神戸大学　数理・データサイエンスセンター
加葉田大志朗

データサイエンス基礎学 4-7 回

神戸大学　数理・データサイエンスセンター
特命准教授　加葉田大志朗

神戸大学　数理・データサイエンスセンター
加葉田大志朗

目次

講義概要 4
講義の目標 4
講義でやること 4
講義でやらないこと 4
参考書籍 4
1.　データの種類 6
用語の定義 6
変数の種類 7
2.　データの特徴を図表で捉える 8
離散変数の度数分布表と棒グラフ 10
連続変数の度数分布表とヒストグラム 13
その他のグラフ 16
円グラフ 16
100%積み上げ棒グラフ 17
3.　データの特徴を統計量で捉える 19
分布の位置を示す統計量①：平均値 20
分布の位置を示す統計量②：中央値 21
（1）分布の形状が左右対称でない場合 21
（2）極端な観測値が含まれる場合 22
データのばらつきを捉える統計量①：標準偏差 24
データの加工と偏差値 27
データの加工の具体的な手続き 27
データの加工を利用した偏差値の計算 29
データのばらつきを捉える統計量②：四分位範囲 31
4.　正規分布を使って観測されたデータの珍しさを考える 32
正規分布の特性 34
5.　標本から母集団を推測する 37
ランダムサンプリングの仮定 38
母平均の推定 39
母平均を推定する最も強力な方法は標本平均を計算すること 39
複数の観測値の標本平均は母集団に近い可能性が高い 39
標本平均のばらつきと母平均の区間推定 43
正規母集団における標本平均のばらつきは標準誤差で示す 43
標準誤差に基づいて標本平均のばらつきの範囲を求める 43

母平均を区間推定する 44
95%信頼区間 44
母分散と標準偏差の推定 47
標準誤差と信頼区間における標準偏差の推定値の役割 48
6. 二種類の変数の関係を評価する 51
相関係数の計算方法 53
相関係数と偏差積 53
相関係数と標準偏差 55
相関の強さ 58
7. 回帰モデルの基本 61
線形回帰モデルの表現方法 61
線形回帰モデルの実践 63
標準化したデータを利用すると相関係数が得られる 64
説明変数が離散変数の場合 65
8. 疑似相関と多変量回帰モデル 67
交絡因子の存在を疑う 70
サブグループを作ることで交絡の影響を抑える 71
サブグループ内で得られた結果を統合する「層別化法」 72
多変量回帰モデルの活用 74
多変量回帰モデルによって調整済みの関連を求める 74
多変量回帰モデルを利用する際に注意すること 77
（1）モデルの過適合（overfitting） 77
（2）多重共線性 80
9. 回帰モデルによる予測とその良さを測る方法 81
予測モデルの良さの指標 81
（1）決定係数 R^2 81
（2）赤池情報量基準 AIC 83
機械学習と回帰モデルの境界はどこ？ 86

神戸大学　数理・データサイエンスセンター
加葉田大志朗

講義概要

講義の目標

以下のデータリテラシーレベルのスキルを身に付けることを目標とする.

（1） データを読み解き起きている事象の背景や意味合いを理解する（データを読む）
（2） データの特徴を捉えるための表現方法を学ぶ（データを説明する）
（3） 正しくデータを汲み取ることができる手法を学ぶ（データを扱う）

講義でやること

（1） 日常生活および学術の場で必要となるデータサイエンス・統計学の基礎内容を学ぶ
（2） データサイエンス・統計学を将来専門的に学ぶための基盤を整える
（3） Python を利用した分析の基本的な手続きに触れる

講義でやらないこと

（1） データサイエンス・統計学に関連する広範な内容を網羅的に学ぶ
（2） データサイエンス・統計学の数学的背景を詳細に理解する
（3） Python プログラミングを実践する技術を網羅的に学ぶ

参考書籍

この講義では大部分で以下に並べる書籍の内容を引用し，参考にしています．講義では初学者と想定される受講者が「講義の目標」を達成できるように，それぞれの内容を抜粋，整理していますが，興味がある方はそれぞれの書籍をしっかり読み込むことで，この講義で得られる以上の知識が獲得できると思います．

（1） 完全独習　統計学入門（小島寛之，ダイヤモンド社，2006）
講義前半の内容は，この書籍の内容を大きく踏襲しています．統計学に関する基本的な事項が非常にわかりやすく整理されています．簡単な四則演算しか利用しないため，誰でも本当に完全に独学できます．まずは実用性よりも背景を理解の理解をきちんとしたい人におすすめです．

（2） 医学研究のための R による統計解析入門（加葉田大志朗，オーム社，2023）
講義後半の内容は，この書籍の内容を大きく踏襲しています．この講義で紹介する

内容について，より詳細に説明されています．本講義では扱いませんが，理系学部の学生が触れることの多い仮説検定などについても解説しているため，医学部以外の理系学部生にもおすすめできます．R プログラミングも基礎から学習できます．

（3） Python で学ぶあたらしい統計学の教科書　第 2 版（馬場真哉，翔泳社，2022）

統計学の基本的な内容を端的な言葉で解説しながら，Python の基本的な操作方法についても学べる書籍です．この講義の内容を異なる文脈で解説してくれているので，講義内容の理解を深めるために役立ちます．

（4） 計量経済学（西山慶彦，新谷元嗣，川口大司，奥井亮，有斐閣，2019）

計量経済学領域で用いられる統計解析に関して網羅されている，当該分野の標準的教科書のひとつです．幅広い範囲がカバーされているので，経済学領域の学生はもちろん，他領域の学生にとっても参考となる一冊です．経済学分野で使われる用語の用法の参考にもなります．

（5） An Introduction to Statistical Learning with Application in R/Python (accessed 17 January 2025)

機械学習の基本的な考え方からその実践方法まで，難解な数式を使わずに説明してくれています．機械学習を直感的に理解して，まずは使ってみたいという方に非常におすすめです．もともとは R 版が出版されていたが，Python 版も発売されました．R 版については，R による統計的学習入門（訳：落海浩・首藤信通）として日本語翻訳版が出版されています．

（6） 現代数理統計学の基礎（久保川達也，共立出版，2017）

統計学の数理的な側面をもっと深く学びたい方におすすめです．数学的に深入りはし過ぎない程度に，定理などが簡潔にまとめられており，非常にバランスがよい名著．演習も充実しています．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

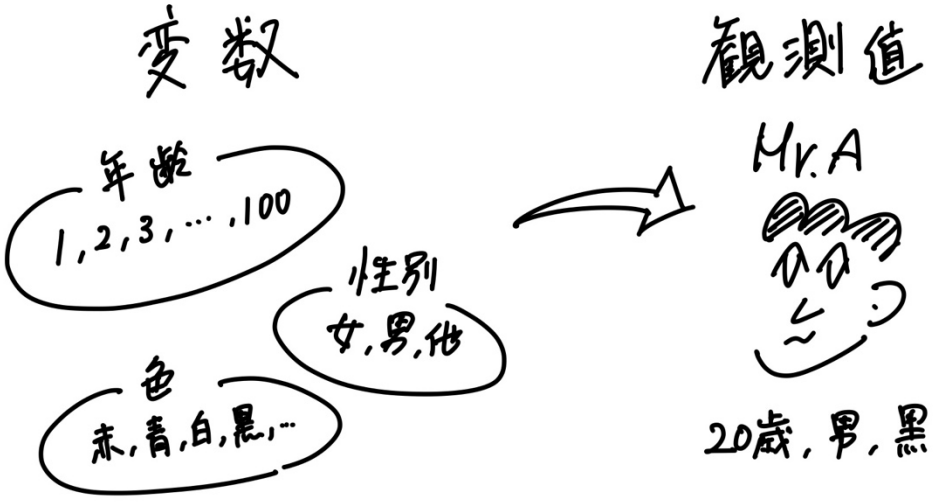
1. データの種類

用語の定義

この講義では，実験や観察の結果として得られた事実や科学的数値を「データ」と呼びます．その中でも特定の特徴と範囲をもつ個々の要素を「変数」と呼びます．例えば年齢は，人の生まれてからの経過年数によってカウントされる 0 歳から 100 歳くらいまでの範囲をもつ変数です．ただし変数そのものは，それぞれデータの値ではなく，それ自体には色々な値が入り得ます．実際にその変数に関する数値が観測されたとき，その数値のことを「観測値」や「実現値」と呼びます．

例えば大学生 A さんの年齢という変数について測定したとき，20 歳であるというデータが観測値になります．この講義では，データの種類などを表現するときには「変数」と表現し，その変数について実際に観測された値であることを強調したい時には「観測値」という用語を使うことにします．またもう少しあいまいに観測値の集合を表現するときなどは「データ」と表現し，観測値などのカウント数を表現するときには「データ数※」と表現することにします．

※　特にデータを取得したヒトやモノの数を明示するときは「サンプルサイズ」という用語を使うことがあります．



これらの用語の定義は，分野や文脈によって大きく異なりますので，この講義ではこのように定義している，というふうに理解しておいてください．特に統計学は幅広い分野で応用されている学問であり，同じ事象について異なる用語が使われていたり，同じ用語で異なる事象を表現していたりすることは珍しくありません．まずは与えられた知見をもとに理解を進めるということに慣れていってもらえれば幸いです．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

変数の種類

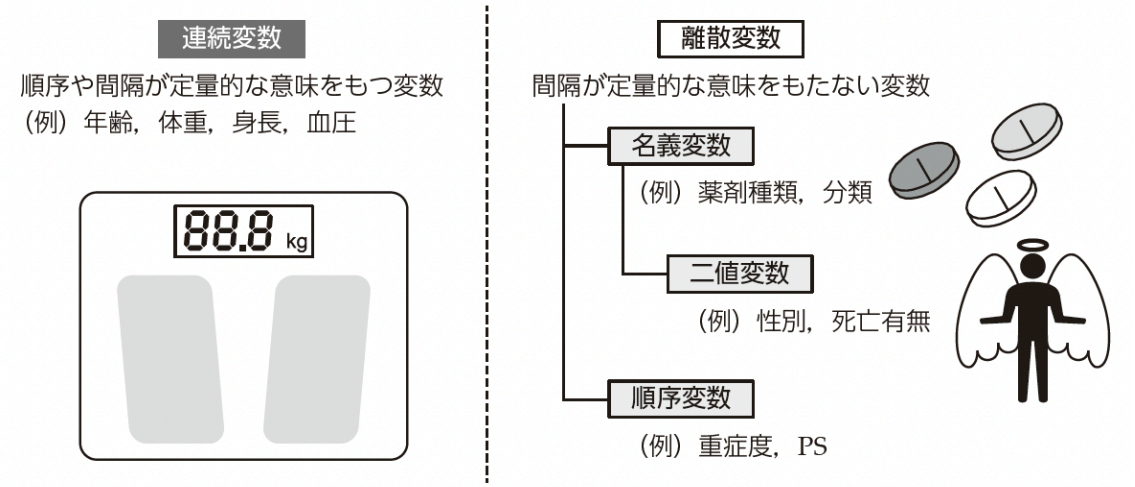
変数の種類は大きく「連続変数」と「離散変数（質的変数，名義変数，あるいはカテゴリー変数とも呼ばれる）」に分けられます．なお以下で示す定義はデータサイエンスや統計学領域で利用されるものであり，数学的な定義や，データ分析のプログラミング言語としてのデータの型とは異なるものであることに注意してください．

（１）　連続変数：間隔や順序が定量的な意味をもつ変数（年齢や体重など）
（２）　離散変数：間隔や順序が定量的な意味をもたない変数（分類やランクなど）

また離散変数については，その内容によって，ふたつにひとつの値をとる「二値変数」，間隔が定量的ではなくとも大小関係などがある「順序変数」，それ以外を単純に「名義変数」と呼んだりします（ここでは大区分の離散変数と区別するために，あえてこのような表現を採用しています）．

また特に二値変数を{0,1}の数値で表現した変数を「ダミー変数」と呼びます．これの扱いは離散変数（二値変数）と変わりませんが，ダミー変数にすることでその順序が明確になり，計算や分析で扱いやすくなります．

図表．変数の種類



神戸大学　数理・データサイエンスセンター
加葉田大志朗

2. データの特徴を図表で捉える

取得されたままのデータの羅列，つまり生のデータから，起きている事象の背景や意味合いを理解することは困難です．データを何らかの基準に基づいて整理し，データのもつ特徴（癖）を汲み取る作業が必要となります．この作業を「統計的記述」と呼び，大きく以下 2 つの手続きがとられます．

（1）図表化する
（2）データを代表する値（統計量）で表す

ここではデータの特徴を捉えるために，（1）図表化する方法について学びます．

今回は例として，大学生 200 名の講義の評点（以下，講義評点と呼びます）のデータを利用します．講義評点は[0,100]の区間の整数で与えられる連続変数です．以下に 200 名分の観測値を，観測した順で並べています．このデータをうまく加工して，データの特徴を捉えてみましょう．

図表．講義評点の生データ

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	64	84	79	74	81	79	73	82	73	84
[2,]	89	58	75	73	83	78	65	69	77	75
[3,]	73	72	81	59	66	72	75	69	80	77
[4,]	71	86	70	78	61	83	72	59	80	81
[5,]	77	71	70	83	83	69	56	85	71	84
[6,]	95	75	69	63	76	72	79	76	80	76
[7,]	60	80	67	70	69	81	89	70	82	76
[8,]	94	63	70	68	70	79	82	81	85	87
[9,]	87	73	79	76	76	59	68	76	72	84
[10,]	77	65	71	70	71	76	71	83	78	61
[11,]	68	72	75	87	86	60	68	80	76	78
[12,]	71	63	73	83	66	79	67	70	78	75
[13,]	81	69	73	84	85	54	91	70	86	80
[14,]	65	85	73	63	65	89	79	81	63	74
[15,]	96	78	68	70	79	84	76	76	72	67
[16,]	67	81	73	68	80	92	80	86	77	67
[17,]	52	92	68	89	80	71	75	70	57	82
[18,]	70	62	74	80	88	69	86	77	74	87
[19,]	78	91	85	76	77	73	73	69	71	74
[20,]	73	68	82	72	56	78	76	71	92	60

神戸大学　数理・データサイエンスセンター
加葉田大志朗

上記の評点に関するデータの特徴を捉えやすくするため，とりあえずデータを小さい値から順に並べ替えておきます．少し見やすくはなりましたが，まだまだ全容の把握には至りません．

図表．小さい順に並べた講義評点の生データ

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	52	54	56	56	57	58	59	59	59	60
[2,]	60	60	61	61	62	63	63	63	63	63
[3,]	64	65	65	65	65	66	66	67	67	67
[4,]	67	67	68	68	68	68	68	68	68	68
[5,]	69	69	69	69	69	69	69	69	70	70
[6,]	70	70	70	70	70	70	70	70	70	70
[7,]	71	71	71	71	71	71	71	71	71	71
[8,]	72	72	72	72	72	72	72	72	73	73
[9,]	73	73	73	73	73	73	73	73	73	73
[10,]	74	74	74	74	74	75	75	75	75	75
[11,]	75	75	76	76	76	76	76	76	76	76
[12,]	76	76	76	76	76	77	77	77	77	77
[13,]	77	77	78	78	78	78	78	78	78	78
[14,]	79	79	79	79	79	79	79	79	80	80
[15,]	80	80	80	80	80	80	80	80	81	81
[16,]	81	81	81	81	81	81	82	82	82	82
[17,]	82	83	83	83	83	83	83	84	84	84
[18,]	84	84	84	85	85	85	85	85	86	86
[19,]	86	86	86	87	87	87	87	88	89	89
[20,]	89	89	91	91	92	92	92	94	95	96

神戸大学　数理・データサイエンスセンター
加葉田大志朗

離散変数の度数分布表と棒グラフ

講義評点は F(-59), C(60-69), B(70-79), A(80-89), S(90-100)の 5 つのカテゴリーを持つ離散変数として学生に提示されます（これを連続変数である講義評点と区別して評点グレードと呼ぶことにします）．ここではまず，この評点グレードの各カテゴリーに属する学生の数や割合について計算し，その分布の特徴を調べてみましょう．

図表．講義評点グレードの生データ

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "F"  "F"  "F"  "F"  "F"  "F"  "F"  "F"  "F"  "F"
 [2,] "F"  "F"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"
 [3,] "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"
 [4,] "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"
 [5,] "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"
 [6,] "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"  "C"
 [7,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
 [8,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
 [9,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
[10,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
[11,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
[12,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
[13,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
[14,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"
[15,] "B"  "B"  "B"  "B"  "B"  "B"  "B"  "B"  "A"  "A"
[16,] "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"
[17,] "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"
[18,] "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"
[19,] "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"
[20,] "A"  "A"  "S"  "S"  "S"  "S"  "S"  "S"  "S"  "S"
```

ここでそれぞれのカテゴリーについて，以下のように並べてまとめてみましょう．すると各カテゴリーに該当する学生の人数の相対的な大小関係がより明白になりますね．

図表．カテゴリーごとに並べた場合

```
FFFFFFFFFFF

CCCC
```

神戸大学　数理・データサイエンスセンター
加葉田大志朗

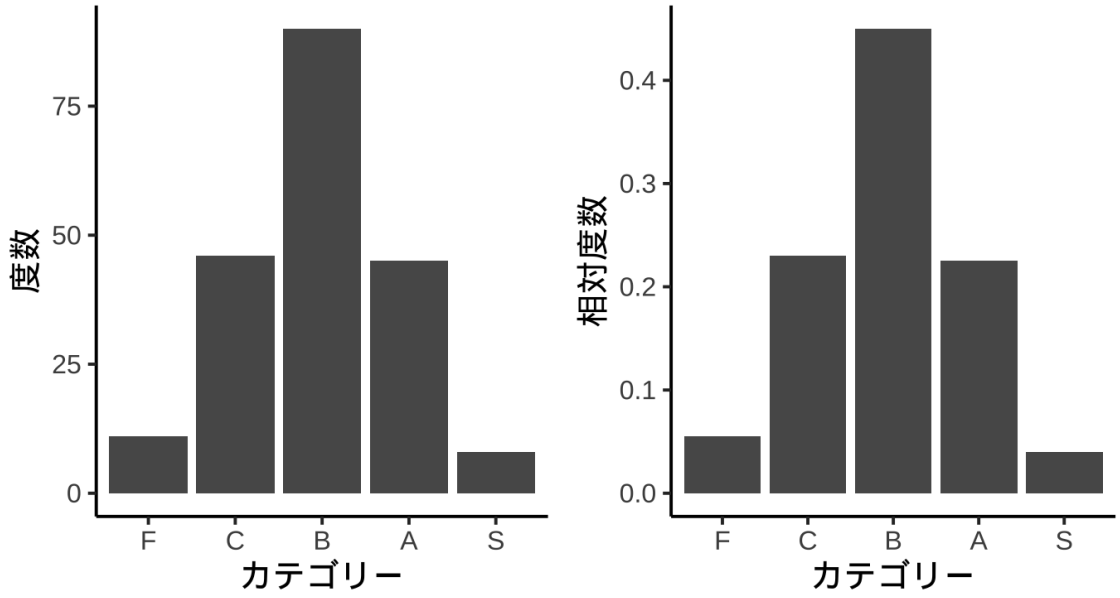
手順１．各カテゴリーに属するデータ数「度数」を算出
手順２．各カテゴリーに属するデータ数の全データ数における割合「相対度数」を算出
上記の手続きに沿って今回のデータで度数分布表を作成すると，以下のようになります．

図表．講義評点グレードの度数分布表

カテゴリー	度数	相対度数
F	12	0.06
C	48	0.24
B	88	0.44
A	44	0.22
S	8	0.04

離散変数についてデータの分布を視覚的にわかりやすくするために，この度数分布表をもとに「棒グラフ」が作成されることがあります．棒グラフは度数を縦軸とする場合と，相対度数を縦軸とする場合があります．縦軸以外グラフの見た目は変わりませんが，相対度数を縦軸とする場合は，全てのグラフの高さの合計は１となります．

図表．棒グラフ（左：度数が縦軸，右：相対度数が縦軸）



神戸大学　数理・データサイエンスセンター
加葉田大志朗

上記の図から,

（1） 講義評点グレードBに属する学生が最も多い

（2） 次いで講義評点グレードAおよびCに属する学生が同程度に多い

（3） さらに次いで講義評点グレードSおよびFに属する学生が同程度に多い

ということが一目瞭然にわかるようになりました.

神戸大学　数理・データサイエンスセンター
加葉田大志朗

連続変数の度数分布表とヒストグラム

　続いて，連続変数である講義評点についても同様に図表を利用した要約を行ってみましょう．基本的な考え方や手順は，離散変数を扱ったときとほとんど変わりません．連続変数について以下の手順で度数分布表を作成し，それを元に棒グラフを作成します．ただしこの手順で作成された棒グラフの集合は「ヒストグラム」と呼ばれます．

手順１．データを任意の区間（これを「階級」と呼ぶ）に切り分ける
手順２．各階級について，度数および相対度数を算出する

今回は例として，[0,100]の範囲をとる連続変数である講義評点について，（１）５点刻みの階級に切り分け，（２）度数および相対度数を計算して度数分布表にまとめましょう．

図表．連続変数の観測値を階級に切り分ける

講義評点	階級
64	(60,65]
84	(80,85]
79	(75,80]
74	(70,75]
81	(80,85]

図表．連続変数を５点刻みで切り分けたときの度数分布表

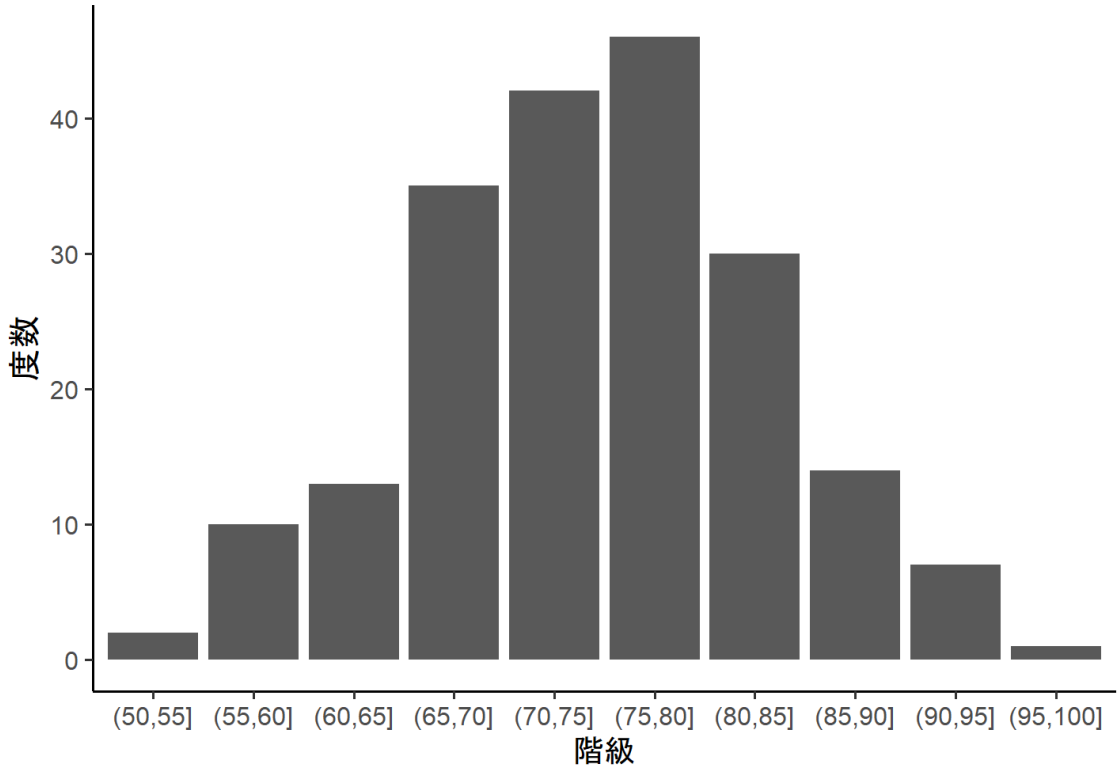
階級	度数	相対度数
(0,5]	0	0
(5,10]	0	0
(10,15]	0	0
(15,20]	0	0
(20,25]	0	0

(25,30]	0	0
(30,35]	0	0
(35,40]	0	0
(40,45]	0	0
(45,50]	0	0
(50,55]	2	0.01
(55,60]	10	0.05
(60,65]	13	0.065
(65,70]	35	0.175
(70,75]	42	0.21
(75,80]	46	0.23
(80,85]	30	0.15
(85,90]	14	0.07
(90,95]	7	0.035
(95,100]	1	0.005

上記の度数分布表に基づくヒストグラムは以下の通りです.

図表. 講義評点のヒストグラム（階級は 5 点刻み）

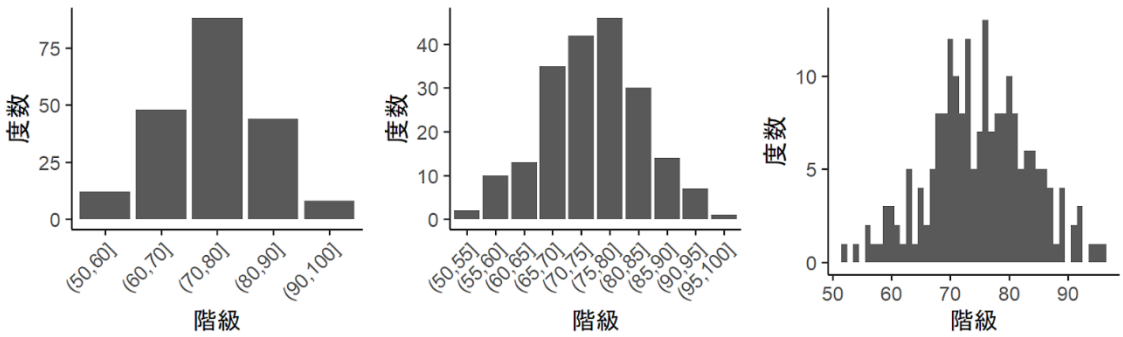
神戸大学　数理・データサイエンスセンター
加葉田大志朗



上記のヒストグラムから，講義評点について以下の特徴が読み取れます.
（１）階級(75-80]点に属する学生が最も多い
（２）階級(75-80]点から離れた階級ほど学生の数がおおむね左右対称に減っていく

ちなみに同じデータでも階級を区切るための区間が異なると，視覚的に異なる印象となる場合があるため注意が必要です．例として，10 点刻み，5 点刻み，1 点刻みの階級を以下に示します.

図表．階級の区間の違いによるヒストグラムの違い



神戸大学　数理・データサイエンスセンター
加葉田大志朗

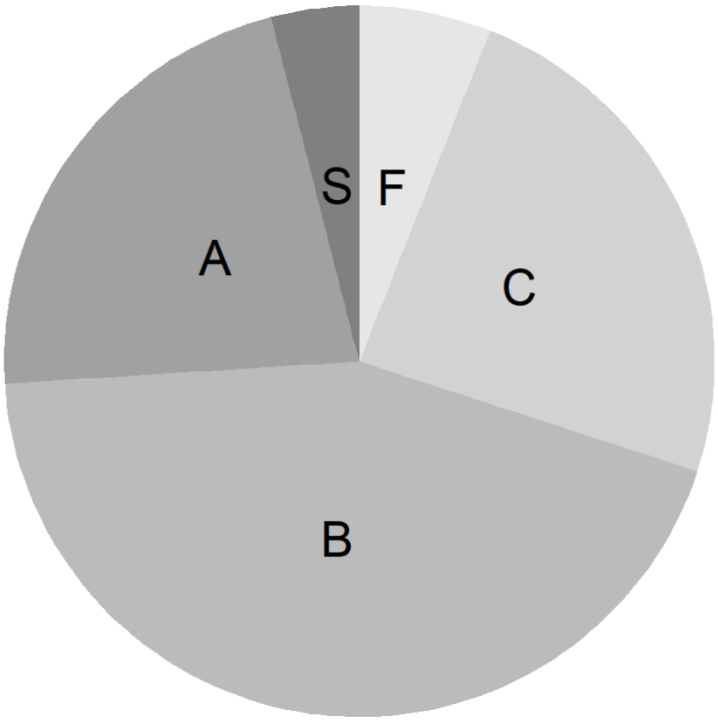
その他のグラフ

上記では，離散変数の各カテゴリーに含まれるデータ数の相対的な大小関係を可視化するための「棒グラフ」と，連続変数についてデータが集中している場所を可視化する「ヒストグラム」を紹介しました．ここでは離散変数の各カテゴリーに含まれる観測数の割合を可視化するためにメディアなどでよく利用される「円グラフ」や「積み上げ棒グラフ」について紹介します．

円グラフ

離散変数について，全観測値に対して各カテゴリーの観測値が占める割合（相対度数）を円の中の扇形の面積で示すグラフです．例えば試験の難易度が適切かどうか検討するために，上述の講義評点グレード F(-59), C(60-69), B(70-79), A(80-89), S(90-100)のバランスを評価する場合などに利用されます．

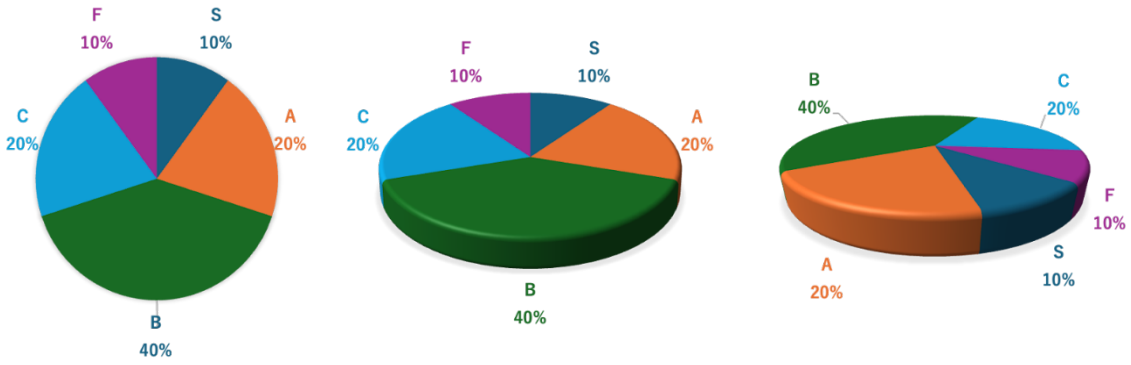
図表．講義評点グレードの円グラフ



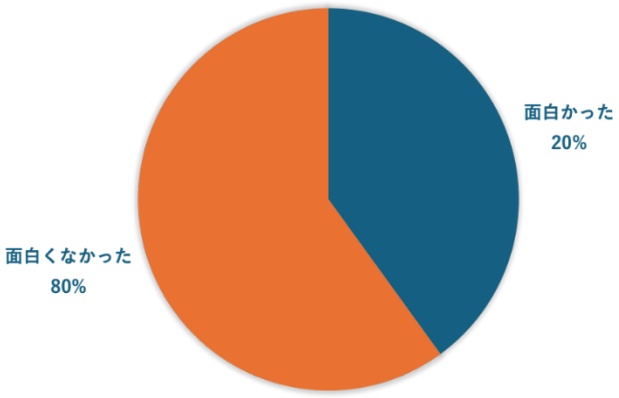
神戸大学　数理・データサイエンスセンター
加葉田大志朗

円グラフはそれ自体は単純なものであるが故か，傾斜をつけたりと様々な装飾を施して表現されることが多い印象です．ただしグラフの過度な装飾は視覚的なミスリードを誘い得るため，利用しないことを推奨します．またメディアなどでは，相対度数の値と円グラフの面積が一致していないことも散見されるため，情報の受け手としても注意が必要です．

図表．視覚的なミスリードを誘う円グラフの例（左：装飾なし，中・右：装飾あり）



図表．相対度数と各カテゴリーの面積が一致していない例



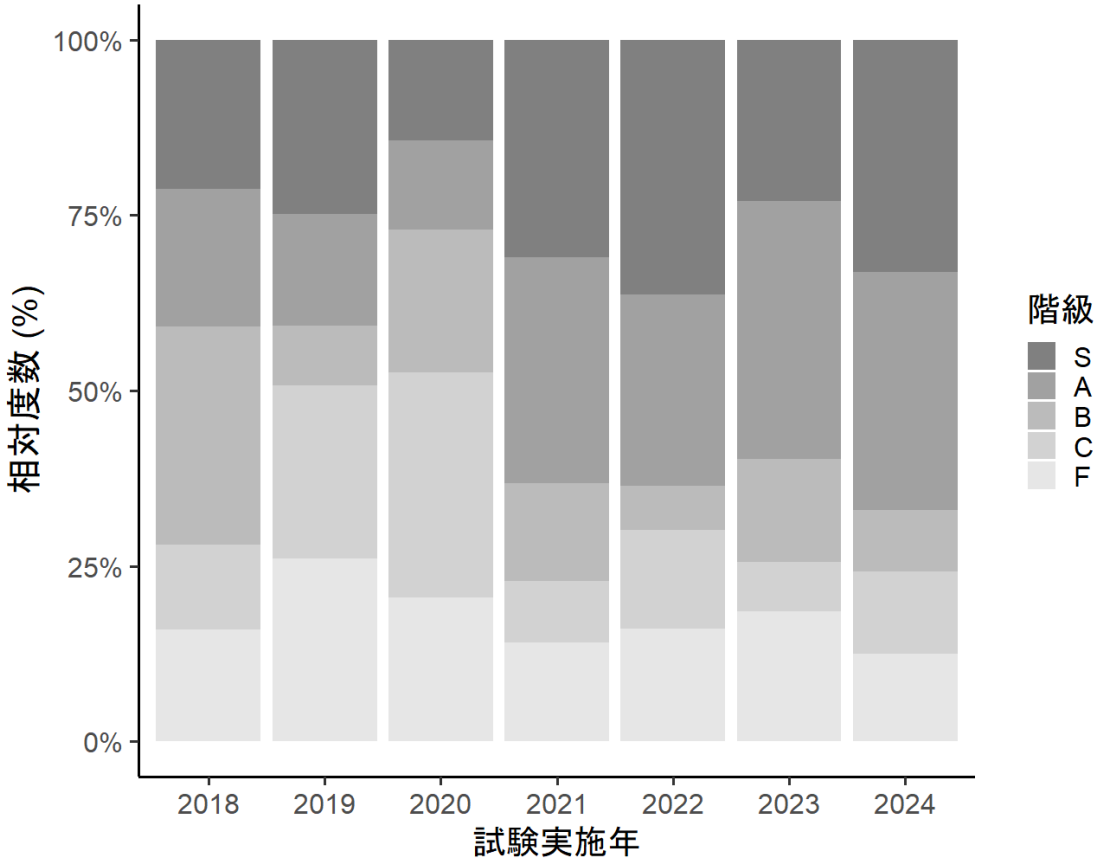
100%積み上げ棒グラフ

離散変数について，全観測値に対して各カテゴリーの観測値が占める割合（相対度数）を1本の棒の中の面積で示すグラフです．特に複数本並べて，各カテゴリーが占める割合の推移を知りたい時などに使われることが多いです．

例えば以下は講義評点グレードの割合を示す100%積み上げ棒グラフです．ここでは試験問題を易化させた2021年からA(80-89)，S(90-100)の講義評点グレードに属する学生の割

合が増加したことが視覚的にわかります.

図表. 講義評点グレードの各カテゴリーが占める割合の年次推移



ここで学んだグラフの見方がわかれば，厚生労働省が毎年実施している国民生活基礎調査の結果などもよく理解できます．ぜひ一度見てみてください※.
https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa22/dl/14.pdf
（統計情報･白書 ＞ 各種統計調査 ＞ 厚生労働統計一覧 ＞ 国民生活基礎調査 ＞ 結果の概要 ＞2022（令和４）年　国民生活基礎調査の概況 ＞ 概況全体版）

※ 上記は 2022 年調査版を利用していますので，視聴時の最新版を参照してみると良いでしょう.

神戸大学　数理・データサイエンスセンター
加葉田大志朗

3. データの特徴を統計量で捉える

取得されたままのデータの羅列，つまり生のデータから，起きている事象の背景や意味合いを理解することは困難です．データを何らかの基準に基づいて整理し，データのもつ特徴（癖）を汲み取る作業が必要となります．この作業を「統計的記述」と呼び，大きく以下 2 つの手続きがとられます．

（1）図表化する
（2）データを代表する値（統計量）で表す

ここではデータの特徴を捉えるために，（2）データを代表する値（統計量）で表す方法について学びます．

離散変数については，各カテゴリーについて，そのデータ数である「度数」と，データの集合内におけるその割合である「相対度数」で表現します．これについては「データの特徴を図表で捉える」を参照してください．

ここでは連続変数について，統計量を使って分布の特徴を表現する方法を学びます．連続変数については基本的に，

（1） 分布の位置を示す統計量
（2） 分布のばらつきを示す統計量

でデータの特徴は表現されます．それぞれの統計量の代表的なものについて解説します．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

分布の位置を示す統計量①：平均値

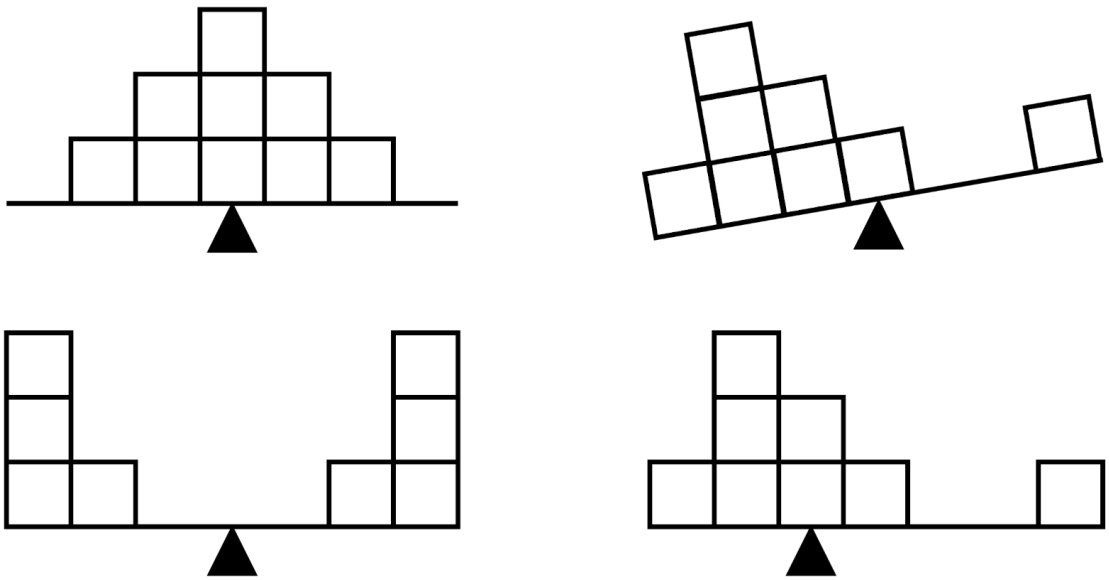
分布の位置を示す統計量の代表的なものとして「平均値」があります．平均値にはいくつかの種類がありますが，ここでは一般的によく利用される「算術平均」について紹介します．

$$算術平均 = \frac{全ての観測値の合計}{データ数}$$

計算式自体は簡単で困ることはありませんが，算術平均がどのような意味をもつのかについて，きちんと説明できるようにしておきたいと思います．

算術平均は「データの分布の左右のバランスがつりあう位置」として考えることができます．もう少し厳密に表現すると，算術平均を支点としたモーメントが左右対称となる位置となります．これは必ずしも，分布の形状が左右対称となっているという意味ではありません．またこの性質はデータの分布の形状にも依存しません．以下の図を見てもらうとよりはっきりとイメージできるでしょう．

図表．ヒストグラムにおける算術平均の位置（左２つ：左右対称，右２つ：左右非対称）



算術平均は「データの分布の左右のバランスがつりあう位置」と表現できますが，その位置は必ずしもデータの分布の中央（下から数えて真ん中）となるとは限りません．上の図の通り，分布が左右対称でないときは，算術平均はデータの分布の中央とは異なる位置にあります．

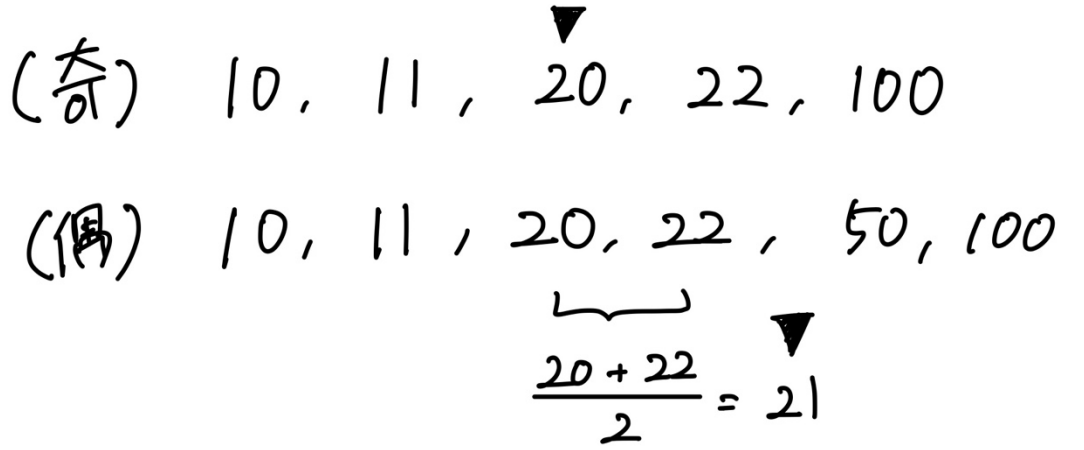
神戸大学　数理・データサイエンスセンター
加葉田大志朗

分布の位置を示す統計量②：中央値

上述の通り，算術平均は必ず「データの分布の左右のバランスがつりあう位置」を示してくれますが，必ずしもデータの分布の中央を捉えられるわけではありません．ここで分布の中央に位置する値に関心がある場合には「中央値」を利用します．中央値は単純に「観測値を最小の値から数えて 50%目に位置する値」と言えますが，データ数が奇数か偶数かによって以下のような数式で表現できます．

$$中央値 = \frac{データ数+1}{2}番目の観測値\left(データ数が奇数\right)$$

$$中央値 = \frac{データ数}{2}番目の観測値と\frac{データ数}{2}+1番目の観測値の算術平均\left(データ数が偶数\right)$$



中央値はデータのちょうど真ん中を示してくれるので，特に以下のような場面で役に立ちます．

（1）分布の形状が左右対称でない場合
（2）極端な観測値が含まれる場合

それぞれの状況について，中央値は算術平均とどのように異なるのか，確認してみます．

（1）分布の形状が左右対称でない場合

分布の形状が左右対称であれば，算術平均は中央値と一致します．しかし以下図のように分布の形状が左右対称でない場合には，平均値は分布の裾がのびる方向に寄っていきます．なお以下の左右対称でない分布のうち，

- 裾が右に伸びる分布：「右裾が長い」「右に歪んだ」「正の歪みをもつ」「左に偏った」分布と呼ぶことがある

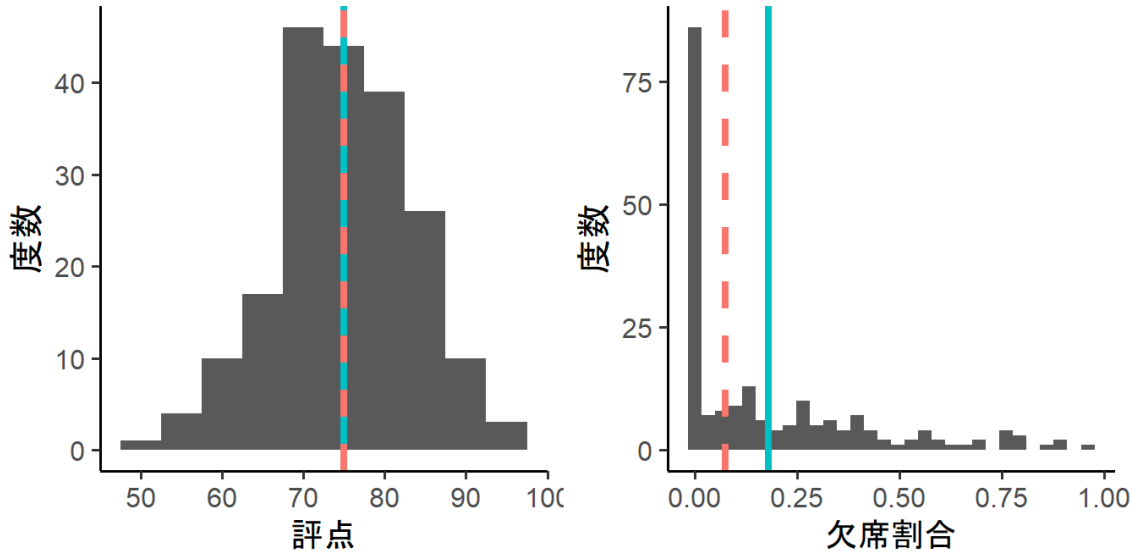
- 裾が左に伸びる分布：「左裾が長い」「左に歪んだ」「負の歪みをもつ」「右に偏った」分布と呼ぶことがある

ことについて触れておきます．裾が右に伸びる分布は，年収や医学分野におけるバイオマーカーなどでよく見られます．

図表．左裾が長い分布（左），左右対称の分布（中），右裾が長い分布（右）



図表．分布における中央値の位置（青実線：平均値，赤破線：中央値）



上述の通り，中央値はデータの分布によらず，データの中央を示すことができます．従ってデータが集中している場所を記述したい時には，中央値を使えば問題はありません．医学領域などで「分布が左右対称のときには平均値，左右非対称のときには中央値を使う」といった慣例を聞くこともありますが，データの真ん中を捉えることが目的であれば中央値を使っておけば問題はありません．

（2）極端な観測値が含まれる場合

算術平均は極端なデータが含まれる場合に，そのデータが存在する方向に中央値から離れていく性質があります．例として，ある高校の卒業生で現在 30 歳の 240 名に年収に関す

るアンケートの結果を考えてみます. 240 名のうち 239 名がアンケートを返送してくれて，その観測値から計算された年収の算術平均と中央値は以下の通りとなりました.

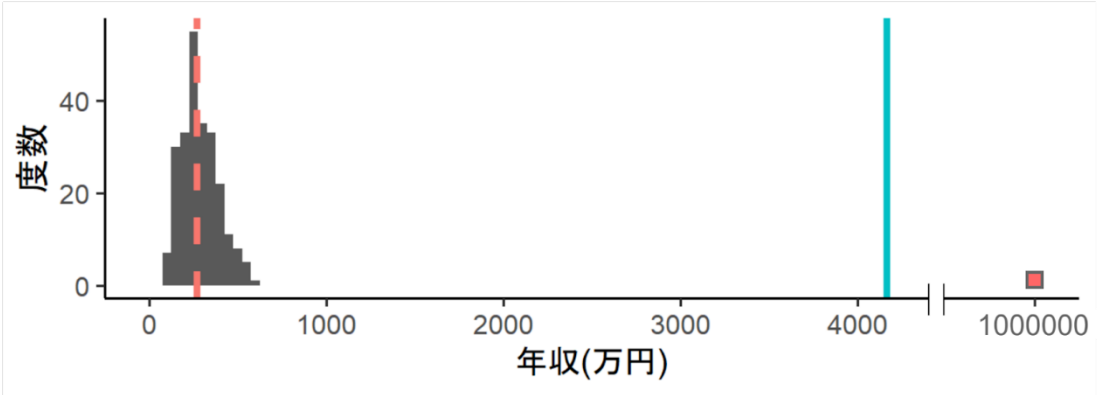
※　この例は令和 5 年賃金構造基本統計調査 結果の概況などを参考に作成しており，実際の人物・団体に関係はありません.

- 算術平均 ≃ 289 万円
- 中央値 ≃ 272 万円

その後，同学年の卒業生である野球選手からアンケートが返送され，彼の年収 100 億円も加えて再計算をすることになりました．すると年収の算術平均と中央値は以下の通りとなりました.

- 算術平均 ≃ 4167 万円 ((289 万円×239 人)+(10,000,000,000 円×1 人))/240 人
- 中央値 ≃ 272 万円

図表．240 名の年収分布（青実線：平均値，赤破線：中央値）



この例から分かるように，極端なデータが含まれることで算術平均は大きく正の方向へ寄った値となります．一方，データの順位は 1 人分しか変わらないため，中央値にはほとんど影響はありませんでした．このように極端なデータが含まれる場合にも，中央値はデータが多く集まる場所を示すことに役立ちます．算術平均と中央値に関する有名な例はたくさんあり，他にも「ビルゲイツが離婚したらワシントン州の独身男性の平均資産が 500 万円増える」や「若者の平均貯蓄額は 200 万円なのに中央値は 0 円（半分以上のひとは貯金がない)」などが挙げられます.

神戸大学　数理・データサイエンスセンター
加葉田大志朗

データのばらつきを捉える統計量①：標準偏差

データの分布の特徴を捉えるためには，データのばらつき具合，つまり分布がどのような広がりをもつかを確認することも重要です．

例えば，講義評点の平均値が 75 点だったときに自分の講義評点が 80 点だったとします．これは喜ぶべきことかでしょうか？このときに考える必要があるのは「自分のデータの平均からの離れ具合（今回であれば 5 点）」が「集団の（平均的な）平均からの離れ具合」よりも相対的に大きいか小さいか？ということです．それぞれの要素について，具体的な手続きを説明します．

（1）各データと平均値の離れ具合「偏差（deviation）」を計算する
（2）各データの偏差の平均的な値「標準偏差（standard deviation, SD）」を計算する
（3）自分の偏差が，標準偏差よりも相対的に大きいか小さいかを評価する

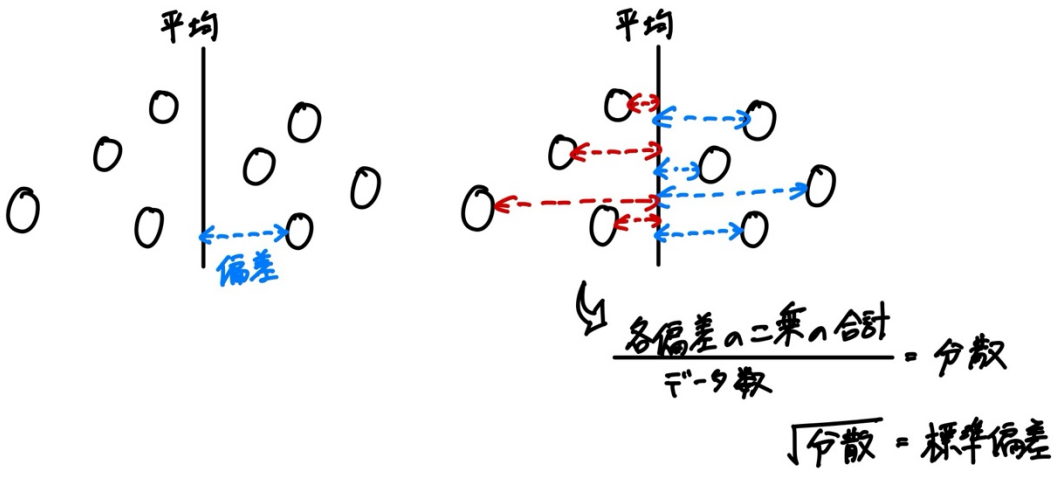
ここで（2）標準偏差の計算について，全員の偏差の算術平均を計算すると必ず 0 になってしまいます．つまりデータにばらつきがないように見えてしまいます．これは避けたいので，以下の手続きで偏差の「二乗平均」を計算することで，標準偏差は計算されます．

手続き 1．すべての観測値の偏差の二乗した値の算術平均「分散（variance）」を計算する

$$分散 = \frac{偏差^2 \text{ の合計}}{データ数}$$

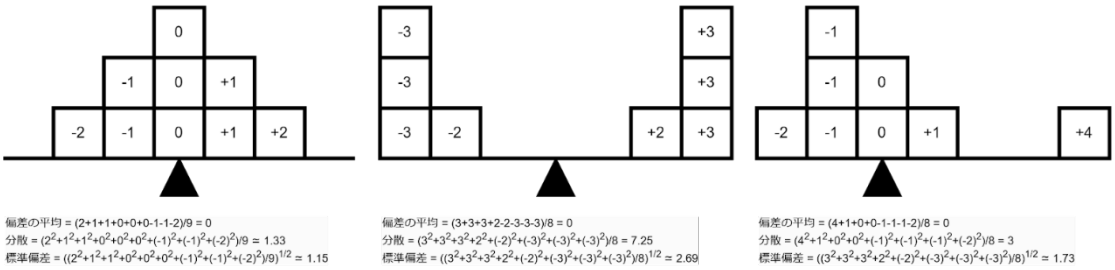
手続き 2．分散のルート（二乗根）をとる

$$標準偏差 = \sqrt{分散}$$



神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．偏差の算術平均，分散，標準偏差（ブロック内の数値は偏差）

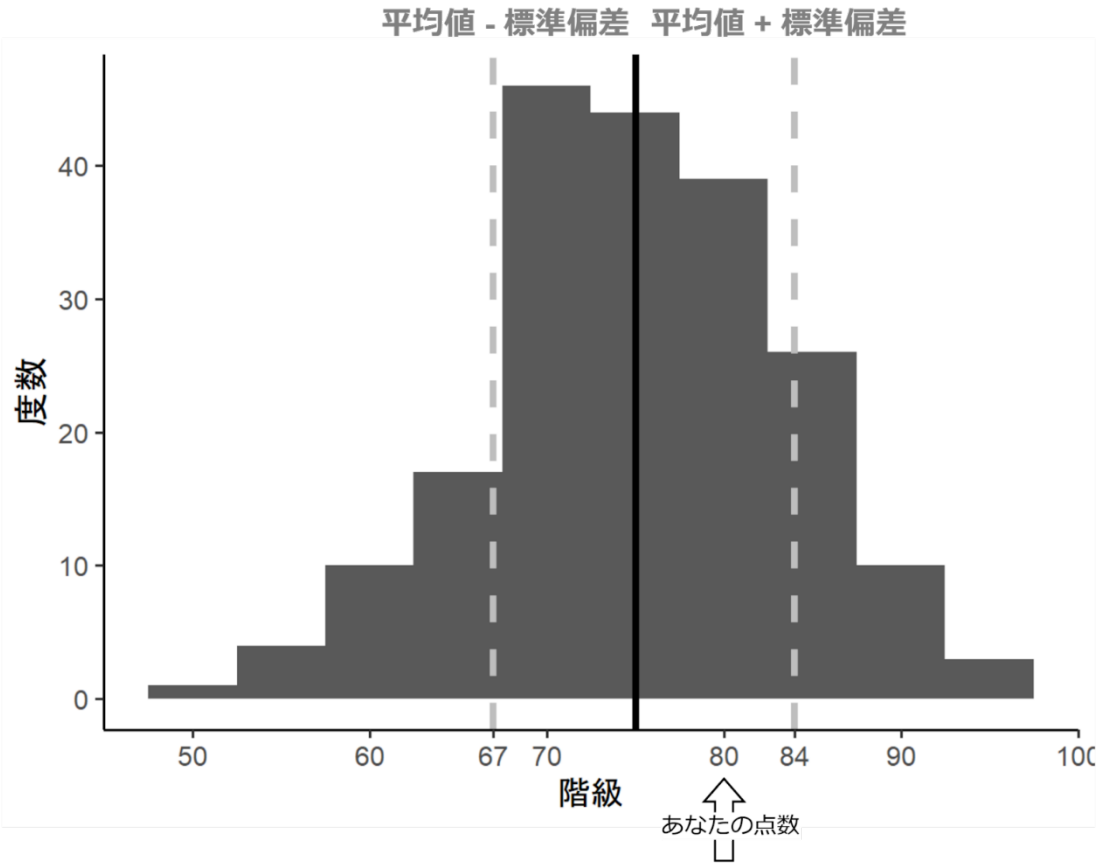


上記の手続きで計算した講義評点の標準偏差は 8.5 点でした．これらより自身の講義評点の偏差（5 点）は，集団の平均からの平均的な離れ具合である標準偏差（8.5 点）と比べて，相対的に小さいことがわかります．

実際の講義評点で表現すれば，評点の平均値は 75 点なので，　75±8.5 点（66.5 点から 83.5 点）の範囲は「集団の平均値付近の平均的なばらつきの範囲内」であると言えそうです．従って自身の講義評点 80 点は，集団の平均的なばらつきの範囲内にあると言えそうです．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．自分のデータが平均的なばらつきの範囲内に含まれている



この自身の評点 80 点が，クラスの講義評点のどの程度の位置（上位何%か等）の具体的な位置を推し量るためには，別途説明する正規分布など特定の分布の性質を使う必要があります．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

データの加工と偏差値

自分がどれだけ特殊な点数をとったのかは，自分の偏差と標準偏差との乖離具合を比較することによって評価できました．ただし異なる試験を受ければ，当然ですが試験ごと平均値も標準偏差も異なります．試験ごとに異なる平均値と標準偏差と自分の点数を相対的に見比べることはとても大変なので，何か便利なツールが欲しいものです．ここで役に立っているのがおなじみの「偏差値」です．

ご存じのとおり，偏差値 50 は試験を受けた集団のちょうど平均的な点数を示します．つまり偏差値は平均値が 50 になるように加工された点数であると考えることができます．これによって試験ごとに平均値が異なっても，自身の偏差値が 50 より上か否かで，平均以上の点数であったかどうかを知ることができます．

ただしここでも，偏差値 50 からどの程度離れていれば，自身の偏差値が特別なものであるか？を考える必要があります．ここで知る必要がある情報として，偏差値は標準偏差が 10 となるように加工された点数であるということです．これまでの説明をなぞると，自身の偏差値が 50±10（40 から 60）の範囲から逸脱していれば，集団の平均的なばらつきの範囲の外にある特殊な偏差値であると言えます．

上記をまとめると，偏差値は「平均値が 50，標準偏差が 10 となるように加工した観測値」であると言えます．この説明だけでは良く分からないと思います．ここで理解が必要なのが「データの加工」の手続きです．

データの加工の具体的な手続き

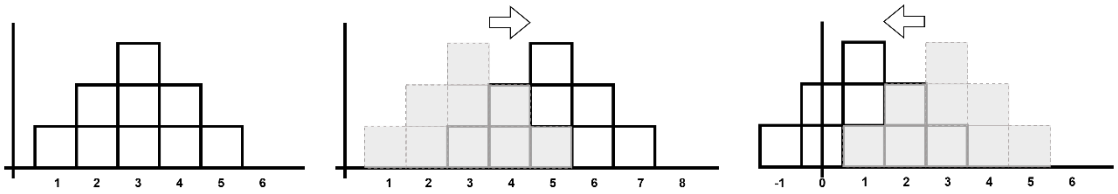
データを分析する際には，それぞれの分析に応じて適切な加工を行うことがあります．偏差値を計算するときにも，ここで紹介するデータの加工が利用されています．ここでは最も基本的な 2 つのデータの加工方法を紹介します．

（1） データに定数を足す（引く）
（2） データに定数を掛ける（割る）

まずは（1）データに定数を足す（引く）加工について紹介します．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．データに定数を足す（引く）加工によって生じる分布の変化



項目	加工前	2 を加える	2 を引く
データ値	$2,3,3,4,4,4,5,5,6$	$4,5,5,6,6,6,7,7,8$	$0,1,1,2,2,2,3,3,4$
平均値	$\frac{2+3+3+4+4+4+5+5+6}{9}=4$	$\frac{4+5+5+6+6+6+7+7+8}{9}=6$	$\frac{0+1+1+2+2+2+3+3+4}{9}=2$
偏差	$-2,-1,-1,0,0,0,1,1,2$	$-2,-1,-1,0,0,0,1,1,2$	$-2,-1,-1,0,0,0,1,1,2$
分散	$\frac{4+1+1+0+0+0+1+1+4}{9}=1.33$	$\frac{4+1+1+0+0+0+1+1+4}{9}=1.33$	$\frac{4+1+1+0+0+0+1+1+4}{9}=1.33$
標準偏差	$\sqrt{1.33}\approx 1.15$	$\sqrt{1.33}\approx 1.15$	$\sqrt{1.33}\approx 1.15$

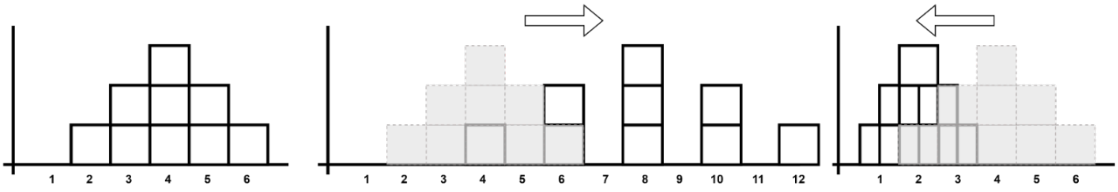
上記を一般化すると，「（1）データに定数 a を足す」という加工を加えると，加工後のデータの統計量には以下の変化が生じます．

- 加工後の算術平均 ＝ 加工前の算術平均 **＋ a**
- 加工後の分散 ＝ 加工前の分散（変化なし）
- 加工後の標準偏差 ＝ 加工前の標準偏差（変化なし）

つまり（1）データに定数を足す（引く）加工は「分布の位置のみをずらす変換」であることが分かります．

続いて（2）データに定数を掛ける（割る）加工について紹介します．

図表．（2）データに定数を掛ける（割る）加工によって生じる分布の変化



項目	加工前	2 を掛ける	2 で割る
データ値	$2,3,3,4,4,4,5,5,6$	$4,6,6,8,8,8,10,10,12$	$1.0,1.5,1.5,2.0,2.0,2.0,2.5,2.5,3.0$
平均値	$\frac{2+3+3+4+4+4+5+5+6}{9}=4$	$\frac{4+6+6+8+8+8+10+10+12}{9}=8$	$\frac{1.0+1.5+1.5+2.0+2.0+2.0+2.5+2.5+3.0}{9}=2$
偏差	$-2,-1,-1,0,0,0,1,1,2$	$-4,-2,-2,0,0,0,2,2,4$	$-1.0,-0.5,-0.5,0.0,0.0,0.0,0.5,0.5,1.0$
分散	$\frac{4+1+1+0+0+0+1+1+4}{9}=1.33$	$\frac{16+4+4+0+0+0+4+4+16}{9}=5.33$	$\frac{1+0.25+0.25+0+0+0+0.25+0.25+1}{9}=0.33$
標準偏差	$\sqrt{1.33}\approx 1.15$	$\sqrt{5.33}\approx 2.31$	$\sqrt{0.33}\approx 0.58$

神戸大学　数理・データサイエンスセンター
加葉田大志朗

上記を一般化すると，「（２）データに定数 a を掛ける」という加工を加えると，加工後のデータの統計量には以下の変化が生じます．

- 加工後の算術平均 ＝ 加工前の算術平均 × **a**
- 加工後の分散 ＝ 加工前の分散 × $\mathbf{a^2}$
- 加工後の標準偏差 ＝ 加工前の標準偏差 × **a**

つまり（２）データに定数を掛ける（割る）加工は「分布の位置をずらしながら，分布のばらつき具合を変える変換」であることが分かる．この加工を分布の尺度（スケール）を変換するという意味で，「スケーリング」などと呼ぶことがあります．

データの加工を利用した偏差値の計算

それでは上記で紹介したデータの加工を利用して，偏差値を計算してみましょう．以下の手続きを踏んで，講義評点を「平均値が 50，標準偏差が 10」となるようにデータを加工してみます．

手続き１．講義評点を「平均値が０，標準偏差が１」に加工する

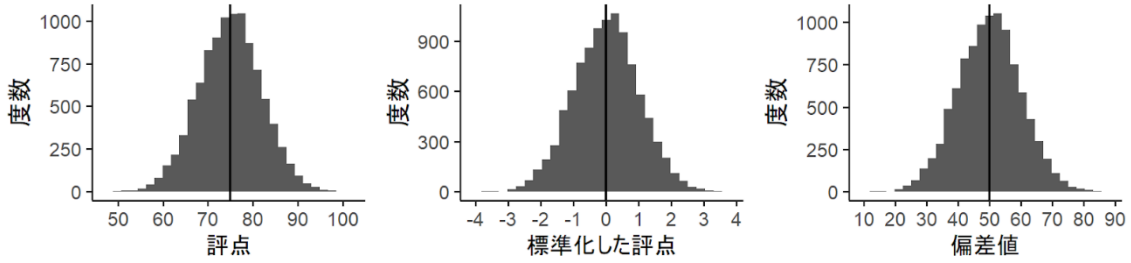
（１）各観測値から平均値を引き（平均値を 0 にして）
（２）標準偏差で割る（標準偏差を 1 にする）
※この手続きを「標準化」と呼びます．ただし標準偏差の代わりに，四分位範囲や範囲（最小値から最大値までの区間）を利用することもあります．

手続き２．手続き１で加工した観測値を「平均値が５０，標準偏差が１０」に加工する

（１） 手続き 1 で加工した各観測値に 10 を掛けて（標準偏差を 10 にして）
（２） 50 を足す（平均値を 50 にする）

以下の図を見るとわかる通り，

図表．観測値が偏差値に変換される流れ

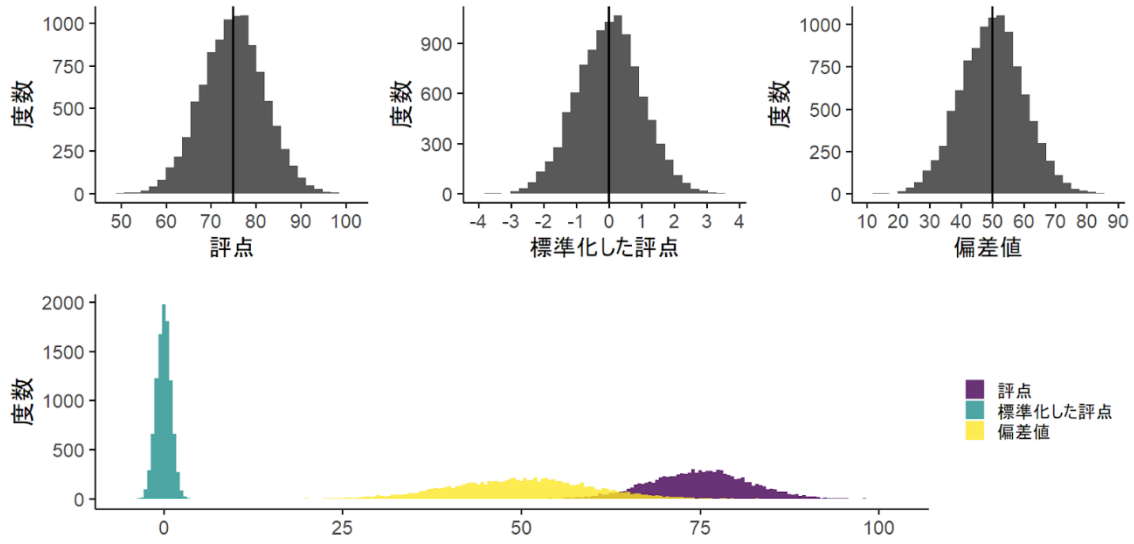


神戸大学　数理・データサイエンスセンター
加葉田大志朗

このような手続きを経て計算された偏差値は，異なる試験であっても必ず「平均値が 50，標準偏差が 10」となっています．従ってどのような試験であっても，相対的な位置を普遍的に示すことができる便利な指標として活用されています．

ちなみに上記の図（以下図の上段に再掲）でデータを加工した後も分布の形状が変わらないように見えるのは，X 軸の目盛りが変化しているためです．実際には以下図の下段の通り，分布の位置もばらつきも変化しています．ただしこれらの加工を施しても，観測値間の相対的な位置は変わらないため，X 軸の目盛りを調整すれば，分布の形状には変化がないように見えます．

図表．データの加工と分布の形状と目盛りの関係



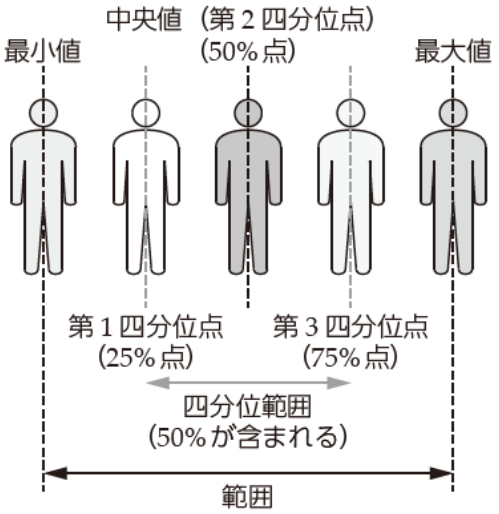
神戸大学　数理・データサイエンスセンター
加葉田大志朗

データのばらつきを捉える統計量②：四分位範囲

データのばらつき具合を表す指標として，平均値との乖離具合を示す標準偏差を紹介しました．ただし分布の位置を示す統計量は算術平均だけでなく，中央値もあります．中央値を利用する場合には，データのばらつきを表す指標には一般的に「四分位範囲（interquartile range, IQR）」が利用されることが多い印象です．

四分位範囲は「観測値を小さい順に並べたときの 25%目の値から 75%目の値までの範囲」と表現できます．つまり四分位範囲には，75%-25%=50%分の観測値が含まれることになります．ちなみに観測値を小さい順に並べたときの 25%目の値を「第 1 四分位点」，75%目の値を「第 3 四分位点」と呼びます．さらに中央値は観測値を小さい順に並べたときの 50%目の値であるため「第 2 四分位点」と呼ぶことができます．

図表．中央値と四分位範囲



中央値（代表値）
→小さい順に数えて 50%目のデータ値

四分位範囲（ばらつき）
→25%目～75%目のデータ値の範囲
　この範囲に 50%が含まれる

神戸大学　数理・データサイエンスセンター
加葉田大志朗

4. 正規分布を使って観測されたデータの珍しさを考える

これまでの例で挙げた講義評点のヒストグラムのように，分布の頂点を中心に左右対称に広がるつり鐘型の分布を「正規分布」と呼びます．正確には以下の数式で表現された関数に従うデータの分布を指します．ちなみに今回の講義では，この式を覚える必要はありません．こんなものなんだな，程度に見ておいてください．

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

π は円周率，σ^2 は分散（従って σ は標準偏差），μ は平均値．

関数とは特定の値を与えたときに計算結果を返す数式と考えてください．特に上記の正規分布に従う変数Xについて，観測値xの出現確率を表現するための特定の関数のことを「確率密度関数」と呼びます．この周辺の知識については本講義の範囲を超えるため，関心のある方は統計学の各論の講義や紹介する専門書で勉強してみてください．

※　実際には連続変数に関する確率密度関数は特定の区間で積分をしないと確率を表現できません．離散変数について観測値xの出現確率を表現するための特定の関数は「確率関数」と呼ぶこともあります．離散変数の場合は，特定の区間で積分をせずとも観測値xの出現確率を算出できます．

図表．正規分布に従うデータのヒストグラム



正規分布のように特定の名前が付けられ，分布を示すための関数が数学的に定義されて

神戸大学　数理・データサイエンスセンター
加葉田大志朗

いる分布はいくつか存在します．その中でも正規分布は最もよく出会い，最もよく利用する，基本となる分布であると言えます．それでは，自身の手元で観察されたデータが，正規分布のような特定の分布に従うと考える（仮定する）ことができると，何が嬉しいのでしょうか？

神戸大学　数理・データサイエンスセンター
加葉田大志朗

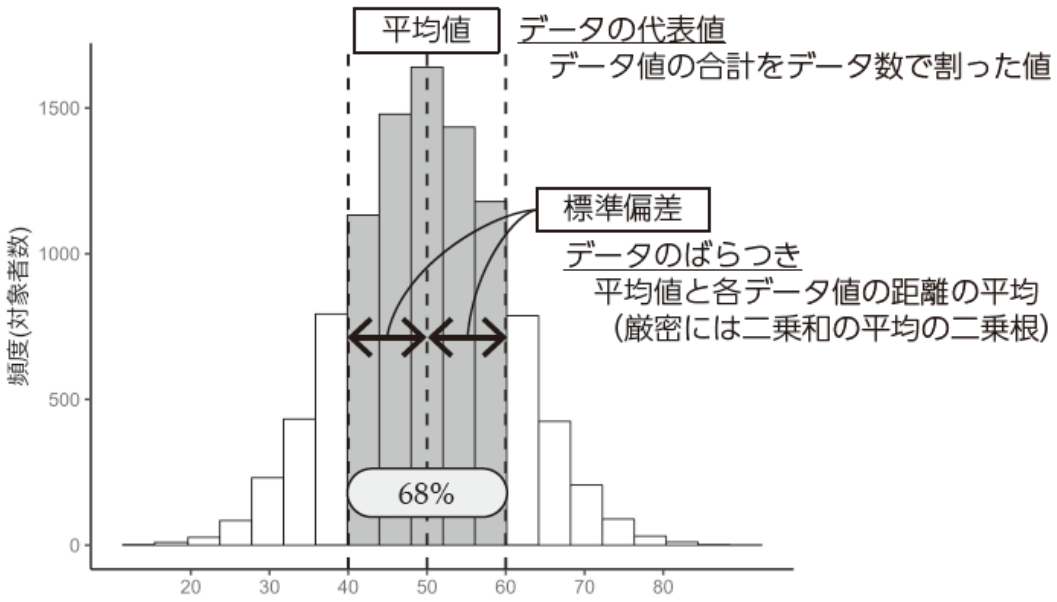
正規分布の特性

例えば「講義評点の平均値が 75 点だったときに，自分の点数は 80 点だった．これは喜ぶべきことか？」という問題について，再度考えてみましょう．これまでは「自分の平均からの離れ具合（偏差，今回であれば 5 点）」が「集団の（平均的な）平均からの離れ具合（標準偏差）」よりも相対的に大きいか小さいかを考えることで，自分の点数が集団のばらつきから逸脱したものかどうかを推し量っていました．ただしこれだけでは「自分の立ち位置が集団の平均的なばらつきの中に収まるかどうか」しか分かりません．このときにデータが従う分布を仮定することができると，集団の中での位置がより詳細に理解できるようになります．

例えば，講義評点のデータが正規分布に従っていると仮定します．すると講義評点の分布の「位置」と「形状」は，「平均値」と「標準偏差」によって決まると言えます．正規分布は平均値と標準偏差に関して，以下のような特徴を持ちます．

（1） 平均値に関する特性：平均値が中央値と一致する（平均値の上下にそれぞれ 50%ずつデータが存在する）．
（2） 標準偏差に関する特性：平均値±標準偏差の範囲にデータの約 68%が含まれる．また 平均値±1.96×標準偏差 の範囲にデータの約 95%が含まれる．

図表．正規分布の特性



神戸大学　数理・データサイエンスセンター
加葉田大志朗

これらの特性を利用すれば，自身の観測値の集団における具体的な位置を知ることができます．例えば観測値が平均値±標準偏差の範囲に収まっていれば，集団の中の 68%に含まれるような平凡な値であると考えることができます．いままで「平均的なばらつきの範囲」としてしか解釈できなかった標準偏差を使って，ある区間からデータが得られる確率を具体的に考えることができるようになりました．

上記の特性を利用して，自分の手元で得られた観測値が具体的にどれだけ特殊なものであるかを考えてみましょう．今回の例の場合，平均値が 75 点，標準偏差が標準偏差が 8.5 点なので，「集団のうち 68%は 75±8.5 点（66.5 点から 83.5 点）に含まれる」と考えることができます．従って自身の評点 80 点は集団の中の 68%に含まれる，あまり珍しい点数ではないと推測できます．また自身の評点が 95 点だった場合についても考えてみます，上記の特性「平均値±1.96×標準偏差の範囲にデータの約 95%が含まれる」を使って考えると，「集団のうち 95%は 75±1.96×8.5 点（58.34 点から 91.66 点）に含まれる」と言えます．従って 95 点という点数は，残りの 5 %に属する非常に珍しい点数であるといえる．また正規分布は平均値を中心に左右対称であるため，残り 5 %を上下方向に 2 分割した上位 2.5%に位置すると考えることができます．上記のように特定の分布に基づいて考えれば，手元のデータが観察された現象に関する珍しさ（確率）についても知見を得ることができるようになりました．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

神戸大学　数理・データサイエンスセンター
加葉田大志朗

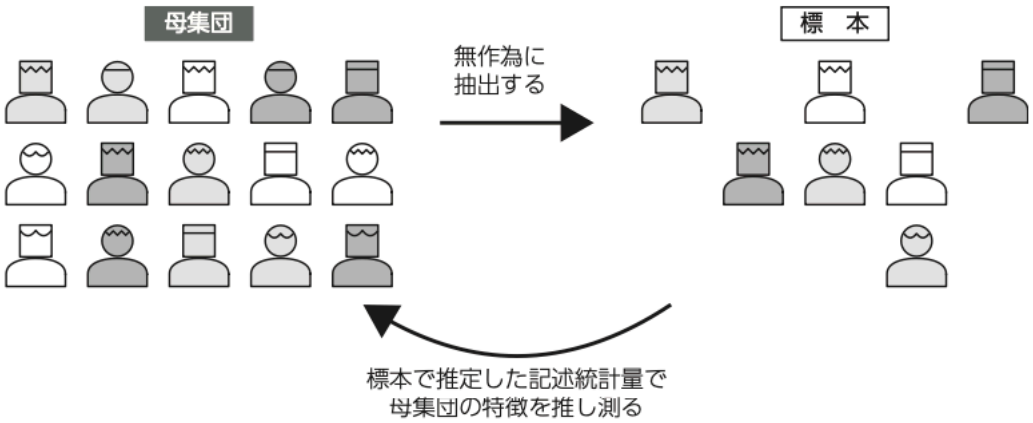
5. 標本から母集団を推測する

観測されたデータから，そのデータの持つ特徴（癖）を汲み取る作業を「統計的記述」と呼びました．私たちが観測できているデータは，あくまでも世の中に存在する全てのデータの一部です．例えば，選挙では全ての票が開かれていないのに，事前の出口調査で一部の票を観測することで，誰が当選しそうか推測することができます．このような全てのデータから抽出された一部のデータを観察することで，全てのデータの特徴を推測する統計学の手続きを「統計的推測」と呼びます．

関心のある全てのデータを「母集団」と呼び，そこから抽出された部分的なデータを「標本」と呼びます．選挙の例で言えば，全ての票が母集団で，出口調査で観測された一部の票が標本です．ここで私の関心の対象は，標本で得られる結論（出口調査上の当選者）ではなく，母集団における結論（全ての投票結果による当選者）であるはずです．しかし母集団が観測される（これを「全数調査」と言います）ことはかなり稀です．したがって統計的推測は，統計学において非常に重要な役割を持ちます．

私たち分析者は，私たちが見ることのない「神のみぞ知る」母集団における特徴を推測する必要があります．あくまでも推測なので，当たり外れもありますし，その推測がどの程度の幅をもって確からしいのかを考える必要もあります．ここでは統計的推測の基本的な考え方に触れるため，母数の平均値を標本から推定する手続きについて紹介します．

図表．母集団と標本



神戸大学　数理・データサイエンスセンター
加葉田大志朗

ランダムサンプリングの仮定

選挙では原則すべての投票が開票されるので，棄権票を含めれば有権者の票をすべて観測できる全数調査と言えます．そのため最終的な結果は全ての票が観測されてから得られるわけですが，出口調査では最終的な結果がでる前に，その一部を「味見」をしていると言えます．この味見が「良い味見」か「悪い味見」になるかは，出口調査で味見した有権者の票が「偏りなく」観測されているかどうかによります．例えば特定の候補者を応援する親族や団体が一同に会して投票する時間だけ出口調査を行ってしまえば，その候補者の票数を過剰に多く見積もってしまう可能性があります．このようなことを防ぐために，標本の抽出はランダムに行う「ランダムサンプリング」されていることが重要になります．

※ 出口調査ってどうやるの？（NHK 北海道 WEB, 2025-01-07 アクセス）にあるように，実際には調査ごとに偏りのないサンプリングを目指した工夫がされています．

選挙の例のほかに，まさに「みそ汁の味見」も日本でよく使われるランダムサンプリングの例として挙げられます．鍋の中のみそ汁の味を知るために，みそ汁をすべて飲み干す（全数調査）するわけにはいきません．そのため私たちはその一部（標本）をおたまにすくって味見することで，鍋の中すべてのみそ汁（母集団）の味を推測するわけです．味見のとき，多くの人はおたまでひとすくいする前にみそ汁をかき混ぜると思います．このかき混ぜることの目的は，鍋の中から偏った部分をすくってしまわないようにすることですよね．味噌が沈殿している味が濃い部分や，上澄みの味が薄い部分をすくって味見してしまっては，鍋全体の味をうまく推測できなくなってしまいます．これを防いで鍋全体の平均的な味を知るために，私たちは一度かき混ぜてからひとすくいするという「ランダムサンプリング」を日常的に行っているのです．

ランダムサンプリングされている場合にも，少ししか味見しない場合と，たくさん味見する場合では，当然たくさん味見した方が，母集団の状態を偏りなく推測できそうです．標本として観測されたデータの数を「標本サイズ」と呼ぶことがありますが，標本サイズが小さいよりも，標本サイズが大きい方が適切な推定ができるということです．これは選挙の例でもみそ汁の例でも，標本サイズが大きくなればなるほど，標本が母集団の状態に近づいていくことは直感的にも違和感はありません．このように標本が母集団からランダムサンプリングされていることが仮定できれば「標本サイズが十分に大きいとき，標本は母集団の状態に非常に近くなる」ことが知られています．そのため「十分に大きい標本」を用意できれば問題はないのですが，それが現実的に難しいので，私たちは「現実的なサイズの標本」からうまく推定する方法を考える必要があります．その方法について，母集団のデータの分布の平均値（これを「母平均」と呼びます）を推測する手続きで学んでいきましょう．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

母平均の推定

母集団のデータの特徴を標本に基づいて推測する手続きを「統計的推測」と呼びました．今回は母集団のデータの特徴のひとつである「母平均」に関心があります．この母平均のような母集団のデータの特徴をそれぞれ個別に「パラメータ（母数）」と呼ぶことがあります．そして特定のパラメータ（今回は母平均）を推定する手続きを「推定」と呼びます．従って今回の私たちの目標は「標本のデータから母平均を推定すること」になります．そして推定に利用する手続きやその計算手法を「推定量」と呼び，実際にデータセットから推定された値そのものを「推定値」や「（推定量の）実現値」などと呼びます．変数と観測値のような関係だと思ってもらって構いません．

母平均を推定する最も強力な方法は標本平均を計算すること

結論から言うと，母平均の良い推定量は「標本平均」です．どのような推定値を良い推定値とするかは様々な基準がありますが，数学的に標本平均は母平均の推定に対して非常に優れた性質を持つことが知られています．なぜ標本平均が母平均の良い推定量となるのか，直感的に理解していきましょう．

複数の観測値の標本平均は母集団に近い可能性が高い

例えば，釣り堀の池の中の魚（母集団）の平均的なサイズ（母平均）を推測したいとします．この推測を良く行うためには，釣り上げた１尾の魚のサイズから推測するか，複数の魚のサイズから推測するか，どちらの方が良さそうでしょうか？多くの方は直感的に「より多くのデータから推測した方が良い推測が可能だろう」と後者を選択すると思います．なぜ私たちがこのような直感を抱くのか，ここでつまびらかにしたいと思います．

まず前者の釣り堀から１尾の魚を標本として抽出する場合について考えます．ここで釣り上げられた１尾の魚は，釣り堀に含まれる魚のうち，どのような魚が釣り上げられたと考えるのが妥当そうでしょうか？釣り堀の中には様々なサイズの魚が泳いでいますが，直感的には釣り堀内により多く含まれるサイズの魚が釣り上げられた可能性が高いと考えられそうです．特に釣り堀内の魚のサイズが正規分布に従う（これを母集団の従う分布が「正規母集団」であると表現します）場合，釣り上げた魚のサイズは母平均に近い値である可能性がより高い，と考えられます．ただし実際には母平均の値は誰にもわからず，今回の関心はこの神のみぞ知る母平均を標本から推定することです．「釣り上げた魚のサイズ（標本の観測値）が釣り堀内の魚のサイズの平均（母集団の母平均）に近い可能性が高い」と考えられるのであれば，逆説的に「釣り堀内の魚のサイズの平均（母集団の平均）」は「釣り上げた魚のサイズ（標本の観測値）に近い可能性が高い」という推論ができます．ただし実際には

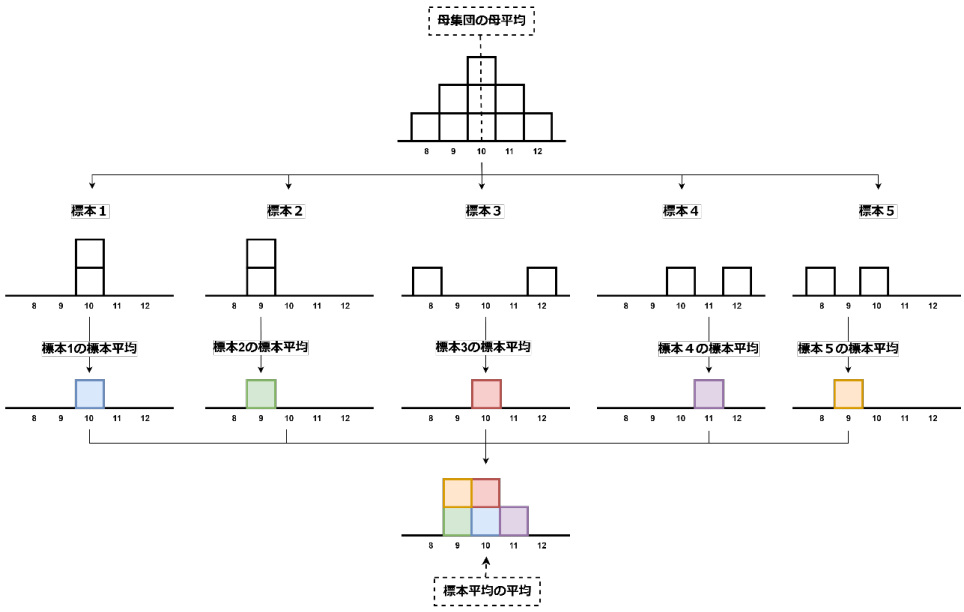
神戸大学　数理・データサイエンスセンター
加葉田大志朗

１尾の魚しか観測していないので，実際には母平均よりもずっと離れたサイズの魚を観測してしまっている可能性は拭えません．

ここでもう少し精度の高い推定を行うために，２尾の魚を釣り上げて標本とする場合について考えます．ここでもやはり，標本として観測した２尾の魚のサイズに近いところに母平均がありそうと考えられそうです．もし２尾の魚が同じようなサイズであれば，やはりその辺りに母平均がある可能性は高いと推測できそうです．反対に２尾の魚のサイズが全く違う場合はどうでしょうか？片方のサイズは母平均に近くとも，もう片方のサイズは母平均からかけ離れているかもしれませんし，あるいは双方ともに母平均から離れている可能性もあります．このようなときには，２つのデータの真ん中（標本平均）の近くに母平均があると考えるのが自然かもしれません．実際に標本平均は，ひとつひとつの観測値よりも，母平均の近くにある可能性が高いことが数学的に知られています．そして 2 つの観測値の標本平均よりも，より多くの観測値から計算した標本平均の方が母平均に近くなります．この事実を視覚的に確認してみましょう．

以下の図では，それぞれの魚のサイズをひとつのブロックとして，母集団である釣り堀の中の魚（1 段目）から，標本として釣り上げた 2 尾の魚のサイズの観測値を並べています（2 段目）．ここでは 2 つの観測値を含む標本が 5 セット並んでいますが，これは 5 人の釣り人がそれぞれ 2 尾ずつ魚を釣り上げているとイメージしてください．それぞれの標本に含まれる魚のサイズについて平均値を算出したもの，つまり標本平均を 3 段目に並べ，それらを 4 段目でまとめて表示しています．

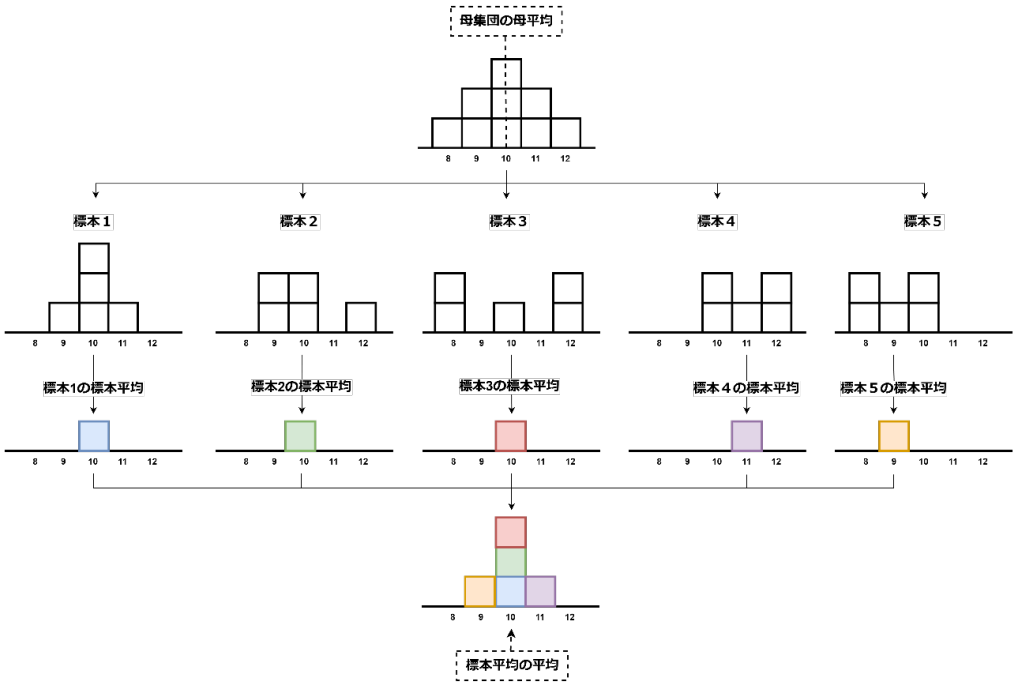
図表．標本平均は母平均に近いところに分布する



神戸大学　数理・データサイエンスセンター
加葉田大志朗

上記を見るとわかるように，それぞれの標本の中の観測値は母平均とかけ離れていても，それらから算出された標本平均は母平均に近い値をとることが多いことが分かります．さらに標本の中の観測値を増やした場合についても見てみます．

図表．観測値の数を増やすと標本平均はより母平均に近いところに分布する



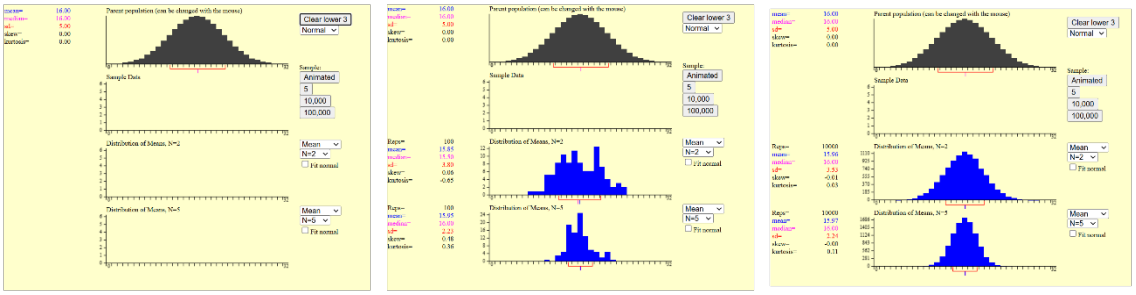
上の図ではそれぞれの標本に含まれる観測値の数が 5 個になっています．つまりそれぞれの釣り人が 5 尾ずつ魚を釣り上げていると考えてください．それぞれの標本に含まれる 5 尾の魚のサイズの標本平均を算出して並べると，標本平均はより母平均に近いところに集中しているように見えます．このように「標本に含まれるデータ数が増えるほど標本平均が母平均に近い値になること」を「大数の法則」と呼びます．従って統計的推定を良く行うためには，取得するデータは多ければ多いほうが良いと言えます．ただし皆さんが研究を進める際には，倫理的な側面（効果があると言えない治療に沢山の患者を曝してよいのか？）や資源的な側面（データをとるにもお金がかかる）から実現可能性の高いサンプルサイズを計画時点で決める必要があることもあります．それらの詳細は各分野の各論で学んでほしいのですが，巷で言われる「データの数は多ければ多いほど良い」や「研究の前にはサンプルサイズ計算が必要」という言葉の背景にはこのような事情があることを知っておくと良いでしょう．

上記の標本のサンプリングはあくまでも説明を簡単にするために作成したものであり，

神戸大学　数理・データサイエンスセンター
加葉田大志朗

人によっては恣意的な事例に感じられるかもしれません．しかし実際に，もっとたくさんの標本（釣り人が釣った魚のセット）からそのサイズの標本平均を算出していくと，やはり標本に含まれる観測値の数が多いほど，標本平均は母平均の近くに集まることが確認できます．Rice Virtual Lab in Statistics というサイトで公開されているシミュレーターでその様子を観察できますので，動画で確認してみましょう．

図表．標本に含まれる観測値の数が大きいほど標本平均が母平均に近くなる可能性が高いことを示すシミュレーター（左：サンプリング前，中：100 回サンプリング後，右：10000 回サンプリング後）※詳細は動画内で解説



今回のシミュレーションのように，複数の標本を観察する（標本を 100 セットや 10000 セット用意する）ことは現実的には不可能です．しかし仮想的にたくさんの標本を用意できて，たくさんの標本平均が計算できたとしたら，それらの標本平均は母平均のまわりに分布すると考えられます．標本平均の分布が母平均の近くに集まっているのであれば，当然皆さんが”たまたま”集めた標本の標本平均も，母平均に近い値である可能性が高くなります．これらが標本平均が母平均の良い推定量であると考えられる背景の直感的な説明です．実際には，「標本平均を無限に集めてその平均を計算すると母平均に一致する」という数学的な性質を確認することなどによって標本平均の推定量としての良さは確認されます．これは数理統計学の講義できちんと学ばれると思いますが，その時も数式や記号だけを追うよりも，この直感的な背景を頭に置いておくことで，少しだけ理解がスムーズになるかもしれません．ちなみにこのような「標本平均を無限に集めてその平均を計算すると母平均に一致する」という性質は，母集団のデータが従う分布が正規分布でなくても成立します．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

標本平均のばらつきと母平均の区間推定

母平均を標本平均というただ一点で推定する手続きを「点推定」と呼びます．これに対して，ある程度の範囲を持って母平均のようなパラメータの真値が存在する区間を推定する手続きを「区間推定」と呼びます．ここでは区間推定に関する基本的な考え方を学んでおきましょう．

正規母集団における標本平均のばらつきは標準誤差で示す

区間推定を行うためには，まず標本平均のような推定量がどの程度ばらつくのかを考える必要があります．私たちがデータのばらつきを表現する方法として最も身近なものに標準偏差があります．もしたくさんの標本を用意できて，たくさんの標本平均を計算できれば，それらのばらつきを「標準偏差」などを使って表現することは容易です．ただし標本平均などの推定量のばらつきを表現するときには「標準誤差」という用語を使うことを覚えておいてください．観測値などのデータそのもののばらつきは「標準偏差」であり，「標準誤差」は推定量のばらつきであるとして区別してください．

標本平均の標準誤差の計算は，ある条件下では非常に簡便に行えます．この条件というのが「母集団が従うデータが正規分布していること」です．このような母集団を「正規母集団」と呼びます．正規母集団においては，そこから抽出された無数の標本における無数の標本平均は正規分布に従い，その算術平均は母平均に等しく，そのばらつきである標準誤差は

$$\textit{標準誤差} = \frac{\textit{母集団の標準偏差}}{\sqrt{\textit{各標本に含まれるデータ数}}}$$

となります．つまり正規母集団のもとでは，標本平均のばらつきは，母集団に含まれるデータのばらつきを$\sqrt{\textit{各標本に含まれるデータ数}}$分の 1 に縮めることで簡単に求められます．したがって各標本に含まれるデータ数が大きいほど，推定値のばらつきである標準誤差は小さくなります．つまり標本に含まれるデータ数が大きいほど，母平均に近い値が出る可能性がより高くなると考えることができます．

標準誤差に基づいて標本平均のばらつきの範囲を求める

ここまでで母集団のデータが正規分布に従っているのであれば，標本平均のばらつきである標準誤差も比較的容易に求められることがわかりました．この知識からどのくらいの範囲にどれくらいの標本平均が含まれるのか？を考えたいと思います．これは以下再掲する正規分布の平均値と標準偏差の特性を利用すれば簡単に求まります．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

（1） 平均値に関する特性：平均値が中央値と一致する（平均値の上下にそれぞれ 50%ずつデータが存在する）.
（2） 標準偏差に関する特性：平均値±標準偏差の範囲にデータの約 68%が含まれる．また 平均値±1.96×標準偏差 の範囲にデータの約 95%が含まれる.

従って正規分布に従う観測値のうち 95%は,

$$母平均 - 1.96 \times 標準偏差 \leq 観測値 \leq 母平均 + 1.96 \times 標準偏差 \ldots (1)$$

の範囲に含まれると考えることができます．ここまでは正規分布から抽出された観測値のばらつきに関する議論ですが，この特性は標本平均についても同じように考えることができます．つまり，無数の標本平均の実現値のうちの 95%を含む範囲は，標本平均のばらつきである標準誤差を利用して以下ように表現することができます.

$$母平均 - 1.96 \times 標準誤差 \leq 標本平均 \leq 母平均 + 1.96 \times 標準誤差 \ldots (2)$$

母平均を区間推定する

ここまでは仮想の「無数の標本平均」を想像して，そのばらつきである標準誤差を考えました．そして標準誤差と，正規分布に従うデータの性質を利用すれば，無数の標本平均のうちの大半が含まれる範囲を（2）式で容易に求められました.

$$母平均 - 1.96 \times 標準誤差 \leq 標本平均 \leq 母平均 + 1.96 \times 標準誤差 \ldots (2)$$

この（2）式の要素のうち，母平均は「神のみぞ知るただひとつの値」であり，私たちはこの値を知り得ません．ただしこの仮に標本が無数にあった場合の標本平均がばらつく範囲から，「母平均がどのあたりにありそうか？」という「区間推定」を行うことが可能になります.

95%信頼区間

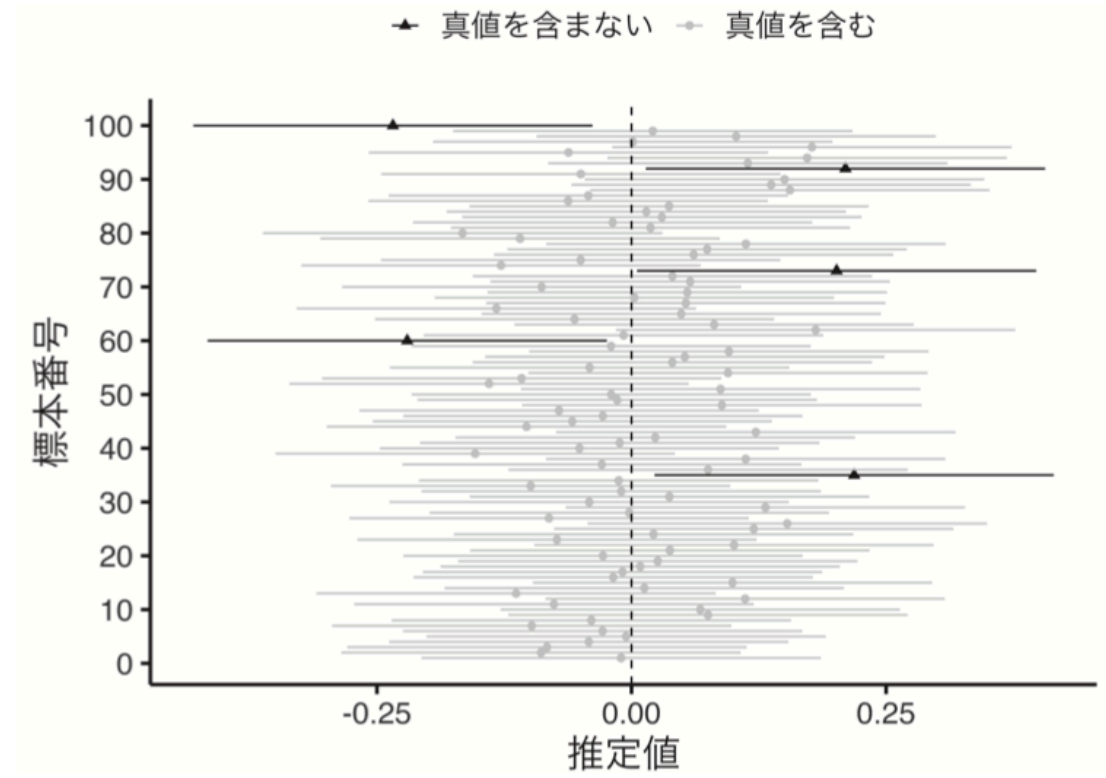
ここから私たちの関心は標本平均のばらつく範囲ではなく，母平均がどのあたりの区間にあるのか？なので，母平均に関する区間を表す不等式に変換してみます.

$$標本平均 - 1.96 \times 標準誤差 \leq 母平均 \leq 標本平均 + 1.96 \times 標準誤差 \ldots (3)$$

この（3）式において，母平均は「神のみぞ知るただひとつの値」であるから変動しません．しかし標本平均は標本が異なれば異なる値に変動します．つまり 100 セットの標本を分析すれば，100 個の標本平均が計算できます．そして 100 個の（3）式で定義される区間を考えることができます.

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．標本が異なれば（３）式は異なる区間をカバーする



ただし上記の 100 個の標本から計算した 100 個の区間が全て母平均を含むわけではありません．（２）式は「母平均±1.96 標準誤差の範囲に 95%の標本平均が含まれる」範囲でした. つまり母平均からかけ離れた標本平均も 5%存在するということです. そのような（２）式を変換して得られた（３）式についても同様に，無数の標本から計算した無数の区間の 95%分が母平均を含んでいると考えることができます．つまり上記のように 100 セットの標本から 100 個の標本平均と区間を計算すると，そのうち 95 個が母平均を含み，5 個は含まないことになります．このような（３）式で定義される区間のことを「95%信頼区間」と呼び，母平均などの推定量がどの程度の範囲にありそうかという区間推定の代表的な指標として使われています.

上記の通り，標本が無数にあり信頼区間が無数に計算されれば，それらの区間は揺らぎを持つので，そのうち 95%は母数を含み，残りの 5%は母数を含まないということが起こり得ます. ただし私たちは実際には無数の標本を観察できず，大抵１セットの標本から計算したひとつの区間しか観察できません（これを「95%信頼区間の実現値」と呼ぶことにします）. ひとつしか実現値がないので当然揺らぎを持ちませんし，そのうちのいずれかは母数を含む・含まないといった議論も出来ません．つまり皆さんが計算した 95%信頼区間の実現値について「この区間に 95%の確率で母数があるだろう」という議論はできないことに注意

神戸大学　数理・データサイエンスセンター
加葉田大志朗

してください．皆さんが計算した 95%信頼区間も母数も，ただひとつに決まってしまったので，皆さんが得られる結論はこの実現値が母数を「含む」か「含まない」のいずれかのみです．ただし実際には私たちは母数は知り得ないので，本当に含まれているかどうかはわかりません．したがって私たちが 95%信頼区間から得られる知識は，「この区間に本当に母数が含まれるかどうかはわからないが，このような区間が無数にあれば，そのうち 95%は母数を含んでいることはわかる」ということです．

なお 95%信頼区間は実際のデータ分析では「この区間が特定の値を含むか含まないか」という視点で使われることが一般的です．例えば痛み止めのお薬を使った時の痛みの増減の母平均について考えてみます．痛み止めの効果が全くない状況を痛みの増減 0 と定義すると，母平均が 0 だと痛み止めには平均的に効果がないということになります．ここで痛みの増減量について標本平均を推定し，その 95%信頼区間を求めたとします．無数にこの区間を計算できれば，たいていの区間は母平均を含んでいるはずです．反対に言えば，この区間の外にある値が母平均が存在するとは考えにくいわけです．このロジックを使って，「95%信頼区間が痛みの増減量 0 を含んでいなければ，痛み止めの効果は 0 ではない」という仮説が妥当だと考える手続きを取ることがあります．このような手続きを区間推定に基づく「仮説検定」と呼ぶことも覚えておくと，後々統計学を学ぶ上で役立つでしょう．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

母分散と標準偏差の推定

ここまでは母平均の推定に焦点を当ててきました．ただし母集団のデータが正規分布に従うことを仮定したとき，その分布の形状は平均値だけでなく，そのばらつき具合である分散および標準偏差にも依存します．この母集団のデータが従う分布の分散を「母分散」と呼びます．母分散の二乗根も「母標準偏差」と呼びたいところですが，あまり一般的ではないためこれまで通り「標準偏差」と呼びます．同時に，標本から算出された分散や標準偏差については，「標本分散」や「標本標準偏差」と呼ぶことで区別することができます．本節では標本に含まれる観測値のばらつきである標本標準偏差に焦点を当てるため，明確に母集団の標準偏差と標本の標本標準偏差を分けて呼称することにします．

これまで単純に母集団のデータのばらつき具合や，標本で観測されたデータのばらつき具合をまとめて標準偏差として，以下の式で計算していました．

$$標準偏差 = \sqrt{\frac{偏差^2\ の合計}{データ数}}$$

この標準偏差を明確に標本に含まれるデータに関するばらつき，つまり「標本標準偏差」であることを表すために，以下のように書き直しておきます．

$$標本標準偏差 = \sqrt{\frac{\left(標本に含まれる各観測値 - 標本平均\right)^2\ の合計}{標本に含まれるデータ数}}$$

母平均と標本平均の関係と同じように，私たちは母集団を観測することはほとんどの場合できないので，母集団におけるデータのばらつき（標準偏差）はわかりませんが，標本に含まれるデータのばらつき（標本標準偏差）は計算できます．そして神のみぞ知る母平均を推定するうえでは，標本平均が良い推定量になることもここまでで説明した通りです．この流れに乗れば，神のみぞ知る母集団の標準偏差も，標本標準偏差でうまく推定できそうです．その考え方自体は悪くないのですが，実は標本標準偏差は，母集団の標準偏差を少しだけ過少推定してしまう（つまりばらつきを小さく見積もってしまう）ことが知られています．このように推定量が算出した値と推定の対象となる真とのズレを「バイアス（偏り）」と呼びます※．つまり標本標準偏差は母集団の標準偏差に対して偏りを持つ推定量であるということです．

※　より厳密には，無数の標本から無数の推定値を算出したときに，その平均値が真の値と一致しない場合に，その推定量にはバイアスが含まれていると表現します．

従って標本のデータのばらつきから母集団のデータのばらつきを「偏りなく」推定するために，私たちは以下のような操作を加えた標本標準偏差を計算する必要があります．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

$$不偏標本標準偏差 = \sqrt{\frac{\left(標本に含まれる各観測値 - 標本平均\right)^2 \text{ の合計}}{標本に含まれるデータ数 - 1}}$$

ここでのポイントは二乗根内の分母が「標本に含まれるデータ数」から「標本に含まれるデータ数－１」に変更されているということです．分母の値が小さくなっているので，標本標準偏差よりも若干大きな値になります．この手続きで計算した標本標準偏差は，母集団の標準偏差に対して「バイアスを持たない（偏りがない）」推定量であることが数学的に知られています．この「バイアスを持たない（偏りがない）」ことを統計学では「不偏」と呼びます．従ってこの推定量は「不偏標本標準偏差」と呼ばれます（単純に不偏標準偏差とも呼ばれます）．従って母集団のデータばらつきを標本から推定するためには，これまで基本としてきた標本標準偏差よりも，不偏標本標準偏差が良い推定量になると考えられます．

標準誤差と信頼区間における標準偏差の推定値の役割

標本平均のような推定量のばらつきは「標準誤差」と呼ばれ，母集団のデータが正規分布に従うのであれば，そこから抽出された無数の標本における無数の標本平均は正規分布に従い，その算術平均は母平均に等しく，そのばらつきである標準誤差は

$$標準誤差 = \frac{母集団の標準偏差}{\sqrt{各標本に含まれるデータ数}}$$

で算出できることを紹介しました．実は上の式では未知の「母集団の標準偏差」を使っているので，現実の世界では私たちは計算がほぼ不可能なことをスルーして話を進めていました（説明のためです，すみません）．この実際には観察できない母集団の標準偏差について，ここで紹介した不偏標本標準偏差で置き換える必要があります．

$$標準誤差 = \frac{不偏標本標準偏差}{\sqrt{各標本に含まれるデータ数}}$$

分子が母集団の標準偏差であったこれまでの標準誤差を利用していた時には，標本平均の95%信頼区間を以下の式で求めることができました．

$$標本平均 - 1.96 \times 標準誤差 \leq 母平均 \leq 標本平均 + 1.96 \times 標準誤差$$

しかし標準誤差の推定に不偏標本標準偏差を利用した場合には，この標準誤差に係る 1.96 という固定の値が利用できません．しかもその係数は標本に含まれるデータ数（厳密にはこのデータ数から 1 を引いた値）によって変化するので，今までのように簡単には求められなくなることを覚えておいてください．これが何故かについては本講義の範囲を超えるので詳細は省きますが，このような現象が生じること，そして皆さんが Python などの

神戸大学　数理・データサイエンスセンター
加葉田大志朗

分析ツールで標本平均に関する標準誤差や 95%信頼区間を求めたときには，このような分析が行われているのだ，ということだけ頭の片隅に置いておいてください．より詳細が知りたい方は，参考書籍や統計学の各論的な講義で学びを深めていただければと思います．

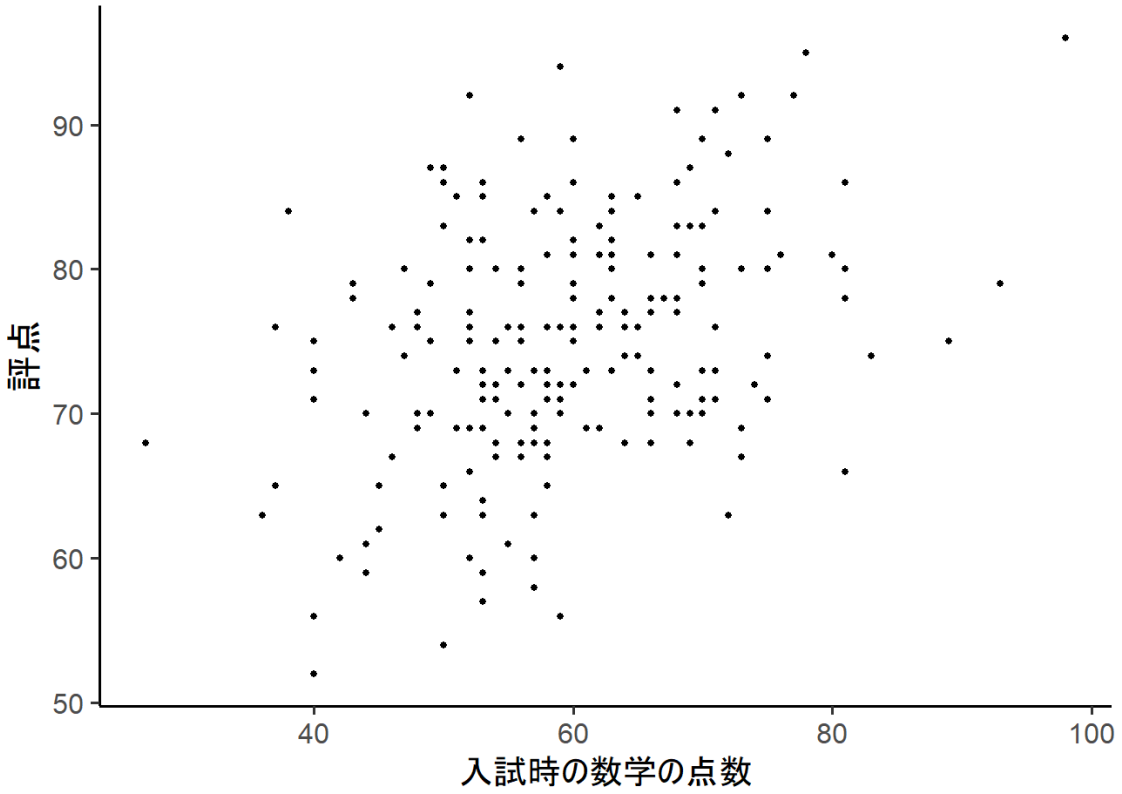
神戸大学　数理・データサイエンスセンター
加葉田大志朗

神戸大学　数理・データサイエンスセンター
加葉田大志朗

6. 二種類の変数の関係を評価する

データ分析の目的が大学の講義の評点（講義評点）と入試時の数学の点数（入試点数）など，二種類の変数間の関係性を評価したいということは珍しくありません．特に二種類の連続変数の関係性を図示する際には以下のような「散布図」が利用されます．

図表．入試点数と講義評点の散布図

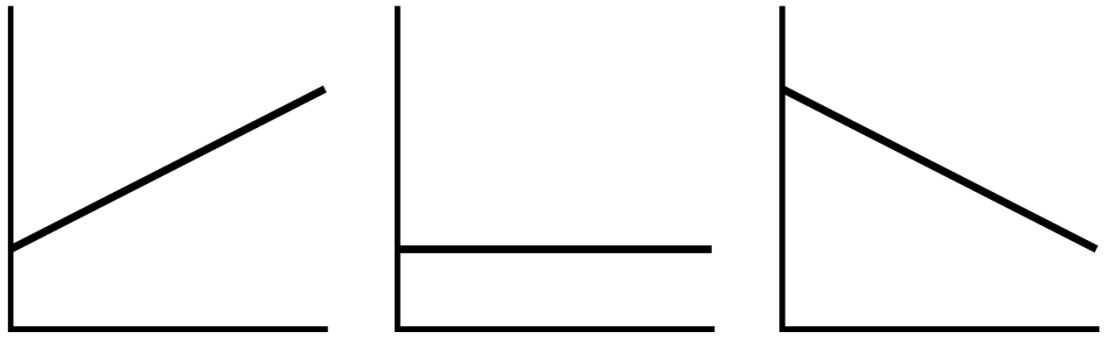


上記の散布図もデータの羅列と同様に，これだけではぼんやりとした傾向しか分かりません．この二種類の変数が連動性を示す統計量として，よく「相関係数」が利用されます．相関係数は－1 から 1 の範囲で二種類の変数間の単調な（増え続ける・減り続けるのような一方向的な）関係性を示します．

相関係数が 1 に近い：正の相関（一方が増えれば他方も増える）
相関係数が 0 に近い：相関なし（双方の変動が連動しない）
相関係数が－1 に近い：負の相関（一方が増えれば他方は減る）

図表．相関を示す直線（左：正の相関，中：相関なし，右：負の相関）

神戸大学　数理・データサイエンスセンター
加葉田大志朗

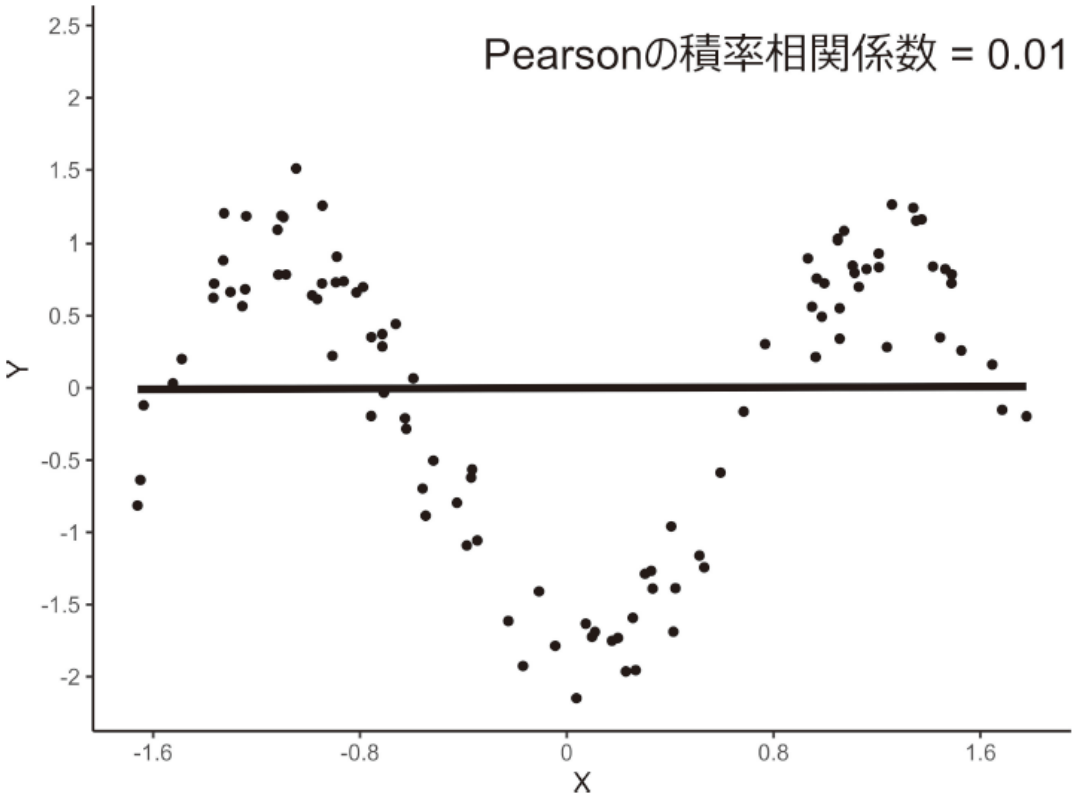


日常生活では慣例的に，いずれかのデータの動きと，もう一方のデータの動きが連動しているように見えるようなとき「相関している」という用語が使われていると思います．ただし，統計学における「相関」は文脈に応じて，もう少し限定的な用語であることに注意してください．例えば以下の図のように，一方のデータの動きに連動して他方のデータも動くような場合でも，相関係数は 0 に近い値となることがあります．これはあくまでも相関係数は「単調な関係の強さ」を示すための指標であるためで，単調な関係性でないときにはその関係性の強さをうまく表現することができないためです．

※　書籍によっては 2 つの変数が何らかの関連した動きをしている状態（非独立な状態）を「相関している」と表現していることもあることに留意してください．多くの専門書ではこの辺りをきちんと定義してくれていると思います．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．データは連動しているように見えるが相関係数は 0 の例



相関係数の計算方法

相関係数にはいくつか種類がありますが，ここでは最も基本的な「ピアソンの相関係数」について紹介します．この講義中に断りなく相関係数と呼んだ時には，ピアソンの相関係数について話していると考えてください．

相関係数と偏差積

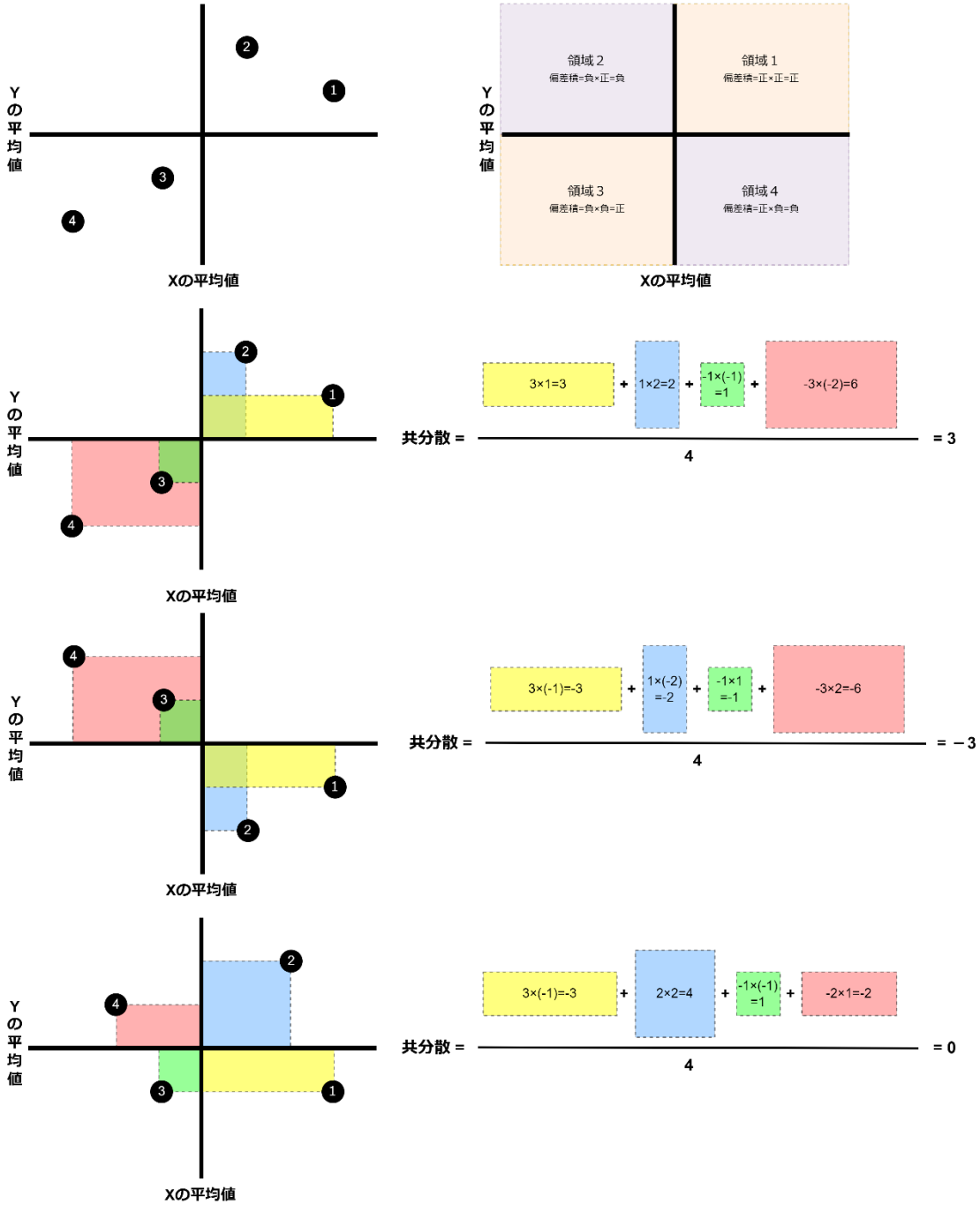
二種類の変数である変数 X と変数 Y について，その相関係数は以下のように示すことができます．

$$X\text{とY}\text{の相関係数} = \frac{X\text{とY}\text{の共分散}}{X\text{の標準偏差} \times Y\text{の標準偏差}}$$

分母の「共分散」は，各データ値の「偏差積」の平均値．偏差積とは名前の通り偏差（平均からの離れ具合）を掛けたものであり，X 軸方向・Y 軸方向にそれぞれどれだけ平均からずれが生じているかを表したものです．直感的には以下図のように，平均からの離れ具合を面積で表していると考えることができます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．偏差積と共分散



上記の通り偏差積はデータの位置によって正負いずれかの値をとります．領域１や領域３にあるデータの偏差積は正，領域２や領域４にデータの偏差積は負の値となります．したがって領域 1 および３に存在する偏差積が作る面積と，領域２および４に存在する偏差積

が作る面積のどちらが大きいかによって，共分散の符号は決まります．また分母の標準偏差は偏差の二乗平均であるため，必ず正の値をとる．したがって共分散の符号が，相関係数の符号として反映されます．

相関係数と標準偏差

相関係数は，変数 X と変数 Y の共分散（偏差の積の平均）を，変数 X と変数 Y それぞれの標準偏差の積で割ることで計算できました．これは言い換えると，変数 X と変数 Y それぞれのデータを標準偏差で割る加工を行ったうえで共分散を計算していることになります．

$$
\begin{aligned}
X\text{とY の相関係数} &= \frac{X\text{とY の共分散}}{SD_X \times SD_Y} = \frac{X\text{とY の偏差積の平均}}{SD_X \times SD_Y} \\
&= \frac{\frac{1}{n}\left(x_1\text{の偏差} \times y_1\text{の偏差} + \cdots + x_n\text{の偏差} \times y_n\text{の偏差}\right)}{SD_X \times SD_Y} \\
&= \frac{1}{n}\left\{\left(\frac{x_1 - \bar{x}}{SD_X} \times \frac{y_1 - \bar{y}}{SD_Y}\right) + \cdots + \left(\frac{x_n - \bar{x}}{SD_X} \times \frac{y_n - \bar{y}}{SD_y}\right)\right\} \\
&= \frac{1}{n}\left\{\left(\frac{x_1}{SD_X} - \frac{\bar{x}}{SD_X}\right) \times \left(\frac{y_1}{SD_Y} - \frac{\bar{y}}{SD_Y}\right) + \cdots + \left(\frac{x_n}{SD_X} - \frac{\bar{x}}{SD_X}\right) \times \left(\frac{y_n}{SD_Y} - \frac{\bar{y}}{SD_Y}\right)\right\}
\end{aligned}
$$

これは「（２）データに定数 a を掛ける」というデータの加工（スケーリング）を行ったときに，標準偏差が a 倍になるという性質に基づいています．

--- 以下参考のため再掲 ---
「（２）データに定数 a を掛ける」という加工を加えると，加工後のデータの統計量には以下の変化が生じます．

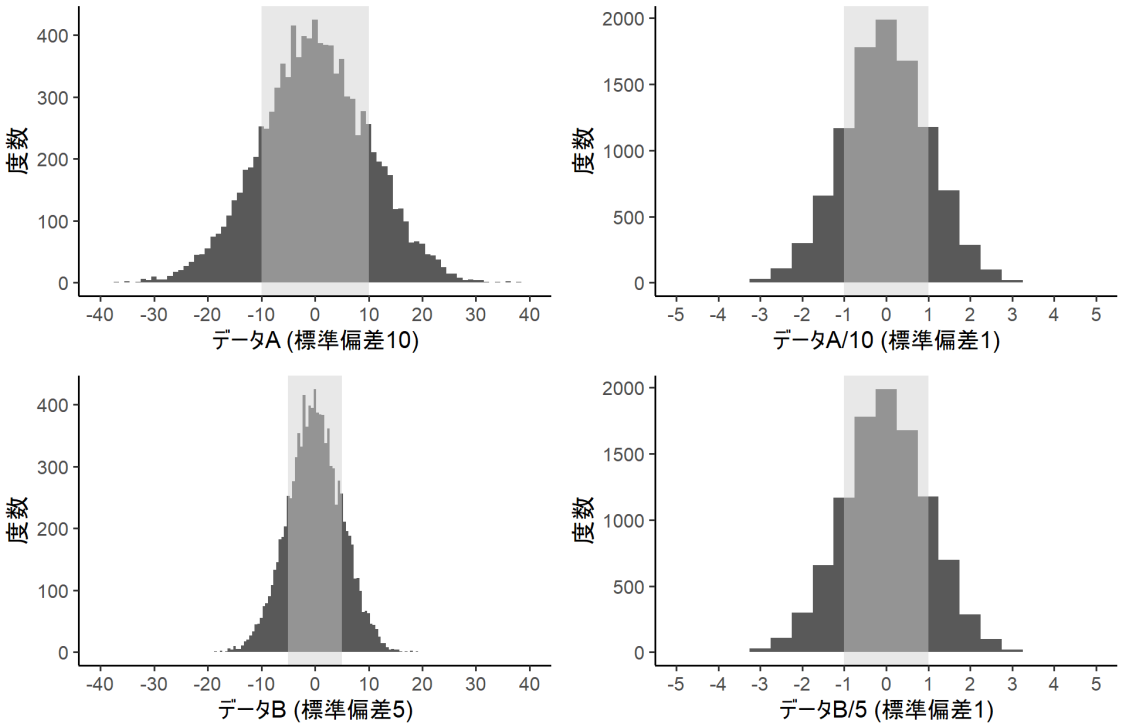
- 加工後の算術平均 ＝ 加工前の算術平均 × **a**
- 加工後の分散 ＝ 加工前の分散 × $\mathbf{a^2}$
- 加工後の標準偏差 ＝ 加工前の標準偏差 × **a**

つまり（２）データに定数を掛ける（割る）加工は「分布の位置をずらしながら，分布のばらつき具合を変える変換」であることが分かる．この加工を分布の尺度（スケール）を変換するという意味で，「スケーリング」などと呼ぶことがあります．

それではなぜ，相関係数を計算するときに元々の未加工のデータではなく，加工済みのデータを利用する（つまり標準偏差で割る）必要があるのでしょうか？これは，データの種類が異なれば，データがばらつく範囲も異なるためです．そしてデータのばらつく範囲が異なると，それぞれの偏差の大きさも異なります．相関係数の大きさは 2 つの変数の偏差が作る面積のバランスによって定まるため，それぞれの変数が作る偏差の単位が異なると，偏差

積のブロックがいずれかの変数の影響を強く受けていびつな形になってしまいます．これによって正しい相関係数が計算できなります．ここで，各データについて標準偏差で割る加工を施せば，分布のばらつきを変化させることができたことを思い出します．特に標準偏差で割れば，加工後のデータの標準偏差は 1 となりました．これにより異なる変数間であっても，ばらつきの範囲が等しい分布に変換でき，それぞれの偏差の大きさも同じ尺度で解釈できるようになります．

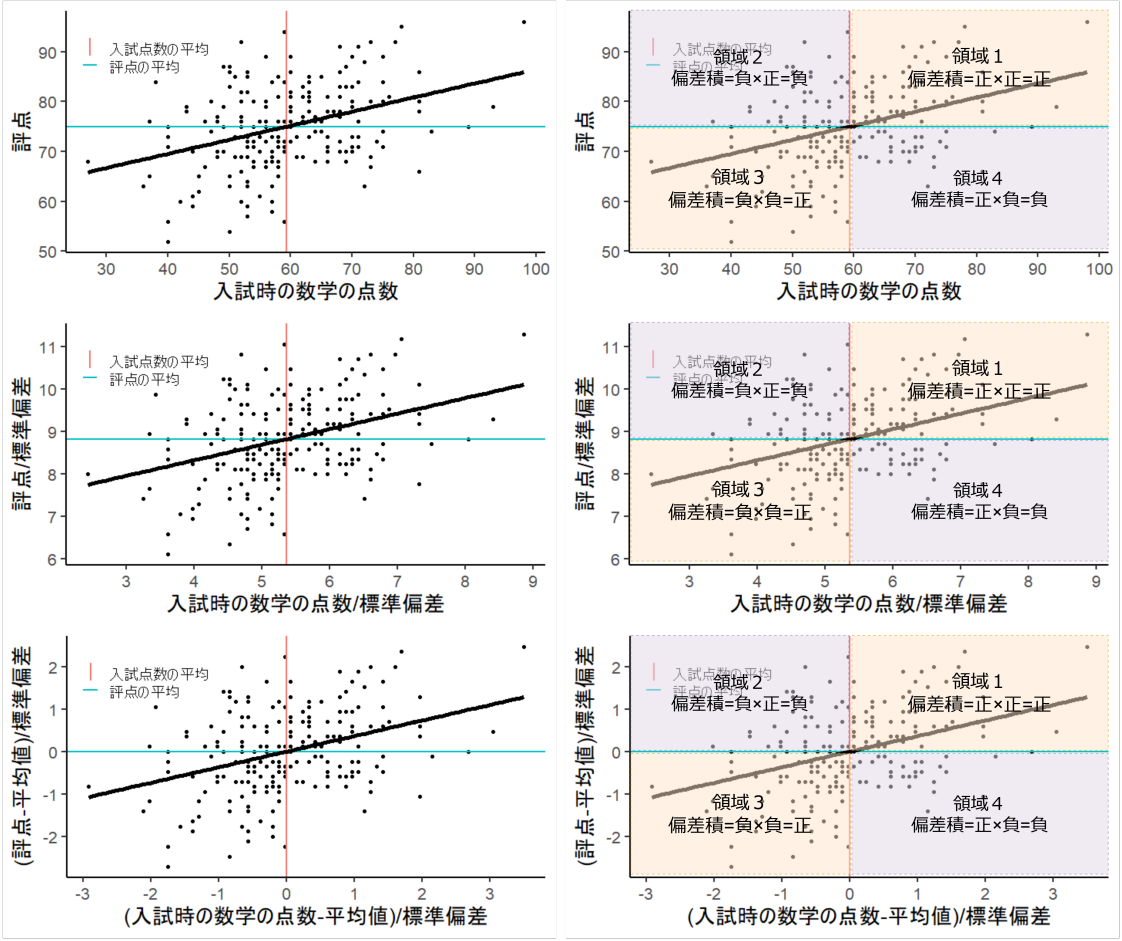
図表．標準偏差で除すことで標準偏差が 1 の分布に変換する（右：変換前，左：変換後）



それでは例に戻り，入試点数と講義評点の相関を評価してみましょう．以下図の通り，入試点数（X 軸）と講義評点（Y 軸）の算術平均を境界に四分割した各領域について注目すると，領域 1 と 3 に存在する点が相対的に多く，また広く分布していることがわかります．したがって，この 2 つの変数は正の相関を持つと考えられます．なお以下には，それぞれの変数をそれぞれの標準偏差で割る加工を行った場合と，標準偏差で割る前に平均値を引く加工を行った場合（つまり，標準化を行った場合）についても併せて表示しておきます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

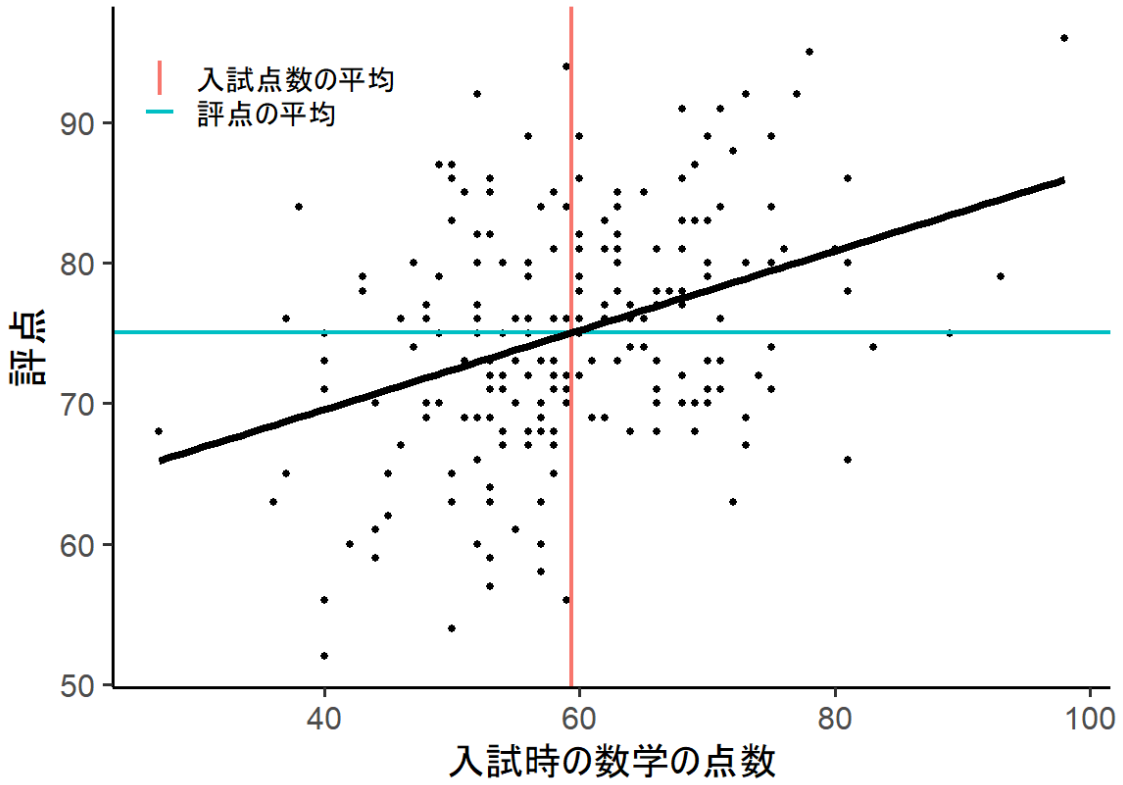
図表．入学時の数学の点数と講義の評点の散布図



実際に先ほどの手続きでピアソンの相関係数を計算すると，約 0.45 であり正の相関を持つことがわかります．以下にその相関係数を示すラインを示しておきます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．相関係数を示すライン



相関の強さ

ここで得られた結果である相関係数約 0.45 から，この相関は強いと判断しますか？弱いと判断しますか？一般的によく見られる基準としては以下があります．ただしあくまでも目安に過ぎません．

－1.0 ～ －0.7：強い負の相関
－0.7 ～ －0.4：中程度の負の相関
－0.4 ～ －0.2：弱い負の相関
－0.2 ～ 0.2：相関なし
0.2 ~ 0.4：弱い正の相関
0.4 ~ 0.7：中程度の正の相関
0.7 ~ 1.0：強い正の相関

上記のように相関係数は-1 から 1 の範囲に収まるため，算出された係数が-1，0，1 のどこに近いかを考えるだけで，二種類の変数がどのような関係にあるのかの目安を簡便に評価することができます．一方，標準偏差を使ってスケーリングされたあとの傾き具合なの

で，実際には「一方がどの程度動いたときに，もう一方の変数もどの程度動くのか？」という説明性に乏しい点も否定できません．実際の研究では，相関係数だけを示すよりも，それぞれの変数をもともとの単位を使って「一方がどの程度動いたときに，もう一方の変数もどの程度動くのか？」をきちんと説明する方が，結果の解釈性が高く，好ましいことも多いと思います．このあたりの手続きについては，回帰モデルの解説で触れたいと思います．

最後に，評価したい相関係数やデータの内容によって，ピアソンの相関係数以外の係数を利用することもあります．例えば，単調性は保たれているが非線形な関係性を持つ場合などにはスピアマンの積率相関係数というものを利用することがあります．ただしこれもピアソンの相関係数の考え方が基本にはあるので，この講義ではまずはピアソンの相関係数の基本的な考え方を押さえておきました．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

神戸大学　数理・データサイエンスセンター
加葉田大志朗

7. 回帰モデルの基本

関心の 2 つの変数間の直線的な関係を表現する際には相関係数を利用することが出来ました．相関係数は各変数の標準偏差で割る加工を施したデータを利用することで，変数がどんな分布を持つ場合でも-1 から 1 の範囲で大小関係を表現することができます．その反面，データが実際にどの程度変化しているのかといった解釈の難しさがあります．

ここではより直感的な解釈が可能で，より汎用性のある方法として「回帰モデル」を紹介します．回帰モデルとは「複数の変数間の関係を説明する関係式」であると言えます．回帰モデルを使えば相関係数のような「ある変数 X と変数 Y の関係性（方向性がある場合には変数 X の変数 Y への影響力）を評価する」ことができたり，「ある変数 X から変数 Y を予測する」ことができたりと，幅広い分析に対応が可能となります．

回帰分析で利用する変数を大きく 2 種類に分けると，(i)結果として生じると仮定する変数「目的変数（従属変数）」と，(ii)目的変数を説明する変数「説明変数（非従属変数）」とに分けられます．回帰モデルには様々な種類があり，目的変数のデータの種類によって内容や名前が異なります．最もよく利用される基本的なものとして，目的変数が連続変数である場合には「線形回帰モデル」，目的変数が二値変数である場合には「ロジスティック回帰モデル」などが挙げられます．本講義では最も基本的で，他の回帰モデルの考え方のベースにもなる，線形回帰モデルについて紹介します．

線形回帰モデルの表現方法

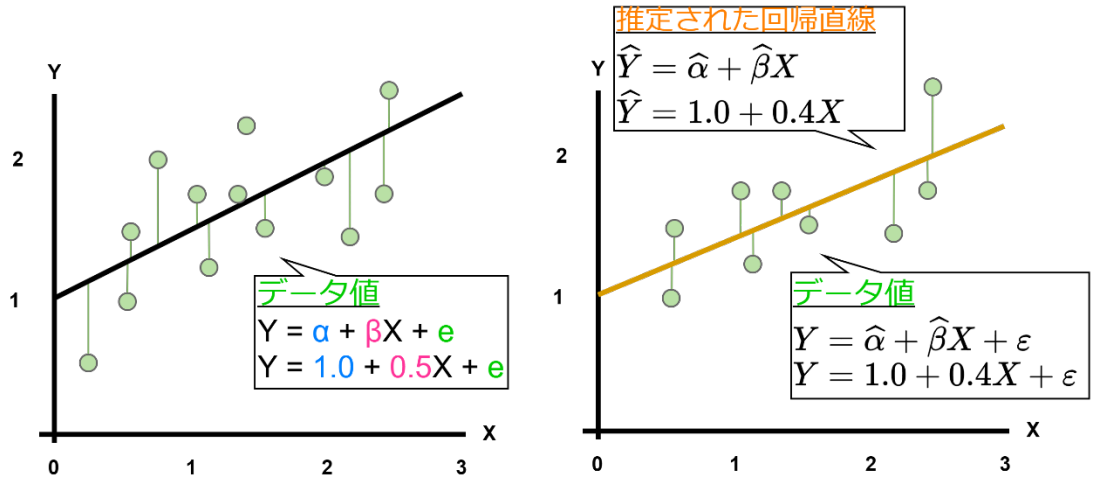
回帰モデルで変数間の関係を数式で表すとき，最も基本的な形は左辺に目的変数 (Y)，右辺に説明変数（X）をおいた，以下のような恒等式を利用します．

$$Y = \alpha + \beta X + e$$

説明変数 X の目的変数 Y への影響力の大きさを係数（β）で示します．特に回帰モデルで利用される係数のことを「回帰係数（β）」と呼びます．さらに説明変数 X が 0 をとるときの値は「切片（α）」と呼ばれます．さらに実際には変数 Y の値は，回帰直線上にぶれずに並んでいるわけではありません．それぞれの Y の値は回帰直線から乖離があり，その乖離具合はランダムにばらついています．この回帰直線とのランダムな乖離を「誤差（e）」と呼び，誤差を加えた $Y = \alpha + \beta X + e$ が実際の変数 Y の値となります．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．目的変数 Y を説明変数 X で説明する線形回帰モデル



また$Y = \alpha + \beta X + e$で表現される目的変数Yは母集団におけるデータ値を表現するものです．私たちは母集団における変数Xと変数Yの真の関係性を私たちは知り得ないので，真のαとβ，そしてeは分かりません．したがって標本のデータから切片αと回帰係数βを推定することが必要です．この標本のデータを使って推定された切片αと回帰係数βは，一般的にハット^をつけて$\hat{\alpha}$，$\hat{\beta}$と表記されます．そのため標本から推定された線形回帰モデルの回帰直線も$\hat{Y} = \hat{\alpha} + \hat{\beta}X$と表現されます．目的変数Yについても，推定された$\hat{\alpha}$と$\hat{\beta}$に基づいて計算されるため，推定値となりハットが付与されていることに注意してください．

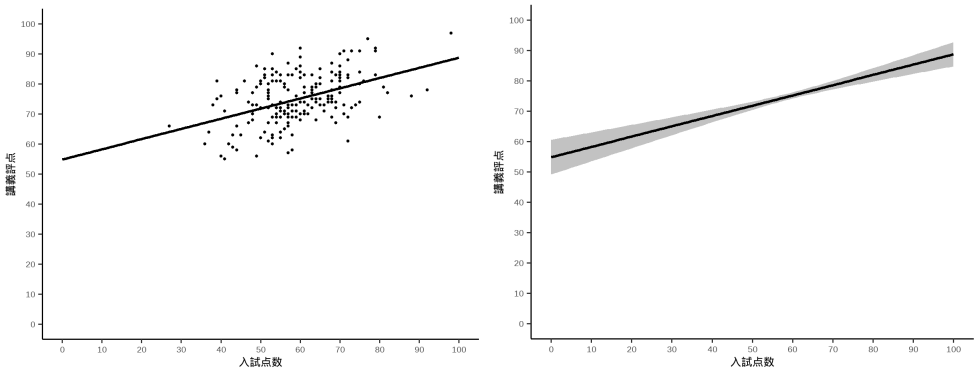
さらに標本における変数 Y の観測値を表すためには，誤差に関する項を加える必要があります．真の回帰直線（$Y = \alpha + \beta X$）と各データの乖離は誤差 e で表現していましたが，推定された回帰直線（$\hat{Y} = \hat{\alpha} + \hat{\beta}X$）と各観測値の乖離は「残差 ϵ」で表現します．したがって標本の各データ値Yは $Y = \hat{\alpha} + \hat{\beta}X + \epsilon$ と表現されます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

線形回帰モデルの実践

線形回帰モデルを利用した分析の例として，講義評点を入試点数で説明してみましょう．このとき目的変数を講義評点，説明変数を入試点数とした線形回帰モデルを推定します．

図表．講義評点と入試点数の関係性を推定する回帰直線



今回のデータで推定された入試点数の回帰係数は約 0.34 であり，入学時の数学の点数が 1 点大きいとき，講義の評点は約 0.34 点高くなる傾向にあるという正の相関があることが示唆されました．また切片の推定値は約 54.86 であり，入試時の数学の点数が 0 点のとき，講義の評点は約 54.86 点であると解釈できます．ただし実際にはそのようなデータは存在せず，この切片はあくまでも他のデータ値から推定されたものにすぎません．実務上，切片はこの後の章で触れる「予測」の文脈では重要になりますが，講義の評点と数学の点数の関係性を示すためには利用しません．

また線形回帰モデルで回帰直線を推定した場合，その推定の精度を表すために 95%信頼区間をともに併記することが一般的です. 回帰直線を図で表す場合には, 以下のように 95%信頼区間を網掛けの範囲で表すことも良くされます．ちなみにこの網掛けの 95%信頼区間の間で回帰直線の傾きは（抽出する標本によって）変化しうるので，この範囲が広いほど，標本によっては傾きが 0（つまり説明変数と目的変数には関連性がない）となり得ると考えることができます.

神戸大学　数理・データサイエンスセンター
加葉田大志朗

標準化したデータを利用すると相関係数が得られる

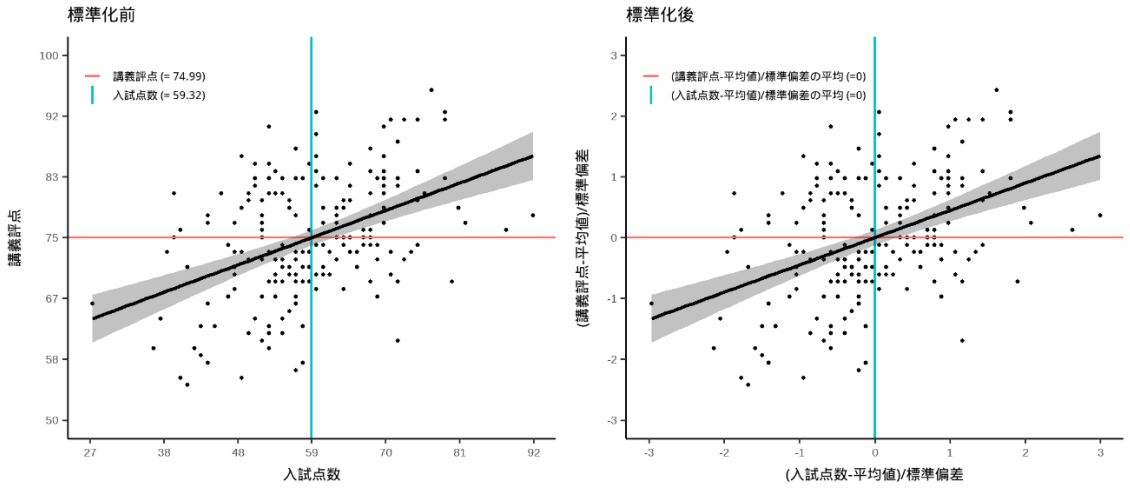
今回は相関係数をより汎用的に表現する方法として，線形回帰モデルを紹介しています．そのためここでは，線形回帰モデルによって相関係数を算出する方法を紹介しておきます．線形回帰モデルを利用して（ピアソンの）相関係数を推定することは非常に簡単で，いかのいずれかの手続きをとれば実行可能です．

（1） 標準化（$\frac{観測値-平均値}{標準偏差}$）したデータを使って回帰係数を推定する

（2） 未加工の観測値から推定した回帰係数に$\frac{説明変数の標準偏差}{目的変数の標準偏差}$をかける

まず（1）の方法で相関係数を推定してみましょう．この標準化後の説明変数（入試点数）の回帰係数推定値は約 0.45 であり，「標準化した説明変数が 1 大きくなると標準化した目的変数が約 0.45 大きくなる」ことがわかります．標準化の過程では，データ値を標準偏差で割っている（つまり 1/標準偏差をかける加工をしている）ため，標準化後のデータの分布の標準偏差は 1 となります．つまり「標準化した説明変数が 1 大きくなると標準化した目的変数が約 0.45 大きくなる」は，「説明変数が 1 標準偏差分だけ大きくなると目的変数が標準偏差の約 0.45 倍大きくなる」と解釈することができます．このことを直感的に理解するために，標準化前後の回帰直線を見てみましょう．

図表．標準化前後のデータから推定した回帰直線（左：標準化前，右：標準化後）



神戸大学　数理・データサイエンスセンター
加葉田大志朗

続いて（2）の方法で相関係数を推定してみましょう．まず標準化しないデータを使って推定した入試点数に係る回帰係数は約 0.34 でした．そして説明変数である入試点数の標準偏差（SD_x）は約 10.91，目的変数である評点の標準偏差（SD_y）は約 8.25 でした．従って相関係数は，

$$\text{未変換の観測値から推定した回帰係数} \times \frac{\textit{説明変数の標準偏差}}{\textit{目的変数の標準偏差}} \simeq 0.34 \times \frac{10.91}{8.25} \simeq 0.45$$

となります．これらの手続きから，線形回帰モデルで推定された回帰係数と相関係数の違いは，データを標準化（もっと厳密に言えば標準偏差を要したスケーリング）しているかどうかにすぎないことが理解できます．

説明変数が離散変数の場合

上記までで紹介した例はいずれも，説明変数が連続変数の場合でした．ここでは説明変数が離散変数，特に二値変数である場合の考え方について説明します．あえて取り上げてはいるものの，基本的な考え方はこれまでと全く同じなので安心してください．

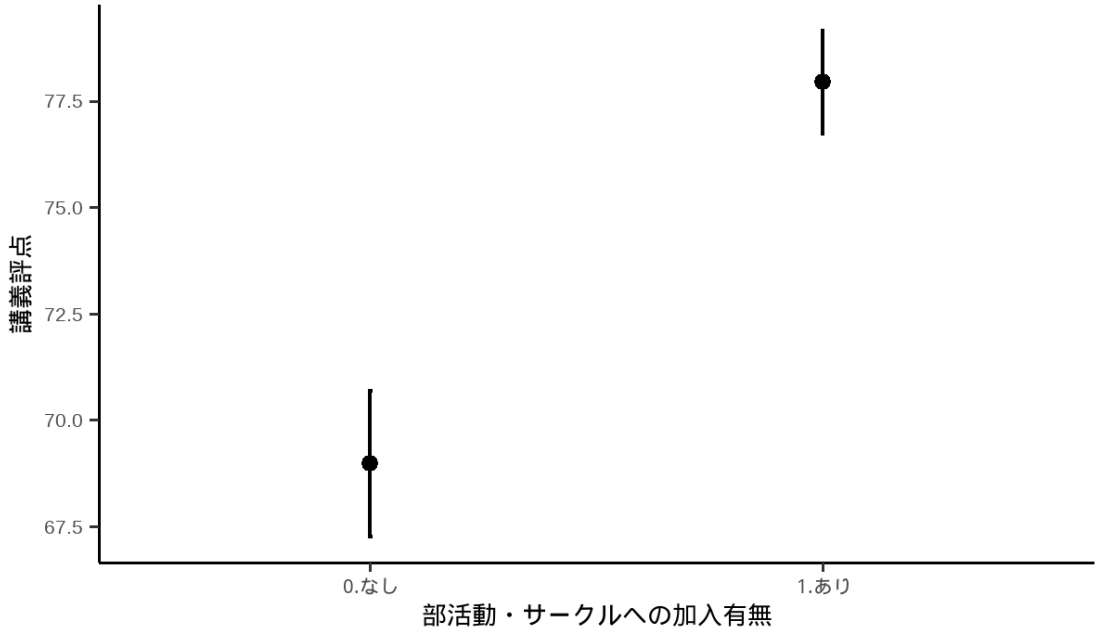
例として，部活動・サークルへの加入有無（あり・なしの二値変数）の講義評点への影響力を評価してみましょう．線形回帰モデルでは講義評点を目的変数，部活動・サークルへの加入有無を説明変数とします．回帰モデルで二値変数を利用する場合，いずれかの値を 0，もう一方の値を 1 として「ダミー変数」として扱うことが一般的です．離散変数である二値変数もダミー変数として扱えば，0 と 1 だけの値を持つ連続変数と考えることができ，これまで連続変数について考えてきた解釈が応用できます．今回の例では，部活動・サークルへの加入有無を，なし＝0，あり＝1 のダミー変数として考えます．

これらのデータを使って推定された部活動・サークルへの加入有無に係る回帰係数は約 8.98 でした．これまでの回帰係数推定値と同様に解釈すると，説明変数が 1 大きくなる（部活動・サークルへの加入有無が 0「なし」から 1「あり」になる）とき，目的変数である大学評点が約 8.98 大きくなると解釈できます．

ちなみにこの回帰係数推定値 8.98 は，部活動・サークルへの加入有無が 0「なし」の集団の講義評点の平均値と，1「あり」の集団の講義評点の平均値の差となることも知っておくと，より分析結果の解釈がしやすくなると思います．実際にそれぞれの平均値は約 77.95 と約 68.97 であり，その差分は約 8.98 となります．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．二値変数を説明変数とした場合の回帰係数の解釈

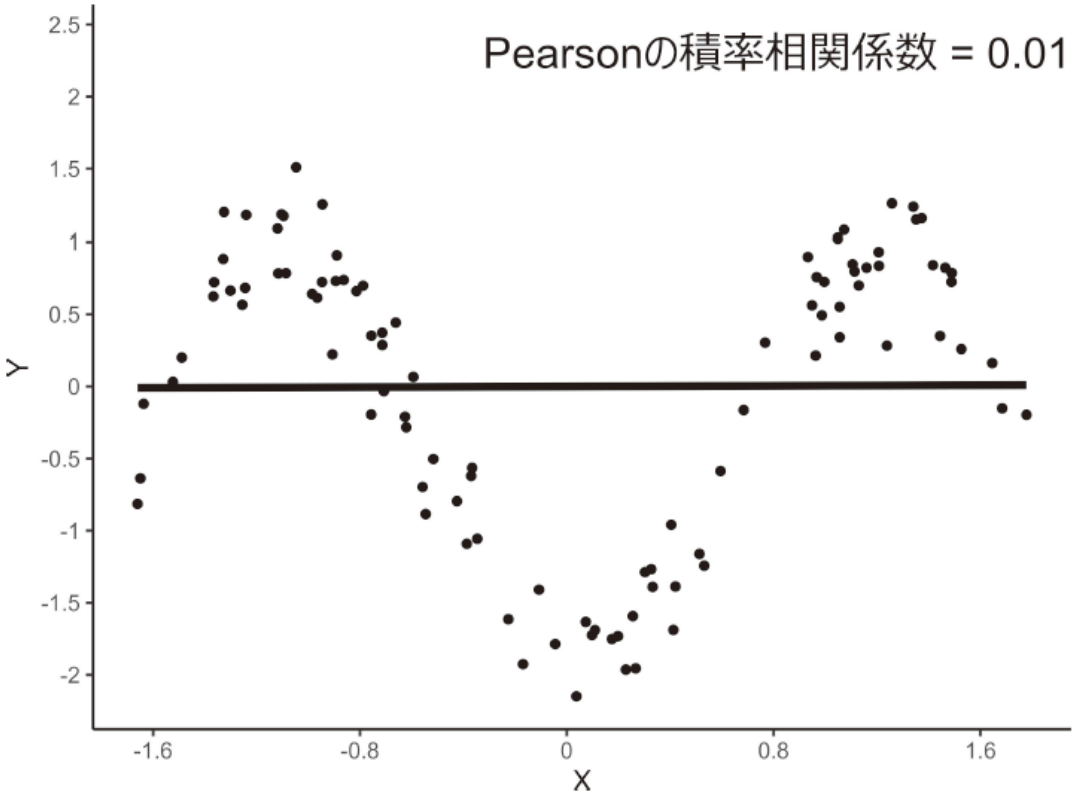


神戸大学　数理・データサイエンスセンター
加葉田大志朗

8. 疑似相関と多変量回帰モデル

相関係数によって，2 つの異なる変数間の直線的な関係の強さが評価できました．ただし直線的な関係にない場合などは，2 つの変数が連動して動いていたとしても，その動き方によっては相関係数は 0 に近い値をとり得ます．

図表．データの分布によっては連動していても相関係数が小さいことがある

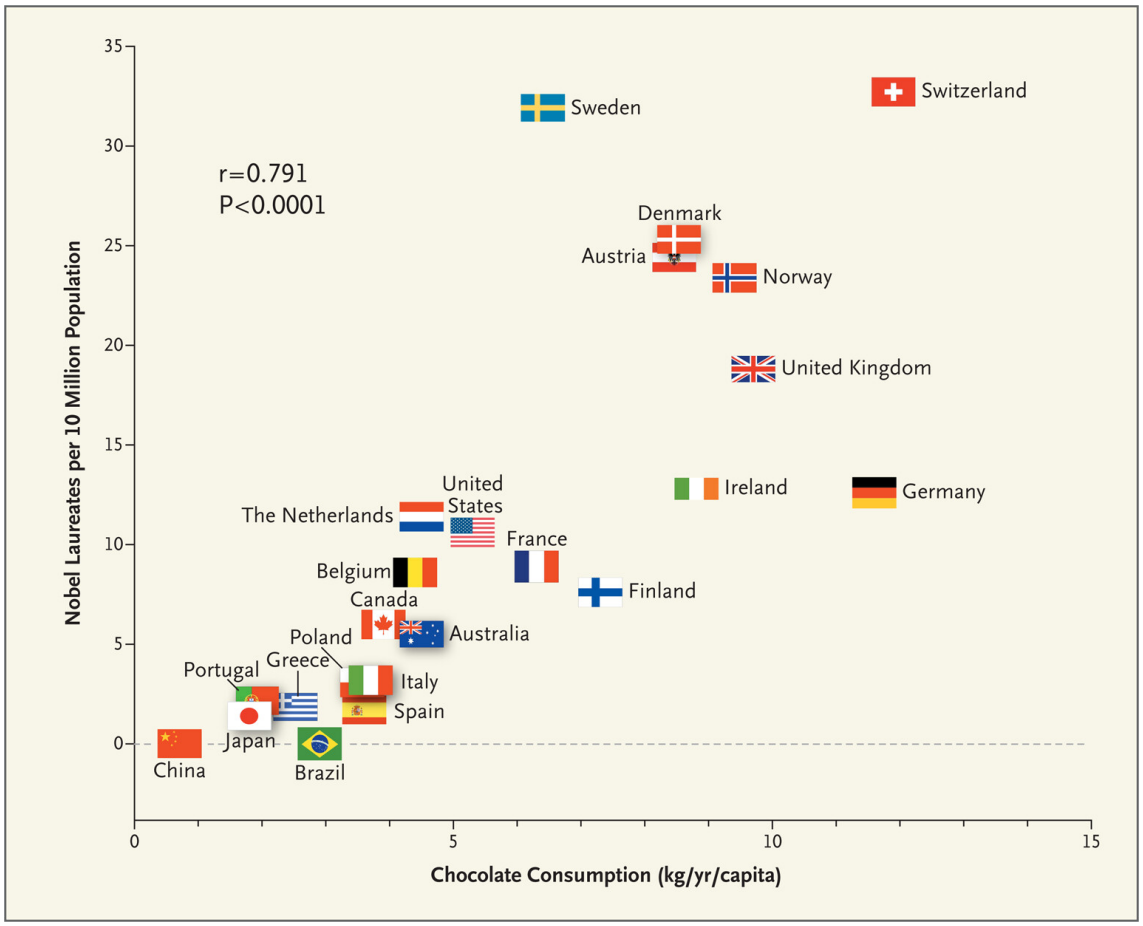


また上記とは異なる理由で，本来であれば全く関係がない（独立な）因子同士でも相関係数が大きくなることがあります．このような見せかけの相関が強くなることを「疑似相関」と呼びます．

疑似相関が生じている可能性がある非常に有名な例として「チョコレートとノーベル賞受賞の相関」があります．以下図は各国の国民 1 人あたりの年間チョコレート消費量と 10 万人あたりのノーベル賞受賞者数の散布図です．相関係数は 0.791 であり，チョコレート消費量とノーベル賞受賞者数はほどほどに強く正に相関しているように見えます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．チョコレート消費量とノーベル賞受賞者数の散布図



（引用）https://www.nejm.org/doi/full/10.1056/NEJMon1211064

上記の「ほどほどに強い正の相関がある」という結果から，「チョコレートを食べるとノーベル賞を受賞しやすくなる」という結論を得られるでしょうか？論文中でも議論がなされていますが，この結果から考えられる仮説の代表的なところは以下の 3 通りあります．

仮説 1．チョコレート消費量（X）がノーベル賞受賞者数（Y）を増加させる

e.g. チョコレートを食べると認知機能が向上してノーベル賞が取りやすくなる？

ただしチョコレート消費量の観測は 2002~2011 年のいずれかの年に観測され，ノーベル賞受賞者数は 1901~2011 年の累積数を利用しています．2002 年以降 10 年間のチョコレート消費量が過去 100 年間のノーベル賞受賞数に影響を及ぼしているかは疑問という反論もありそうです．

仮説 2．ノーベル賞受賞者数（Y）がチョコレートの消費量を増加させる

e.g. 認知機能が高い人はチョコレートの健康効果を意識してチョコレートをよく食べる？

e.g. ノーベル賞を受賞すると祝賀行事でチョコレート消費量が増える？

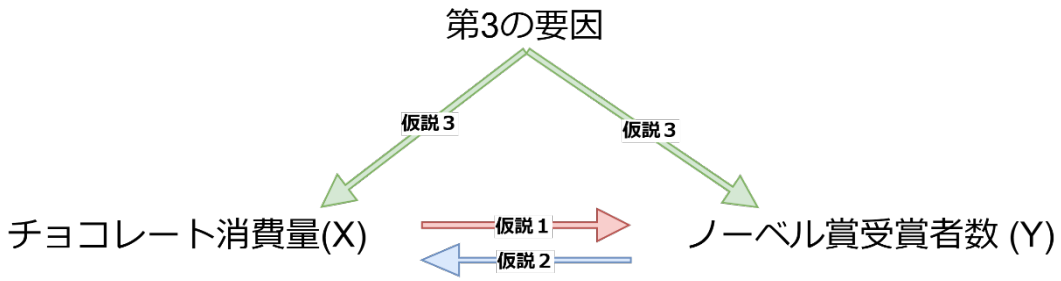
ノーベル賞受賞者はスウェーデン・ストックホルムのノーベル博物館のメダルチョコレートを爆買いするため消費量が増えているとも考えられそうです．ただしその一時的な消費が調査に含まれているかは疑問ですね．

仮説 3．チョコレート消費量（X）とノーベル賞受賞者数（Y）の両方を同時に高める第 3 の因子が存在する．

e.g. 各国の経済水準，文化・歴史的要因，地理・機構的要因など？

この論文でも言及されている通り「相関がある」だけでは，仮説 1 と 2 のようにどちらが原因で，どちらが結果かは判別できません．また仮説 3 のように，両者は直接的な結果と原因の関係にないかもしれません．このような原因と結果の関係性が特定できる関係のことを「因果関係」と呼びます．*つまり*「相関がある」ことと「因果関係がある」ということは別ものであることに注意してください．また冒頭で触れたような複雑で非単調な連動をすることもあるため，因果関係があるからと言って，必ず相関がある（相関係数の絶対値が大きい）とも言えないことにも注意が必要です．

図表．影響の方向性を示すダイアグラム



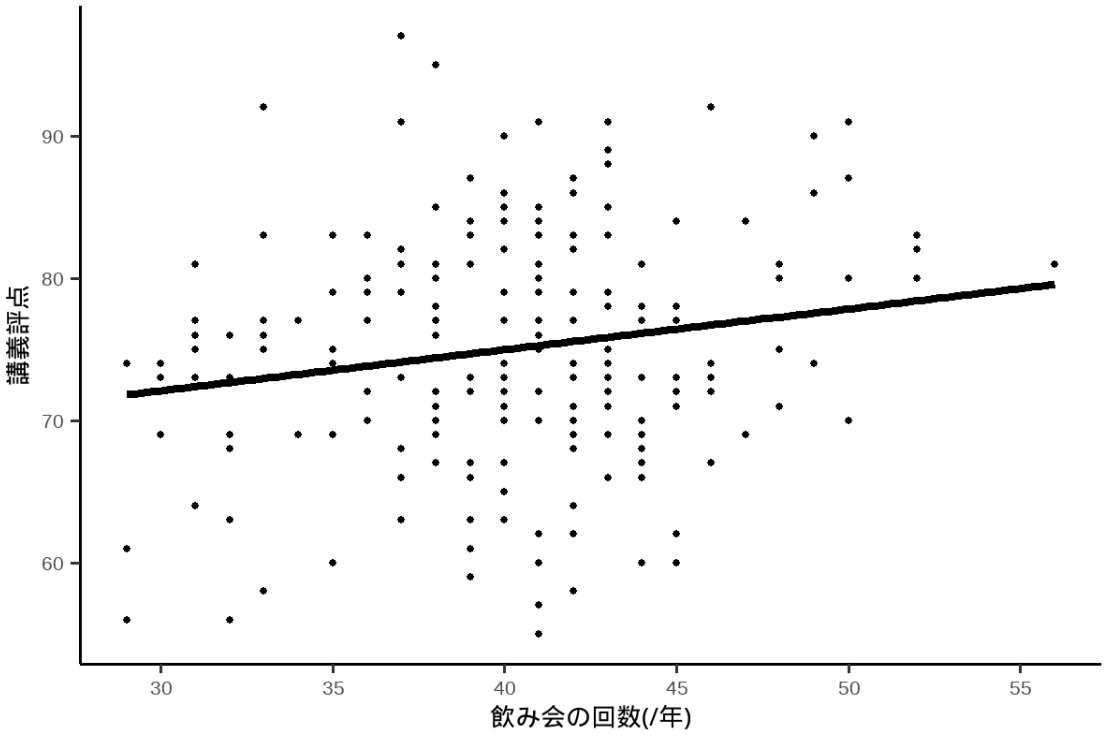
神戸大学　数理・データサイエンスセンター
加葉田大志朗

交絡因子の存在を疑う

上記の仮説 3 で触れたように，2 つの因子に同時に影響を及ぼすことで，疑似的に両者が関連しているように見せる第 3 の因子を「交絡因子」と呼びます．交絡因子の存在によって見せかけの関係が生じる現象を「交絡」と表現します．経済学領域などでは「内生」といった用語が利用されています．

今回は例として，講義評点と年間の飲み会の回数の関係について評価する．この 2 変数の相関係数を算出すると 0.18 であり，大きな値ではないものの正の相関があるという結果となりました．つまり因果性は分からないものの，よく飲み会に参加する学生ほど良い評点を得ているという，一見直感に反する結果が得られています．

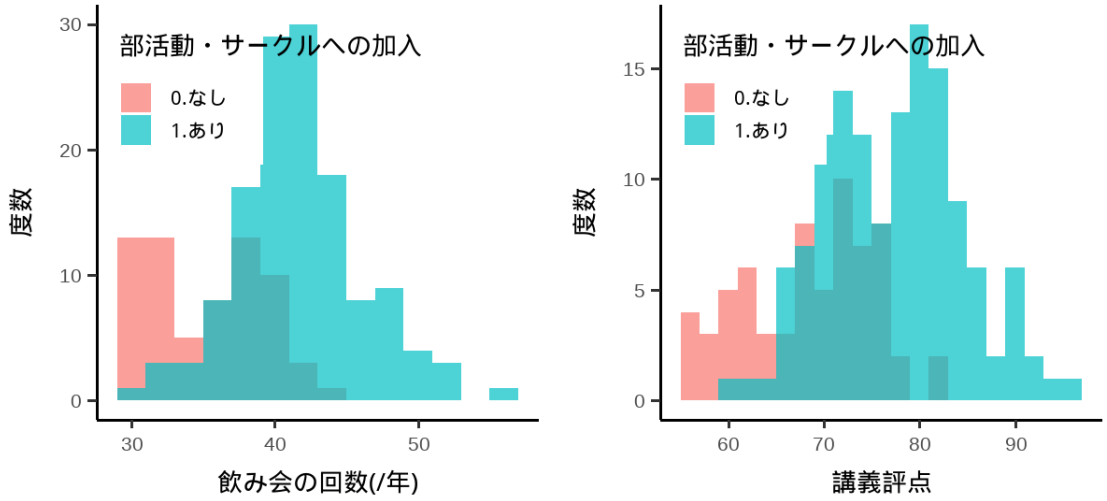
図表．講義の評点と年間の飲み会の回数の散布図



ここで第 3 の因子の存在を疑ってみましょう．今回は「部活動・サークルへの加入有無」に注目してみます．まずは今回の容疑者である部活動・サークルへの加入有無について，説明変数（飲み会への参加回数）と目的変数（講義評点）に同時に影響を及ぼしている可能性があるか確認してみましょう．ここで部活動・サークルへの加入有無ごとに，飲み会への参加回数および講義評点の分布を目視してみます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．部活動・サークルへの参加状況別の飲み会回数および評点の分布



上の図の通り，部活動・サークルへ加入している学生は，飲み会への参加回数が多いことが分かります．これは経験的に納得できる分布ですね．同時に部活動・サークルへ加入している学生は，評点も高い傾向にあることが分かります．大学生の成績には過去問や講義に関する情報を共有してくれる先輩・知人の数が影響するという報告も散見され，これも先験的な感覚と合致しますね．これらから「部活動・サークルへの加入有無」は飲み会の回数と講義評点を同時に上昇させている，つまり飲み会への参加回数と講義評点の交絡因子となっている疑いが強くなりました．

サブグループを作ることで交絡の影響を抑える

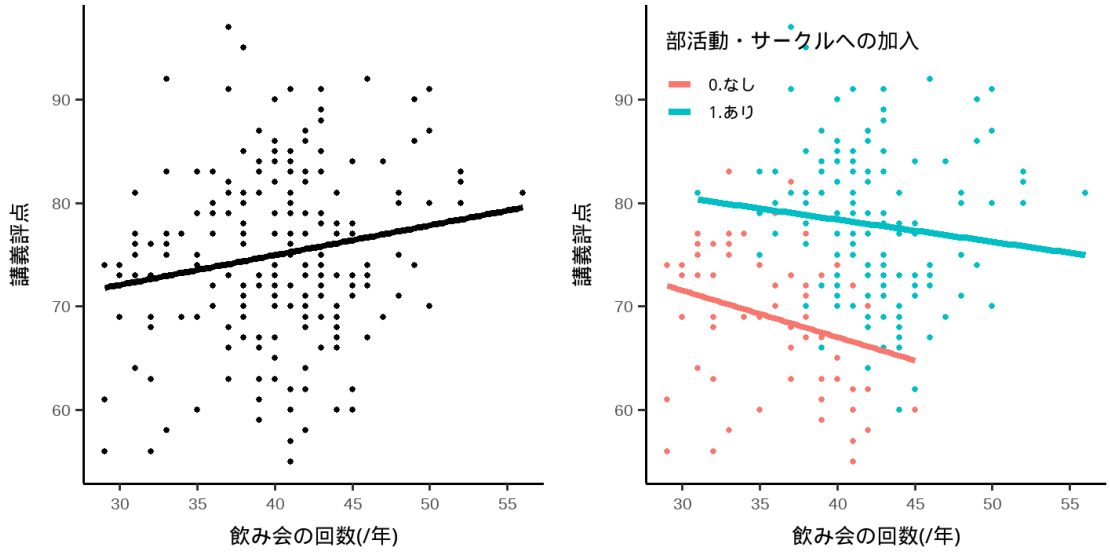
上記より「部活動・サークルへの加入有無」は，飲み会への参加回数と評点に見せかけの関連性を生じさせる交絡因子である可能性が考えられました．それでは，この「部活動・サークルへの加入有無」の影響がなければ，飲み会への参加回数と評点はどのような関係性になっているのでしょうか？つまり交絡の影響を除いた関係性を知ることが次の目標です．

交絡の影響を除く最もシンプルな方法は，同じ状態の「サブグループ」を作ることです．つまり交絡因子の要素（部活動・サークルへの加入あり・なし）ごとにデータを分割したサブグループを作成して，それぞれのサブグループ内で関心のある変数間の評価（飲み会への参加回数と評点）を行うものです．なぜサブグループ化すると，対象の交絡の影響が取り除けるのでしょうか？部活動・サークルへの加入あり・なしでサブグループ化すると，それぞれのサブグループには当然，部活動・サークルへの加入あり・なしいずれかの学生しか存在しないことになります．つまりサブグループ内では部活動・サークルに関する状態はみんな

同じなので，「部活動・サークルへの加入有無による違い」は生じません．従って，それぞれのサブグループ内での評価には交絡の影響が介在しないことになります．

今回の例でも，部活動・サークルへの加入有無でサブグループ化し，それぞれのサブグループの中で相関係数を算出してみます．以下図の通り，すべての集団における散布図は右肩上がり（正の相関）の傾向があるのに対して，それぞれのサブグループ内では散布図が右肩下がり（負の相関）の傾向に変わっています．実際にサブグループ化後の相関係数は，加入ありの場合は約-0.12，加入なしの場合は約-0.27 でした．つまり交絡が生じている状態の相関係数は，交絡が生じていない状態と正反対の結果を示していたと考えると恐ろしいですね．

図表．サブグループ化前後の散布図



サブグループ内で得られた結果を統合する「層別化法」

上記では各サブグループ内で評価された結果は，全ての集団内で評価された結果と異なる場合があることを説明しました．そして今回の例のように，各サブグループ内で得られた結果の方向性が類似している場合には，それらの結果を統合したひとつの値として評価することがあります．このようなサブグループ化（層別化）して求めた結果を統合して扱う手続きを「層別化法」と呼びます．層別化というと，それぞれのサブグループ内で別々に結果を求めるだけのように聞こえてしまいますが，層別化法はその後の統合がセットになっていることに注意してください※．

※　このあたりの混乱を回避するために日本語では「層化統合解析」などと呼んだ方が分かりやすいかもしれないという提案もあります（宇宙怪人しまりす 統計よりも重要なこ

とを学ぶ，佐藤俊哉，2024)．

実際の層別化法では，それぞれのサブグループで得られた結果のデータ数で重みをつけた算術平均を計算することが一般的です．データ数で重みをつけるというのは，100 名の対象者を 60 名と 40 名のサブグループに分けて解析した場合，60 名のサブグループから得られた結果と 40 名のサブグループから得られた結果を 6:4 の比率で考慮するということです．例としてデータ数n_1と n_2のサブグループから得られた結果 $_1$ と結果 $_2$ の重みづけ平均は以下の手続きで計算が可能です．

$$重みづけ平均 = \frac{n_1 \times 結果_1 + n_2 \times 結果_2}{全データ数(n_1 + n_2)}$$

ただし相関係数については単純な平均化ではなく，一度 Fisher の Z 変換と呼ばれる変換が必要です．実際の分析ではソフトウエアが計算してくれますし，この講義の範囲を超えるため詳細は省きますが，大きな流れとしては上記の重みづけ平均と同様だと考えてください．

ここでは交絡が生じている状況下で，交絡因子の影響を低減するシンプルな方法としてサブグループ化法を紹介しました．実はこれとほとんど同じ仕組みで交絡の影響を低減する方法として，「多変量回帰モデル」を利用する方法があるので，以下で紹介します．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

多変量回帰モデルの活用

回帰モデルは複数の変数間の関係性を説明する関係式でした．基本的な構造として，説明変数に回帰係数を掛けることによって目的変数を表現します．実はこの説明変数は，複数個考慮することができます．つまり今まで考えていた

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

とひとつの説明変数 X だけで目的変数 Y を予測（$\hat{Y}$を算出）していたものについて，

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

のように複数の説明変数$X_1, X_2, \ldots, X_p$で目的変数 Y を予測（$\hat{Y}$を算出）する恒等式で表現することができます．このように複数の説明変数を持つ回帰モデルを「多変量回帰モデル」や「重回帰モデル」と呼びます（対照的にひとつの説明変数を持つ回帰モデルを「単変量回帰モデル」や「単回帰モデル」と呼ぶことがあります）．これによって以下のような恩恵が得られます．

（１） 良い回帰係数の推定が可能になる（擬似相関を見破れる）

（２） 良い目的変数の予測が可能になる（より多くのデータで予測できる）

今回は交絡を低減して疑似相関を見破ることを目的とした文脈で多変量回帰モデルを扱うため，（１）に焦点を当ててその内容を解説していきます．

多変量回帰モデルによって調整済みの関連を求める

「講義の評点と飲み会の回数」のように，因果関係がないにも関わらず，見た目上は相関して見える関係性を「擬似相関」と呼びました．擬似相関を生じさせる要因の一つとして，2 つの因子に同時に影響を及ぼすことで，疑似的に両者が関連しているように見せる「交絡因子」があります．この交絡因子によって生じる擬似的な相関を，多変量回帰モデルによって見破ることを目指します．

飲み会への参加回数の講義評点への影響を評価するとき，単変量の回帰モデルでは説明変数は飲み会への参加回数だけでした．

$$\text{講義評点の予測値} = \hat{\alpha} + \hat{\beta}\text{飲み会への参加回数}$$

ここでは部活動・サークルへの加入有無による交絡を取り除くため，部活動・サークルへの加入についても説明変数に加えます．つまり，

$$\text{講義評点の予測値} = \hat{\alpha} + \hat{\beta}_1\text{飲み会への参加回数} + \hat{\beta}_2\text{部活動・サークルへの加入有無}$$

というモデルを考えることになります．このように多変量回帰モデルでは，飲み会への参加回数と部活動･サークルへの加入有無の影響を，それぞれ$\hat{\beta}_1$と$\hat{\beta}_2$という係数で考えています．なぜこのような形にすると，交絡の影響を小さくすることができるのか見てみましょう．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

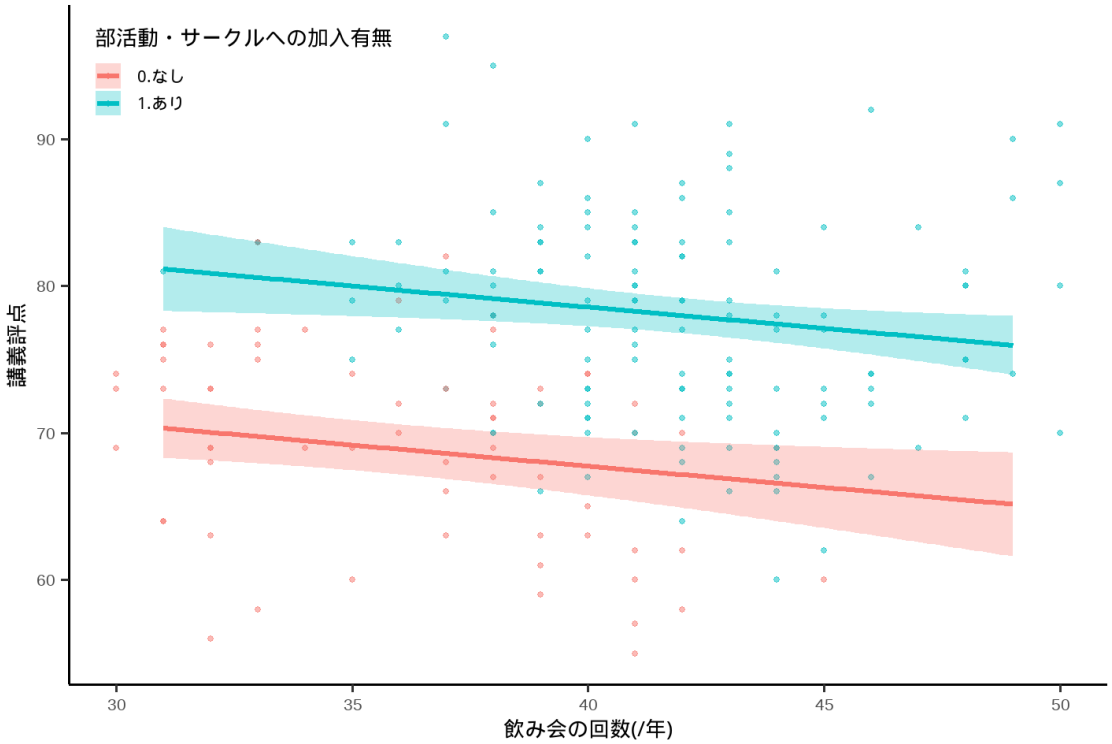
モデルから得られる結果をわかりやすくするため，部活動・サークルへの加入有無の有無をダミー変数として考え，その値が 0 の場合と 1 の場合について上記の多変量回帰モデルの推定式を書き出してみます．

部活動・サークルへの加入ありの講義評点の予測値 $= \hat{\alpha} + \hat{\beta}_1$飲み会への参加回数 $+ \hat{\beta}_2 \times 0$

部活動・サークルへの加入なしの講義評点の予測値 $= \hat{\alpha} + \hat{\beta}_1$飲み会への参加回数 $+ \hat{\beta}_2 \times 1$

上記それぞれのサブグループにおける講義評点の予測値は，それぞれ以下の回帰直線上に存在することになります．

図表．サブグループ内の回帰直線



上記の赤ラインは「部活動・サークルへの加入ありの学生の飲み会の講義評点への影響」であり，青ラインは「部活動・サークルへの加入なしの学生の飲み会の講義評点への影響」です．図を見てわかる通り，赤ラインと青ラインは平行になっています．このことから「飲み会の回数」の回帰係数の推定値は「部活動・サークルへの加入有無」には依存しないことが分かります．これは回帰モデル内の「部活動・サークルへの加入有無」に何が入ろうと $\hat{\beta}_1$ の値は変わらないことからも明らかです．今回の結果は，部活動・サークルへの加入有無によらず，飲み会の回数は講義評点に負の影響を与える可能性があることが分かります．つまり多変量回帰モデルで交絡因子である「部活動・サークルへの加入有無」の影響を受けない「飲み会の講義評点への影響」を評価できていると言えます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

反対に「部活動・サークルへの加入有無の影響」は２つの回帰直線の高さの違いとして反映されており，２つの回帰直線は平行なので，その影響力は飲み会への参加回数に依存しないことが分かります．従って「部活動・サークルへの加入有無の影響」は飲み会の回数によって変わらず$\hat{\beta}_2$のままです．このようにそれぞれの説明変数の影響を独立に考えることで，それぞれの回帰係数推定値を，それぞれの独立した影響力としてみなすことができるわけです．このような交絡因子の影響を取り除いた独立な回帰係数を求めることを「調整する」と表現したり，そこで得られた回帰係数推定値を「調整済み回帰係数」などと呼ぶことがあります．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

多変量回帰モデルを利用する際に注意すること

交絡の調整のための非常に強力なツールとなる多変量回帰モデルですが，実際の分析ではいくつか注意が必要な点があります．特に代表的な以下の２つについて紹介しておきます．

（１） モデルの過適合（overfitting）
（２） 多重共線性

（１）モデルの過適合（overfitting）

「モデルの過適合」とは「自分の推定したモデルが自分の標本でしか再現性がない状態」を指します．つまり同じ母集団から得られた標本であっても，少しデータの内容が異なってしまうと異なる結果が導かれてしまいます．これを「外的妥当性」や「汎用性」が低い状態などと呼びます．モデルの過適合はデータ数に対して過剰に多くの説明変数が考慮されているモデルや，過度に複雑な組み合わせや非線形性を考慮したモデルなどで起こります．

あくまでも直感的なイメージの助けとして例を出すと，皆さんは株式会社 ZOZO が提供していた「ZOZOSUITE」をご存じでしょうか？あれはサービス利用者の体の各部位のサイズを精緻に計測し，その結果をもとに個々の利用者にジャストフィットする服を提供してくれるという革新的なサービスです．個々の利用者のデータを詰め込んでジャストフィットさせた服のメリットは，やはりその個人の体にはピッタリ適合してくれるという点です．一方で服は同じ身長・体重の方であっても，肩幅や腕の長さが少しでも違っていれば，その服はむしろ着心地の悪いものになってしまいそうです．つまりこの服は他の誰にも着こなせない汎用性の低い服になってしまっている可能性があります．むしろある程度体格が似た複数人が着用することを前提とすれば，UNIQLO などの M サイズの方が多くの方にとってある程度着心地が良いものとなるでしょう．つまり個人のデータを詰め込むことで，各個人へのフィットが高まり過ぎて，少しでも異なる他の人への汎用性が低下してしまう過適合の問題が生じてしまうと言えます．ZOZOSUIT は過適合した服を作ることが目的ですが，統計解析では目の前の標本についてのみ説明するのみでなく，今後同じ母集団から得られる他の標本にも応用できる推測を行う必要があります．従って特定の標本のデータを詰め込み過ぎた分析結果は，他の標本の結果の推測に利用できないため不適切なものとなってしまいます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

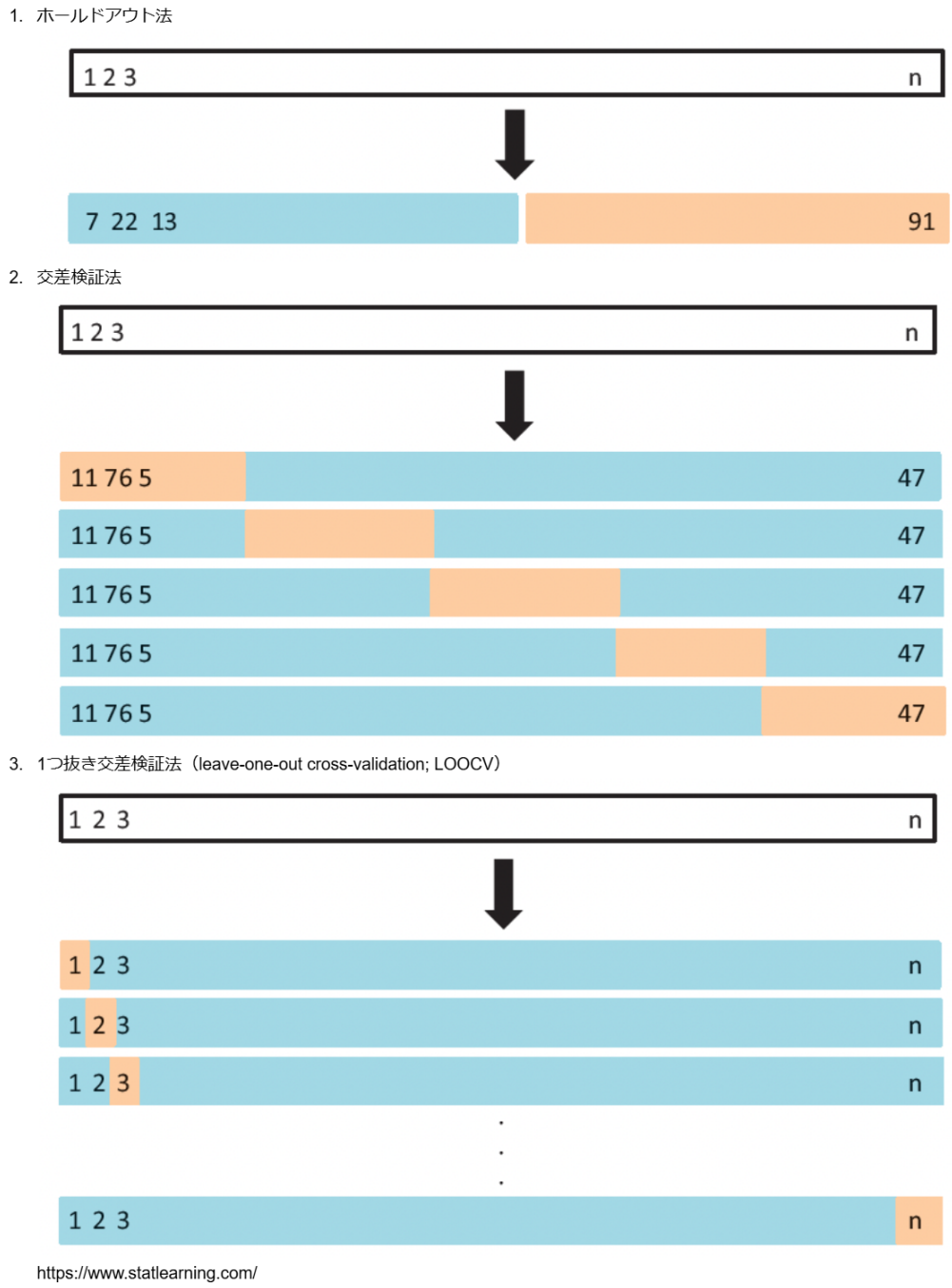
図表．モデルの過適合のイメージ



この過適合を防ぐために，１つの標本のデータから模擬的に複数の標本を作り出して（１つの標本を分割して）過適合が起こっていないか確認する方法があります．代表的なものとして（１）ホールドアウト法，（２）交差検証法，（３）１つ抜き交差検証法（leave-one-out cross-validation, LOOCV）などが挙げられます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．過適合の有無を検証するための代表的な方法



ただし上記のような検証法を実際の分析で実装するのは，結構大変です．そこで多変量回帰モデルにおいて過適合が起こらない範囲で含めることができる説明変数の数の簡単な目安（rule of thumb）なども提案されています．全ての例を説明することは控えますが，例えば目的変数が連続変数である線形回帰モデルの場合は，(標本に含まれるデータ数/15)個までの説明変数を含めることができます※．これを見ても分かる通り，標本に含まれるデータ

数が多いほど，過適合は起こりにくく，多くの説明変数を考慮することが可能です．

※　３つ以上のカテゴリーをもつ離散変数を説明変数とする場合は，(カテゴリー数-1)個分の説明変数としてカウントします．

（２）多重共線性

「多重共線性」とは，相関が過度に強い説明変数を同時にモデルに含めることによって，それらに係る回帰係数の推定値が安定しなくなる現象です．例えば体重と BMI (body mass index ＝ 体重 kg/身長 m^2) を同時に説明変数として考慮とします．BMI は体重を身長で調整した指標なので，体重と BMI は当然ながら非常に強い相関があります．この場合，回帰モデルは体重と BMI のどちらが目的変数に影響を与えているかを明確に区別することが難しくなります．すると以下のような影響が生じ得ます．

1. **回帰係数の不安定性**
   体重と BMI がどちらもほぼ同じ情報を持っているため，モデルはその影響を適切に分配できず，回帰係数が過剰に大きくなったり，予測には重要でない係数に大きな値が割り当てられることがあります．
2. **解釈の困難さ**
   体重の回帰係数が正で，BMI の回帰係数が負といった矛盾する結果が得られる場合があります．このような場合，どちらの変数が目的変数にどのような影響を与えているのかが不明瞭になります．
3. **推定精度の低下**
   多重共線性があると，回帰係数の推定の精度が低下するため，標準誤差が大きくなることがあります．なお多重共線性が生じていることで推定精度が極端に悪くなっている（推定量の分散が増大している）ことを測る指標として VIF (variation inflation factor)が利用されます．VIF の二乗根は「説明変数の回帰係数推定値の標準誤差が，説明変数同士が独立である場合の何倍か？」を示します．多重共線性が起こっていると判断する基準としては，文献によって異なりますが３，５，１０などの値が利用されます．

上記の通り多重共線性があるとモデルの信頼性や解釈可能性に悪影響を及ぼすため，分析者はこれを避ける努力をする必要があります．具体的にはモデルに含める説明変数を減らしたりすることで多重共線性の影響を軽減することが可能です．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

9. 回帰モデルによる予測とその良さを測る方法

これまでは変数間の関連性を評価することを目的として，それぞれの説明変数の回帰係数の推定値に着目して回帰モデルを扱ってきました．ここでは回帰分析のもう一つの大きな役割として「予測モデル」としての側面を紹介します．

回帰モデルは「説明変数と目的変数の関係を説明する関係式」であり，説明変数から目的変数を予測する目的でも利用することができます．つまり予測を目的とする場合，私たちの関心は個々の説明変数に係る回帰係数ではなく，目的変数の予測値になります．例えば以下の多変量線形回帰モデルにおいては，説明変数説明変数$X_1, X_2, \ldots, X_p$から目的変数 Y を予測（$\hat{Y}$を算出）していると言えます．

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

予測を目的とした場合も，得られた結果に対する焦点が異なるだけで，回帰モデルの推定の手続き自体はこれまでと大きく変わりません．そこで今回は「どのような場合に良い予測ができるのか」という点に焦点を当てて解説します．

予測モデルの良さの指標

予測モデルの良さを測る代表的な指標として，
（1） 決定係数 R^2
（2） 赤池情報量基準 AIC（Akaike Information Criteria）
を紹介します．

（1）決定係数 R^2

R^2: 決定係数は「回帰係数が目的変数のばらつきを説明できている割合」と表現できます．実際の計算式は以下の通りです．

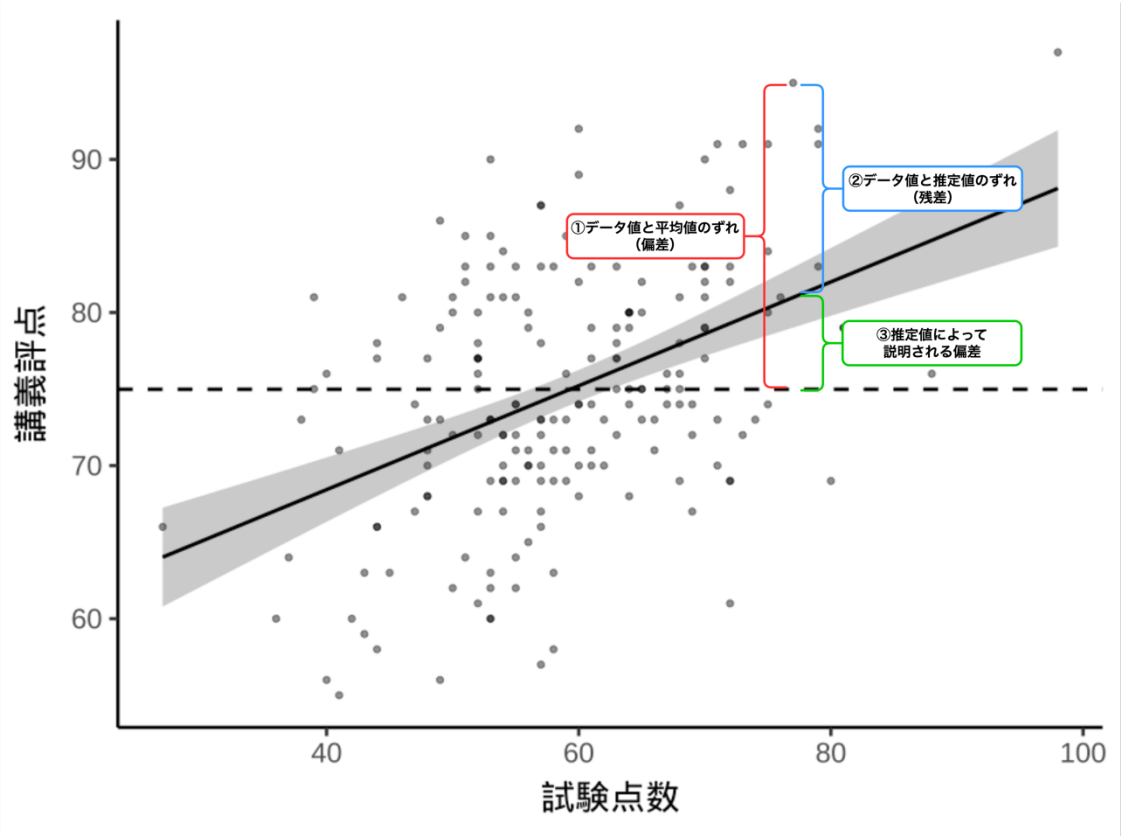
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - \frac{②データ値と推定値のずれ\left(残差\right)の二乗和}{①データ値と平均値のずれ\left(偏差\right)の二乗和}$$

$$= \frac{③推定値によって説明される偏差の二乗和}{①データ値と平均値のずれ\left(偏差\right)の二乗和}$$

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．決定係数の各要素



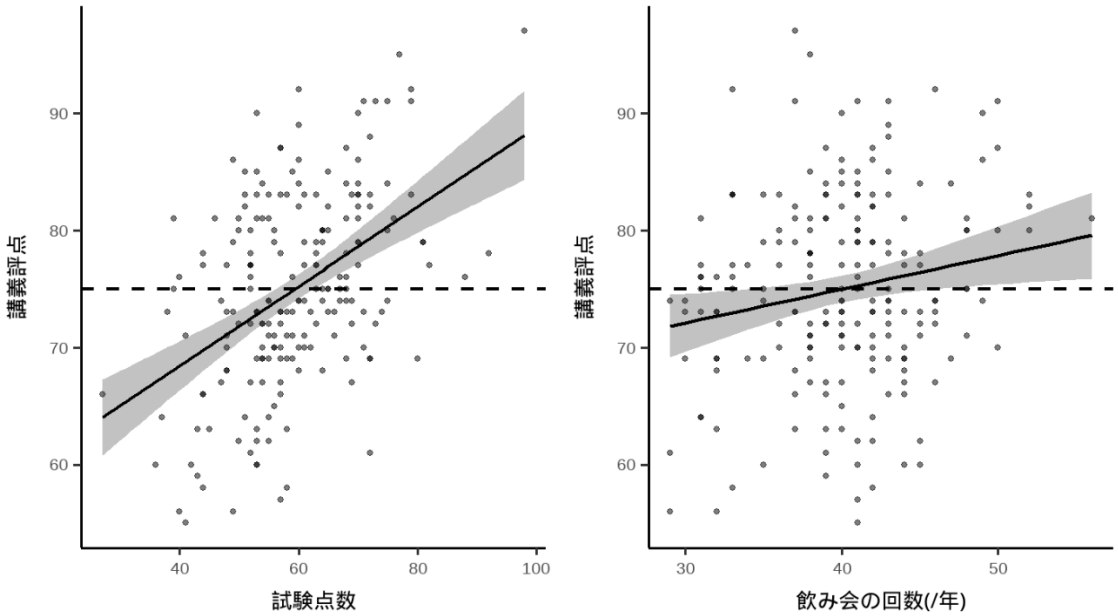
まず第二項の分母にあたる①データ値と平均値のずれ（偏差）は，これまで見てきた通り「データのばらつき具合」を示します．そして第二項の分子にあたる②データ値と推定値のずれ（残差）も，これまで見てきたように「モデルによって推定された予測値と実際に観測されたデータ値のずれ =推定値によって説明できなかったデータのばらつき」を示しています．従って決定係数の第二項は「データのばらつきのうち，推定値によって説明されなかった割合」と解釈できます．これを 1 から引いた決定係数は「データのばらつきのうち，推定値によって説明された割合」と解釈することができます．

ここで推定値が完全に観測されたデータ値と一致した場合，第二項の分子にあたる②残差は 0 になるため，第二項「推定値によって説明されなかったばらつきの割合」は 0 になります．従って決定係数は 1 となり，推定値によってデータのばらつきはすべて説明されていると解釈することが可能です．反対に，推定値が平均値と全く同じだった場合，第二項の分子にあたる②は，分母の①偏差と等しくなります．従って第二項は 1 になり，決定係数は 0，推定値はデータのばらつきを説明するのに全く役に立たないことになります．

例えば，講義評点を目的変数とした単変量の線形回帰モデルについて，
① 入試時の数学の点数を説明変数としたモデル
② 1 年間の飲み会の回数を説明変数としたモデル
の 2 つのモデルについて，どちらの R^2 がより大きくなるか見比べてみます．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

図表．説明変数が異なる 2 つのモデル（左：①入試点数，右：②飲み会回数）



上記より，①入試点数のほうが，平均値よりもデータのばらつきを捉えることができているように見えます．実際それぞれの R^2 は①では 0.201，②では 0.033 であり，①の方がデータのばらつきをよく説明できていると解釈することができます．

上記は目的変数のデータのばらつきをひとつの説明変数で説明する単変量の回帰モデルに関して，説明変数の違いによる R^2 の違いを評価しました．しかし 1 種類のデータ（説明変数）だけで予測するよりも，複数の種類のデータ（説明変数）を使って予測を行った方が，より良い予測に繋がりそうです．この直感の通り，多変量の回帰モデルでは説明変数の数が増えれば増えるほど，R^2 の値は単調増加していきます．つまり R^2 のみを予測モデルの良さの基準としてしまうと，単純に説明変数の数が多いモデルが良いモデル，という結論に至ってしまいます．このような懸念を低減するために，モデル内に考慮した説明変数の数（厳密にはそこから計算される自由度）に応じてペナルティを加えることで R^2 が過剰に大きくなることを抑えた「自由度調整済み R^2」というものも使われています．ただし実際には自由度調整済み R^2 も，程度の違いはあれど，説明変数の数が増加すれば大きくなることは覚えておいてください．

また実際の標本に含まれるデータ数は有限なので，多くのケースでモデルの過適合への配慮が必要になります．そのためモデルの過適合が起こらない範囲で，R^2 を高く保つような予測モデルを構築することが必要になります．

（2）赤池情報量基準 AIC

AIC（Akaike Information Criterion）は，予測モデルがどれくらい「うまくデータを説明できているか」を評価しながら，モデルの「複雑さ」にもペナルティを与える指標です．平

神戸大学　数理・データサイエンスセンター
加葉田大志朗

たく言うと,**「良い予測ができるシンプルなモデル」を選ぶための基準**だと考えてください.

AIC の定義を正確に理解するためには，尤度などの統計学に関する踏み込んだ知識が必要となります．従ってこの講義では，定義について簡単な概要だけ説明するに留めます．AIC は以下の数式で与えられます.

$$AIC = -2logL + 2k$$

ここで L は「推定したモデルがどれだけ標本で観察された現象を尤もらしく説明できているか」を示す指標（尤度）であり，これが大きいほどモデルの標本のデータへの適合度が高いことを意味します．この尤度に負の符号がついているので，モデルが標本のデータへ適合しているほど，第一項は小さくなっていきます．k はモデルで推定しているパラメータの数であり，これは説明変数の数やその複雑さに依存し，これが大きいほどモデルが複雑であることを表します．従って考慮する説明変数の数が多いほど，第二項は大きくなっていきます.

モデルの適合度が高いほど良いモデルと考えられるため，AIC は小さいほど良いモデルであると言えます．ただし適合度を高めるために無暗に説明変数の数を増やして複雑なモデルを組んでしまうと，第二項の値が大きくなるため AIC は反対に大きくなってしまいます．従って，推定するパラメータの数によって決定される第二項は，モデルを複雑にすることに対するペナルティであると考えることができます．このペナルティを受けつつも AIC を最小にする落としどころが,「適合度が高く、かつシンプルなモデル」として選び出されます.

上記で紹介する R^2 は説明変数の数が大きくなるほど大きくなる特性があるため，複雑なモデルを良いモデル，としてしまう傾向があると紹介しました．そしてそれにペナルティを与えるために推定するパラメータの数でペナルティを加える自由度調整済み R^2 であっても，単調増加関数であることには変わりありません．これらと対比して，AIC がどのようなふるまいを見せるのか，実際のデータ分析で比べてみます．今回は例として，講義の評点を目的変数として,

① 入試点数，部活動・サークルへの加入を説明変数としたモデル
② 入試点数，部活動・サークルへの加入有無，飲み会の回数を説明変数としたモデル

についてそれぞれ，R^2，自由度調整済み R^2，AIC を計算してみました．結果は以下の通りです.

図表．それぞれのモデルの R^2，自由度調整済み R^2，AIC

モデルの説明変数	R^2	自由度調整済み R^2	AIC
① 入試点数，部活動	0.478	0.473	1288.457
② 入試点数，部活動，飲み会	0.481	0.473	1289.457

上記の通り，説明変数の数を増やすと R^2 は大きくなりました．ただし単調増加関数であるものの推定するパラメータの数に応じてペナルティを加える自由度調整済み R^2 はほとん

神戸大学　数理・データサイエンスセンター
加葉田大志朗

ど変わりませんでした．そして AIC については，考慮する説明変数を増やした（モデルを複雑にした）にも関わらず大きくなって（モデルの適合が悪くなって）しまっています．これは飲み会への参加回数というデータを加えて得られる説明力が小さく，それよりも推定するパラメータが増えたことによるペナルティが勝ったと解釈できます．平たく言えば「この程度の貢献なら飲み会への参加回数は考慮しない方がマシ」と解釈できそうです．

このように評価に利用する指標が異なればどれを「良いモデル」とするかも変わってきます．どの指標に基づいてモデルを選択するかは，実際には研究の目的などによっても変わってきます．パラメータの数やモデルの複雑さなど度外視してよければ R^2 がとにかく高いモデルを作ればいいですし，考慮するデータを無闇矢鱈に増やさずに当てはまりの良いモデルを探したいということであれば AIC を使えばよいと言えます．多くの場合には実現可能性や汎用性の側面から，予測モデルに使われるデータはある程度選定されることが多いので，AIC を使ったモデル選択がより多く使われている印象です．

神戸大学　数理・データサイエンスセンター
加葉田大志朗

機械学習と回帰モデルの境界はどこ？

予測を題材とすると，統計学と機械学習の違いについての議論を時折見かけます．よく観測される主張としては「①モデルの内容の説明性を重視する場合は統計学を使う」「②モデルの内容の説明性は重視せず，得られた予測や分類結果だけに興味がある場合は機械学習を使う」といったものがあります．この両者の違いは①回帰係数に着目しているか，②目的変数の予測値に着目しているかだけの違いであり，どちらも同じモデルから得られる結果であることは，これまで学んできた通りです．つまり基本的にはやっていることに大差はなく，我々がどこに焦点を当てているかの違いしかない，と言えます．

もちろん機械学習を利用すると，膨大な説明変数の様々な組み合わせに応じた目的変数の予測値が得られるため，その中身にある説明変数の影響力について人間が解釈できなくなることは現実的にあり得ます．ただしだからと言って全く異なる理論に基づいて推定が行われているわけではなく，多くの機械学習モデルはこれまでに学んできた推定の手続きと同様の理論に基づいています．機械学習の書籍でも，基本的な手法としてロジスティック回帰モデルが冒頭で紹介されることが多いのですが，これは統計学の教科書でも良く扱う，目的変数が二値変数の場合の回帰モデルです．このように，前提とする条件によって意見が異なるところもあると思いますが，基本的には機械学習は統計学の概念の中に納まるものと考えて良いと思います．