命題に「～ではない（not）」という表現が含まれている場合は、命題の上全体に横線を引くか（$\overline{\text{釣り}}$）、命題の前に－を書きます（ － 釣り）。

逆、裏、対偶

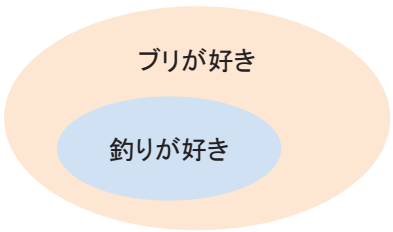
「1. 釣りが好きな人は、ブリが好きである。」を論理式にしたものは、下記のようになりました。

釣り→ ブリ

この命題の「**逆**」とは、前言と結言の方向を反対にしたものです。

ブリ→ 釣り

命題が真のとき、逆は必ずしも真とはなりません。下の図の通り、ブリが好きな人の全てが釣りが好きとは限らないからです。



逆、裏、対偶

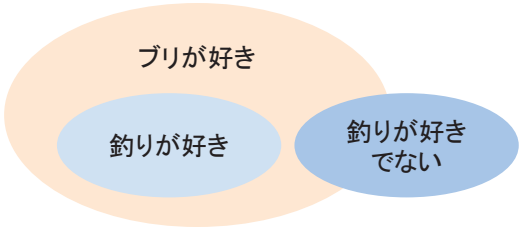
「1. 釣りが好きな人は、ブリが好きである。」を論理式にしたものは、下記のようになりました。

釣り→ ブリ

この命題の「裏」とは、前言と結言を否定したものです。

$\overline{釣り} \rightarrow \overline{ブリ}$

命題が真のとき、裏は必ずしも真とはなりません。下の図の通り、釣りが好きでない人が、ブリが好きな場合がありえるからです。



逆、裏、対偶

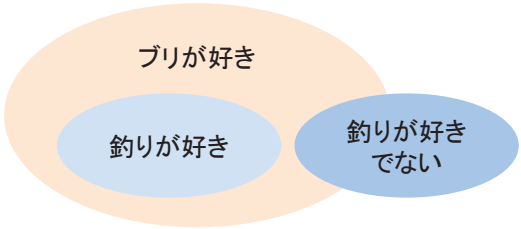
「1. 釣りが好きな人は、ブリが好きである。」を論理式にしたものは、下記のようになりました。

釣り→ ブリ

この命題の「対偶」とは、前言と結言を否定したものの、方向を反対にしたものです。

$\overline{ブリ} \rightarrow \overline{釣り}$

命題が真のとき、対偶は必ず真となります。下の図の通り、（釣りが好きな人は必ずブリが好きなので、）ブリが好きでない人に釣りが好きな人は含まれないからです。



三段論法

命題の1と2を利用すると、

2. 男の人は、釣りが好きである。

1. 釣りが好きな人は、ブリが好きである。

なので、

男の人はブリが好きである。

（男→ブリ）

となります。

このような推論方法を三段論法といいます。

命題の分解

命題の3には、「かつ」という表現が含まれています。

3. お酒が好きな人は、イカの塩辛が好きであり、かつ男である。（お酒→ イカの塩辛 ∧男）

これは以下の2つの命題に分解できます。

お酒が好きな人は、イカの塩辛が好きである。（お酒→イカの塩辛）

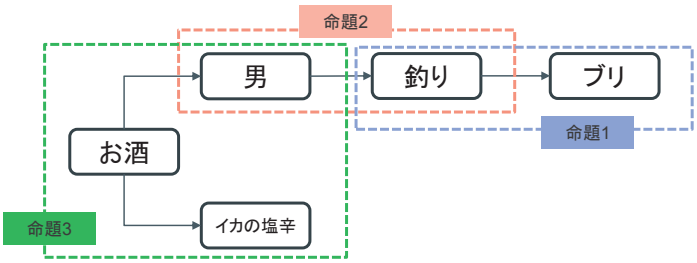
お酒が好きな人は、男である。（お酒→男）

「お酒→イカの塩辛」、「お酒→男」のようにこれ以上分解できない命題を、要素命題といいます。

命題の図示

命題1～3を、有向グラフで表現してみます。

1. 釣りが好きな人は、ブリが好きである。（釣り→ ブリ）
2. 男の人は、釣りが好きである。（男→ 釣り）
3. お酒が好きな人は、イカの塩辛が好きであり、かつ男である。（お酒→ イカの塩辛 ∧男）



命題論理の例の答え

次のア～オの中で確実にいえることを全て選んでください。

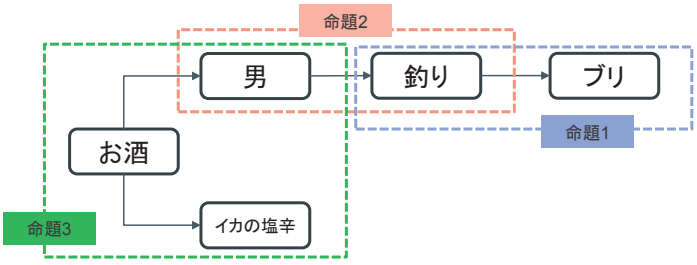
~~ア. 男の人は、イカの塩辛が好きである。~~

イ. お酒が好きな人は、男である。

~~ウ. イカの塩辛が好きな人は、ブリが好きである。~~

~~エ. イカの塩辛が好きでない人は、釣りが好きでない。~~

オ. ブリが好きでない人は、お酒が好きでない。（「お酒→ブリ」の対偶なので真）



一階述語論理

述語論理とは

- 述語論理とは、命題の中身まで見て推論の正しさを確かめていく学問です。
- 命題論理では、命題を基本単位として取り扱いました。述語論理では、命題を個体名とそれに関係する述語に分解し、それらの関係について探求します。

述語論理と命題論理の違い

命題論理の章で扱った命題1は、以下のように表現できました。

1. 釣りが好きな人は、ブリが好きである。

「釣り」という記号で表現した箇所　「ブリ」という記号で表現した箇所

「釣り」と「ブリ」という2つの記号を、「 → 」という記号でつなぎ、

釣り→ ブリ

と表現しました。

命題理論では、「釣りが好きな人は、ブリが好きである。」という粒度の表現は取り扱えますが、以下のように2つの文に分けてしまうと、取り扱うことができません。

1-1. 釣りが好きな人

1-2. ブリが好き

述語論理は1-1、1-2のような表現も取り扱える表現方法です。

一階述語論理で使用する表現

- 一階述語論理はオブジェクト（モノ）を単位として構成されます。
- 一階述語論理はオブジェクトそのものを表現する個体記号、個体の性質や状態を表す述語記号、個体間の関係を表す関係記号、個体の量を指定する量化子などで構成されます。

記号の種類	表記	意味
個体記号	A（特定の個体）、x（任意の個体）など	特定の個体や任意の個体を表す
述語記号	like(a)、run(a)など	個体の性質や状態を表す
関数記号	MOTHER(a)など	個体間の関係を表す
量化子	∀（任意の）、∃（ある～が存在する）など	個体の量を指定する

個体記号と述語記号

- 太郎は男性です。
  - Man(太郎)
- 太郎は釣りが好きです。
  - Like(太郎, 釣り)
- 男性である太郎は釣りが好きです。
  - Like(Man(太郎), 釣り)

関数記号

- 太郎と花子は友達です。
  - FRIEND(太郎, 花子)
- 三郎は太郎の父親です。
  - FATHER(三郎, 太郎)
- 太郎の父親です。(※「三郎」というオブジェクト名がない場合)
  - FATHER(太郎)
- 四郎と太郎の父親は友達です。
  - FRIEND(四郎, FATHER(太郎))

量化子

- 花子と友達なのが太郎の場合、以下のように表現できました。
  - FRIEND(太郎, 花子)

- 「全ての人が花子と友達です。」と表現したい場合、任意のオブジェクトをxとし、また「全ての～」を意味する∀を使用して、以下のように表現します。
  - ∀x FRIEND(x, 花子)

- 「ある人が花子と友達です（花子と友達な人が存在する）。」と表現したい場合、「ある～が存在する」を意味する∃を使用して、以下のように表現します。
  - ∃x FRIEND(x, 花子)

一階述語論理による複雑な表現

- ハルクは人間の友達を持っている。
  - ∃x Human(x) ∧ FRIEND(x, Hulk)

- 釣りが好きな人は、ブリが好きである。
  - ∀x Like(x, 釣り) → Like(x, ブリ)

背理法

背理法

- 背理法とは、以下の流れで証明を行う手法のことです。
  - 命題が正しくないと仮定する、
  - その結果、矛盾してしまう、
  - よって、命題は正しい。

背理法の例

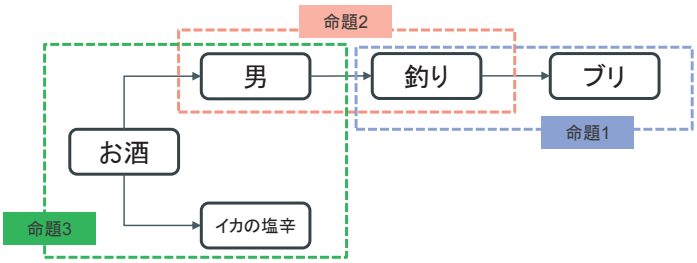
- √2 が無理数であることを証明してください。
  - √2 が有理数だと仮定します。
  - 上記のように仮定すると、√2 = q / p（p、qともに素な整数）で表現できる。
  - $2p^2 = q^2$となる。
  - 左辺が偶数となるため、右辺q^2も偶数となる。
  - qは整数のため、q2が偶数となり得るのは4 (=$(2*1)^2$), 16 (=$(2*2)^2$), 36(=$(2*3)^2$),,,のいずれかとなり、右辺は4の倍数になる。右辺を素因数分解した場合、2は偶数個存在する。
  - 左辺を因数分解した場合、2は奇数個存在することになる。
  - よって右辺と左辺を因数分解した場合の2の数に矛盾が発生するため、√2は有理数となり得ない。

演習問題

演習1：命題論理

命題1～3に従った場合、「お酒が好きな人は釣りが好き」は確実に言えるでしょうか？

1. 釣りが好きな人は、ブリが好きである。（釣り→ ブリ）
2. 男の人は、釣りが好きである。（男→ 釣り）
3. お酒が好きな人は、イカの塩辛が好きであり、かつ男である。（お酒→ イカの塩辛 ∧男）



演習2：対偶

命題1～3の対偶をそれぞれ述べてください。

1. 釣りが好きな人は、ブリが好きである。（釣り→ ブリ）
2. 男の人は、釣りが好きである。（男→ 釣り）
3. お酒が好きな人は、イカの塩辛が好きであり、かつ男である。（お酒→ イカの塩辛 ∧男）

演習3：一階述語論理

命題2を一階述語論理で表現してください。

2. 男の人は、釣りが好きである。（男→ 釣り）

第3回：基礎数学3

アジェンダ

- 集合論とは
- 集合の表記
- 様々な集合
- 集合のパラドックス
- 写像

集合論とは

- [Wikipediaより]集合論（しゅうごうろん、英: set theory, 仏: théorie des ensembles, 独:Mengenlehre）は、集合とよばれる数学的対象をあつかう数学理論である。

- [Wikipediaより]数学における集合（しゅうごう、英: set, 仏: ensemble, 独: Menge）とは、大雑把に言えばいくつかの「もの」からなる「集まり」である。集合を構成する個々の「もの」のことを元（げん、英: element; 要素）という。集合は、集合論のみならず現代数学全体における最も基本的な概念の一つであり、現代数学のほとんどが集合と写像の言葉で書かれていると言ってよい。

集合の表記

集合は「xに関する命題P(x)が真となるようなxの集まり」という表現がされ、下記のように表記されます。

$$\{x|P(x)\}$$

例えば、「10以上の整数の集合」は以下のように表記されます。

$$\{x|x \in \mathbb{Z}, x \geq 10\}$$

ここで、「$x \in \mathbb{Z}$」は「xは整数の集合$\mathbb{Z}$に属する」という意味です。
このようにある集合を構成するxを要素、または元といいます。

「xは整数の集合$\mathbb{Z}$に属さない」と表記したい場合は、以下のようにします。

$$x \notin \mathbb{Z}$$

集合の表記

集合の要素を列挙して集合を定義することもでます。この場合、要素がx1、x2、x3、であれば下記のように表記します。

$$\{x1, x2, x3\}$$

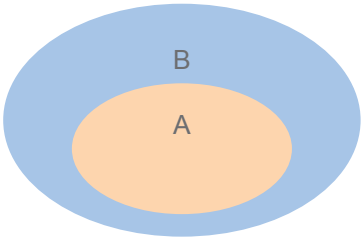
例えば10以下の素数の集合は次の表に表記されます。

$$\{n | n\text{は素数}, n \leq 10\} = \{2, 3, 5, 7\}$$

部分集合

2つの集合AとBに対して「x∈A⇒x∈B」が成り立つ場合、AはBの部分集合といい、以下のように表記します。

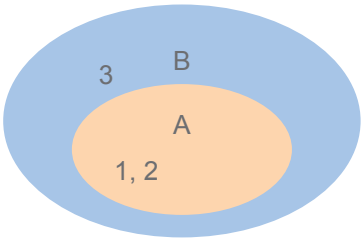
A⊂BもしくはB⊃A



真部分集合

A⊂Bであり、かつx∈Aでありx∉Bの要素が存在する場合、AはBの真部分集合であるといいます。
例えばA= $\{1, 2\}$、B= $\{1, 2, 3\}$であるときAはBの部分集合といい、下記のように表記します。

$A \subsetneq B$

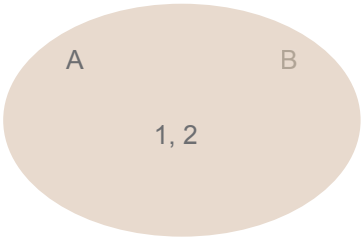


等価な集合

A⊂BかつB⊂Aである場合、「x∈A⇔x∈B」であり、2つの集合は等しくなり、以下のように表記します。

$A = B$

例えばA= $\{1, 2\}$、B= $\{1, 2\}$であるときAとBは等しくなります。



空集合

要素を1つも含まない集合を空集合といい、∅で表記します。この場合、任意の要素xは「$x \notin \emptyset$」となります。

共通部分

2つの集合A、Bに対して

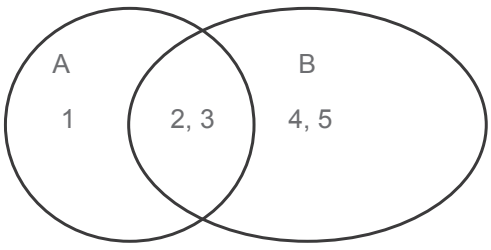
$A \cap B = \{x | x \in A, x \in B\}$：AとBに同時に属している要素の集合

をAとBの共通部分といいます。

$A \cap B = \phi$

であるとき「AとB共通部分がない」、もしくは「AとBは互いに素である」といいます。

例えばA= $\{1, 2, 3\}$、B= $\{2, 3, 4, 5\}$とした場合、$A \cap B = \{2, 3\}$となります。



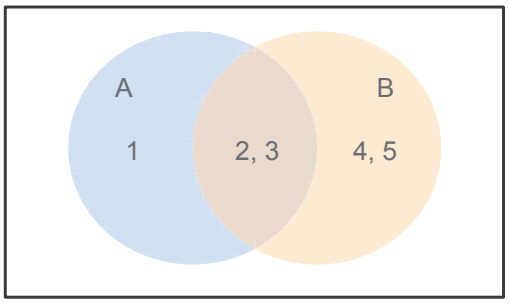
和集合

2つの集合A、Bに対して

$A \cup B = \{x | x \in A$または$x \in B\}$:AもしくはBに属している要素の集合

をAとBの和集合といいます。

例えばA= $\{1, 2, 3\}$、B= $\{2, 3, 4, 5\}$とした場合、$A \cup B = \{1, 2, 3, 4, 5\}$となります。



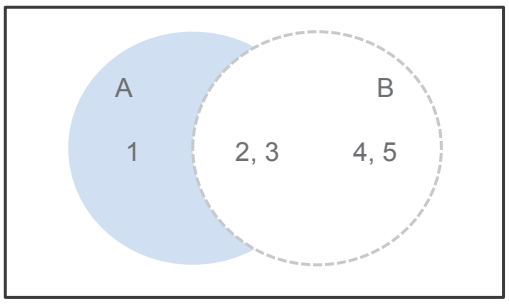
差集合

2つの集合A、Bに対して

$A - B = \{x | x \in A$または$x \notin B\}$:Aに属していてBには属していない要素の集合

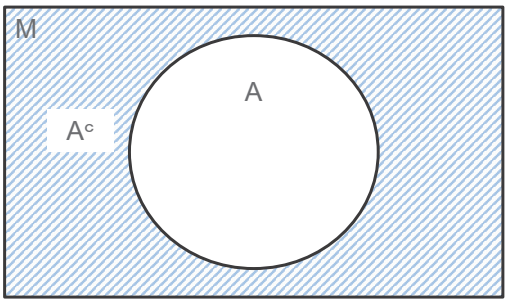
をAとBの差集合といいます。

例えばA= $\{1, 2, 3\}$、B= $\{2, 3, 4, 5\}$とした場合、$A - B = \{1\}$となります。



補集合

A⊂M（集合Aは集合Mの部分集合）のとき、差集合$M - A$を「AのMにおける補集合」といいます。
Aが集合Mの部分集合であることが明らかな場合、$M - A$をA^cと表記します。



直積集合

二つの集合AとBに対し、「Aの要素とBの要素を1つずつ取ってきて作ったペアを全て集めた集合」を直積集合（または、デカルト積）といいます。

$$\{(a, b)|a \in A, b \in B\}$$

例えばA= $\{1, 2\}$、B= $\{3, 4, 5\}$とした場合は以下のようになります。

$$A \times B = \{(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5)\}$$

ここで$(1, 2)$と$(2, 1)$のように同じ数字でも順番がことなる要素は別の要素として扱います。
したがって$A \times B$と$B \times A$は違ったものになります。

べき集合

集合Aに対して、Aの部分集合を全て集めたものをAのべき集合といい、2^Aと表記します。

例えば

$A = \{a, b\}$

のべき集合は

$2^A = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$

となります。

※空集合$\emptyset$と元の集合$\{a, b\}$も部分集合として扱います。

集合の演算

Mを集合としA、B、Cはその部分集合とします。また、補集合はMで考えることとします。このとき次が成り立ちます。

1. $A \cap A^c = \emptyset$
2. $A \cup A^c = M$
3. $(A \cap B)^c = A^c \cup B^c$:ド・モルガンの法則
4. $(A \cup B)^c = A^c \cap B^c$:ド・モルガンの法則
5. $A \cap (A \cup B) = A$
6. $A \cup (A \cap B) = A$

集合論のパラドックス

集合論には素朴集合論(naive set theory)と公理的集合論(axiomatic set theory)があります。
素朴集合論では、以下のようなパラドックス([Wikipediaより]正しそうに見える前提と、妥当に見える推論から、受け入れがたい結論が得られる事を指す言葉)が生じてしまいます。

例えば「床屋のパラドックス」という話があります。
以下のような床屋がいるとします。

- 自分でひげをそらない人全員のひげをそる。
- 自分でひげをそる人のひげはそらない。

この場合、床屋は自分のひげは自分でそるのでしょうか？
床屋が自分でひげをそるとした場合も、そらないとした場合もどちらも矛盾が生じてしまいます。

これは自分自身を要素として含まない集合すべての集まり$A = \{X | X \notin X\}$という集合を考えてしまうため生じるパラドックスです。
現代数学では「$A = \{X | X \notin X\}$を集合ではない」と考えるのが主流です。

写像

集合Aの各要素に対して集合Bの要素がただ1つ対応する規則fが定まっているとき、この対応をAからBへの写像といい、以下のように表記します。

$$f : A \to B$$

2つの写像

$$f : A \to B$$
$$g : A \to B$$

が等しいとは、Aのすべての要素aについて、f(a) = g(a)が成立することを意味し、f = g で表します。

全単射

集合Pはある学校の先生の名前、集合Qは担当教科だとします。
下図のように、2つの集合の要素がすべて1対1で対応している場合を全単射といいます。
例えば「A先生は理科の担当」であることを、以下のように表記します。

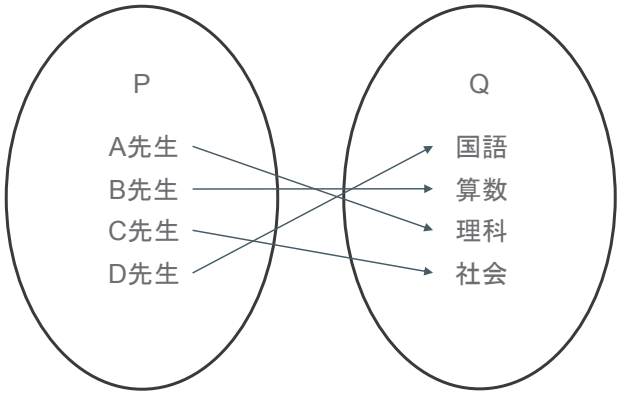
f：A先生 → 理科

逆にQからAへの写像も成り立ちます。

g：理科 → A先生

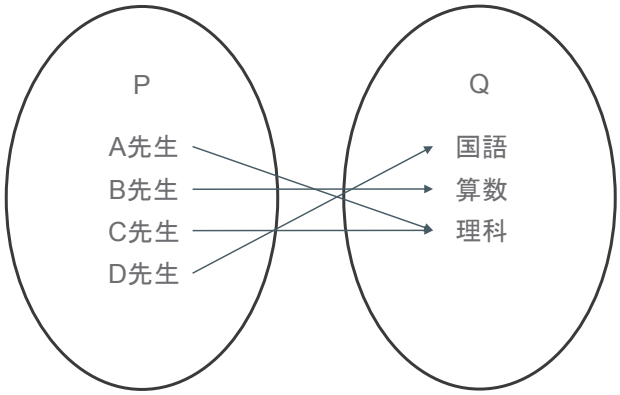
全単射は、双方向に写像が成り立ちます。

$P \underset{g}{\overset{f}{\leftrightarrow}} Q$



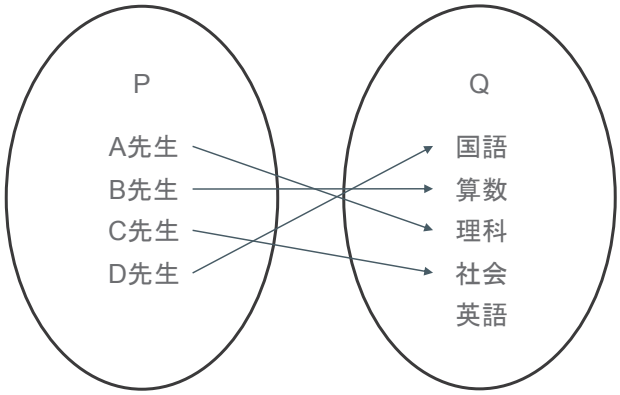
全射

Qの要素を定めても、Pの要素が1つに定まりません。
このように$P \to Q$の写像が可能でも、$Q \to P$の写像が不可能な場合、全射といいます。



単射

$P \to Q$の写像は要素が対応していますが、$Q \to P$の写像において「英語」に対応する先生が存在しません。
このような場合を単射といいます。

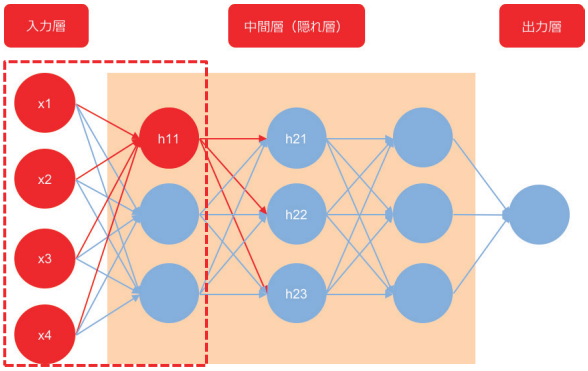


参考：ベクトル演算と写像

ニューラルネットワークにおいて、「入力層の値」と「入力層−中間層間の重み」を行列計算し、中間層への出力値を算出します。
入力層の値の集合をX、出力層の値をHi、重みを掛け合わせてXからHiを算出する計算式をwとすると、

$$w : \mathrm{X} \to Hi$$

の写像だとみなせます。



演習問題

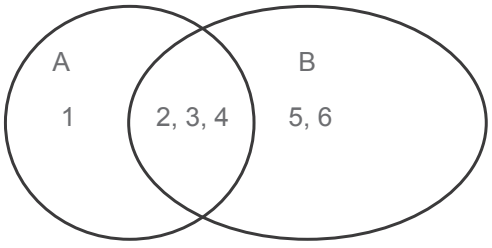
演習1：共通部分

2つの集合A、Bに対して

$A \cap B = \{x | x \in A, x \in B\}$：AとBに同時に属している要素の集合

をAとBの共通部分といいます。

A= $\{1, 2, 3, 4\}$、B= $\{2, 3, 4, 5, 6\}$とした場合、$A \cap B$を求めてください。



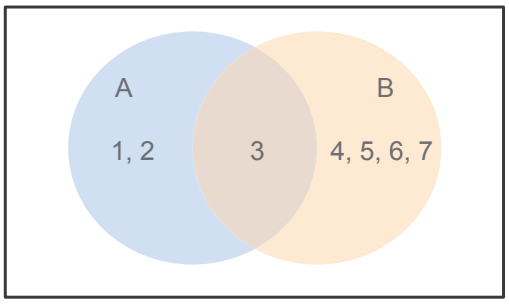
演習2：和集合

2つの集合A、Bに対して

$A \cup B = \{x | x \in A$または$x \in B\}$：AもしくはBに属している要素の集合

をAとBの和集合といいます。

例えばA= $\{1, 2, 3\}$、B= $\{3, 4, 5, 6, 7\}$とした場合、$A \cup B$を求めてください。



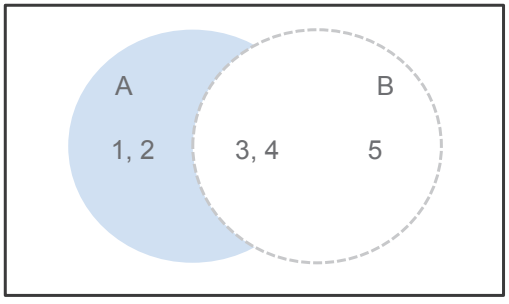
演習3：差集合

2つの集合A、Bに対して

$A - B = \{x | x \in A$または$x \notin B\}$：Aに属していてBには属していない要素の集合

をAとBの差集合といいます。

例えばA= $\{1, 2, 3, 4\}$、B= $\{3, 4, 5\}$とした場合、$A - B$を求めてください。



演習4：直積集合

二つの集合AとBに対し、「Aの要素とBの要素を1つずつ取ってきて作ったペアを全て集めた集合」を直積集合（または、デカルト積）といいます。

$$\{(a, b)|a \in A, b \in B\}$$

A= $\{1, 2\}$、B= $\{3, 4, 5\}$とした場合、$A \times B$と$B \times A$をそれぞれ求めてください。

第4回：基礎数学4

アジェンダ

- 関数
- 初等関数
- 初等関数の例

関数とは

- [Wikipediaより]数学における関数（かんすう、英: function、仏: fonction、独: Funktion、 蘭: functie、羅: functio、函数とも書かれる）とは、かつては、ある変数に依存して決まる値あるいはその対応を表す式の事であった。この言葉はライプニッツによって導入された。その後定義が一般化されて行き、現代的には数の集合に値をとる写像の一種であると理解される。

初等関数とは

- [Wikipediaより]初等関数（しょとうかんすう、英: Elementary function）とは、実数または複素数の1変数関数で、代数関数、指数関数、対数関数、三角関数、逆三角関数および、それらの合成関数を作ることを有限回繰り返して得られる関数のことである。
- [Wikipediaより]初等関数のうちで代数関数でないものを初等超越関数という。双曲線関数やその逆関数も初等関数である。

初等関数の例：定数関数

定数関数とは、それが取りうる値が変数の値によって変動しない定数値となる関数のことです。
例えば、

$$f(x) = 2$$

はxがどのような値でも2に写像する定数関数です。

初等関数の例：指数関数

指数関数とは、a＞0 かつ a≠1 のとき「$y=a^x$で表される関数」のことです。
また、この関数$y=a^x$のことを「aを底とするxの指数関数」と呼びます。

a>1の場合、xが大きくなるに従って指数関数yは大きくなります。
例えば以下のようなケースがあります。

$$y = a^x = 2^3 = 2*2*2 = 8$$

1>a>0の場合、xが大きくなるに従って指数関数yは小さくなります。
例えば以下のようなケースがあります。

$$y = a^x = (1/2)^3 = (1/2)*(1/2)*(1/2) = 1/8$$

底をネイピア数e (= 2.718281828…)とする指数関数のことをexp(x)と表記することがあります。

初等関数の例：対数関数

関数$y=a^b$のbのことを冪（べき）指数といい、bのことを「aを底とするyの対数」といいます。

$$b = \log_a y$$

と表記します。
底aがネイピア数e (= 2.718281828…)の場合、

$$b = \ln y$$

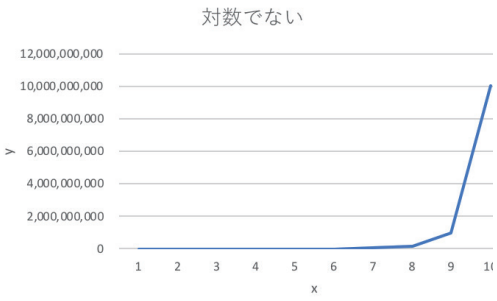
と表記することがあります。

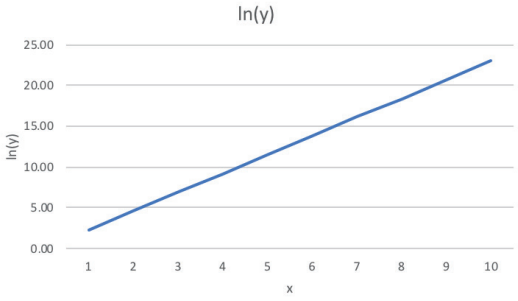
参考：対数を利用した正規化

線形で増加する変数xに対して、変数yが指数関数的に増加する場合を考えます。
変数x、yともにそのままの数値をグラフにすると、真ん中のグラフのようになります。
変数yについて対数をとったものをグラフにすると、右のグラフのようになります。

線形の機械学習器を使用する場合、右のグラフのように対数で変数を加工した場合のほうが良い精度を得られることがあります。

x	y	ln(y)
1	10	2.30
2	100	4.61
3	1,000	6.91
4	10,000	9.21
5	100,000	11.51
6	1,000,000	13.82
7	10,000,000	16.12
8	100,000,000	18.42
9	1,000,000,000	20.72
10	10,000,000,000	23.03

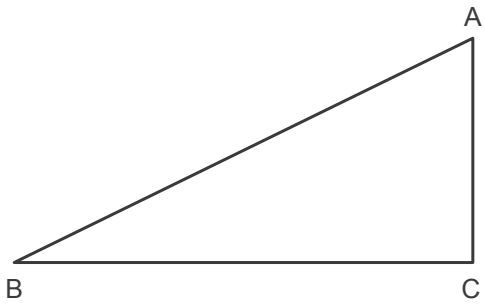




初等関数の例：三角関数：sin（サイン）

以下のような直角三角形を考えた場合、サインは以下のように表記されます。

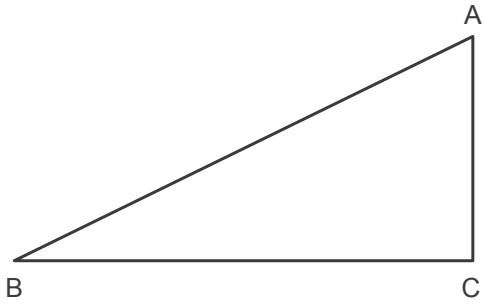
sin = AC ／ AB



初等関数の例：三角関数：cos（コサイン）

以下のような直角三角形を考えた場合、コサインは以下のように表記されます。

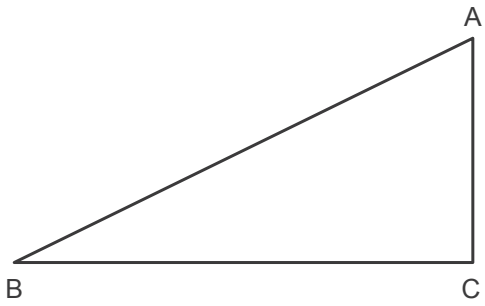
cos = BC ／ AB



初等関数の例：三角関数：tan（タンジェント）

以下のような直角三角形を考えた場合、タンジェントは以下のように表記されます。

tan = AC ／ BC



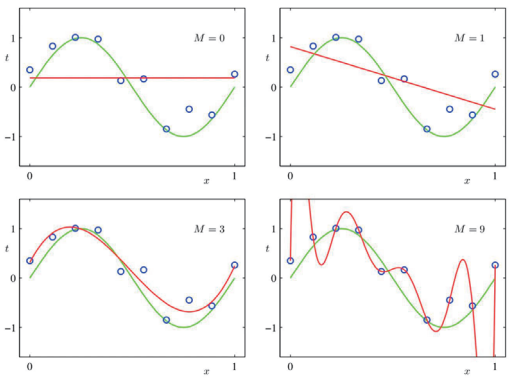
初等関数の例：多項式関数

多項式関数とは、変数xに対して以下のような関数のことをいいます。

$$f : x \mapsto a_m x^m + a_{m-1} x^{m-1} + , , , + a_1 x^1 + a_0 x^0$$

多項式は次数を上げることによって表現力が増していきます。
例えば、サイン波に対して次数mを増やしていった際のグラフを右に示します。

回帰に多項式を使用した場合、次数を増やしすぎると過剰にデータにフィッティングしてしまい、過学習が発生してしまいます（右図m=9の場合）。



参照：https://qiita.com/snow67675476/items/8aa72e5a0711d2afa754

演習問題

演習1：定数関数

- $f(x) = 3$となる定数関数のグラフを作成してください。また、$f(x) = 0.5$となる定数関数のグラフも作成してください。

演習2：指数関数

- $y = 3^4$を計算してください。
- $y = (2/3)^4$を計算してください。

演習3：対数関数

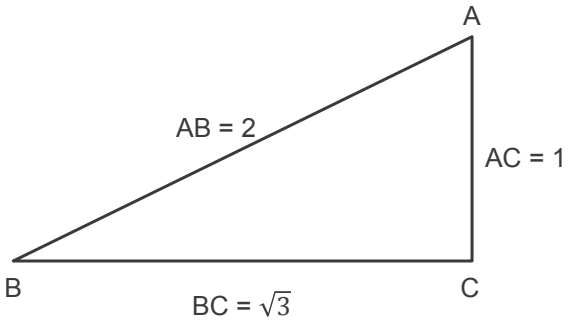
- 下図のようなxとyの組み合わせについて、グラフを作成してください。
- yから底をネイピア数とする数値（ln(y)）を計算し、グラフを作成してください。
- yから底を10とする数値（$\log_{10}y$）を計算し、グラフを作成してください。

x	y	ln(y)	log 10 y
1	10	2.30	1
2	100	4.61	2
3	1,000	6.91	3
4	10,000	9.21	4
5	100,000	11.51	5
6	1,000,000	13.82	6
7	10,000,000	16.12	7
8	100,000,000	18.42	8
9	1,000,000,000	20.72	9
10	10,000,000,000	23.03	10

演習4：三角関数

以下のような直角三角形があるとします。

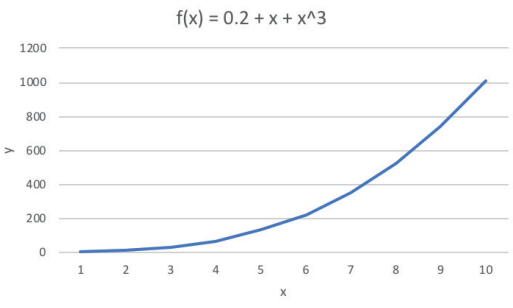
- sinBを計算してください。
- cosBを計算してください。
- tanBを計算してください。



演習5：多項式関数

- x = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}についてf(x) = 0.2 + x + x^3を計算し、グラフを作成してください。

x	y
1	2.2
2	10.2
3	30.2
4	68.2
5	130.2
6	222.2
7	350.2
8	520.2
9	738.2
10	1010.2



第5回：基礎数学5

アジェンダ

- 微分とは
- 微分の例
- 微分の定義
- 可微分性

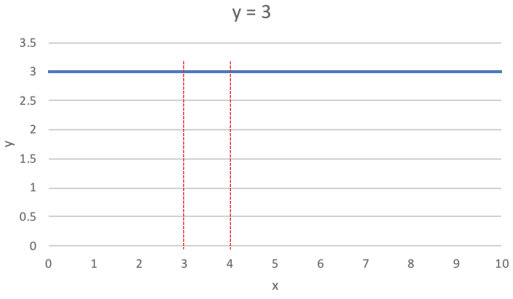
微分とは

- 微分とは、任意の関数の各点における変化の割合（傾き）を求めることです。
- 傾きは以下の式で定義されます。
  - 変化の割合（傾き）$= \frac{\text{yの増加量}}{\text{xの増加量}}$
- 上式の変化の割合のことを、導関数といいます。

微分の例：傾きがない関数

- 下図のような「y = 3」という関数を考えます。
- この関数ではxの値によらずyの値は3のため、xの変化に対するyの変化の割合は0となります。

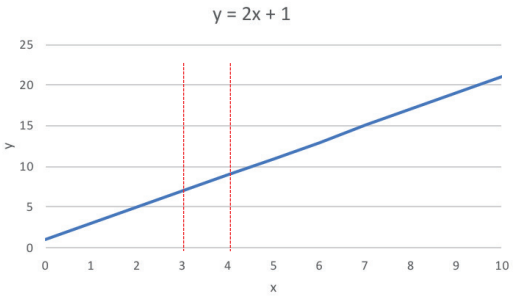
- 例えばxが3から4に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{0}{4-3} = 0$



微分の例：1次関数

- 下図のような「y = 2x + 1」という関数を考えます。
- この関数ではxの値が1増加すると、yの値は2増加します。

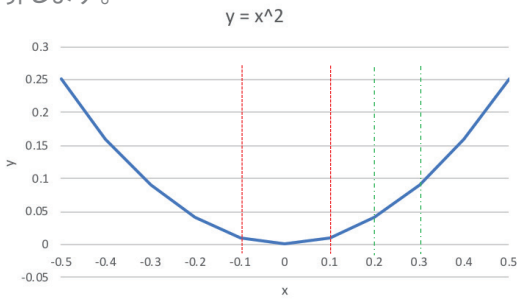
- 例えばxが3から4に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{9-7}{4-3} = 2$



微分の例：2次関数

- 下図のような「$y = x^2$」という関数を考えます。
- この関数ではxの増加量に対するyの増加量は、xの場所によって異なります。

- 例えばxが−0.1から0.1に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{0.01-0.01}{0.1-(-0.1)} = 0$
- 次に、xが0.2から0.3に増えたときの傾きを計算します。
  - $\frac{\Delta y}{\Delta x} = \frac{0.09-0.04}{0.3-0.2} = 0.5$



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

微分の定義

微分は、以下の式のように変化の割合だと述べました。

$$\text{変化の割合（傾き）} = \frac{\text{yの増加量}}{\text{xの増加量}}$$

また、変化の割合はxの場所によって変化することも学習しました。

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

微分の定義

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

$f(x)=x^2$のとき、上式にしたがって微分を実施すると、以下のようになります。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}=\lim_{h\to 0}\frac{(x+h)^2-x^2}{h}=\lim_{h\to 0}\frac{x^2+2hx+h^2-x^2}{h}=\lim_{h\to 0}(2x+h)=2x$$

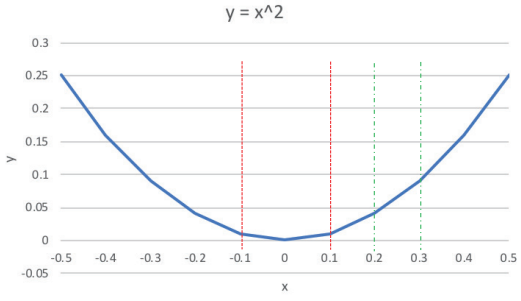
ここで求めた導関数を基に計算すると、

x=0.2のとき0.4

x=0.3のとき0.6

となります。「微分の例：2次関数」ではxが0.2から0.3に増えるときの傾きを計算し0.5となりました。

これは上の数字のちょうど真ん中にあることが確認できます。



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

可微分性

関数$f(x)$の任意の場所aにおける微分は、以下の式で表せました。

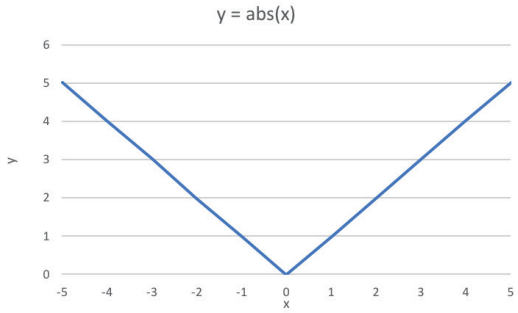
$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

絶対値関数$f(x)=|x|$の、x=0における傾きは

h＞0のときは1

h＜0のときは−1

となり、x=0で微分可能ではありません。



演習問題

演習1：定数関数の微分

- $f(x) = 4$となる定数関数の導関数を求めてください。

演習2：1次関数の微分

- y = 5x + 3の導関数を求めてください。

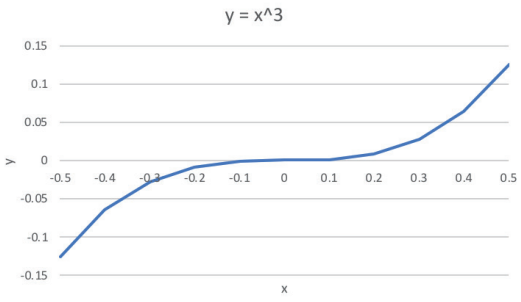
演習3：2次関数の微分

- y = 2x^{2}の導関数を求めてください。
- 求めた導関数より、x=1とx=2の傾きをそれぞれ求めてください。

演習4：3次関数の微分

- y = x3の導関数を求めてください。
- 求めた導関数より、x=0とx=0.3の傾きをそれぞれ求めてください。

x	y
-0.5	-0.125
-0.4	-0.064
-0.3	-0.027
-0.2	-0.008
-0.1	-0.001
0	0
0.1	0.001
0.2	0.008
0.3	0.027
0.4	0.064
0.5	0.125



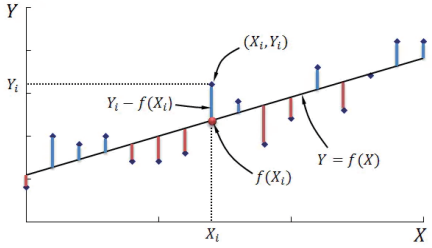
演習問題：機械学習への応用

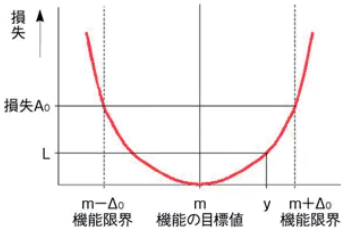
微分と機械学習

線形回帰モデルを構築することを考えているとします。
全てのデータに対し、線形回帰モデルの出力値$f(x_i)$が実際の値y_iに近いほど良いモデルだといえます。
例えば、「出力値$f(x_i)$と実際の値y_iの差の2乗の平均」を損失関数として（平均二乗誤差（MSE）といいます）、線形回帰モデルの重みの数値を変えながら、損失関数が最小値をとる重みを探索していきます。

$$\mathrm{MSE}\big(y - f(x_i)\big) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 : \text{平均二乗誤差}$$

その際には、MSEを重みで微分し、傾きを0に近づけていく操作を繰り返します。





参照：https://qiita.com/mine820/items/f8a8c03ef1a7b390e372

演習5：平均二乗誤差の算出

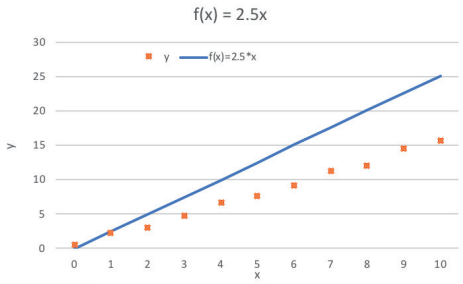
以下のようなデータセットがあります。yは、xを1.5倍して乱数を足して計算しています。
このyを近似する線形回帰モデルf(x)=w*xを求めていきます。

- 重みを2.5としてf(x)=2.5xを各xについて計算してください。
- 重みが2.5のときの平均二乗誤差を計算してください。
- 答えであるyと、計算したf(x)を図示してください。

x	y
0	0.44230588
1	2.224020809
2	3.068219681
3	4.671535779
4	6.729566858
5	7.584718889
6	9.144549573
7	11.27023797
8	12.11023737
9	14.46032741
10	15.75178752

x	y	f(x)=2.5*x	(y-f(x))^2
0	0.44230588	0	0.19563449
1	2.224020809	2.5	0.07616451
2	3.068219681	5	3.7317752
3	4.671535779	7.5	8.00020985
4	6.729566858	10	10.6957329
5	7.584718889	12.5	24.1599884
6	9.144549573	15	34.2862997
7	11.27023797	17.5	38.809935
8	12.11023737	20	62.2483543
9	14.46032741	22.5	64.6363354
10	15.75178752	25	85.5294341

30.2154422 平均二乗誤差

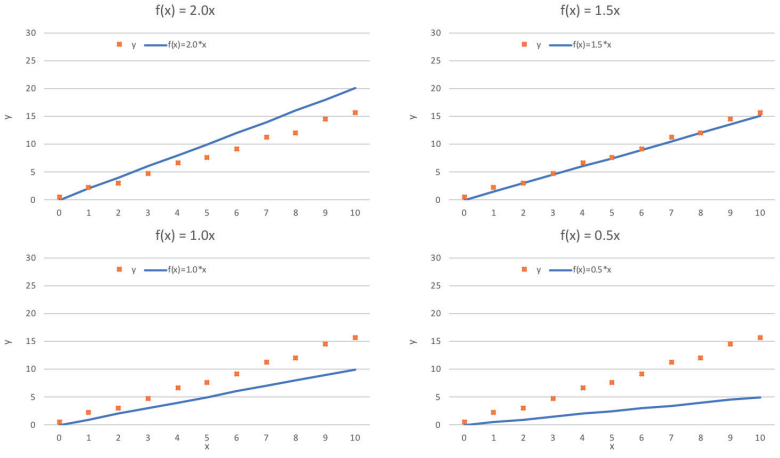


演習6：平均二乗誤差の算出（続き）

演習5のデータについて、以下を算出してください。

- 重みを2.5、2.0、1.5、1.0、0.5として各f(x)を計算してください。
- 上記の重みについて、それぞれ平均二乗誤差を計算してください。
- 上記の重みについて、それぞれ答えであるyと、計算したf(x)を図示してください。

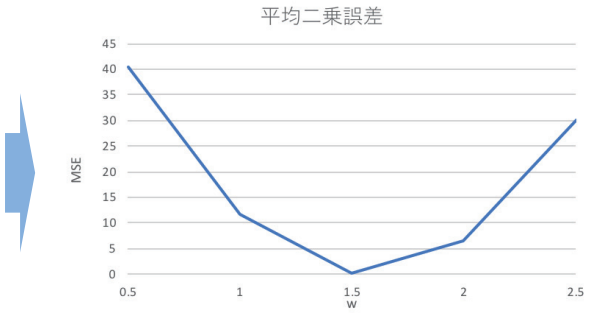
x	y
0	0.44230588
1	2.224020809
2	3.068219681
3	4.671535779
4	6.729566858
5	7.584718889
6	9.144549573
7	11.27023797
8	12.11023737
9	14.46032741
10	15.75178752



演習7：誤差関数の変動

演習5～6で算出した平均二乗誤差を、重み0.5～2.5を変数として図示してください。

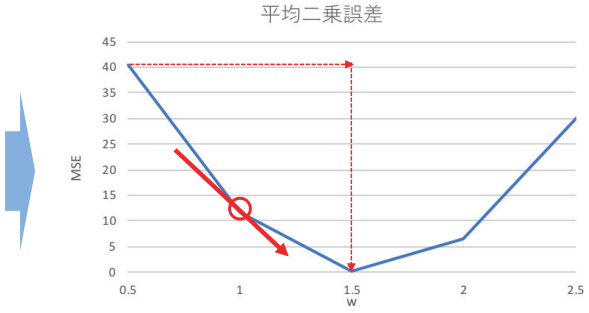
w	平均二乗誤差
0.5	40.4040289
1	11.6068822
1.5	0.30973552
2	6.51258885
2.5	30.2154422



演習8：誤差関数の微分

演習8で図示した平均二乗誤差について、重み1.0～2.0の範囲で傾きを算出して下さい。
yはもともと、xを1.5倍して乱数を足して計算した数値でした。
このyを近似する線形回帰モデルf(x)=w*xの導関数を考えた場合、重みwが1.5の場合に最も傾きが小さくなる（平均二乗誤差が小さくなる）ことが確認できます。

w	平均二乗誤差	傾き
0.5	40.4040289	
1	11.6068822	-40.09429335
1.5	0.30973552	-5.094293355
2	6.51258885	29.90570665
2.5	30.2154422	



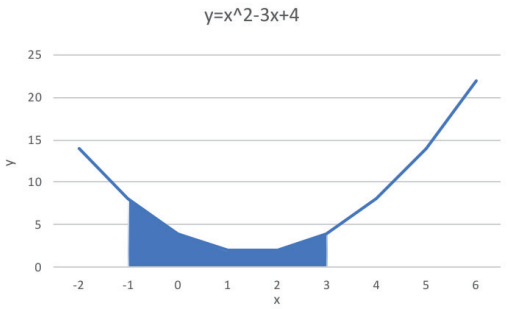
第6回：基礎数学6

アジェンダ

- 積分とは
- 積分の例
- 複雑な関数の定義
- 微分と積分の関係

積分とは

- 積分とは、任意の関数f(x)で囲まれた部分の面積を求めることを意味しています。
  - $\int_a^b f(x)dx$
- 例えば$f(x) = x^2 - 3x + 4$、a=−1、b=3の場合、下図の青い部分の面積を求めることができます。

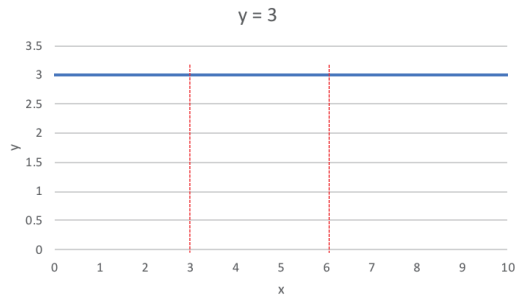


積分の例：傾きがない関数

- 下図のような「y = 3」という関数を考えます。
- この関数ではxの値によらずyの値は3のため、長方形の面積を求めることと同じになります。

- 例えばxが3から6の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_3^6 3dx = [3x]_3^6 = 3 \times 6 - 3 \times 3 = 9$

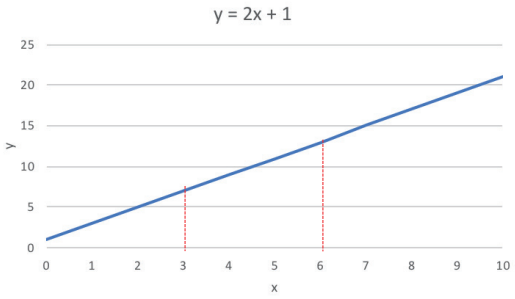
ここでは「3」を導関数とする原始関数「3x」を求めています。



積分の例：1次関数

- 下図のような「y = 2x + 1」という関数を考えます。
- この関数ではxの値が1増加すると、yの値は2増加します。

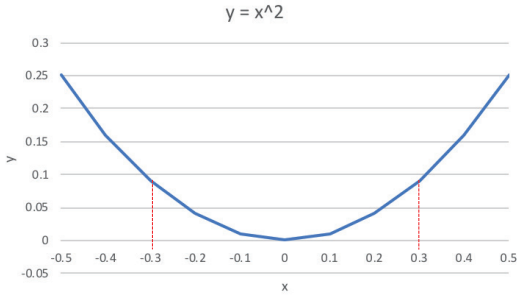
- 例えばxが3から6の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_3^6 (2x+1)dx = [x^2+x]_3^6 = (6\times 6+6)-(3\times 3+3) = 42-12 = 30$



積分の例：2次関数

- 下図のような「$y = x^2$」という関数を考えます。
- この関数ではxの増加量に対するyの増加量は、xの場所によって異なります。

- 例えばxが−0.3から0.3の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_{-0.3}^{0.3} x^2 dx = \left[\frac{1}{3}x^3\right]_{-0.3}^{0.3} = \left(\frac{0.3\times 0.3\times 0.3}{3}\right) - \left(\frac{(-0.3)\times(-0.3)\times(-0.3)}{3}\right) = 0.009 + 0.009 = 0.018$



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

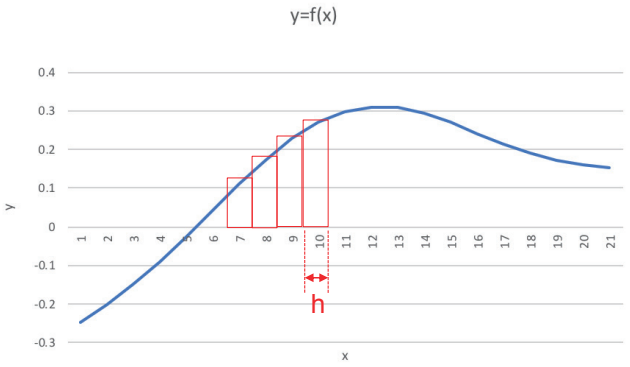
複雑な関数の積分

これまでの例では、導関数の原始関数を解析的に求めることができました。

解析的に求めることができない関数に対して面積を算出する際は、コンピュータプログラムなどで以下のようにして面積の近似値を求めます。

$$\sum_{i=a}^{b} f(i)h$$

hの間隔を徐々に狭くしていけば、上式の値は真の値に近づいていきます。



微分と積分の関係

「複雑な関数の積分」のページでは、面積の近似値をプログラムで求める方法を記載しましたが、より厳密に面積を求めていきます。

下図のオレンジ色の面積は、青い部分より大きく、青＋赤より小さいことがわかります。

$$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$$

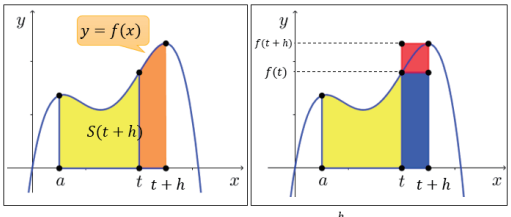
$$f(t) < \frac{S(t+h)-S(t)}{h} < f(t+h)$$

hの極限をとると、

$$f(t) < \lim_{h\to 0}\frac{S(t+h)-S(t)}{h} < \lim_{h\to 0} f(t+h) = f(t)$$

$$\lim_{h\to 0}\frac{S(t+h)-S(t)}{h} = f(t)$$

最後の式は、f(x)の導関数を求める式と同じであることが確認できます。



$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$

参照：https://atarimae.biz/archives/22721

演習問題

演習1：定数関数の積分

- $f(x) = 4$となる定数関数の原始関数を求めてください。
- 1<=x<=3の範囲で面積を求めてください。

演習2：1次関数の積分

- y = 5x + 3の原始関数を求めてください。
- 1<=x<=3の範囲で面積を求めてください。

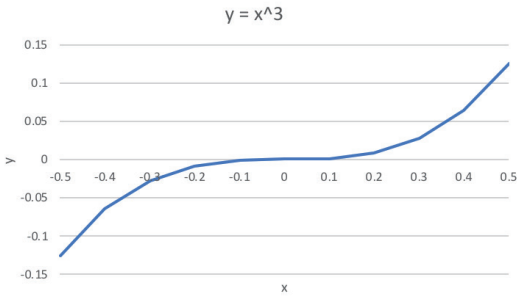
演習3：2次関数の積分

- y = $2x^2$の原始関数を求めてください。
- 1<=x<=3の範囲で面積を求めてください。

演習4：3次関数の積分

- y = x3の原始関数を求めてください。
- −0.4<=x<=0.4の範囲で面積を求めてください。

x	y
-0.5	-0.125
-0.4	-0.064
-0.3	-0.027
-0.2	-0.008
-0.1	-0.001
0	0
0.1	0.001
0.2	0.008
0.3	0.027
0.4	0.064
0.5	0.125



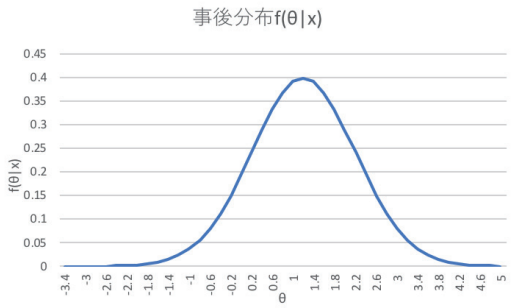
演習問題：機械学習への応用

積分と機械学習

ベイズ推定において下図のように事後分布が得られたとします。
事後分布の解釈を行うために、点推定量としてEAP（Expected a Posterioi：事後期待値）推定量というものを計算することがあります。

$$EAP\text{推定量} = \int f(\theta|x) \cdot \theta \, d\theta$$

本演習では、右表のデータを基にEAP推定量を計算します。



θ	f(θ|x)
-3.4	1.0141E-05
-3.2	2.4942E-05
-3	5.8943E-05
-2.8	0.00013383
-2.6	0.00029195
-2.4	0.0006119
-2.2	0.00123222
-2	0.00238409
-1.8	0.00443185
-1.6	0.00791545
-1.4	0.01358297
-1.2	0.02239453
-1	0.03547459
-0.8	0.05399097
-0.6	0.07895016
-0.4	0.11092083
-0.2	0.14972747
0	0.19418605
0.2	0.24197072
0.4	0.28969155
0.6	0.3332246

θ	f(θ|x)
0.8	0.36827014
1	0.39104269
1.2	0.39894228
1.4	0.39104269
1.6	0.36827014
1.8	0.3332246
2	0.28969155
2.2	0.24197072
2.4	0.19418605
2.6	0.14972747
2.8	0.11092083
3	0.07895016
3.2	0.05399097
3.4	0.03547459
3.6	0.02239453
3.8	0.01358297
4	0.00791545
4.2	0.00443185
4.4	0.00238409
4.6	0.00123222
4.8	0.0006119
5	0.00029195

演習5：事後分布とパラメータθの積

EAP推定量を求めるため、$f(\theta|x) \cdot \theta$を各データについて計算してください。

$$EAP\text{推定量} = \int f(\theta|x) \cdot \theta \, d\theta$$

θ	f(θ|x)	f(θ|x)*θ
-3.4	1.0141E-05	-3.44789E-05
-3.2	2.4942E-05	-7.98159E-05
-3	5.8943E-05	-0.000176829
-2.8	0.00013383	-0.000374725
-2.6	0.00029195	-0.000759062
-2.4	0.0006119	-0.001468565
-2.2	0.00123222	-0.002710882
-2	0.00238409	-0.004768176
-1.8	0.00443185	-0.007977327
-1.6	0.00791545	-0.012664723
-1.4	0.01358297	-0.019016157
-1.2	0.02239453	-0.026873436
-1	0.03547459	-0.035474593
-0.8	0.05399097	-0.043192773
-0.6	0.07895016	-0.047370095
-0.4	0.11092083	-0.044368334
-0.2	0.14972747	-0.029945493
0	0.19418605	0
0.2	0.24197072	0.048394145
0.4	0.28969155	0.115876621
0.6	0.3332246	0.199934762

演習6：事後分布とパラメータθとdθの積

EAP推定量を求めるために計算した$f(\theta|x)\cdot\theta$に、$d\theta$（右表の場合は0.2刻みなので0.2）をかけてください。

$$EAP\text{推定量} = \int f(\theta|x)\cdot\theta\ d\theta$$

θ	f(θ|x)	f(θ|x)*θ	f(θ|x)*θ*⊿θ
-3.4	1.0141E-05	-3.44789E-05	-6.89578E-06
-3.2	2.4942E-05	-7.98159E-05	-1.59632E-05
-3	5.8943E-05	-0.000176829	-3.53658E-05
-2.8	0.00013383	-0.000374725	-7.49449E-05
-2.6	0.00029195	-0.000759062	-0.000151812
-2.4	0.0006119	-0.001468565	-0.000293713
-2.2	0.00123222	-0.002710882	-0.000542176
-2	0.00238409	-0.004768176	-0.000953635
-1.8	0.00443185	-0.007977327	-0.001595465
-1.6	0.00791545	-0.012664723	-0.002532945
-1.4	0.01358297	-0.019016157	-0.003803231
-1.2	0.02239453	-0.026873436	-0.005374687
-1	0.03547459	-0.035474593	-0.007094919
-0.8	0.05399097	-0.043192773	-0.008638555
-0.6	0.07895016	-0.047370095	-0.009474019
-0.4	0.11092083	-0.044368334	-0.008873667
-0.2	0.14972747	-0.029945493	-0.005989099
0	0.19418605	0	0
0.2	0.24197072	0.048394145	0.009678829
0.4	0.28969155	0.115876621	0.023175324
0.6	0.3332246	0.199934762	0.039986952

演習7：事後分布とパラメータθの積の積分

EAP推定量を求めるために計算した$f(\theta|x)\cdot\theta d\theta$全データについて合計（積分）してEAP推定量を計算してください。

$$EAP\text{推定量} = \int f(\theta|x)\cdot\theta\ d\theta$$

※EAP推定量は1.2に近い値となります。

θ	f(θ|x)	f(θ|x)*θ	f(θ|x)*θ*⊿θ
-3.4	1.0141E-05	-3.44789E-05	-6.89578E-06
-3.2	2.4942E-05	-7.98159E-05	-1.59632E-05
-3	5.8943E-05	-0.000176829	-3.53658E-05
-2.8	0.00013383	-0.000374725	-7.49449E-05
-2.6	0.00029195	-0.000759062	-0.000151812
-2.4	0.0006119	-0.001468565	-0.000293713
-2.2	0.00123222	-0.002710882	-0.000542176
-2	0.00238409	-0.004768176	-0.000953635
-1.8	0.00443185	-0.007977327	-0.001595465
-1.6	0.00791545	-0.012664723	-0.002532945
-1.4	0.01358297	-0.019016157	-0.003803231
-1.2	0.02239453	-0.026873436	-0.005374687
-1	0.03547459	-0.035474593	-0.007094919
-0.8	0.05399097	-0.043192773	-0.008638555
-0.6	0.07895016	-0.047370095	-0.009474019
-0.4	0.11092083	-0.044368334	-0.008873667
-0.2	0.14972747	-0.029945493	-0.005989099
0	0.19418605	0	0
0.2	0.24197072	0.048394145	0.009678829
0.4	0.28969155	0.115876621	0.023175324
0.6	0.3332246	0.199934762	0.039986952

第7回：統計学の基本事項

アジェンダ

- 質的変数と量的変数
- 変数の尺度
- データの次元
- 時系列データ/クロスセクションデータ/パネルデータ

質的変数と量的変数

質的変数

- 質的変数とは、下記のようにカテゴリなどデータ間の「質」が異なる情報を保持するデータです。
  - 性別
  - 名前
  - 等級
  - 曜日
- 質的データは数値データではないため、そのままでは統計分析や機械学習に利用することができません。

質的変数のダミー変数化

- 質的変数を、下表のように0または1に置き換えて数値に変換する操作をダミー変数化といいます。
- 下の例では、曜日を7列のダミー変数に変換しています。また、投薬有無を2列のダミー変数に変換しています。機械学習データとして使用するためには、多重共線性を回避する上で、ダミー変数の列数を減らす（曜日であれば6列など）ほうが望ましいです。

店名称	定休日
A商店	日曜日
B商店	月曜日
C商店	火曜日
D商店	水曜日
E商店	木曜日
F商店	金曜日
G商店	土曜日

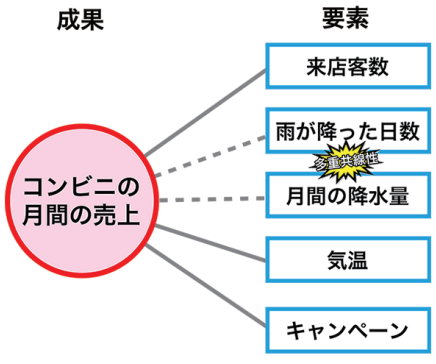
店名称	定休日_日	定休日_月	定休日_火	定休日_水	定休日_木	定休日_金	定休日_土
A商店	1	0	0	0	0	0	0
B商店	0	1	0	0	0	0	0
C商店	0	0	1	0	0	0	0
D商店	0	0	0	1	0	0	0
E商店	0	0	0	0	1	0	0
F商店	0	0	0	0	0	1	0
G商店	0	0	0	0	0	0	1

患者番号	投薬
001	あり
002	なし
003	あり

患者番号	投薬あり	投薬なし
001	1	0
002	0	1
003	1	0

多重共線性

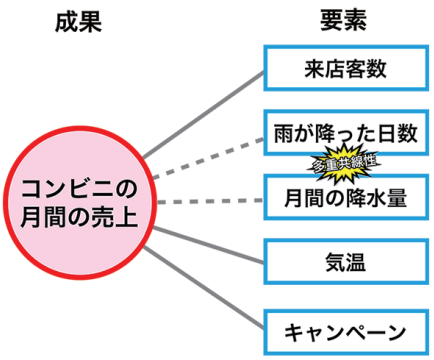
- 説明変数間で相関係数が高い時に多重共線性（multicollinearity）という問題が発生します。
- 多重共線性とは、モデル式の係数が不安定（符号と大きさが安定しない）になり、モデルの予測結果に対する係数の寄与度を正しく評価することができなくなってしまいます。



出展：https://xica.net/vno4ul5p/

多重共線性

- 多重共線性の回避策としては、相関が高い係数のどちらか一方をモデルから外すことが一般的です。



出展：https://xica.net/vno4ul5p/

量的変数

- 量的変数とは、下記のようにデータの数値が「量」の情報を保持するデータです。
  - 身長・体重
  - 面積
  - 年収
  - 年齢
- 量的データは数値データなので、そのまま統計分析や機械学習に利用することができます。

変数の尺度

変数の尺度

- 変数を性質に応じて下記の4つの尺度に分けて考えることがあります。
  - 名義尺度
  - 順序尺度
  - 間隔尺度
  - 比例尺度

変数の尺度：名義尺度

- 区分値のように、分類するための尺度のことです。以下のような例があります。
  - 男女
  - 都道府県
  - 学科

変数の尺度：順序尺度

- 値の順序、大小には意味がありますが、値の間隔には意味がない（計算ができない）尺度のことです。
  - 学年（小学1年生、2年生など）
  - 級位（1級、2級など）

変数の尺度：間隔尺度

- 値の間隔が等間隔になっており、その間隔に意味がある尺度のことです。ただし、数値間に比例関係はないものです。
  - 西暦。各年ともに等間隔だが、西暦2000年は西暦1000年の2倍とは言えない。
  - 気温。20℃は10℃の2倍ではない。

変数の尺度：比例尺度

- 0の概念があり、値の間隔と比率に意味がある尺度のことです。
  - 利益。利益0を中心として、利益−100万円（赤字）と利益100万円（黒字）は比例の関係。
  - 速度。速度0を中心として、速度10m/sは速度5m/sの2倍。

データの次元

データの次元

- 1つの観測対象に対して、1つの数値で表現したデータを1次元データ、2つの数値で表現したデータを2次元データ、N個の数値で表現したデータをN次元データといいます。

就職先	通勤時間
A株式会社	1時間10分
B株式会社	50分
C合同会社	1時間20分

1次元データ
「就職先」という対象を表現するために、「通勤時間」という1つのデータを使用している。

就職先	通勤時間	平均給与
A株式会社	1時間10分	500万円
B株式会社	50分	450万円
C合同会社	1時間20分	550万円

2次元データ
「就職先」という対象を表現するために、「通勤時間」、「平均給与」という2つのデータを使用している。

就職先	通勤時間	平均給与	退職金
A株式会社	1時間10分	500万円	なし
B株式会社	50分	450万円	あり
C合同会社	1時間20分	550万円	あり

3次元データ
「就職先」という対象を表現するために、「通勤時間」、「平均給与」、「退職金」という3つのデータを使用している。

就職先	通勤時間	平均給与	退職金	海外赴任
A株式会社	1時間10分	500万円	なし	あり
B株式会社	50分	450万円	あり	あり
C合同会社	1時間20分	550万円	あり	なし

4次元データ
「就職先」という対象を表現するために、「通勤時間」、「平均給与」、「退職金」、「海外赴任」という4つのデータを使用している。

データの次元と尺度

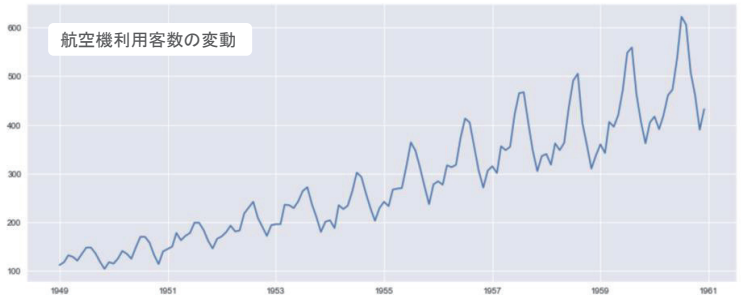
- 「就職先」という対象を表現するために、「通勤時間」、「平均給与」、「退職金」、「海外赴任」という4つのデータを使用しています。それぞれの次元の数値の尺度は、必ずしも同じものでなくても構いません。
- 下記の例でいうと、通勤時間と平均給与は比例尺度の数値で、退職金と海外赴任は名義尺度の数値です。

就職先	通勤時間	平均給与	退職金	海外赴任
A株式会社	1時間10分	500万円	なし	あり
B株式会社	50分	450万円	あり	あり
C合同会社	1時間20分	550万円	あり	なし

時系列データ/クロスセクションデータ/パネルデータ

時系列データ

- [Wikipediaより]時系列（じけいれつ、Time Series）とは、ある現象の時間的な変化を、連続的に（または一定間隔をおいて不連続に）観測して得られた値の系列[1]（一連の値）のこと。
- ある1つの項目を時間によって計測したデータのことです。



クロスセクションデータ

- ある時点での複数項目のデータをクロスセクションデータといいます。
- 同一時点での複数項目間の分析ができます。

日付	日経平均	ドル円
1月5日	15,100円	123円
1月6日	15,300円	121円
1月7日	15,400円	120円

クロスセクションデータ

時系列データ

パネルデータ

- クロスセクションデータを時間方向に拡張したデータをパネルデータといいます。
- 単一項目の時系列分析のみでなく、多項目間の連動も加味した分析が可能となります。

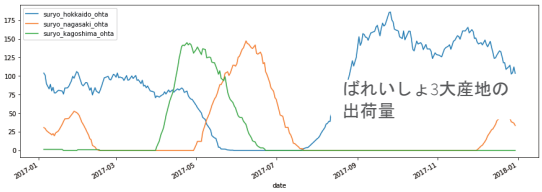
日付	日経平均	ドル円
1月5日	15,100円	123円
1月6日	15,300円	121円
1月7日	15,400円	120円

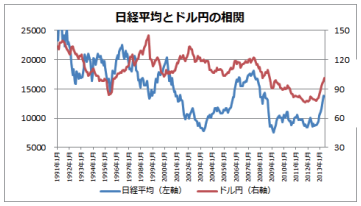
クロスセクションデータ

時系列データ

時系列データ、パネルデータの事例

- 時間とともに変動する時系列データには、様々な事例があります。

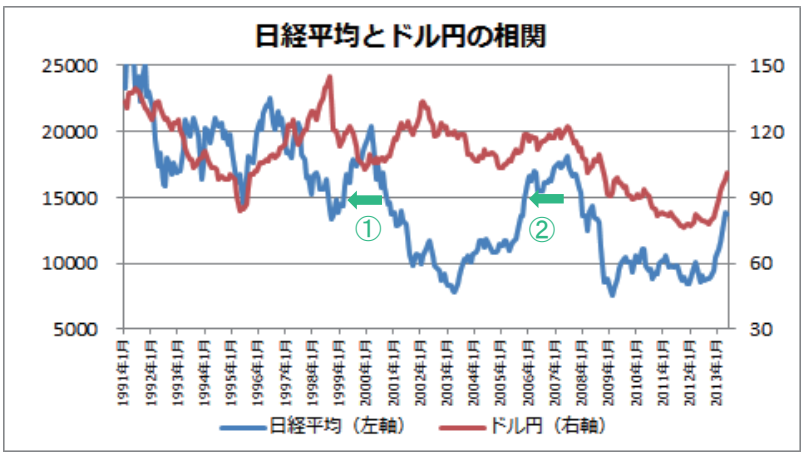




株価と為替レートの変動

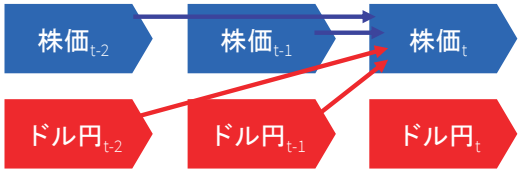
パネルデータの特徴

- ①と②は日経平均が同じ15,000円ですが、反発直後の①と、しばらく上げ続けている②とでは、その後の動きが異なります。
- ①と②でドル円の傾きも異なっています。



パネルデータの分析

- 過去の同じ変数からの影響を考慮する必要があります（自己相関）。
- 違う変数からの影響を考慮する必要があります（多変量）。
- その他、周期性やトレンドも考慮する必要があります。



演習問題

演習1：質的変数と量的変数

- 質的変数と量的変数の事例を、それぞれ3つずつ挙げてください。

演習2：変数の尺度

- 名義尺度、順序尺度、間隔尺度、比例尺度の事例を、それぞれ3つずつ挙げてください。

演習3：データの次元

- 住宅価格を対象とし、住宅価格を表現する3次元データの事例を挙げてください。

演習4：時系列/クロスセクション/パネルデータ

- 時系列データ、クロスセクションデータ、パネルデータの事例を挙げてください。

第8回：度数分布表と各種代表値

アジェンダ

- 度数分布表とヒストグラム
- 各種代表値

度数分布表とヒストグラム

度数分布表とは

- データの分布を観察するため、データを任意の範囲で区切り、その範囲に含まれるデータ数を見ることがあります。そのような表を度数分布表といいます。
- 下の例は、2019年の日本在住外国人の年齢ごと人数を度数分布表にしたものです。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：階級

- 度数を集計するための区間を表します。下の例では、[0–9歳]など10歳ごとに区切った年齢の幅が階級です。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：階級値

- 階級の代表値のことで、階級の下限値と上限値の平均値を表します。下の例では、[0–9歳]の階級に対する階級値は4.5となります。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：度数

- 各階級に含まれるデータ数のことです。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：相対度数

- 各階級の度数が全体に占める割合のことです。
- 下の例では、[30−39歳]の相対度数は572,874 / 2,667,139 = 0.21478971となります。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

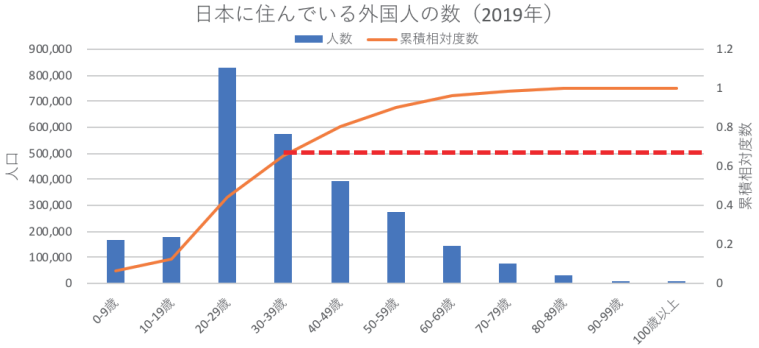
度数分布表：累積相対度数

- 相対度数を、その階級まで上から累積した値です。
- 一番最後の階級では、全ての相対度数を累積することになるので、累積相対度数1となります。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

ヒストグラム

- ヒストグラムは度数分布表をグラフにしたものです。
- 横軸が階級、縦軸が度数のグラフです。
- 累積度数とセットで表示すると、データ全体の分布がよりわかりやすくなります。下の例では、30代までの外国人人口が、全ての外国人人口の7割弱であることがわかります。



各種代表値

平均値

平均は、全てのデータを足してデータの数で割った値です。n個のデータの平均値$\bar{x}$（エックスバー）は以下の式で計算します。

$$\bar{x} = \frac{(x_1+x_2+\cdots+x_n)}{n}$$

データが度数分布表の場合は階級値と度数を使って、以下のように平均値の近似値を計算します。

$$\bar{X} = \frac{(f_1v_1+f_2v_2+\cdots+f_nv_n)}{n}$$

中央値（メディアン Median）

データを小さい順（もしくは大きい順）に並べたときに、ちょうど真ん中に来る値のことです。
例えば「1, 2, 2, 3, 5, 7, 9, 10, 12」というデータの場合、中央値は「5」です。

データの数が偶数の場合は、中央にある2つの値の平均が中央値となります。
例えば「1, 2, 2, 3, 5, 7, 9, 10, 12, 14」というデータの場合、中央値は(5 + 7) / 2 = 6です。

最頻値（モード Mode）

最もデータ数の多い値のことを最頻値といいます。
例えば「1, 2, 2, 3, 5, 7, 9, 10, 1」というデータの場合、モードは「2」です。

データが度数分布表の場合は、最も度数が大きい階級の階級値が最頻値となります。

分散

データAとデータBがあるとします。どちらも合計、平均が等しいとします。
合計と平均だけでは差がないデータAとBですが、同じ傾向を持ったデータとしてよいでしょうか？

例えば、データAの平均値から各点を引いた差と、データBの平均値から各点を引いた差は、それぞれ違う傾向を示しています。
ただし、データAとBの平均値からの差は、合計値は0になります。平均も0になってしまいます。
データAとBの傾向の違いを表現する方法はないでしょうか。

	データA	データB
	1	4
	2	4
	3	4
	4	4
	5	4
	6	4
	7	4
合計	28	28
平均	4	4

	データA	平均からの差
	1	3
	2	2
	3	1
	4	0
	5	-1
	6	-2
	7	-3
合計	28	0
平均	4	0

データB	平均からの差
4	0
4	0
4	0
4	0
4	0
4	0
4	0
28	0
4	0

分散

データAの平均値から各点を引いた差を二乗して合計し、データ数で割って平均を求めます。
データBに関しても同様の操作を実施します。
その結果算出した数値は、データAは4であり、データBは0となります。ばらつきの大きいデータAの方が数値が大きくなりました。

このように算出した「平均からの差の二乗の平均」を分散といい、データのばらつきの指標として用いられています。

	データA	平均からの差	差の二乗
	1	3	9
	2	2	4
	3	1	1
	4	0	0
	5	-1	1
	6	-2	4
	7	-3	9
合計	28	0	28
平均	4	0	4

データB	平均からの差	差の二乗
4	0	0
4	0	0
4	0	0
4	0	0
4	0	0
4	0	0
4	0	0
28	0	0
4	0	0

標準偏差

- 分散は「平均からの差の二乗の平均」で算出しました。
- この分散の平方根をとったものを、標準偏差といいます。

- 分散は元の数値を二乗しているので、単位は元の数値と合いません。平方根をとることで単位が元の数値と同じになり、平均値などと足し引きしてデータを表現することができるようになります。

演習問題

演習1：度数分布表

- オープンデータを検索し、度数分布表を作成してください。

※適切なオープンデータが見つからない場合「演習/第8回：度数分布表と各種代表値.txt 」に掲載されているデータを使用してください。2019年の日本在住外国人の人数を、都道府県ごと年齢層ごとに集計したデータです。

演習2：ヒストグラム

- 演習1で作成した度数分布表を基に、ヒストグラムを作成してください。

演習3：各種代表値

- 演習1で収集したデータに対し平均値、中央値、最頻値、分散、標準偏差を求めてください。

第9回：順列と組み合わせ、標本空間

アジェンダ

- 順列と組み合わせ
- 標本空間

順列と組み合わせ

順列とは

異なるn個の中から異なるr個を取り出し、かつ1列に並べた場合のパターン数のことです。

例えば3つのもの{A, B, C}から2つを取り出す順列を考えると、
AB, BA, AC, CA, BC, CA
の6つのパターンがあります。
ポイントは、ABとBAのように順序が違う場合も違うパターンとして考えることです。

順列の公式

先程の例 {A, B, C}から2つを取り出す順列で考えると、
　　1回目はA, B, Cの3つから選択可能、
　　2回目は残りの2つから選択可能、
なので、パターン数としては3×2=6となります。

これを一般化して「異なるn個の中から異なるrを取り出し並べる順列の数」は、以下のように計算できます。

$$nPr = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

「！」に記号は階乗と読み、1~nまでの積を表します。
例えば4!=4×3×2×1=24となります。

組み合わせとは

順列とは「異なるn個の中から異なるr個を取り出し、かつ1列に並べた場合のパターン数」のことでした。
例えば3つのもの{A, B, C}から2つを取り出す順列を考えると、
　　AB, BA, AC, CA, BC, CA
の6つのパターンがあり、ABとBAのように順序が違う場合も違うパターンとして考えていました。

組み合わせは、ABとBAは同じものとして考え、パターン数にはカウントしません。
2つのもの（例えばAとB）から構成される、カウントしないパターンの数は、${}_2P_2$=2で計算できます。

組み合わせの公式

例えば3つのもの{A, B, C}から2つを取り出す順列を考えると、

AB, BA, AC, CA, BC, CA

の6つのパターンがありましたが、カウントしないパターン数は2でした。

よって

6 ／ 2 = 3

で組み合わせのパターン数を計算することができます。

これを一般化して「異なるn個の中から異なるrを取り出し並べる組み合わせの数」は、以下のように計算できます。

$$\mathrm{nCr} = \frac{nPr}{rPr} = \frac{1}{r!}nPr = \frac{n!}{r!(n-r)!}$$

標本空間

標本空間

試行（実験）の結果として起こり得るすべての場合を要素とした集合を、標本空間といいます。
起こり得るすべての場合を、$\omega_i(i = 1, 2, , , n)$とするとき、標本空間は以下のように定義されます。

$$\Omega = \{\omega; \omega = (\omega_1, \omega_2, , , \omega_n)\}$$

例えば、サイコロを振ることを考えた場合、標本空間は以下のようになります。

$$\Omega = \{\omega; \omega = (1, 2, 3, 4, 5, 6)\} = \{1, 2, 3, 4, 5, 6\}$$

事象

標本空間の部分集合を事象(event)といいます。

例えば、サイコロを1回振って偶数が出る事象Aは以下のようになります。

A = {2, 4, 6}

サイコロを2回振って、合計が4になる事象Bは以下のようになります。

B = {(1,3), (2,2), (3,1)}

和事象

サイコロを例とした標本空間Ωの中で、事象AとBを以下のように定義します。

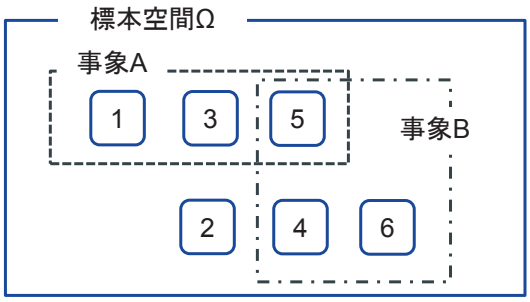
A = {1, 3, 5}：奇数が出る事象

B = {4, 5, 6}：4～6が出る事象

このとき事象AもしくはBが出る事象を和事象といいます。

A ∪ B = {1, 3, 4, 5, 6}

※和事象の記号「∪」は、「カップ」といいます。



積事象

サイコロを例とした標本空間Ωの中で、事象AとBを以下のように定義します。

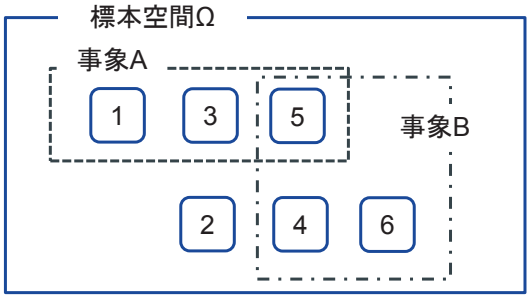
A = {1, 3, 5}：奇数が出る事象

B = {4, 5, 6}：4～6が出る事象

このとき事象AとBが同時に起こる事象を積事象といいます。

A ∩ B = {5}

※和事象の記号「∩」は、「キャップ」といいます。



余事象

サイコロを例とした標本空間Ωの中で、事象AとBを以下のように定義します。

A = {1, 3, 5}：奇数が出る事象

B = {4, 5, 6}：4〜6が出る事象

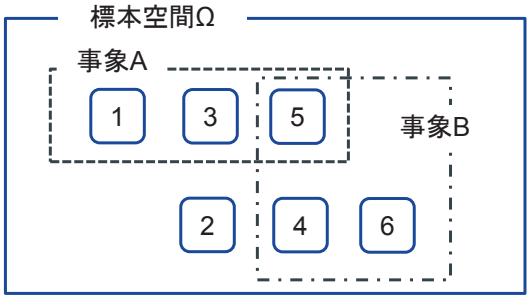
このとき事象Aが起きない事象を余事象といいます。

A^c = {2, 3, 6}

和事象や積事象にも余事象を考えることができます。

$(A \cup B)^c$ = {2}

$(A \cap B)^c$ = {1, 2, 3, 4, 6}



演習問題

演習1：順列

- 5の階乗を求めてください。
- 3の階乗を求めてください。
- {A, B, C, D, E}から異なる3つを取り出して並べる順列のパターン数を求めてください。

演習2：組み合わせ

- {A, B, C, D, E}から異なる3つを取り出す組み合わせのパターン数を求めてください。

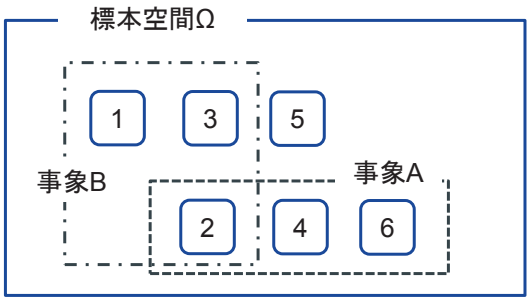
演習3：和事象

サイコロを例とした標本空間Ωの中で、事象AとBを以下のように定義します。

A = {2, 4, 6}：偶数が出る事象

B = {1, 2, 3}：1〜3が出る事象

このとき事象AもしくはBが出る和事象A ∪ Bを求めてください。



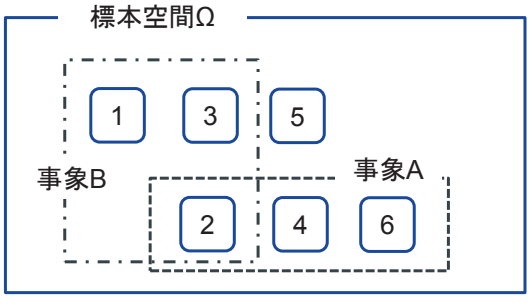
演習4：積事象

サイコロを例とした標本空間Ωの中で、事象AとBを以下のように定義します。

A = {2, 4, 6}：偶数が出る事象

B = {1, 2, 3}：1〜3が出る事象

このとき事象AとBが同時に起こる積事象A ∩ Bを求めてください。



演習5：余事象

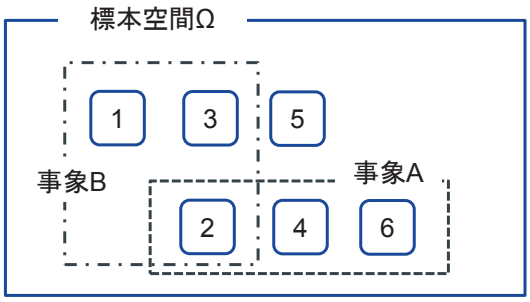
サイコロを例とした標本空間Ωの中で、事象AとBを以下のように定義します。

A = {2, 4, 6}：偶数が出る事象

B = {1, 2, 3}：1～3が出る事象

このとき事象AもしくはBが出る和事象の余事象$(A \cup B)^c$を求めてください。

このとき事象AとBが同時に起こる積事象の余事象$(A \cap B)^c$を求めてください。



演習6

- 自身で標本空間と事象を定義し、和事象、積事象、余事象を求めてください。

第10回：確率変数

アジェンダ

- 確率変数
- 確率分布
  - 離散型確率分布
  - 連続型確率分布
- 期待値
- チェビシェフの不等式

確率変数

確率変数とは

ある現象がいろいろな値を取り得るとき、取り得る値全体を確率変数といいます。

例えば、サイコロを振ったときに出る目は[1, 2, 3, 4, 5, 6]のいずれかとなります。
この場合、確率変数Xは

$$\mathrm{X} = 1, 2, 3, 4, 5, 6$$

と表します。

確率変数をXと置くことで、サイコロの目が取りうる値の確率を、以下のように記載することができます。

$$\mathrm{P}(x) = \frac{1}{6}(\mathrm{X} = 1, 2, 3, 4, 5, 6)$$

サイコロを振って4が出る確率は以下のように書きます。

$$\mathrm{P}(x = 4) = \frac{1}{6}$$

離散型の確率変数

離散型確率変数は、「とびとびの値」を指します。
隣り合った数値の間には、数値は存在しません。
例えばサイコロの目、コインの裏表、ルーレットの番号などが該当します。

連続型の確率変数

連続型確率変数は、「連続した値」を指します。
例えば速度であれば、5km/hと6km/hの間には5.1km/hや5.01km/h、5.0001km/hなど無数の値が存在します。

その他の連続確率変数には温度、湿度、高度、体重などがあります。

確率分布

確率分布とは

確率変数のそれぞれの値に対し、その確率変数をとる確率の分布のことです。

離散型確率変数に対する確率分布として、以下のような確率分布があります。

- ポアソン分布
- 二項分布
- 幾何分布
- 一様分布

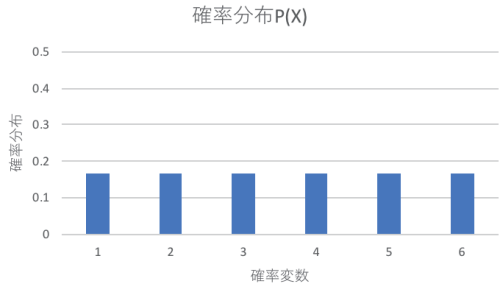
連続型確率変数に対する確率分布として、以下のような確率分布があります。

- 正規分布
- 指数分布
- 一様分布

離散型確率分布

確率変数が離散型の場合の確率分布を、離散型確率分布といいます。
サイコロの例だと、以下のようになります。
確率変数Xを横軸、Xが起きる確率をP(X)とすると、値は全て1/6となります。

確率変数X	1	2	3	4	5	6
確率分布P(X)	1/6	1/6	1/6	1/6	1/6	1/6



離散型確率分布と確率質量関数

確率変数Xが離散型の場合の確率分布P(X)を、離散型確率分布といいました。
確率分布P(x)を関数f(x)で表現した場合、f(x)を「確率質量関数」といいます。

サイコロの例だと、以下のようになります。

$$\sum_{i=1}^{6} P(X = x_i) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

連続型確率分布

確率変数が連続型の場合の確率分布を、連続型確率分布といいます。

サイコロの例は離散型確率変数のため、値は全て1/6となりました。
仮に、1～6までの連続確率変数を考えてみます
1～6の間には、1.1、1.01、1.001、、、と無数の数が存在します。

サイコロの目のように6つの値を持つ離散型確率変数であれば、目が1になる確率は

$$P(X=1)=\frac{1}{6}$$

となります。ですが、連続型確率変数であれば値は無数にあるため、

$$P(X=1)=\frac{1}{\infty}=0$$

となります。

連続型確率分布と確率密度関数

確率変数Xが連続型の場合の確率分布P(X)を、連続型確率分布といいました。
前出の様に、確率変数が連続型の場合には、確率変数が特定の値をとる確率は0になることから、縦軸は確率ではなく「確率密度」という考え方を使います。
確率密度は、連続型確率変数が取りうる範囲内での、特定の値の「相対的な出やすさ」を表しています。

連続型確率分布P(x)を関数f(x)で表現した場合、f(x)を「確率密度関数」といいます。

連続型確率分布と確率密度関数

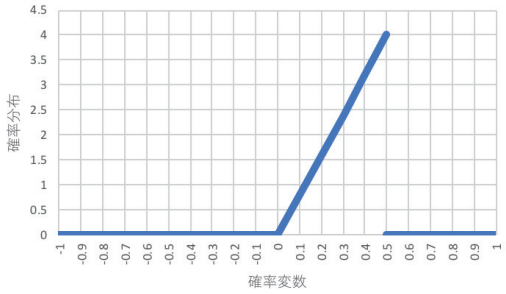
例えば、以下のような確率変数Xが$0 \leq X \leq 0.5$の範囲で8xとなる確率密度関数を考えます。

$$f(x) = \begin{cases} 8x(0 \leq X \leq 0.5) \\ 0(X < 0, X > 0.5) \end{cases}$$

確率変数Xが0.1、0.4のときの値は下記のようになります。

$$f(0.1) = 0.8$$
$$f(0.4) = 3.2$$

この確率密度関数です、X=0.4の状態は、X=0.1の状態より4倍起こりやすいということが言えます。



期待値

期待値とは

期待値とは、1回の試行で得られる値の平均値のことです。
得られうるすべての値（すべての確率変数）とそれが起こる確率の積を足し合わせて計算できます。

離散型確率変数の期待値

離散型確率変数の期待値は、確率変数がとり得る値に対応する確率を掛け、掛けた結果を全て足します。

$$\mathrm{E}(X) = \sum_{i=1}^{n}(x_i \cdot P_i)$$

サイコロの例だと、期待値は以下のようになります。

$$\mathrm{E}(X) = \sum_{i=1}^{n}(x_i \cdot P_i)$$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

確率変数X	1	2	3	4	5	6
確率分布P(X)	1/6	1/6	1/6	1/6	1/6	1/6
X・P(X)	1/6	1/3	1/2	2/3	5/6	1
					ΣX・P(X)	3.5

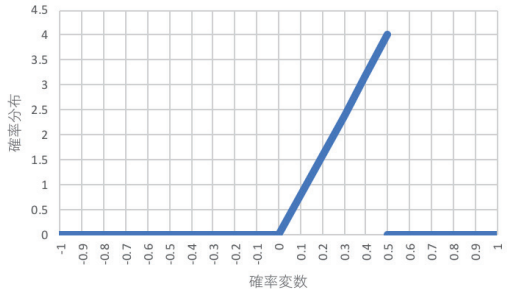
連続型確率変数の期待値

連続型確率変数の期待値は、積分によって計算します。

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

例えば確率密度関数f(x)=8x、確率変数Xが0から0.5の値を取る場合は、以下のようになります。

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$
$$= \int_{0}^{0.5} x \cdot 8xdx$$
$$= \left[\frac{8}{3}x^3\right]_0^{0.5}$$
$$= \frac{1}{3}$$



チェビシェフの不等式

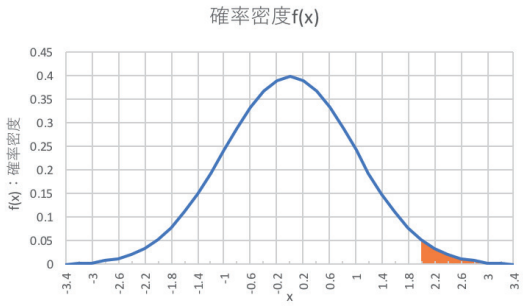
チェビシェフの不等式とは

チェビシェフの不等式とは、確率変数Xの平均値をμ、標準偏差をσとしたときに、以下で与えられる不等式のことです。

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

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

任意の確率分布において、μ +kσ 以上、またはμ −kσ 以下の確率がどれくらいなのか見当をつけるのに用います。



チェビシェフの不等式とは

例えば標準正規分布（平均μ=0、標準偏差σ=1）を考えます。
Xが2σ以上となる場合は以下のようになります。

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

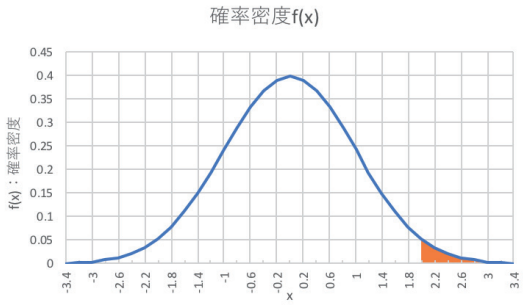
$$P(X \geq 0 + 2 \cdot 1) \leq \frac{1}{2^2}$$

$$P(X \geq 2) \leq \frac{1}{4}$$

標準正規分布においてXが2以上となる部分の面積は約2.3%となります。

$$P(X \geq 2) \cong 0.023 \leq \frac{1}{4}$$

となり、チェビシェフの不等式が成り立っています。



演習問題

演習1：離散型の確率変数

- 離散型確率変数の例を挙げてください。
- 確率変数と確率分布の表を作成してください。
- 期待値を求めてください。

演習2：連続型の確率変数

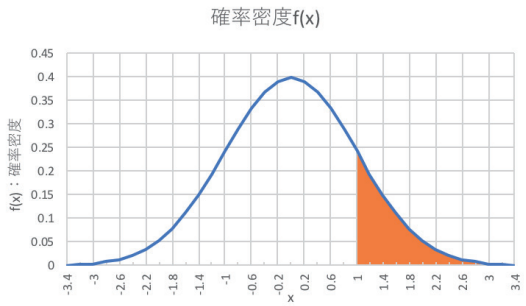
- 連続型確率変数の例を挙げてください。
- 確率変数と確率分布の表を作成してください。
- 期待値を求めてください。

演習3：チェビシェフの不等式

標準正規分布（平均 μ =0、標準偏差 σ =1）において、k=1の場合にチェビシェフの不等式が成り立つことを確認してください。

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

※標準正規分布における確率は、[https://to-kei.net/distribution/normal-distribution/table/]などを参照してください。



第11回：代表的な確率分布

アジェンダ

- 幾何分布
- 超幾何分布
- ベルヌーイ分布と二項分布
- ガウス分布
- ポアソン分布
- 一様分布

幾何分布

幾何分布の前提知識：ベルヌーイ試行

例えば、コイン投げのように結果が2通りにしかならない確率実験のことをベルヌーイ試行といいます。
試行を繰り返したとき、何度試行を繰り返しても結果が起こる確率は同じです（各試行の結果は互いに独立）。

幾何分布とは

成功確率をpとしたベルヌーイ試行を繰り返すとします。初めて成功するまでの試行回数Xが従う確率分布を「幾何分布（きかぶんぷ）」といいます。

成功確率がpの試行において、k回目で初めて成功する確率は次の式で計算できます。

$$P(X = k) = (1 - p)^{k-1}p : k = 1, 2, 3, , ,$$

幾何分布の例

さいころを投げて3が出る確率は1/6です。4投目で初めて3が出る確率は次のように計算できます。

$$P(X = 4) = \left(1 - \frac{1}{6}\right)^3 \frac{1}{6} \cong 0.096$$

初めて3が出るまでに投げる回数	確率
1	0.167
2	0.139
3	0.116
4	0.096
5	0.080
6	0.067
7	0.056
8	0.047
9	0.039
10	0.032

無記憶性

「ある事象の発生確率は、その事象が発生する前の情報の影響を受けない」という性質を、無記憶性といいます。
コイン投げの例であれば、n回目で表が出る確率は過去のコイン投げの影響を受けることが無く、独立した結果となる、ということです。

超幾何分布

超幾何分布とは

[Wikipediaより]超幾何分布（ちょうきかぶんぷ、英: hypergeometric distribution）とは、成功状態をもつ母集団から非復元抽出したときに成功状態がいくつあるかという確率を与える離散確率分布の一種である。

例えば、
・箱の中に玉がN個あり、M個が赤い玉、N−M個が白い玉とします。
・玉を箱から取り出して色を調べ、玉を元に戻さない非復元抽出でn回調べるとします。
・このときの赤い玉の個数を確率変数Xとします。
X=k個となる確率は超幾何分布に従います。

超幾何分布の確率質量関数

母集団の要素数をN（箱の中のN個の玉）、属性Aの要素数をM（M個の赤い玉）、標本の要素数をn（n回調べる）とします。
標本を非復元抽出した場合、標本に含まれる属性Aの要素数k（n回調べたうち、k個の赤い玉が出た）とすると、確率変数Xは超幾何分布HG(N,M,n)に従い、その確率質量関数は、以下の式で表されます。

$$f(k) = P(X = k) = \frac{{}_M C_k \times {}_{N-M} C_{n-k}}{{}_N C_n}$$

超幾何分布の確率質量関数の導出

前ページの確率質量関数は、以下のようにして導出されます。

「母集団の要素数をN（箱の中のN個の玉）、属性Aの要素数をM（M個の赤い玉）、標本の要素数をn（n回調べる）とします。」なので、N個のものからn個のものを取り出す組合せ数は以下の式となります。

$$ {}_N C_n = \frac{N!}{n!(N-n)!} $$

n個のうち、k個がA（赤い玉）となる組合せ数は、M個からk個のAを取り出す組合せ数と、N−M個からn−k個のB（白い玉）を取り出す組合せ数の積となります。

$$ {}_M C_k \times {}_{N-M} C_{n-k} = \frac{M!}{k!(M-k)!} \times \frac{(N-M)!}{(n-k)!\{(N-M)-(n-k)\}!} $$

下の式を上の式で割ると、前ページの確率質量関数となります。

ベルヌーイ分布と二項分布

ベルヌーイ分布とは

ベルヌーイ分布とは、「成功/失敗」、「表/裏」「勝ち/負け」のように2種類の結果が得られる実験（ベルヌーイ試行）の結果を0と1で表した分布のことです。
1である確率をpとした場合、0である確率は1−pとなります。

ベルヌーイ分布の確率質量関数

ベルヌーイ分布の確率質量関数は、以下の式となります。

$$f(k) = P^k(1-P)^{(1-k)}$$

「k」は成功か失敗を表します（1が成功で0が失敗）。「P」は成功確率を表します。
「k=1（成功）」のときにP（成功確率）、
「k=0（失敗）」のときに1−P（失敗確率）、
となります。

二項分布

ベルヌーイ試行をn回行い、成功する回数Xが従う確率分布を「二項分布」といいます。
n回のベルヌーイ試行を行いk回成功する確率は次の式から計算できます。

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

例えばコイン投げを例にとります。表が出る確率p=0.5なので、
10回中4回表が出る確率は以下のように計算できます。

$$P(X = 4) = {}_{10}C_4 0.5^4 (1-0.5)^{10-4} = 0.205$$

表の出る回数	確率
0	0.001
1	0.010
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.010
10	0.001

ガウス分布

ガウス分布とは

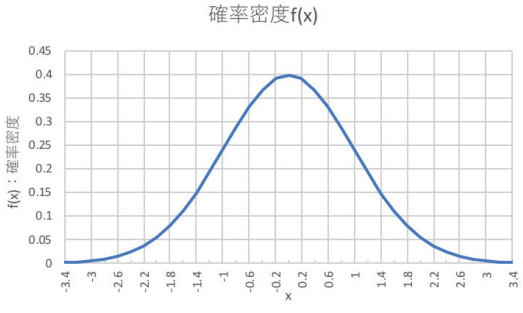
[Wikipediaより]確率論や統計学で用いられる正規分布（せいきぶんぷ、英: normal distribution）またはガウス分布（英: Gaussian distribution）は、平均値の付近に集積するようなデータの分布を表した連続的な変数に関する確率分布である。

ガウス分布の確率密度関数

ガウス分布に従う確率変数Xの確率密度関数は以下の式で表されます。

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

「σ」は標準偏差、「μ」は平均値を表します。
標準偏差σが1、平均μが0のガウス分布を、標準正規分布といいます。



ポアソン分布

二項分布とポアソン分布の関係

ベルヌーイ試行をn回行い、成功する回数Xが従う確率分布を「二項分布」といいます。
n回のベルヌーイ試行を行いk回成功する確率は次の式から計算できました。

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

ここで

$$\lambda = np$$

として λ は一定とします。試行回数nを大きくしていく（pが0に近づく）と以下のようになります。

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

このように「単位時間あたりに平均 λ 回起こる現象が、単位時間にk回起きる確率」のことをポアソン分布といいます。

ポアソン分布の例

鵜戸神宮には、岩穴に運玉を投げ入れるという運試しスポットがあります。
運試しは、運玉を200個投げると1回岩穴に入るとします。このとき、運玉を10回投げたときに、岩穴に運玉が1個入る確率はいくつでしょうか？（成功確率はポアソン分布に従うとします）

画像：Wikipediaより
https://ja.wikipedia.org/wiki/%E9%B5%9C%E6%88%B8%E7%A5%9E%E5%AE%AE

ポアソン分布の例

鵜戸神宮には、岩穴に運玉を投げ入れるという運試しスポットがあります。
運試しは、運玉を200個投げると1回岩穴に入るとします。このとき、運玉を10回投げたときに、岩穴に運玉が1個入る確率はいくつでしょうか？（成功確率はポアソン分布に従うとします）

岩穴に入る確率はp=1/200、運玉を投げた回数はn=10回なので、

$$\lambda = np = 10 \times \frac{1}{200} = 0.05$$

となります。したがって

$$P(X = 1) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{0.05 \times e^{-0.05}}{1!} = 0.0475$$

となり、約4.8%であることがわかります。

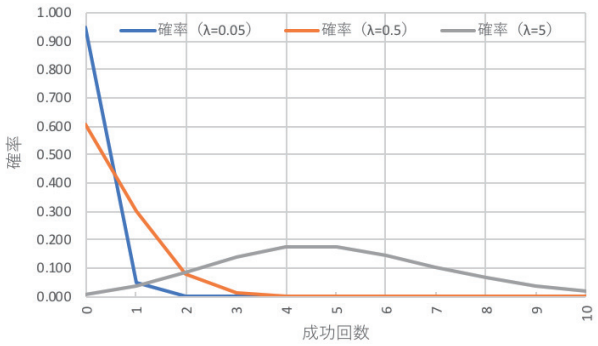
画像：Wikipediaより
https://ja.wikipedia.org/wiki/%E9%B5%9C%E6%88%B8%E7%A5%9E%E5%AE%AE

成功確率が変動するとどうなるか

岩穴に入る確率が200個に1個の場合（λ=0.05）の成功確率を扱ってきました。
もし成功確率が20個に1個、2個に1個であれば、成功確率は下のグラフの様に変動します。

成功確率が大きくなるにしたがって、確率のピークが右に移動し、左右対称の形（正規分布）に近づいていきます。

ポアソン分布の例



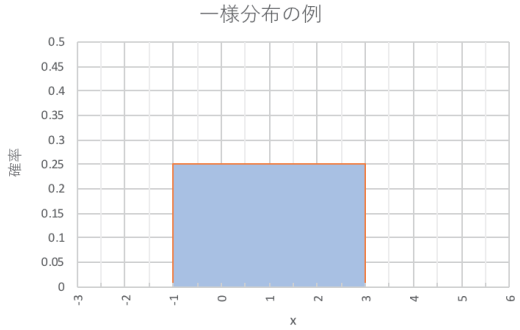
一様分布

一様分布とは

[Wikipediaより]一様分布（いちようぶんぷ）は、離散型あるいは連続型の確率分布である。サイコロを振ったときの、それぞれの目の出る確率など、すべての事象の起こる確率が等しい現象のモデルである。

区間[a, b]上の確率分布は以下のように表されます。

$$f(x) = f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x\text{が上記以外} \end{cases}$$



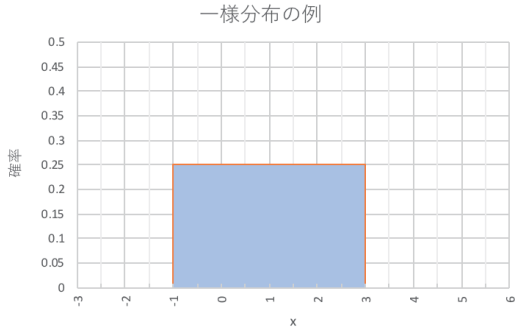
一様分布の期待値

区間[a, b]で期待値を計算します。

$$\mathrm{E}[X] = \int_a^b x f(x)\,dx = \int_a^b x \frac{1}{b-a}\,dx$$

$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2}$$

$$= \frac{a+b}{2}$$



一様分布の分散

区間[a, b]で分散を計算します。

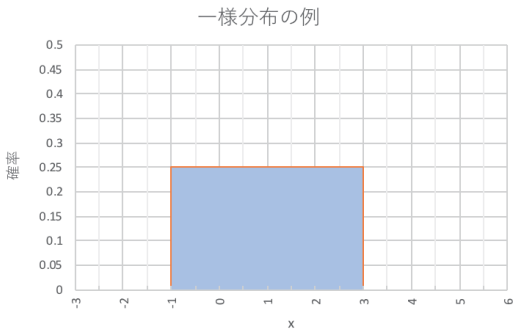
$$\mathrm{V}[X] = \int_a^b x^2 f(x)\,dx - E[X]^2$$

$$= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b - \left(\frac{a+b}{2}\right)^2$$

$$= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4}$$

$$= \frac{a^2-2ab+b^2}{12}$$

$$= \frac{(a-b)^2}{12}$$



演習問題

演習1：幾何分布

- さいころを投げて3が出る確率は1/6です。7投目で初めて3が出る確率を計算してください。
- 8投目で初めて3が出る確率を計算してください。

演習2：ベルヌーイ分布

- ベルヌーイ分布の例を挙げてください（ベルヌーイ分布とは、「成功/失敗」、「表/裏」「勝ち/負け」のように2種類の結果が得られる実験（ベルヌーイ試行）の結果を0と1で表した分布のことです。）

演習3：二項分布

- コイン投げで表が出る確率p=0.5とし、10回中7回表が出る確率を計算してください。
- 10回中8回表がでる確率を計算してください。

演習4：ポアソン分布

- 鵜戸神宮の運試しにおいて、運玉を20個投げると1回岩穴に入るとします。このとき、運玉を10回投げたときに、岩穴に運玉が1個入る確率はいくつでしょうか？（成功確率はポアソン分布に従うとします）

第12回：多次元の確率分布

第12回の目的

- 確率変数が2つ以上ある場合に、それぞれの確率変数がとる値とその確率の分布を「同時確率分布」といいます。
- 確率変数が離散型の場合には「離散型同時確率分布」といい、確率変数が連続型の場合には「連続型同時確率分布」といいます。

- 第12回講義においては確率変数が2つの場合の同時確率分布について学習します。

アジェンダ

- 離散型同時確率分布
- 連続型同時確率分布
- 確率変数の独立性
- 2変数のガウス分布

離散型同時確率分布

離散型同時確率分布とは

2つの離散型確率変数XとYが、それぞれある値をとるときの確率を表したものを「離散型同時確率分布」といいます。

例えば、男子20名、女子20名のあるクラスがあるとします。生徒の居住地区を表にしてみました。
性別をX、居住地区をYとすると、2つの離散型確率変数とみなせます。

	A地区	B地区	C地区	D地区	計
男子	4	6	6	4	20
女子	6	8	4	2	20
				総合計	40

全生徒40人に対する各マスの数字の割合を計算してみました。これは2つの離散型確率変数XとYがそれぞれの値を同時にとる、離散型同時確率分布となります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

離散型同時確率分布とは

2つの確率変数からなる同時確率分布は、以下のように表記します。

$$f(x_i, y_j) = P(X = x_i, Y = y_j) \quad i = 1,2,3,,,;j = 1,2,3,,,$$

例えば、男子でD地区に住む生徒の確率は、以下のようになります。

$$P(X = \text{男子}, Y = \text{D地区}) = 0.1$$

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

ここで$f(x_i, y_j)$のことを同時確率関数といいます。各iとjについて全ての確率を足すと総和は1になります。

$$\sum_i \sum_j f(x_i, y_j) = 1$$

周辺確率分布

性別X、居住地区Yのそれぞれの値について、確率の合計を計算してみます。
男子の割合は0.5、A地区に居住する生徒の割合は0.25であることがわかります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5
計	0.25	0.35	0.25	0.15	1.00

このようにある1つの確率変数を固定し、別の確率変数が取りうる全ての確率を合計したものを周辺確率分布といいます。

$$f_x(x_i) = \sum_j f(x_i, y_j) = P(X = x_i) \quad i = 1,2,3,,,$$
$$f_y(y_j) = \sum_i f(x_i, y_j) = P(Y = y_j) \quad j = 1,2,3,,,$$

ここで、$f_x(x_i)$と$f_y(y_j)$をそれぞれXとYの周辺確率関数といいます。

連続型同時確率分布

連続型同時確率分布とは

XとYが連続型確率変数であるとき、それぞれある値をとるときの確率を表したものを「連続型同時確率分布」といいます。

XとYの同時確率分布を表す関数を「同時確率密度関数」といいます。

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$$

確率の総和は1になるため、同時確率密度関数に関して以下の式が成り立ちます。

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

連続型確率変数XとYの周辺確率密度関数は、以下の式で求めることができます。

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$
$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y) = \begin{cases} x+y \ (0 \leq x \leq 1, 0 \leq y \leq 1) \\ 0 \ (\text{上記以外の} x,y \text{の場合}) \end{cases}$$

同時確率変数XとYの全範囲についての確率を求めてみます。

$$P(0 \leq x \leq 1, 0 \leq y \leq 1) = \int_0^1 \int_0^1 (x+y)\,dx\,dy = \int_0^1 \left[\frac{x^2}{2} + yx\right]_0^1 dy = \int_0^1 \left(\frac{1}{2} + y\right) dy$$

$$= \left[\frac{y}{2} + \frac{y^2}{2}\right]_0^1 = 1$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\le x\le 1, 0\le y\le 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

Xの周辺確率密度関数を求めてみます。

$$f_x(x)=\int_0^1 (x+y)\,dy=\left[x+\frac{y^2}{2}\right]_0^1=x+\frac{1}{2}$$

確率変数の独立性

独立な確率変数とは

2つの確率変数XとYの同時確率分布（同時確率密度関数）f(x, y)が、それぞれの確率変数の周辺確率分布（周辺確率密度関数）g(x)とh(y)の積に分解できる時、その2つの確率変数は独立(independent)であると言います。

$$f(x,y) = g(x)h(y)$$

直感的な理解としては、「XとYの動きは、お互いに影響を及ぼさない」ということです。

共分散とは

共分散とは、二組の対応するデータの関係を表す指標です。

例えば、各地区における男女別の生徒数の関係を考えてみます。
共分散を参照すると、「男子生徒の人数が多い地区は、女子生徒の人数も多いのか？」などの傾向を分析することができます。

生徒数

	A地区	B地区	C地区	D地区	平均
男子	10	60	50	20	35
女子	6	40	30	10	21.5

共分散の計算

共分散の定義は「[Xの偏差 × Yの偏差]の平均」です。
また、偏差とは「平均との差」のことです。

男子生徒の各地区の平均人数は35人、女子生徒は21.5人です。
A地区における男子生徒の偏差と女子生徒の偏差は、それぞれ

10 − 35 = −25 : 男子生徒の偏差

6 − 21.5 = −15.5 : 女子生徒の偏差

各地区について男女の偏差を計算し、それらの
平均を取ります。

(387.5 + 462.6 + 127.5 + 172.5) / 4 = 287.5

生徒数

	A地区	B地区	C地区	D地区	平均
男子	10	60	50	20	35
女子	6	40	30	10	21.5

偏差

	A地区	B地区	C地区	D地区	平均
男子	-25	25	15	-15	
女子	-15.5	18.5	8.5	-11.5	
地区ごと偏差の積	387.5	462.6	127.5	172.5	287.5

共分散の意味

共分散の定義は「[Xの偏差 × Yの偏差]の平均」ですので、

共分散が大きい　->Xが大きいとYも大きい傾向がある。

共分散が0付近　->XとYにあまり関係はない。

共分散が小さい　->Xが大きいとYは小さくなる傾向がある。

今回の例では共分散が287.5ですので、「男子生徒が多い地区は女子生徒も多い傾向がある」といえます。

共分散は「Covariance」と言いますので、XとYの
共分散のことを

$Cov(X, Y)$

と書くことがあります。もしくは

σ_{XY}

と書くこともあります。また、期待値の記号では

$E[(X - \mu_X)(Y - \mu_Y)]$

と書きます。

生徒数

	A地区	B地区	C地区	D地区	平均
男子	10	60	50	20	35
女子	6	40	30	10	21.5

偏差

	A地区	B地区	C地区	D地区	平均
男子	-25	25	15	-15	
女子	-15.5	18.5	8.5	-11.5	
地区ごと偏差の積	387.5	462.6	127.5	172.5	287.5

共分散の簡単な計算方法

分散の式を展開します。

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y]$$

「和の期待値」は「期待値の和」ですので、以下のようになります。

$$\mathrm{Cov}(X,Y) = E[XY] - E[X\mu_Y] - E[Y\mu_X] + E[\mu_X\mu_Y]$$

定数倍は期待値の外に出して、

$$\mathrm{Cov}(X,Y) = E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y$$

$E[X] = \mu_X$、$E[Y] = \mu_Y$なので、

$$\mathrm{Cov}(X,Y) = E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y$$

となります。

共分散の欠点

先程の地区ごと生徒数の人数を、各マスともに単純に10倍します。
10倍した人数で共分散を計算すると、28750になります。

人数は10倍していますが、「男子生徒が多い地区は女子生徒も多い傾向がある」という関係性においては本質的に何も差がありません。それなのに、共分散の数値は大きくなってしまいます。

この様に、共分散には
スケール変換に対して不変でない
という欠点があります。

生徒数

	A地区	B地区	C地区	D地区	平均
男子	100	600	500	200	350
女子	60	400	300	100	215

偏差

	A地区	B地区	C地区	D地区	平均
男子	-250	250	150	-150	
女子	-155	185	85	-115	
地区ごと偏差の積	38750	46250	12750	17250	28750

相関係数

共分散は「スケール変換に対して不変でない」という欠点がありました。
この欠点を解消するため、共分散を規格化して相関係数という指標にします。

二組の対応するデータ(X, Y)に対し、相関係数 ρ は以下のように定義されます。

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$Cov(X,Y)$: 共分散
σ_X : Xの標準偏差
σ_Y : Yの標準偏差

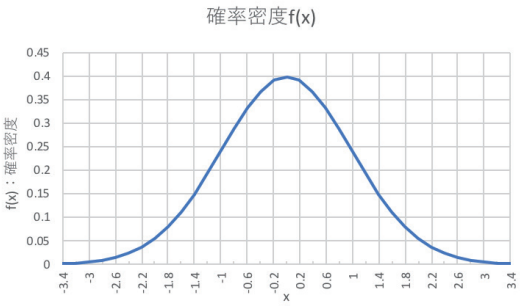
2変数のガウス分布（正規分布）

[第11回講義資料より抜粋]ガウス分布の確率密度関数

ガウス分布に従う確率変数Xの確率密度関数は以下の式で表されます。

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

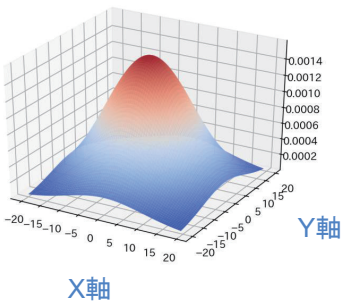
「σ」は標準偏差、「μ」は平均値を表します。
標準偏差σが1、平均μが0のガウス分布を、標準正規分布といいます。

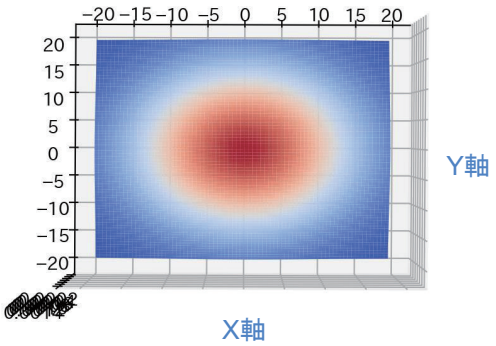


2次元のガウス分布

変数XとYの平均値、それぞれの分散、X-Y間の共分散を操作した際に、2次元のガウス分布の形状がどの様に変わるのかを観察します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix} = \begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}$の例

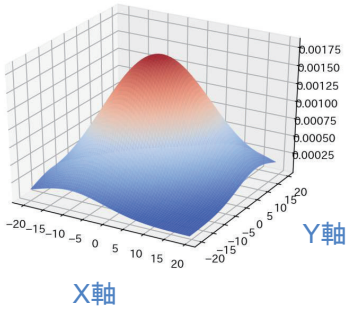


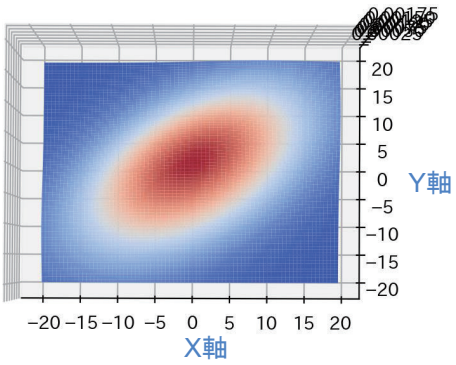


2次元のガウス分布

X-Y間の共分散を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}=\begin{pmatrix}100 & 50\\ 50 & 100\end{pmatrix}$の例

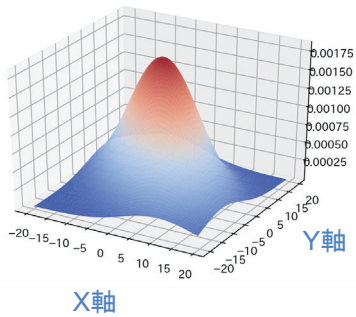


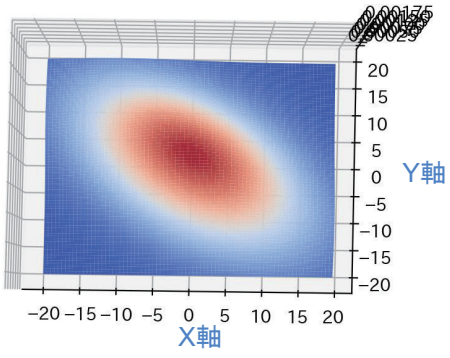


2次元のガウス分布

X-Y間の共分散を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}=\begin{pmatrix}100 & -50\\ -50 & 100\end{pmatrix}$の例

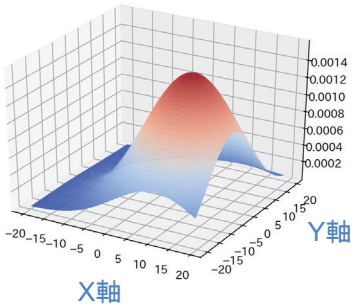


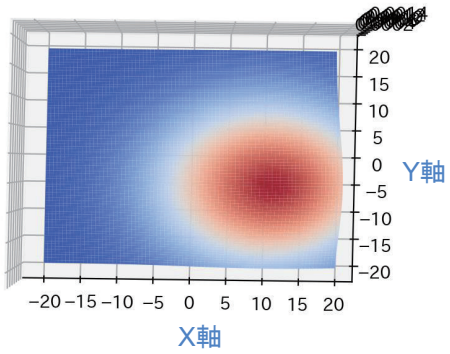


2次元のガウス分布

X、Yの平均値を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}=\begin{pmatrix}10\\ -5\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}=\begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}$の例





演習問題

演習1：離散型同時確率分布

- 2変数の離散型同時確率分布の例を挙げてください。
- 上記で挙げた例の2つの変数それぞれに対し、周辺化を実施してください。

演習2：連続型同時確率分布

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\leq x\leq 1, 0\leq y\leq 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

Yの周辺確率密度関数を求めてください。

演習3：確率変数の独立性

- 本資料で扱った男女別地区別生徒居住数のデータに対し、各セルの人数を変更して共分散を計算してください。
- 上記で作成した表に対し、各セルの人数を100倍して共分散を計算し、共分散値がスケールの変更に不変でないことを確認してください。

演習4：2変数のガウス分布

- [演習/Chapter12_Multidimensional_probability_distribution.ipynb]の平均値、共分散値を変更して実行し、2次元ガウス分布の形状がどのように変化するかを確認してください。

第13回：大数の法則

アジェンダ

- 大数の法則
- 大数の法則のコンピュータ・シミュレーション（サイコロの目）
- 大数の法則のコンピュータ・シミュレーション（コインの表裏）

大数の法則

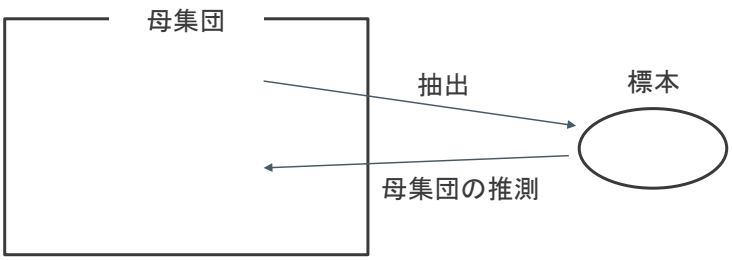
母集団と標本

日本に住む20～60代の人々の平均ボーナス額を調べたいとします。
2019年時点で20～60代の人口はおよそ7千7百万人います。これだけ多くの人々全員に、ボーナス額を聞いて回るのは現実的ではありません。

このような場合、7千7百万人から一部の人を選び出してボーナス額を調査し、その結果から7千7百万人全体のボーナス額平均を推定するという方法が取られます。

母集団と標本

ボーナス額平均を知りたいと思っている対象の7千7百万人の集団のことを「母集団」といいます。
母集団のボーナス額平均を推測するために選ばれた一部の集団を「標本」といいます。
母集団から一部のサンプルを選んで標本とすることを「抽出」といいます。



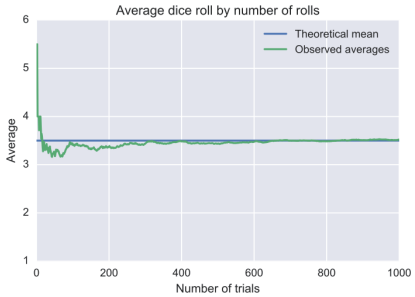
大数の法則とは

大数の法則とは、「ある独立した試行において、試行回数が大きくなるにつれて標本平均は母平均（期待値）に収束する」ということを意味します。

サイコロを何度も投げ続けることを考えます。サイコロの目の期待値は

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

なので、試行を繰り返すと標本平均は3.5に近づいていきます。

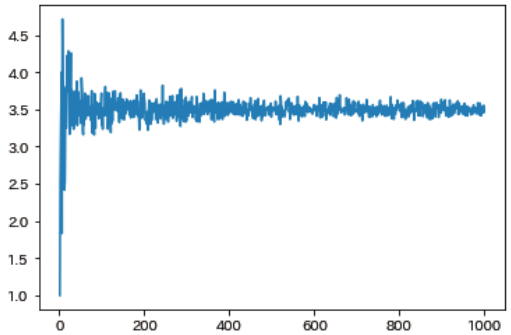


参照（Wikipedia）:
https://ja.wikipedia.org/wiki/%E5%A4%A7%E6%95%B0%E3%81%AE%E6%B3%95%E5%89%87

大数の法則のコンピュータ・シミュレーション
サイコロの目の平均値

本演習のゴール

サイコロを1回、2回、、、N回と多数投げ、出た目の平均値を計算していきます。
1回目からN回目までの平均値を図示することにより、試行回数を増やすにしたがってサイコロの目の平均値が3.5に収束することを確認します。



演習1：サイコロの目の平均値の計算

- 任意の回数サイコロを振り、出た目の平均値を求める関数を実装してください。
- サイコロを任意の回数振り、出た目の平均値を確認してください。

演習2：平均値の図示

- サイコロを1回振ったときの値、2回振ったときの平均値、3回振ったときの平均値、、、N回振ったときの平均値を図示する関数を実装してください。

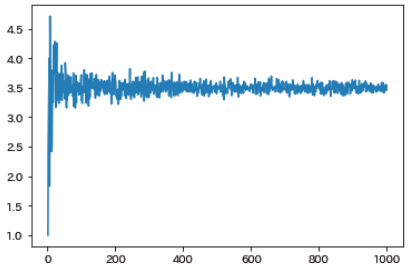
演習3：平均値の収束の確認

- サイコロを1,000回まで投げる試行を3回繰り返し、いずれも平均値が3.5に近づくことを確認してください。

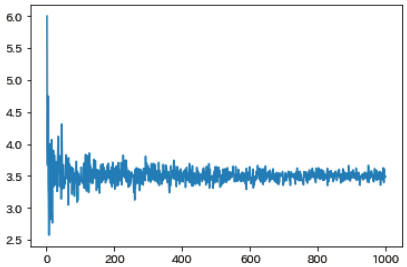
サイコロの目の平均値の収束

- サイコロを1,000回投げるまでの平均値の変化を図示した結果です。
- 3回テストを繰り返し、いずれも平均値が3.5に収束していることが確認できます。

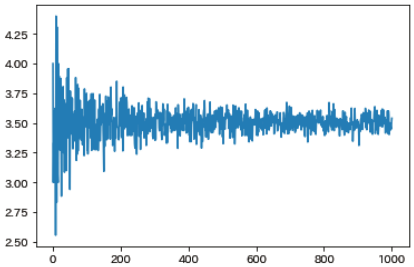
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目

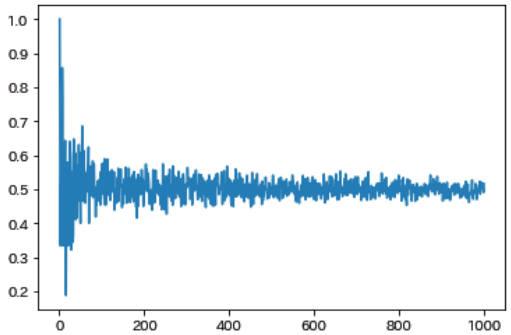


3回目

大数の法則のコンピュータ・シミュレーション
コインの裏表

本演習のゴール

コインを1回、2回、、、N回と多数投げ、表が出た場合を1、裏が出た場合を0として、表裏の平均値を計算していきます。
1回目からN回目までの平均値を図示することにより、試行回数を増やすにしたがってコインの裏表の平均値が0.5に収束することを確認します。



演習1：コインの裏表の平均値の計算

- 任意の回数コインを投げ、表が出た時を1、裏が出た時を0として、表裏の平均値を求める関数を実装してください。
- コインを任意の回数投げ、表裏の平均値を確認してください。

演習2：平均値の図示

- コインを1回投げたときの値、2回投げたときの平均値、3回投げたときの平均値、、、N回投げたときの平均値を図示する関数を実装してください。

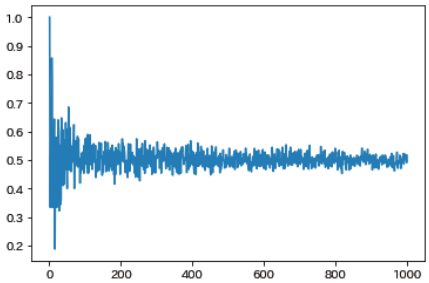
演習3：平均値の収束の確認

- コインを1,000回まで投げる試行を3回繰り返し、いずれも平均値が0.5に近づくことを確認してください。

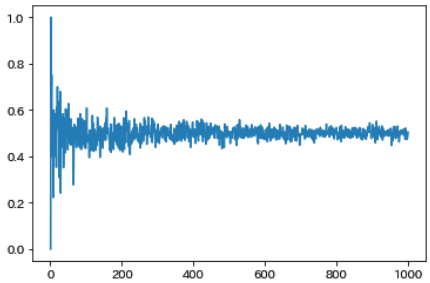
コインの裏表の平均値の収束

- コインを1,000回投げるまでの平均値の変化を図示した結果です。
- 3回テストを繰り返し、いずれも平均値が0.5に収束していることが確認できます。

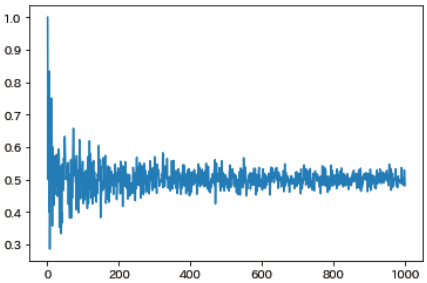
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

第14回：中心極限定理

アジェンダ

- 中心極限定理
- 中心極限定理のコンピュータ・シミュレーション（一様分布からのサンプリング）
- 中心極限定理のコンピュータ・シミュレーション（任意の分布からのサンプリング）

中心極限定理

中心極限定理とは

母集団の確率分布によらず、標本の大きさが十分に大きければ和や標本平均の分布は正規分布に従うという定理です。

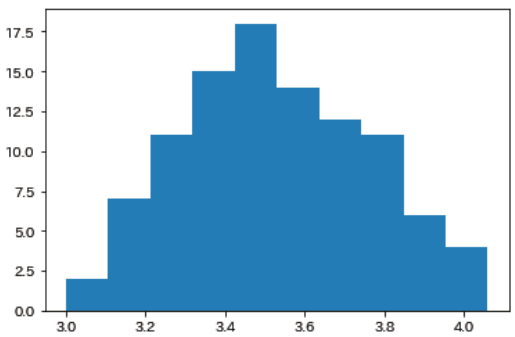
サンプル数をn、母集団の平均（母平均）をμ、分散（母分散）をσ^2とすると、$N(\mu, \sigma^2/n)$ という正規分布になります。

中心極限定理のコンピュータ・シミュレーション
一様分布からのサンプリング

中心極限定理のコンピュータ・シミュレーション

サイコロをN回振ったときの平均値の算出をM回繰り返し、平均値の分布を図示することにより、平均値の分布が正規分布に近づくことを確認します。

サイコロの目なので、母集団は一様分布となります。



演習1：サイコロの目の平均値の計算

- 任意の回数サイコロを振り、出た目の平均値を求める関数を実装してください。
- サイコロを任意の回数振り、出た目の平均値を確認してください。

演習2：平均値の分布の図示

- サイコロをN回振ったときの平均値の算出をM回繰り返し、その結果の分布を図示する関数を実装してください。

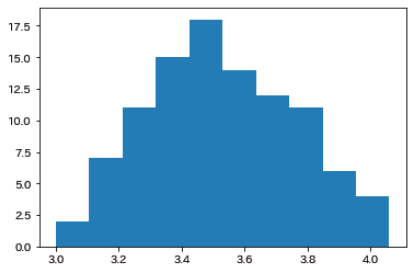
演習3：平均値の分布の形状の確認

- 「サイコロを50回投げて平均値を求めることを100回繰り返し散布図を図示する」ことを3回繰り返し、標本平均の分布が正規分布に近づくことを確認してください。

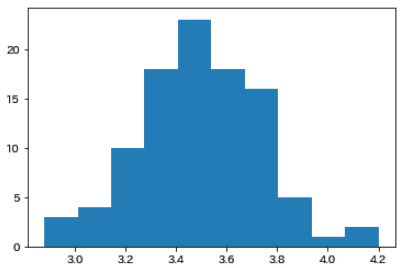
サイコロの目の平均値の分布

- サイコロを50回投げて平均値を求めることを100回繰り返したときの平均値の分布を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

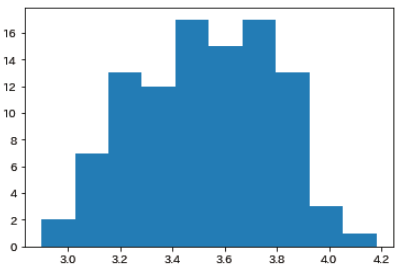
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

中心極限定理のコンピュータ・シミュレーション
任意の分布からのサンプリング

中心極限定理のコンピュータ・シミュレーション

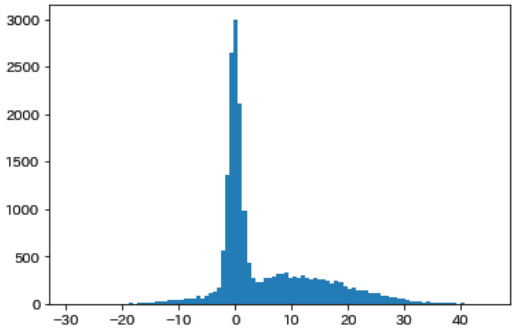
中心極限定理が成立する条件は、「サンプリング元の母集団の分布によらない」とされています。
前ページまでのシミュレーションではサイコロの目を扱ったため、母集団は一様分布にしたがっていました。

以降のシミュレーションでは、正規分布を2つ重ね合わせた分布を母集団とし、中心極限定理が成り立つことを確認します。

※中心極限定理が成立しない例外的な分布も存在しますが、ここでは取り扱いません。
https://ja.wikipedia.org/wiki/中心極限定理

演習1：母集団の生成

- 正規分布を2つ重ねた母集団データを作成してください。



演習2：抽出データの平均値の計算

- 任意の回数サンプルを抽出し、抽出したデータの平均値を求める関数を実装してください。
- 抽出したデータの平均値を確認してください。

演習3：平均値の分布の図示

- N回サンプリングしたデータの平均値の算出をM回繰り返し、その結果の分布を図示する関数を実装してください。

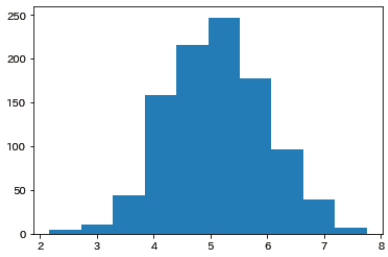
演習4：平均値の分布の形状の確認

- 「100回サンプリングして平均値を求めることを1,000回繰り返し散布図を図示する」ことを3回繰り返し、標本平均の分布が正規分布に近づくことを確認してください。

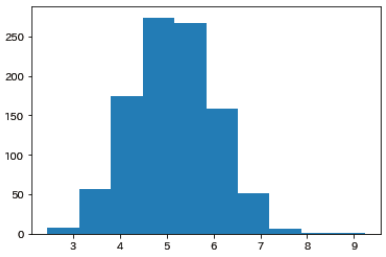
抽出データの平均値の分布

- 100回サンプリングして平均値を求めることを1,000回繰り返し散布図を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

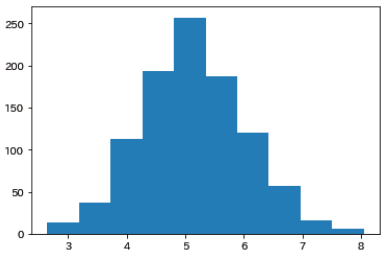
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

第15回：統計学Ⅰ総復習

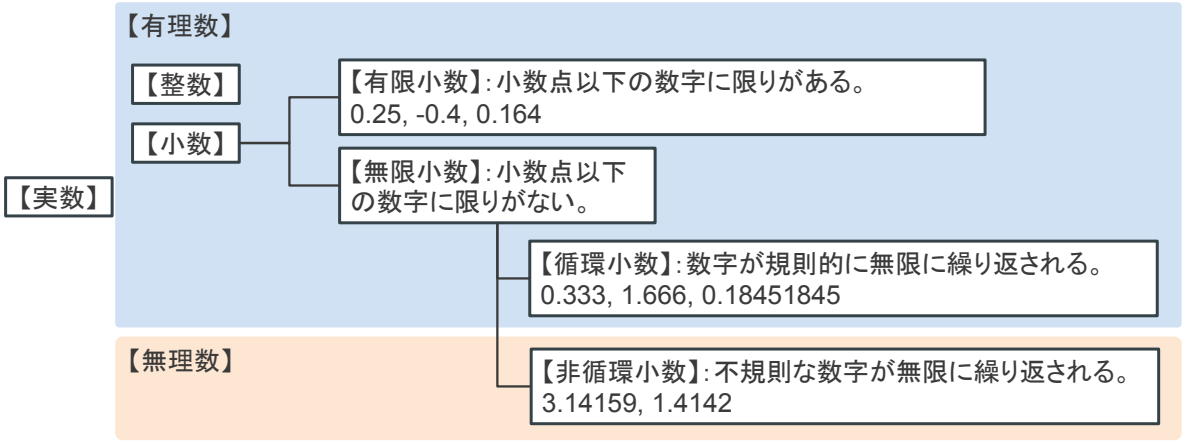
全15回の講義について

- 統計学および人工知能を学ぶ上で必須となる基礎数学を学習し、データ処理の基本知識である記述統計学について学習します。

第1回：基礎数学1

有理数/無理数とは

- 実数のうち、非循環小数を無理数といいます。
- 無理数以外の実数を有理数といいます。



約数・倍数・素数

- 約数：ある数を割ることができる整数のことです。
  - 12の約数は1, 2, 3, 4, 6, 12
- 素数：約数が1とその数しかない整数のことです。
  - 例えば2, 3, 5, 7, 11,などです。
- 公約数：複数の整数に共通する約数のことです。
  - 12の約数は1, 2, 3, 4, 6, 12
  - 18の約数は1, 2, 3, 6, 9, 18
    - 12と18の公約数は1, 2, 3, 6
- 最大公約数：最大の公約数のことです。
  - 12と18の最大公約数は6

約数・倍数・素数

- 倍数：ある整数を整数倍した整数のことです。
  - 3の倍数は3, 6, 9, 12など
- 公倍数：複数の整数に共通した倍数のことです。
  - 3の倍数は3, 6, 9, 12など
  - 4の倍数は4, 8, 12, 16など
    - 3と4の公倍数は12、24など
- 最小公倍数：最小の倍数のことです。
  - 3と4の最小公倍数は12

奇数と偶数

- 2で割り切れる自然数を偶数、割り切れない自然数を奇数といいます。
  - 偶数：2, 4, 6,,,
  - 奇数：1, 3, 5,,,
  - ※定義を「2で割り切れる整数」とした場合は、0は偶数に含まれます。

- 偶数と奇数の演算は、以下のようになります。
  - 偶数 ± 偶数 = 偶数
  - 偶数 ± 奇数 = 奇数
  - 奇数 ± 奇数 = 偶数
  - 偶数 × 偶数 = 偶数
  - 偶数 × 奇数 = 偶数
  - 奇数 × 奇数 = 奇数

第2回：基礎数学2

数理論理学とは

- [Wikipediaより]論理学（ろんりがく、英: logic）とは、「論理」を成り立たせる論証の構成やその体系を研究する学問である。ここでいう論理とは、思考の形式及び法則である。これに加えて、思考のつながり、推理の仕方や論証のつながりを指す。**論理学は、伝統的には哲学の一分野である。数学的演算の導入により、数理論理学（記号論理学）という分野ができた。**

- [Wikipediaより]数理論理学（独: mathematische Logik、英: mathematical logic）は、**論理学（形式論理学）の数学への応用の探求ないしは論理学の数学的な解析を主たる目的とする**、数学の関連分野である。局所的には数理論理学は超数学、数学基礎論、理論計算機科学などと密接に関係している。[1]数理論理学の共通な課題としては形式体系の表現力や形式証明系の演繹の能力の研究が含まれる。

- 数理論理学とは、「推論のやり方と、その正しさ」を扱う学問です。

命題論理とは

- [Wikipediaより]命題論理（めいだいろんり、英: propositional logic）とは、数理論理学（記号論理学）の基礎的な一部門であり[1]、命題全体を1つの記号に置き換えて単純化し、論理演算を表す記号（論理記号・論理演算子）を用いて、その命題（記号）間の結合パターンを表現・研究・把握することを目的とした分野のこと。

命題論理の例

次の1～3（命題といいます）は、全て事実だとします。

1. 釣りが好きな人は、ブリが好きである。
2. 男の人は、釣りが好きである。
3. お酒が好きな人は、イカの塩辛が好きであり、かつ男である。

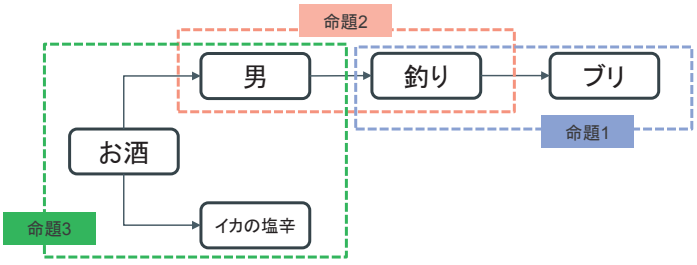
このとき、次のア～オの中で確実にいえることを全て選んでください。

ア. 男の人は、イカの塩辛が好きである。
イ. お酒が好きな人は、男である。
ウ. イカの塩辛が好きな人は、ブリが好きである。
エ. イカの塩辛が好きでない人は、釣りが好きでない。
オ. ブリが好きでない人は、お酒が好きでない。

命題論理の例の答え

次のア～オの中で確実にいえることを全て選んでください。

~~ア. 男の人は、イカの塩辛が好きである。~~
イ. お酒が好きな人は、男である。
~~ウ. イカの塩辛が好きな人は、ブリが好きである。~~
~~エ. イカの塩辛が好きでない人は、釣りが好きでない。~~
オ. ブリが好きでない人は、お酒が好きでない。（「お酒→ブリ」の対偶なので真）



第3回：基礎数学3

集合論とは

- [Wikipediaより]集合論（しゅうごうろん、英: set theory, 仏: théorie des ensembles, 独:Mengenlehre）は、集合とよばれる数学的対象をあつかう数学理論である。

- [Wikipediaより]数学における集合（しゅうごう、英: set, 仏: ensemble, 独: Menge）とは、大雑把に言えばいくつかの「もの」からなる「集まり」である。集合を構成する個々の「もの」のことを元（げん、英: element; 要素）という。集合は、集合論のみならず現代数学全体における最も基本的な概念の一つであり、現代数学のほとんどが集合と写像の言葉で書かれていると言ってよい。

集合の表記

集合は「xに関する命題P(x)が真となるようなxの集まり」という表現がされ、下記のように表記されます。

$$\{x|P(x)\}$$

例えば、「10以上の整数の集合」は以下のように表記されます。

$$\{x|x \in \mathbb{Z}, x \geq 10\}$$

ここで、「$x \in \mathbb{Z}$」は「xは整数の集合$\mathbb{Z}$に属する」という意味です。
このようにある集合を構成するxを要素、または元といいます。

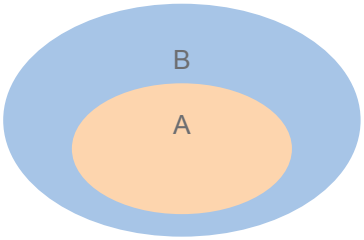
「xは整数の集合$\mathbb{Z}$に属さない」と表記したい場合は、以下のようにします。

$$x \notin \mathbb{Z}$$

部分集合

2つの集合AとBに対して「x∈A⇒x∈B」が成り立つ場合、AはBの部分集合といい、以下のように表記します。

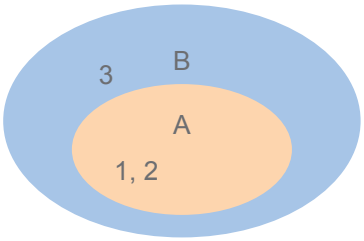
A⊂BもしくはB⊃A



真部分集合

A⊂Bであり、かつx∈Aでありx∉Bの要素が存在する場合、AはBの真部分集合であるといいます。
例えばA= $\{1, 2\}$、B= $\{1, 2, 3\}$であるときAはBの部分集合といい、下記のように表記します。

$A \subsetneq B$

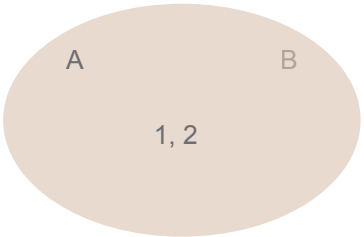


等価な集合

A⊂BかつB⊂Aである場合、「x∈A⇔x∈B」であり、2つの集合は等しくなり、以下のように表記します。

$A = B$

例えばA= $\{1, 2\}$、B= $\{1, 2\}$であるときAとBは等しくなります。



集合論のパラドックス

集合論には素朴集合論(naive set theory)と公理的集合論(axiomatic set theory)があります。
素朴集合論では、以下のようなパラドックス([Wikipediaより]正しそうに見える前提と、妥当に見える推論から、受け入れがたい結論が得られる事を指す言葉)が生じてしまいます。

例えば「床屋のパラドックス」という話があります。
以下のような床屋がいるとします。

- 自分でひげをそらない人全員のひげをそる。
- 自分でひげをそる人のひげはそらない。

この場合、床屋は自分のひげは自分でそるのでしょうか？
床屋が自分でひげをそるとした場合も、そらないとした場合もどちらも矛盾が生じてしまいます。

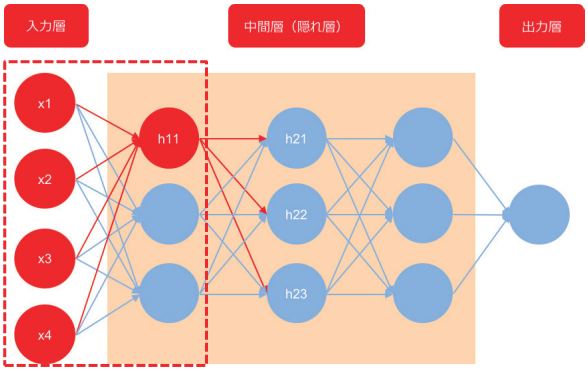
これは自分自身を要素として含まない集合すべての集まり$A = \{X | X \notin X\}$という集合を考えてしまうため生じるパラドックスです。
現代数学では「$A = \{X | X \notin X\}$を集合ではない」と考えるのが主流です。

参考：ベクトル演算と写像

ニューラルネットワークにおいて、「入力層の値」と「入力層−中間層間の重み」を行列計算し、中間層への出力値を算出します。
入力層の値の集合をX、出力層の値をHi、重みを掛け合わせてXからHiを算出する計算式をwとすると、

$$w : \mathrm{X} \rightarrow Hi$$

の写像だとみなせます。



第4回：基礎数学4

関数とは

- [Wikipediaより]数学における関数（かんすう、英: function、仏: fonction、独: Funktion、 蘭: functie、羅: functio、函数とも書かれる）とは、かつては、ある変数に依存して決まる値あるいはその対応を表す式の事であった。この言葉はライプニッツによって導入された。その後定義が一般化されて行き、現代的には数の集合に値をとる写像の一種であると理解される。

初等関数とは

- [Wikipediaより]初等関数（しょとうかんすう、英: Elementary function）とは、実数または複素数の1変数関数で、代数関数、指数関数、対数関数、三角関数、逆三角関数および、それらの合成関数を作ることを有限回繰り返して得られる関数のことである。
- [Wikipediaより]初等関数のうちで代数関数でないものを初等超越関数という。双曲線関数やその逆関数も初等関数である。

初等関数の例：定数関数

定数関数とは、それが取りうる値が変数の値によって変動しない定数値となる関数のことです。
例えば、

$$f(x) = 2$$

はxがどのような値でも2に写像する定数関数です。

初等関数の例：指数関数

指数関数とは、a＞0 かつ a≠1 のとき「$y=a^x$で表される関数」のことです。
また、この関数$y=a^x$のことを「aを底とするxの指数関数」と呼びます。

a>1の場合、xが大きくなるに従って指数関数yは大きくなります。
例えば以下のようなケースがあります。

$$y = a^x = 2^3 = 2*2*2 = 8$$

1>a>0の場合、xが大きくなるに従って指数関数yは小さくなります。
例えば以下のようなケースがあります。

$$y = a^x = (1/2)^3 = (1/2)*(1/2)*(1/2) = 1/8$$

底をネイピア数e (= 2.718281828…)とする指数関数のことをexp(x)と表記することがあります。

初等関数の例：対数関数

関数$y=a^b$のbのことを冪（べき）指数といい、bのことを「aを底とするyの対数」といいます。

$$b = \log_a y$$

と表記します。
底aがネイピア数e (= 2.718281828…)の場合、

$$b = \ln y$$

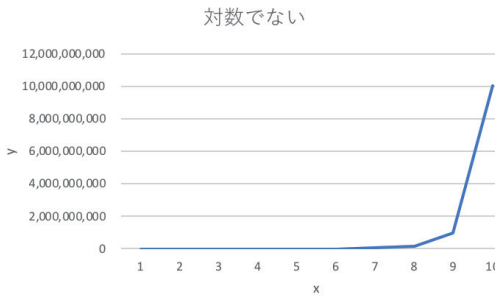
と表記することがあります。

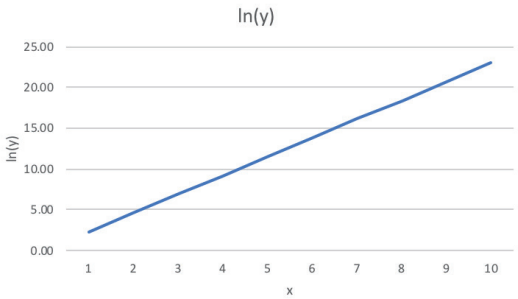
参考：対数を利用した正規化

線形で増加する変数xに対して、変数yが指数関数的に増加する場合を考えます。
変数x、yともにそのままの数値をグラフにすると、真ん中のグラフのようになります。
変数yについて対数をとったものをグラフにすると、右のグラフのようになります。

線形の機械学習器を使用する場合、右のグラフのように対数で変数を加工した場合のほうが良い精度を得られることがあります。

x	y	ln(y)
1	10	2.30
2	100	4.61
3	1,000	6.91
4	10,000	9.21
5	100,000	11.51
6	1,000,000	13.82
7	10,000,000	16.12
8	100,000,000	18.42
9	1,000,000,000	20.72
10	10,000,000,000	23.03





第5回：基礎数学5

微分とは

- 微分とは、任意の関数の各点における変化の割合（傾き）を求めることです。
- 傾きは以下の式で定義されます。
  - 変化の割合（傾き）$=\frac{\text{yの増加量}}{\text{xの増加量}}$
- 上式の変化の割合のことを、導関数といいます。

微分の定義

微分は、以下の式のように変化の割合だと述べました。

変化の割合（傾き）$=\frac{\text{yの増加量}}{\text{xの増加量}}$

また、変化の割合はxの場所によって変化することも学習しました。

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

微分の定義

関数$f(x)$の任意の場所aにおける微分は、以下の式で表します。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

$f(x)=x^2$のとき、上式にしたがって微分を実施すると、以下のようになります。

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}=\lim_{h\to 0}\frac{(x+h)^2-x^2}{h}=\lim_{h\to 0}\frac{x^2+2hx+h^2-x^2}{h}=\lim_{h\to 0}(2x+h)=2x$$

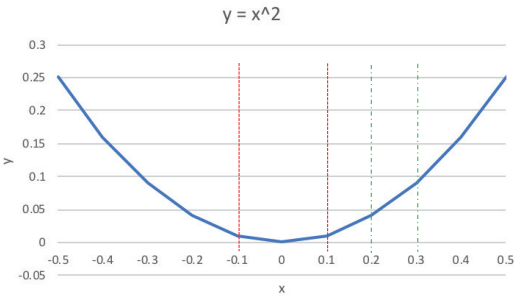
ここで求めた導関数を基に計算すると、

x=0.2のとき0.4

x=0.3のとき0.6

となります。「微分の例：2次関数」ではxが0.2から0.3に増えるときの傾きを計算し0.5となりました。

これは上の数字のちょうど真ん中にあることが確認できます。



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

可微分性

関数$f(x)$の任意の場所aにおける微分は、以下の式で表せました。

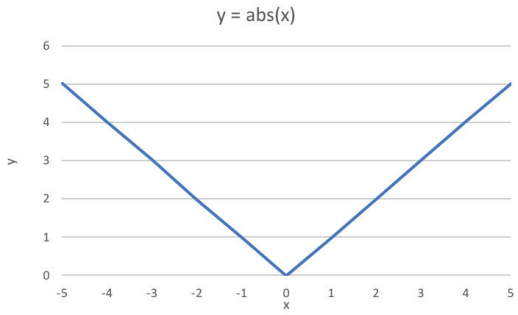
$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

絶対値関数$f(x)=|x|$の、x=0における傾きは

h＞0のときは1

h＜0のときは−1

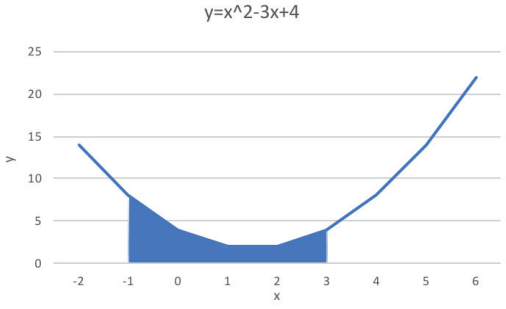
となり、x=0で微分可能ではありません。



第6回：基礎数学6

積分とは

- 積分とは、任意の関数f(x)で囲まれた部分の面積を求めることを意味しています。
  - $\int_a^b f(x)dx$
- 例えば$f(x) = x^2 - 3x + 4$、a=−1、b=3の場合、下図の青い部分の面積を求めることができます。

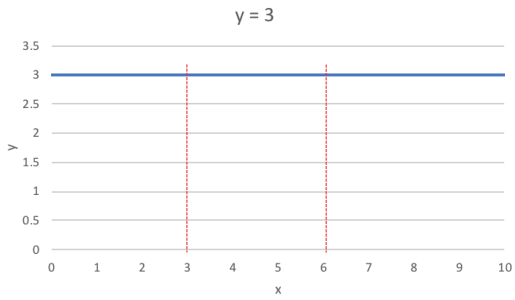


積分の例：傾きがない関数

- 下図のような「y = 3」という関数を考えます。
- この関数ではxの値によらずyの値は3のため、長方形の面積を求めることと同じになります。

- 例えばxが3から6の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_3^6 3dx = [3x]_3^6 = 3 \times 6 - 3 \times 3 = 9$

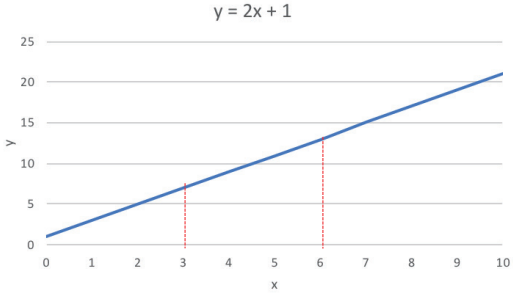
ここでは「3」を導関数とする原始関数
「3x」を求めています。



積分の例：1次関数

- 下図のような「y = 2x + 1」という関数を考えます。
- この関数ではxの値が1増加すると、yの値は2増加します。

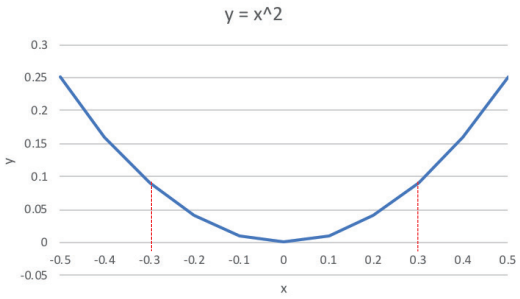
- 例えばxが3から6の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_3^6 (2x+1)dx = [x^2 + x]_3^6 = (6 \times 6 + 6) - (3 \times 3 + 3) = 42 - 12 = 30$



積分の例：2次関数

- 下図のような「y = x²」という関数を考えます。
- この関数ではxの増加量に対するyの増加量は、xの場所によって異なります。

- 例えばxが−0.3から0.3の範囲の面積は以下のように計算できます。
  - $\int_a^b f(x)dx = \int_{-0.3}^{0.3} x^2 dx = \left[\frac{1}{3}x^3\right]_{-0.3}^{0.3} = \left(\frac{0.3\times0.3\times0.3}{3}\right) - \left(\frac{(-0.3)\times(-0.3)\times(-0.3)}{3}\right) = 0.009 + 0.009 = 0.018$



x	y
-0.5	0.25
-0.4	0.16
-0.3	0.09
-0.2	0.04
-0.1	0.01
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25

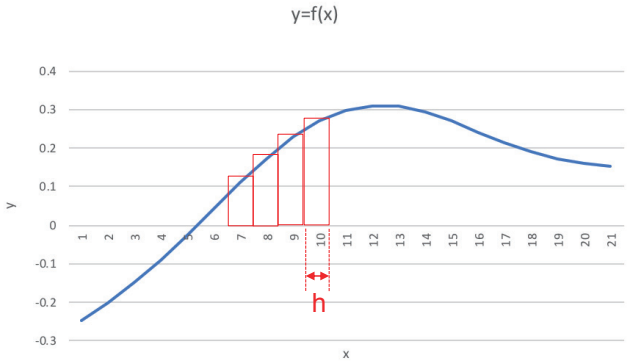
複雑な関数の積分

これまでの例では、導関数の原始関数を解析的に求めることができました。

解析的に求めることができない関数に対して面積を算出する際は、コンピュータプログラムなどで以下のようにして面積の近似値を求めます。

$$\sum_{i=a}^{b} f(i)h$$

hの間隔を徐々に狭くしていけば、上式の値は真の値に近づいていきます。



微分と積分の関係

「複雑な関数の積分」のページでは、面積の近似値をプログラムで求める方法を記載しましたが、より厳密に面積を求めていきます。
下図のオレンジ色の面積は、青い部分より大きく、青＋赤より小さいことがわかります。

$$f(t)\cdot h < S(t+h) - S(t) < f(t+h)\cdot h$$

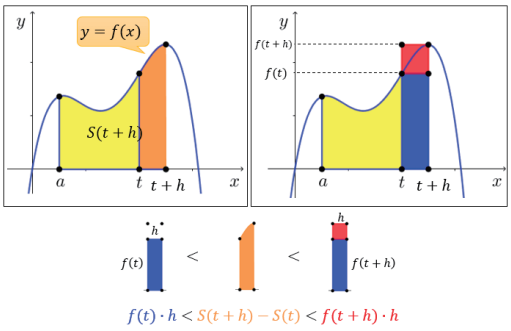
$$f(t) < \frac{S(t+h)-S(t)}{h} < f(t+h)$$

hの極限をとると、

$$f(t) < \lim_{h\to 0}\frac{S(t+h)-S(t)}{h} < \lim_{h\to 0} f(t+h) = f(t)$$

$$\lim_{h\to 0}\frac{S(t+h)-S(t)}{h} = f(t)$$

最後の式は、f(x)の導関数を求める式と同じであることが確認できます。



参照：https://atarimae.biz/archives/22721

第7回：統計学の基本事項

質的変数

- 質的変数とは、下記のようにカテゴリなどデータ間の「質」が異なる情報を保持するデータです。
  - 性別
  - 名前
  - 等級
  - 曜日
- 質的データは数値データではないため、そのままでは統計分析や機械学習に利用することができません。

質的変数のダミー変数化

- 質的変数を、下表のように0または1に置き換えて数値に変換する操作をダミー変数化といいます。
- 下の例では、曜日を7列のダミー変数に変換しています。また、投薬有無を2列のダミー変数に変換しています。機械学習データとして使用するためには、多重共線性を回避する上で、ダミー変数の列数を減らす（曜日であれば6列など）ほうが望ましいです。

店名称	定休日
A商店	日曜日
B商店	月曜日
C商店	火曜日
D商店	水曜日
E商店	木曜日
F商店	金曜日
G商店	土曜日

店名称	定休日_日	定休日_月	定休日_火	定休日_水	定休日_木	定休日_金	定休日_土
A商店	1	0	0	0	0	0	0
B商店	0	1	0	0	0	0	0
C商店	0	0	1	0	0	0	0
D商店	0	0	0	1	0	0	0
E商店	0	0	0	0	1	0	0
F商店	0	0	0	0	0	1	0
G商店	0	0	0	0	0	0	1

患者番号	投薬
001	あり
002	なし
003	あり

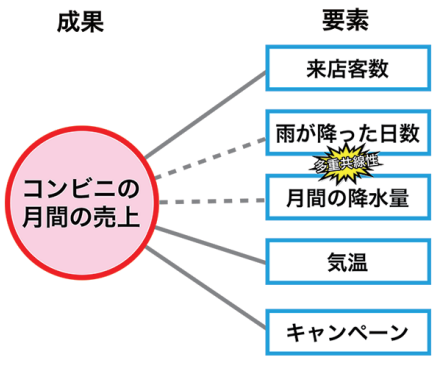
患者番号	投薬あり	投薬なし
001	1	0
002	0	1
003	1	0

量的変数

- 量的変数とは、下記のようにデータの数値が「量」の情報を保持するデータです。
  - 身長・体重
  - 面積
  - 年収
  - 年齢
- 量的データは数値データなので、そのまま統計分析や機械学習に利用することができます。

多重共線性

- 説明変数間で相関係数が高い時に多重共線性（multicollinearity）という問題が発生します。
- 多重共線性とは、モデル式の係数が不安定（符号と大きさが安定しない）になり、モデルの予測結果に対する係数の寄与度を正しく評価することができなくなってしまいます。



出展：https://xica.net/vno4ul5p/

変数の尺度

- 変数を性質に応じて下記の4つの尺度に分けて考えることがあります。
  - 名義尺度
  - 順序尺度
  - 間隔尺度
  - 比例尺度

データの次元

- 1つの観測対象に対して、1つの数値で表現したデータを1次元データ、2つの数値で表現したデータを2次元データ、N個の数値で表現したデータをN次元データといいます。

就職先	通勤時間
A株式会社	1時間10分
B株式会社	50分
C合同会社	1時間20分

1次元データ
「就職先」という対象を表現するために、「通勤時間」という1つのデータを使用している。

就職先	通勤時間	平均給与
A株式会社	1時間10分	500万円
B株式会社	50分	450万円
C合同会社	1時間20分	550万円

2次元データ
「就職先」という対象を表現するために、「通勤時間」、「平均給与」という2つのデータを使用している。

就職先	通勤時間	平均給与	退職金
A株式会社	1時間10分	500万円	なし
B株式会社	50分	450万円	あり
C合同会社	1時間20分	550万円	あり

3次元データ
「就職先」という対象を表現するために、「通勤時間」、「平均給与」、「退職金」という3つのデータを使用している。

就職先	通勤時間	平均給与	退職金	海外赴任
A株式会社	1時間10分	500万円	なし	あり
B株式会社	50分	450万円	あり	あり
C合同会社	1時間20分	550万円	あり	なし

4次元データ
「就職先」という対象を表現するために、「通勤時間」、「平均給与」、「退職金」、「海外赴任」という4つのデータを使用している。

第8回：度数分布表と各種代表値

度数分布表とは

- データの分布を観察するため、データを任意の範囲で区切り、その範囲に含まれるデータ数を見ることがあります。そのような表を度数分布表といいます。
- 下の例は、2019年の日本在住外国人の年齢ごと人数を度数分布表にしたものです。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：階級

- 度数を集計するための区間を表します。下の例では、[0–9歳]など10歳ごとに区切った年齢の幅が階級です。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：階級値

- 階級の代表値のことで、階級の下限値と上限値の平均値を表します。下の例では、[0–9歳]の階級に対する階級値は4.5となります。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：度数

- 各階級に含まれるデータ数のことです。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

度数分布表：相対度数

- 各階級の度数が全体に占める割合のことです。
- 下の例では、[30−39歳]の相対度数は572,874 / 2,667,139 = 0.21478971となります。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

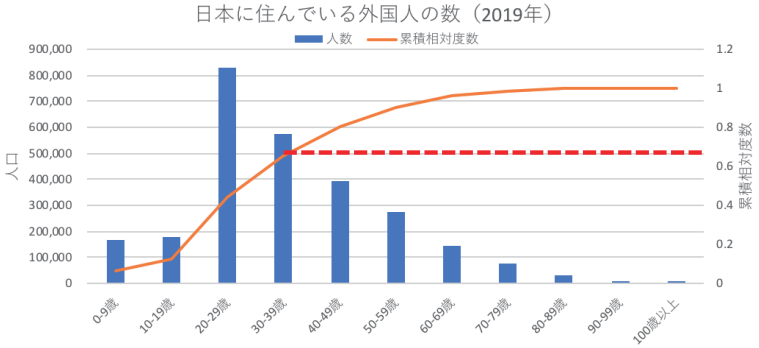
度数分布表：累積相対度数

- 相対度数を、その階級まで上から累積した値です。
- 一番最後の階級では、全ての相対度数を累積することになるので、累積相対度数1となります。

階級 年齢	階級値	度数 人数	相対度数	累積相対度数
0-9歳	4.5	164,468	0.06166458	0.06166458
10-19歳	14.5	176,176	0.0660543	0.12771888
20-29歳	24.5	828,034	0.31045776	0.43817664
30-39歳	34.5	572,874	0.21478971	0.65296634
40-49歳	44.5	394,333	0.14784869	0.80081503
50-59歳	54.5	275,921	0.10345205	0.90426708
60-69歳	64.5	145,164	0.05442686	0.95869394
70-79歳	74.5	74,752	0.02802704	0.98672098
80-89歳	79.5	29,491	0.01105717	0.99777814
90-99歳	84.5	5,749	0.00215549	0.99993364
100歳以上	-	177	6.6363E-05	1
合計	-	2,667,139	-	-

ヒストグラム

- ヒストグラムは度数分布表をグラフにしたものです。
- 横軸が階級、縦軸が度数のグラフです。
- 累積度数とセットで表示すると、データ全体の分布がよりわかりやすくなります。下の例では、30代までの外国人人口が、全ての外国人人口の7割弱であることがわかります。



第9回：順列と組み合わせ_標本空

順列とは

異なるn個の中から異なるr個を取り出し、かつ1列に並べた場合のパターン数のことです。

例えば3つのもの{A, B, C}から2つを取り出す順列を考えると、
AB, BA, AC, CA, BC, CA
の6つのパターンがあります。
ポイントは、ABとBAのように順序が違う場合も違うパターンとして考えることです。

順列の公式

先程の例 {A, B, C}から2つを取り出す順列で考えると、
　　1回目はA, B, Cの3つから選択可能、
　　2回目は残りの2つから選択可能、
なので、パターン数としては3×2=6となります。

これを一般化して「異なるn個の中から異なるrを取り出し並べる順列の数」は、以下のように計算できます。

$$nPr = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

「！」に記号は階乗と読み、1~nまでの積を表します。
例えば4!=4×3×2×1=24となります。

組み合わせとは

順列とは「異なるn個の中から異なるr個を取り出し、かつ1列に並べた場合のパターン数」のことでした。
例えば3つのもの{A, B, C}から2つを取り出す順列を考えると、
　　AB, BA, AC, CA, BC, CA
の6つのパターンがあり、ABとBAのように順序が違う場合も違うパターンとして考えていました。

組み合わせは、ABとBAは同じものとして考え、パターン数にはカウントしません。
2つのもの（例えばAとB）から構成される、カウントしないパターンの数は、${}_2P_2$=2で計算できます。

組み合わせの公式

例えば3つのもの{A, B, C}から2つを取り出す順列を考えると、

AB, BA, AC, CA, BC, CA

の6つのパターンがありましたが、カウントしないパターン数は2でした。

よって

6 ／ 2 = 3

で組み合わせのパターン数を計算することができます。

これを一般化して「異なるn個の中から異なるrを取り出し並べる組み合わせの数」は、以下のように計算できます。

$$\mathrm{nCr} = \frac{nPr}{rPr} = \frac{1}{r!}nPr = \frac{n!}{r!(n-r)!}$$

標本空間

試行（実験）の結果として起こり得るすべての場合を要素とした集合を、標本空間といいます。

起こり得るすべての場合を、$\omega_i(i = 1, 2, , , n)$とするとき、標本空間は以下のように定義されます。

$$\Omega = \{\omega; \omega = (\omega_1, \omega_2, , , \omega_n)\}$$

例えば、サイコロを振ることを考えた場合、標本空間は以下のようになります。

$$\Omega = \{\omega; \omega = (1, 2, 3, 4, 5, 6)\} = \{1, 2, 3, 4, 5, 6\}$$

第10回：確率変数

確率変数とは

ある現象がいろいろな値を取り得るとき、取り得る値全体を確率変数といいます。

例えば、サイコロを振ったときに出る目は[1, 2, 3, 4, 5, 6]のいずれかとなります。
この場合、確率変数Xは

$$\mathrm{X} = 1, 2, 3, 4, 5, 6$$

と表します。

確率変数をXと置くことで、サイコロの目が取りうる値の確率を、以下のように記載することができます。

$$\mathrm{P}(x) = \frac{1}{6}(\mathrm{X} = 1, 2, 3, 4, 5, 6)$$

サイコロを振って4が出る確率は以下のように書きます。

$$\mathrm{P}(x = 4) = \frac{1}{6}$$

離散型の確率変数

離散型確率変数は、「とびとびの値」を指します。
隣り合った数値の間には、数値は存在しません。
例えばサイコロの目、コインの裏表、ルーレットの番号などが該当します。

連続型の確率変数

連続型確率変数は、「連続した値」を指します。
例えば速度であれば、5km/hと6km/hの間には5.1km/hや5.01km/h、5.0001km/hなど無数の値が存在します。

その他の連続確率変数には温度、湿度、高度、体重などがあります。

確率分布とは

確率変数のそれぞれの値に対し、その確率変数をとる確率の分布のことです。

離散型確率変数に対する確率分布として、以下のような確率分布があります。

- ポアソン分布
- 二項分布
- 幾何分布
- 一様分布

連続型確率変数に対する確率分布として、以下のような確率分布があります。

- 正規分布
- 指数分布
- 一様分布

第11回：代表的な確率分布

幾何分布とは

成功確率をpとしたベルヌーイ試行を繰り返すとします。初めて成功するまでの試行回数Xが従う確率分布を「幾何分布（きかぶんぷ）」といいます。

成功確率がpの試行において、k回目で初めて成功する確率は次の式で計算できます。

$$P(X = k) = (1 - p)^{k-1}p: k = 1, 2, 3, , ,$$

超幾何分布とは

[Wikipediaより]超幾何分布（ちょうきかぶんぷ、英: hypergeometric distribution）とは、成功状態をもつ母集団から非復元抽出したときに成功状態がいくつあるかという確率を与える離散確率分布の一種である。

例えば、
・箱の中に玉がN個あり、M個が赤い玉、N−M個が白い玉とします。
・玉を箱から取り出して色を調べ、玉を元に戻さない非復元抽出でn回調べるとします。
・このときの赤い玉の個数を確率変数Xとします。
X=k個となる確率は超幾何分布に従います。

ベルヌーイ分布とは

ベルヌーイ分布とは、「成功/失敗」、「表/裏」「勝ち/負け」のように2種類の結果が得られる実験（ベルヌーイ試行）の結果を0と1で表した分布のことです。
1である確率をpとした場合、0である確率は1−pとなります。

二項分布

ベルヌーイ試行をn回行い、成功する回数Xが従う確率分布を「二項分布」といいます。
n回のベルヌーイ試行を行いk回成功する確率は次の式から計算できます。

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

例えばコイン投げを例にとります。表が出る確率p=0.5なので、
10回中4回表が出る確率は以下のように計算できます。

$$P(X = 4) = {}_{10}C_4 0.5^4 (1-0.5)^{10-4} = 0.205$$

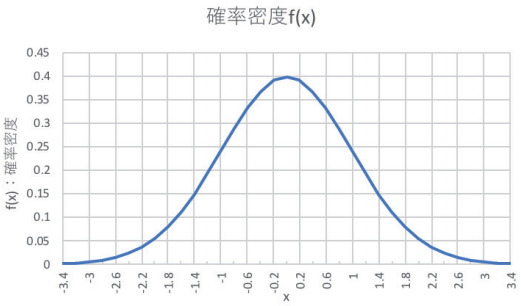
表の出る回数	確率
0	0.001
1	0.010
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.010
10	0.001

ガウス分布の確率密度関数

ガウス分布に従う確率変数Xの確率密度関数は以下の式で表されます。

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

「σ」は標準偏差、「μ」は平均値を表します。
標準偏差σが1、平均μが0のガウス分布を、標準正規分布といいます。



ポアソン分布の例

鵜戸神宮には、岩穴に運玉を投げ入れるという運試しスポットがあります。
運試しは、運玉を200個投げると1回岩穴に入るとします。このとき、運玉を10回投げたときに、岩穴に運玉が1個入る確率はいくつでしょうか？（成功確率はポアソン分布に従うとします）

岩穴に入る確率はp=1/200、運玉を投げた回数はn=10回なので、

$$\lambda = np = 10 \times \frac{1}{200} = 0.05$$

となります。したがって

$$P(X = 1) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{0.05 \times e^{-0.05}}{1!} = 0.0475$$

となり、約4.8%であることがわかります。

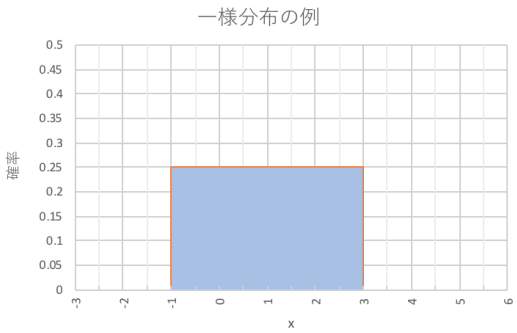
画像：Wikipediaより
https://ja.wikipedia.org/wiki/%E9%B5%9C%E6%88%B8%E7%A5%9E%E5%AE%AE

一様分布とは

[Wikipediaより]一様分布（いちようぶんぷ）は、離散型あるいは連続型の確率分布である。サイコロを振ったときの、それぞれの目の出る確率など、すべての事象の起こる確率が等しい現象のモデルである。

区間[a, b]上の確率分布は以下のように表されます。

$$f(x) = f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x\text{が上記以外} \end{cases}$$



第12回：多次元の確率分布

離散型同時確率分布とは

2つの離散型確率変数XとYが、それぞれある値をとるときの確率を表したものを「離散型同時確率分布」といいます。

例えば、男子20名、女子20名のあるクラスがあるとします。生徒の居住地区を表にしてみました。
性別をX、居住地区をYとすると、2つの離散型確率変数とみなせます。

	A地区	B地区	C地区	D地区	計
男子	4	6	6	4	20
女子	6	8	4	2	20
				総合計	40

全生徒40人に対する各マスの数字の割合を計算してみました。これは2つの離散型確率変数XとYがそれぞれの値を同時にとる、離散型同時確率分布となります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

離散型同時確率分布とは

2つの確率変数からなる同時確率分布は、以下のように表記します。

$$f(x_i, y_j) = P(X = x_i, Y = y_j) \quad i = 1,2,3,,,;j = 1,2,3,,,$$

例えば、男子でD地区に住む生徒の確率は、以下のようになります。

$$P(X = \text{男子}, Y = \text{D地区}) = 0.1$$

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5

ここで$f(x_i, y_j)$のことを同時確率関数といいます。各iとjについて全ての確率を足すと総和は1になります。

$$\sum_i \sum_j f(x_i, y_j) = 1$$

周辺確率分布

性別X、居住地区Yのそれぞれの値について、確率の合計を計算してみます。
男子の割合は0.5、A地区に居住する生徒の割合は0.25であることがわかります。

	A地区	B地区	C地区	D地区	計
男子	0.10	0.15	0.15	0.10	0.5
女子	0.15	0.20	0.10	0.05	0.5
計	0.25	0.35	0.25	0.15	1.00

このようにある1つの確率変数を固定し、別の確率変数が取りうる全ての確率を合計したものを周辺確率分布といいます。

$$f_x(x_i) = \sum_j f(x_i, y_j) = P(X = x_i) \quad i = 1,2,3,,,$$
$$f_y(y_j) = \sum_i f(x_i, y_j) = P(Y = y_j) \quad j = 1,2,3,,,$$

ここで、$f_x(x_i)$と$f_y(y_j)$をそれぞれXとYの周辺確率関数といいます。

連続型同時確率分布とは

XとYが連続型確率変数であるとき、それぞれある値をとるときの確率を表したものを「連続型同時確率分布」といいます。

XとYの同時確率分布を表す関数を「同時確率密度関数」といいます。

$$P(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$$

確率の総和は1になるため、同時確率密度関数に関して以下の式が成り立ちます。

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

連続型確率変数XとYの周辺確率密度関数は、以下の式で求めることができます。

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$
$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\le x\le 1, 0\le y\le 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

同時確率変数XとYの全範囲についての確率を求めてみます。

$$P(0\le x\le 1, 0\le y\le 1)=\int_0^1\int_0^1 (x+y)\,dx\,dy=\int_0^1\left[\frac{x^2}{2}+yx\right]_0^1 dy=\int_0^1\left(\frac{1}{2}+y\right)dy$$

$$=\left[\frac{y}{2}+\frac{y^2}{2}\right]_0^1=1$$

連続型同時確率分布の計算例

次のような同時確率密度関数を考えます。

$$f(x,y)=\begin{cases} x+y\ (0\le x\le 1, 0\le y\le 1) \\ 0\ (\text{上記以外の}x,y\text{の場合}) \end{cases}$$

Xの周辺確率密度関数を求めてみます。

$$f_x(x)=\int_0^1 (x+y)\,dy=\left[x+\frac{y^2}{2}\right]_0^1=x+\frac{1}{2}$$

独立な確率変数とは

2つの確率変数XとYの同時確率分布（同時確率密度関数）f(x, y)が、それぞれの確率変数の周辺確率分布（周辺確率密度関数）g(x)とh(y)の積に分解できる時、その2つの確率変数は独立(independent)であると言います。

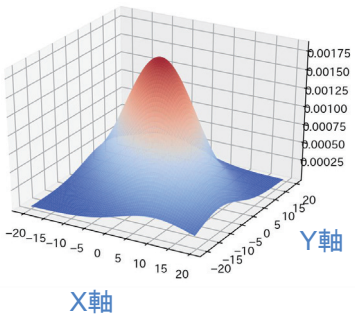
$$f(x,y) = g(x)h(y)$$

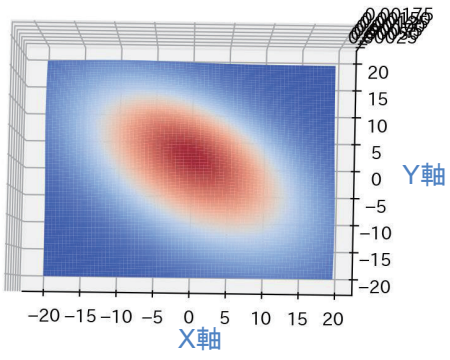
直感的な理解としては、「XとYの動きは、お互いに影響を及ぼさない」ということです。

2次元のガウス分布

X-Y間の共分散を操作します。

$\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix} = \begin{pmatrix}100 & -50\\ -50 & 100\end{pmatrix}$の例





第13回：大数の法則

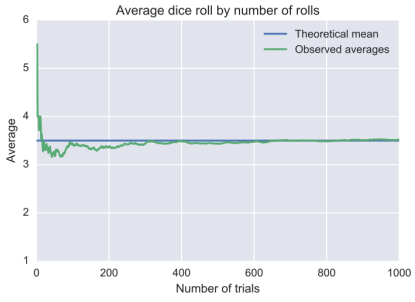
大数の法則とは

大数の法則とは、「ある独立した試行において、試行回数が大きくなるにつれて標本平均は母平均（期待値）に収束する」ということを意味します。

サイコロを何度も投げ続けることを考えます。サイコロの目の期待値は

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

なので、試行を繰り返すと標本平均は3.5に近づいていきます。



参照（Wikipedia）：
https://ja.wikipedia.org/wiki/%E5%A4%A7%E6%95%B0%E3%81%AE%E6%B3%95%E5%89%87

第14回：中心極限定理

中心極限定理とは

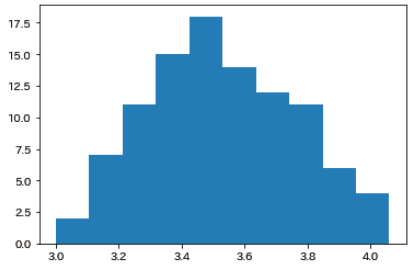
母集団の確率分布によらず、標本の大きさが十分に大きければ和や標本平均の分布は正規分布に従うという定理です。

サンプル数をn、母集団の平均（母平均）をμ、分散（母分散）をσ^2とすると、$N(\mu, \sigma^2/n)$ という正規分布になります。

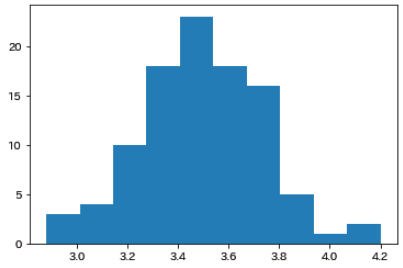
サイコロの目の平均値の分布

- サイコロを50回投げて平均値を求めることを100回繰り返したときの平均値の分布を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

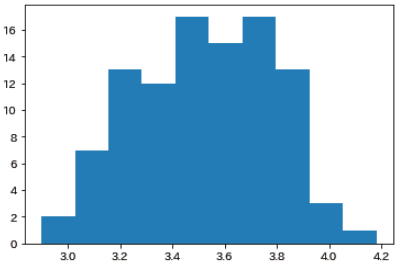
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目

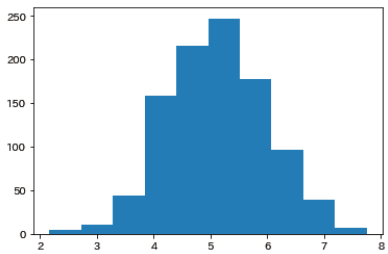


3回目

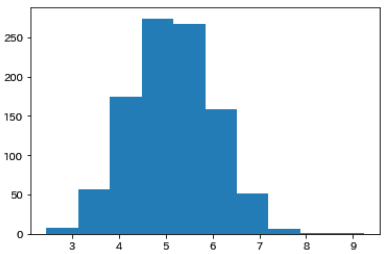
重ね合わせた正規分布からの抽出データの平均値の分布

- 100回サンプリングして平均値を求めることを1,000回繰り返し散布図を図示した結果です。
- 3回テストを繰り返し、いずれも平均値の分布が正規分布に近づいていることが確認できます。

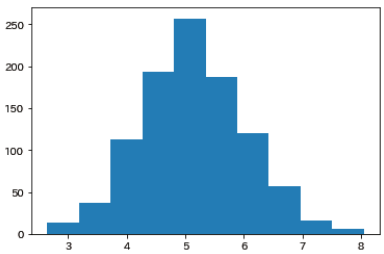
※プログラムには乱数を使用しているため、皆さんのテスト結果と以下の図は完全には一致しません。



1回目



2回目



3回目

令和２年度「専修学校による地域産業中核的人材養成事業」
Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

令和２年度「専修学校による地域産業中核的人材養成事業」
Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

統計学Ⅰ

令和３年２月

学校法人電子学園（日本電子専門学校）
〒169-8522　東京都新宿区百人町 1-25-4
TEL 03-3369-9333　FAX 03-3363-7685