予測に必要な
統計の基礎知識

北海道　DX人材育成研究会

目次

- 統計を学ぶ意義、目標（ゴール）について
- データの代表値
- データの種類
- 主要なグラフとその説明
- データの集計
- データの散らばり
- 相関関係
- 混同行列

統計を学ぶ意義、
目標（ゴール）について

統計を学ぶにあたって

統計学は、あるデータ群に対し、性質を調べたり、未来のデータを推測するための学問です。
つまり、統計とはデータを解析してデータがどのようなものであるかを説明する手法と言えます。

例えば、近年注目の集まるビッグデータも、統計で扱えるデータの1つです。
マーケティング手法や企画案を策定する際には、すでにビッグデータを統計学で分析する方法が一般的になっています。
また、自社アンケートなどの結果を統計的に分析することで、顧客のニーズを把握できます。

営業や提案のプレゼンでは、製品・サービスを勧める際の根拠として統計を示すことがあります。
また、生産過程において、商品の品質管理のために統計を取ることも一般的です。
さらに、会社の経営判断や投資の予想においても統計学が重視されています。

このように、ビジネスで統計が活用されているシーンは少なくありません。
予測の話を抜きにしても、社会人として統計を学ぶメリットは大きいのではないでしょうか。

統計を学ぶ意義、目標（ゴール）について

pythonやAIによる予測を学習していく上で、先に統計のみを学ぶ意義について簡単に説明します。

統計は上記を学習する流れの中でも学ぶことができますが、それだけでは理解を深めるには厳しい場合もあります。

例えば、ヒストグラムの作成にあたり、実際はpython（matplotlib）が自動的に描いてくれるため、その過程である度数分布表については自分で作る必要はありません。

しかし、pythonの中で何が行われているのかについて知識があった方が、今後の予測の学習の理解が深まると考えられます。

ここでは、統計学を通して、実際にpythonの中で何が行われているのか理解し、今後の学習の理解を深めることを目標に学習していきましょう。

データの代表値

代表値

データ全体の特徴を代表して表すような値のこと。

主に下記の3つがある。

- **平均値**
- **中央値**
- **最頻値**

通常は、最も使い勝手がよいため平均値を用いる。

しかし、極端な外れ値があり、それに平均値が引っ張られる場合は、平均値の代わりに中央値を使った方がよいこともある。

平均値と外れ値

平均値

- データの値の合計を、データ数で割ることで求められる値。
- 外れ値の影響を受ける 。

外れ値

- 測定された値の中で他のデータとかけ離れている値。
- 他のデータの分布とは明らかに異なる場所に数値が出現したりする値。

販売店	価格
A	103
B	102
C	98
D	108
E	105
F	110
G	105
H	97
I	100

平均値

合計÷販売店数＝

103.11

外れ値があると…

販売店	価格
A	103
B	102
C	98
D	108
E	105
F	110
G	105
H	97
I	99999→**外れ値**

平均値

合計÷販売店数＝

11,939.00

最頻値・中央値

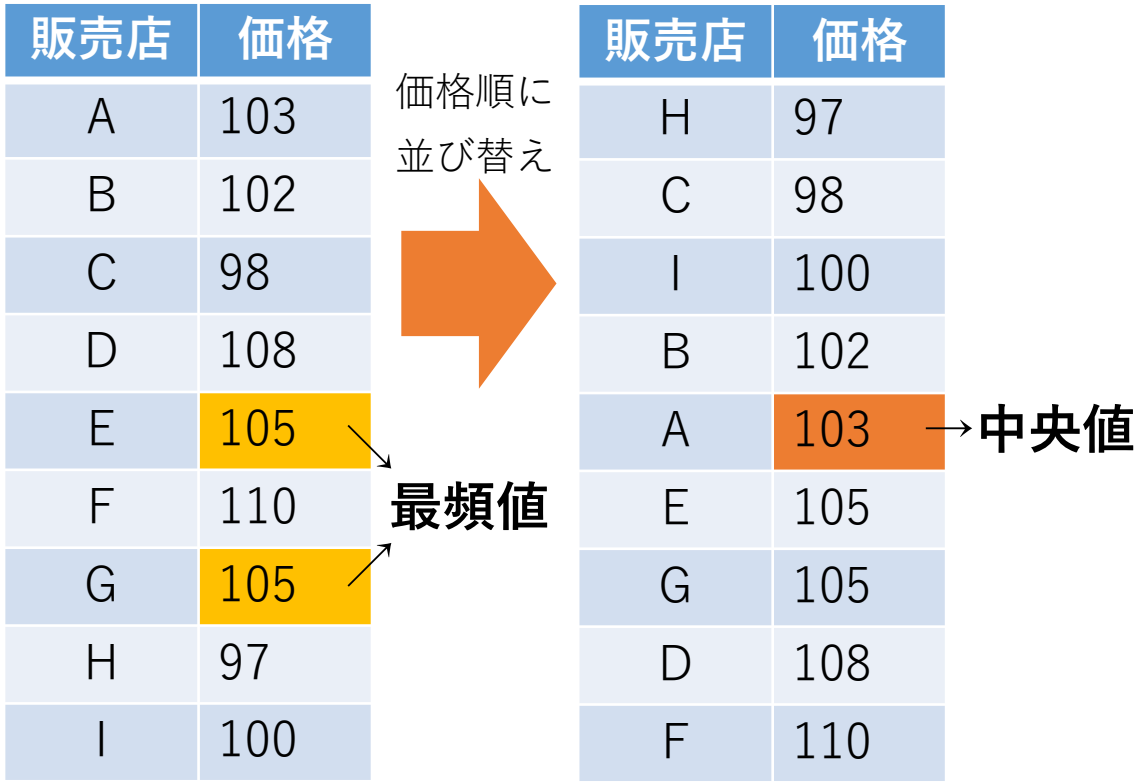
最頻値

- データの出現率が最大の値。

中央値

- 全てのデータを順に並べた時に真ん中に来る値。
- データ数が偶数の場合は、n番目のケースとn+1番目のケースの値を足して2で割った値となる。

- 歪んだ分布や外れ値が多い分布では中央値を用いることが多い。



データの種類

量的変数

計算できる値のこと。

- 比例尺度　四則演算が全て成り立つ（売上など）
- 間隔尺度　足し算・引き算のみ成り立つ（偏差値など）

Pythonの主要な型でいうと・・・

- int型（整数値）
- float型（小数点を含む実数値）
- datetime型（日付や時間の値）

量的変数の例

・気温（10℃、25℃）
・湿度（30%、50%）
・値段（110円、500円）

質的変数

計算できない値のこと。

- 順序尺度　値のあいだに順序がある（アンケートの評価など）
- 名義尺度　値は、他の値との区別があるだけ（天気、性別など）

Pythonの主要な型でいうと・・・

- bool型（TrueかFalseかの真偽値）
- str型（文章・テキストを形成する文字列値）
- int型（文字として扱う場合）

※例えば「学年」というカラムに　1, 2, 2, 3, 1, 3, といった値が入っていた場合は、int型で取り込まれる

質的変数の例

・天気（晴れ、曇り）
・背番号（44番、100番）
・好物（寿司、ステーキ）

テキスト：量的変数と質的変数

【問題】下記の変数は量的変数、質的変数のどちらか。

問題
好きな色
身長
順位
部屋の数
性別
車のナンバー
名前
くじ引きの結果
学年

具体的な変数で
考えてみると…

具体例
赤、青、緑
160cm、175cm、180cm
1位、2位、10位
3部屋、4部屋、5部屋
男、女
7777、0001
太郎、花子、一郎
１等、2等、はずれ
1年生、2年生、3年生

次ページに
解答があるので、
答えを考えてから
進んでください！

解答：量的変数と質的変数

【問題】下記の変数は量的変数、質的変数のどちらか。

問題	具体例	答え
好きな色	赤、青、緑	質的変数
身長	160cm、175cm、180cm	量的変数
順位	1位、2位、10位	**質的変数**
部屋の数	3部屋、4部屋、5部屋	量的変数
性別	男、女	質的変数
車のナンバー	7777、0001	**質的変数**
名前	太郎、花子、一郎	質的変数
くじ引きの結果	１等、2等、はずれ	**質的変数**
学年	1年生、2年生、3年生	**質的変数**

具体例が数字でも、
質的変数となる場合がある。

“数値”と“数字”の違いについて

int型,float型（数字）でも量的変数（数値扱い）ではなく
質的変数（文字扱い）になる場合がある。

例えば学年のクラス分けで1組2組と分ける場合、
Excelの表にすると「組」の列の値には 1や2 が入るが、
合計で1組＋2組＝3組という計算には意味がない。

	組	人数
	1	35
	2	34
合計	3	69

計算が成りたたない

量的変数として扱うには数値として計算が成り立つ必要があるため、
量的変数として扱えない。
よって、量的変数ではなく質的変数となる。

解答：量的変数と質的変数

【問題】下記の変数は量的変数、質的変数のどちらか。

問題	具体例	答え	数字が含まれる場合、計算可能か（数値として扱えるか）
好きな色	赤、青、緑	質的変数	
身長	160cm、175cm、180cm	量的変数	160cm×1.1＝176cmとして成立する。
順位	1位、2位、10位	**質的変数**	**1位+2位の計算結果に意味はない。**
部屋の数	3部屋、4部屋、5部屋	量的変数	3部屋+4部屋＝7部屋として成立する。
性別	男、女	質的変数	
車のナンバー	7777、0001	**質的変数**	**7777＋0001の計算結果に意味はない。**
名前	太郎、花子、一郎	質的変数	
くじ引きの結果	１等、2等、はずれ	**質的変数**	**1等+2等の計算結果に意味はない。**
学年	1年生、2年生、3年生	**質的変数**	**1年生+2年生の計算結果に意味はない。**

主要なグラフとその説明

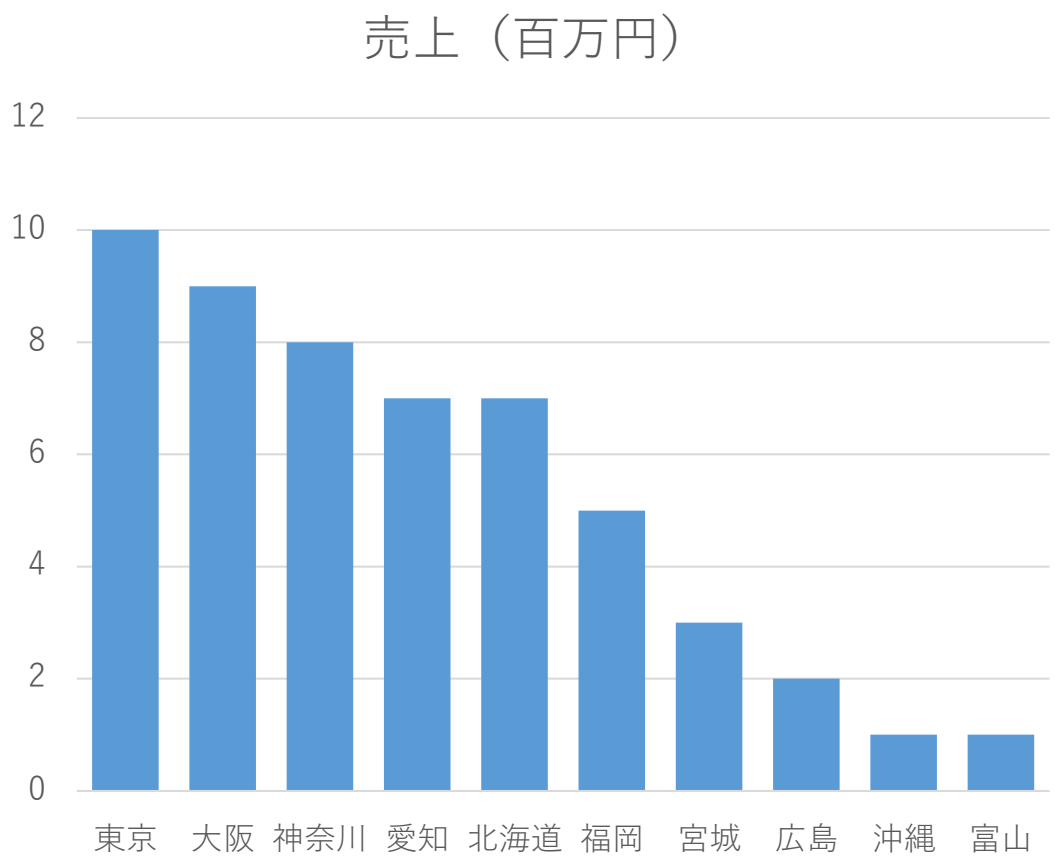
棒グラフ

縦軸にデータ量をとり、棒の高さでデータの大小を表したグラフ。

値の高い項目や低い項目を判別するのに有効で、データの大小を比較するのに適している。

縦軸：量的変数

横軸：質的変数（量的変数の場合もあり）



折れ線グラフ

横軸に年や月といった時間を、縦軸にデータ量をとり、それぞれのデータを折れ線で結んだグラフ。

主に時系列などの連続的変化を捉えるときに使用する。

データの増減を見るのに適しており、グラフの傾きから、変化の大きさを読み取ることができる。

縦軸：量的変数

横軸：日付型（量的変数）

データの集計

単純集計

アンケート等の設問ごとの集計のこと。

回答者	性別	評価
1	男	良い
2	女	普通
3	男	普通
⋮	⋮	⋮
50	女	普通

「性別」を単純集計

性別	人数	比率（%）
男	20	40
女	30	60

「評価」を単純集計

評価	人数	比率（%）
良い	20	40
普通	20	40
悪い	10	20

クロス集計

単純集計をかけ合わせた集計のこと。

・質的変数同士の関係性を検証するのに有効。

性別	人数	比率（%）
男	20	40
女	30	60

評価	人数	比率（%）
良い	20	40
普通	20	40
悪い	10	20

クロス集計

	良い	普通	悪い	合計
男	12	2	6	20
女	8	18	4	30
合計	20	20	10	50

度数分布表

データを任意の範囲ごとに分割し、
それぞれの範囲内に存在するデータ数を表にまとめたもの

度数分布表

階級	階級値	度数	相対度数	累積相対度数
0以上20未満	10	28	0.096551724	0.096551724
20以上40未満	30	71	0.244827586	0.34137931
40以上60未満	50	86	0.296551724	0.637931034
60以上80未満	70	67	0.231034483	0.868965517
80以上100未満	90	38	0.131034483	1

階級：度数を集計するための区間。

階級値：その階級を代表する値、階級の中央値。

度数：各階級に含まれるデータ数。

相対度数：各階級の度数が全体に占める割合。
度数÷総データ数

累積相対度数：その階級までの相対度数の全ての和。
（累積和）

度数分布表の作り方

78	26	84	90	47	6	90	6	83	43	77	57
69	34	97	82	70	89	26	52	69	69	99	5
57	15	31	5	99	67	37	99	42	23	68	79
55	58	6	65	80	61	67	8	44	14	2	85
36	85	42	2	28	13	30	15	71	23	25	76
61	50	5	70	22	78	14	35	89	16	33	72
26	38	61	73	77	22	60	63	44	74	33	48
35	59	56	64	33	30	75	53	35	64	25	60
31	41	75	45	35	65	73	32	65	64	37	66
36	25	36	69	47	31	77	51	64	62	58	63
70	72	72	63	24	53	25	36	75	78	28	52
44	41	57	44	47	43	47	48	56	49	50	56
52	51	50	52	47	57	54	53	51	48	53	47
49	41	56	47	53	49	42	59	53	51	59	59
48	46	41	54	44	56	48	53	44	46	42	44
90	37	88	56	79	47	33	93	33	79	9	20
61	97	38	24	18	31	6	39	90	84	9	15
18	36	53	27	91	91	49	69	49	91	74	38
3	73	97	79	32	99	37	71	34	77	18	76
70	28	91	56	18	44	70	93	98	80	61	21
98	38	62	69	53	26	53	23	71	25	92	27
60	80	3	96	87	88	92	35	98	29	35	71
16	29	35	25	43	52	87	58	25	3	33	28
62	90	80	78	99	31	66	70	77	33	30	40
28	45	60	11	90	53	34	50	25	66	73	38

元データ：的当てゲームのスコア

小さい順に並べ替え、
階級ごとに区切る

2	18	27	34	40	47	52	56	63	70	78	90
2	18	28	34	41	47	52	57	63	70	78	90
3	18	28	34	41	47	52	57	64	71	78	90
3	20	28	35	41	47	52	57	64	71	78	90
3	21	28	35	41	47	52	57	64	71	79	90
5	22	28	35	42	47	53	58	64	71	79	91
5	22	29	35	42	47	53	58	65	72	79	91
5	23	29	35	42	48	53	58	65	72	79	91
6	23	30	35	42	48	53	59	65	72	80	91
6	23	30	35	43	48	53	59	66	73	80	92
6	24	30	36	43	48	53	59	66	73	80	92
6	24	31	36	43	48	53	59	66	73	80	93
8	25	31	36	44	49	53	60	67	73	82	93
9	25	31	36	44	49	53	60	67	74	83	96
9	25	31	36	44	49	53	60	68	74	84	97
11	25	31	37	44	49	53	60	69	75	84	97
13	25	32	37	44	49	54	61	69	75	85	97
14	25	32	37	44	50	54	61	69	75	85	98
14	25	33	37	44	50	55	61	69	76	87	98
15	25	33	38	44	50	56	61	69	76	87	98
15	26	33	38	45	50	56	61	69	77	88	99
15	26	33	38	45	51	56	62	70	77	88	99
16	26	33	38	46	51	56	62	70	77	89	99
16	26	33	38	46	51	56	62	70	77	89	99
18	27	33	39	47	51	56	63	70	77	90	99

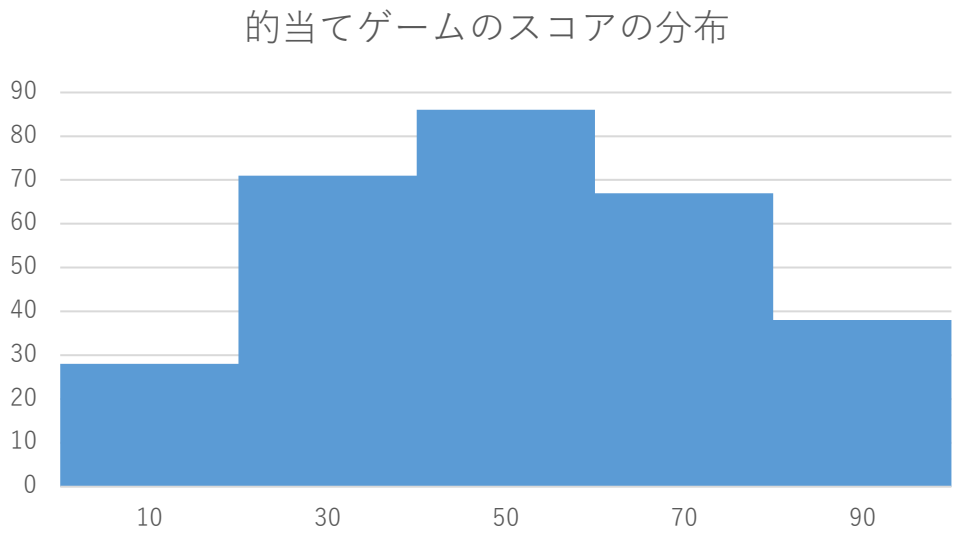
各項目を計算

階級	階級値	度数	相対度数	累積相対度数
0以上20未満	10	28	0.096551724	0.096551724
20以上40未満	30	71	0.244827586	0.34137931
40以上60未満	50	86	0.296551724	0.637931034
60以上80未満	70	67	0.231034483	0.868965517
80以上100未満	90	38	0.131034483	1

ヒストグラム

- 横軸に階級、縦軸に度数をとり、各区間の個数や数値のばらつきを表現するグラフ。
- 分布の可視化に適している。

- 棒グラフと似ているが、棒グラフは別カテゴリーのデータ同士を比較するため、ヒストグラムは同種のデータ内でどの階級が多くどの階級が少ないかを把握ために用いられることが多い。



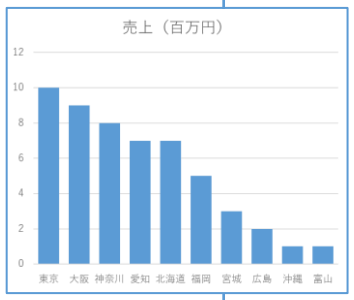
縦軸：量的変数の度数

横軸：量的変数の階級

ヒストグラムと他のグラフとの違い

- 棒グラフや折れ線グラフ等、ヒストグラム以外のグラフでは縦軸と横軸に2種類のデータを用いる。
- 一方でヒストグラムでは、縦軸も横軸も同一のデータをもとに算出されるため、用いるデータは1種類のみとなる。

棒グラフ



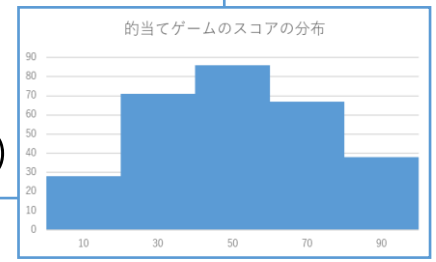
縦軸：売上額
横軸：都道府県名

【使うデータ】
・売上額のデータ（量的変数）
・都道府県のデータ（質的変数）

ヒストグラム

縦軸：度数
（的当てゲームのスコアのデータから算出）
横軸：階級
（的当てゲームスコアのデータから算出）

【使うデータ】
・スコアのデータ（量的変数）



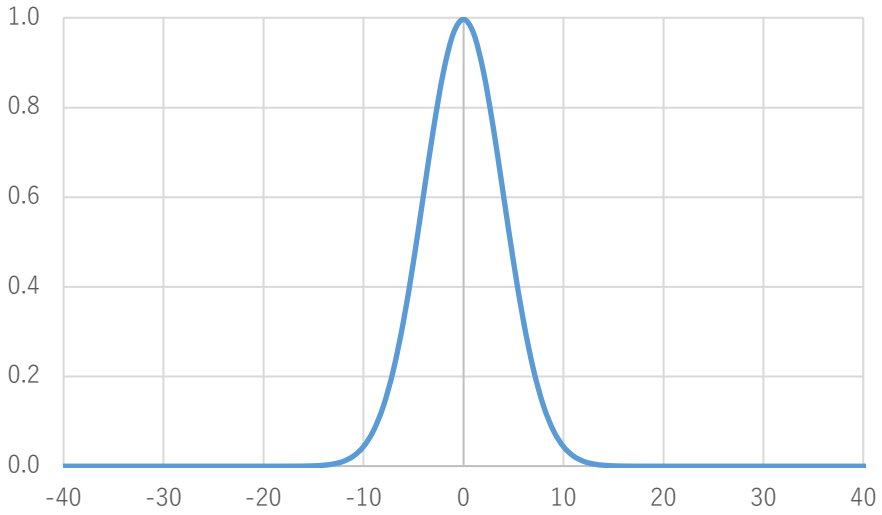
データの散らばり

分散

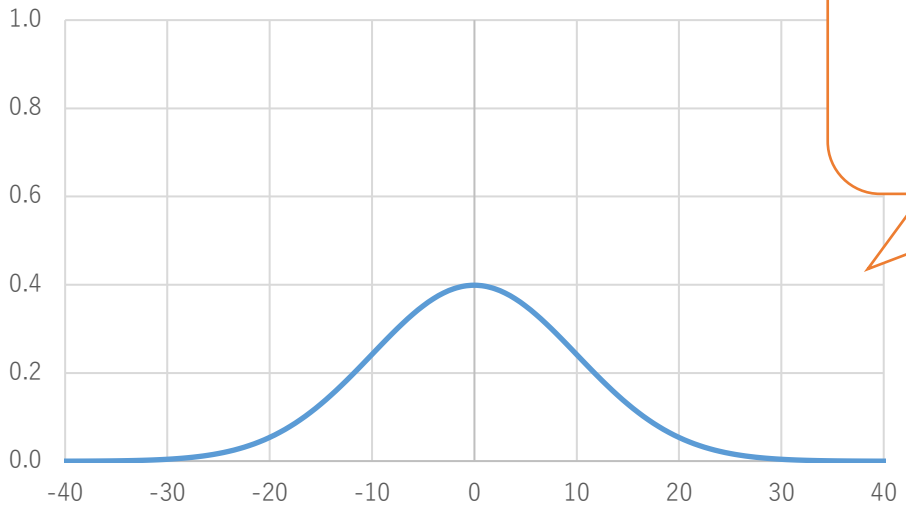
数値データのばらつき具合を表すための指標。

平均値と個々のデータの差の２乗の平均を求めることによって計算される。
平均値から離れた値をとるデータが多ければ多いほど、大きくなる。

平均0、分散16



平均0、分散100



平均が同じデータでも
分散が異なると
分布の見た目が
大きく変わる。

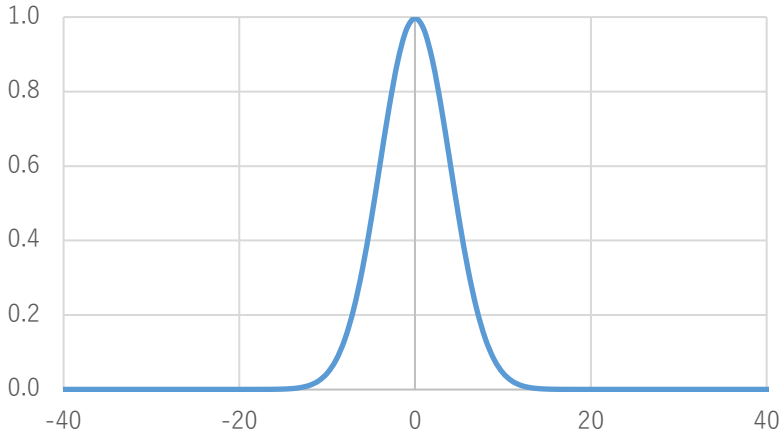
標準偏差

分散の正の平方根のこと。

分散は2乗の値になり、そのままでは平均と単位を揃えることができないため、標準偏差が用いられる。

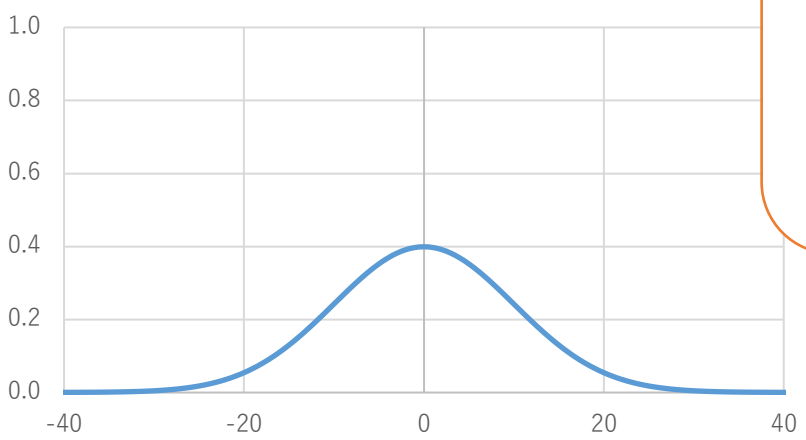
平均0、分散16

標準偏差4



平均0、分散100

標準偏差10



平均が〇〇cmに対し、
分散は〇〇cm²となってしまい
このままでは
比較ができないので
標準偏差を計算し
単位を〇〇cmに戻す。

標準化

スケールが異なるデータ同士を比較する方法のこと。

偏差（各データと平均値の差）を標準偏差で割ることで求められる。
データを標準化すると、標準化したデータの平均は0に、分散と標準偏差は1になる。

例）テストの偏差値
※偏差値とはデータを標準化し、平均を50、標準偏差を10になるように変換した指標

$$偏差値 = \frac{得点 - 平均点}{標準偏差} \times 10 + 50$$

標準化

	国語の点数	数学の点数
生徒A	60	50
生徒B	85	65
生徒C	90	80
平均	78	65
Cの偏差値	59	62

Cの点数は
国語の方が高いが
偏差値を見ることで
成績は
数学の方が高い
ことが判った

相関関係

相関関係

一方が他方との関係を離れては意味をなさないようなものの間の関係。
どちらかの事象がもう片方の事象の直接的な原因かどうかは不明。

相関係数
あるデータ同士の関係性を数値化したもので、「-1〜+1」の間で表現する手法。

1に近いほど正の相関が強い、-1に近いほど負の相関が強い、0に近いほど相関がないことを意味する。

散布図

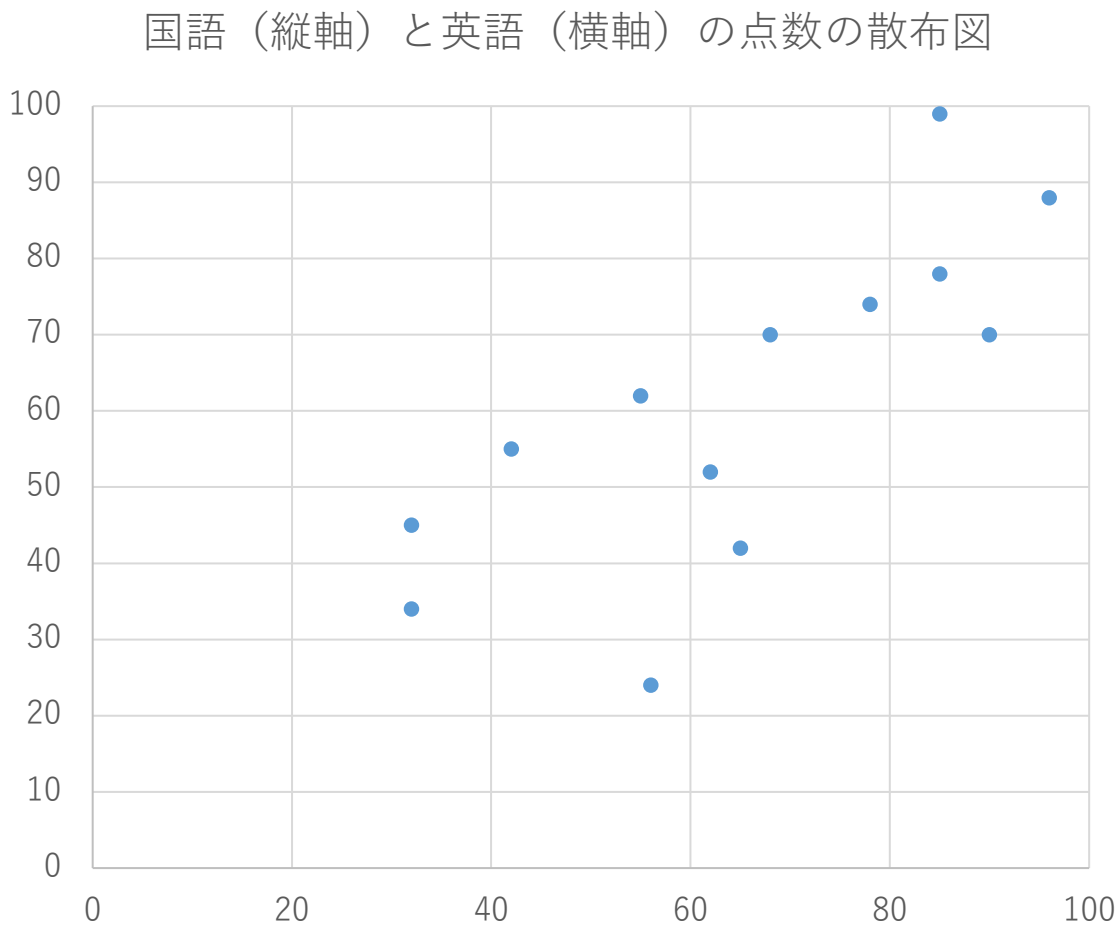
横軸と縦軸にそれぞれ別の量をとり、データが当てはまるところに点を打って示す（プロットする）グラフ。

2つの量の間に関係があるかを示すだけであり、どちらかが原因となってもう一方が起こるといった、後述の因果関係を示すものではない。

縦軸：量的変数

横軸：量的変数

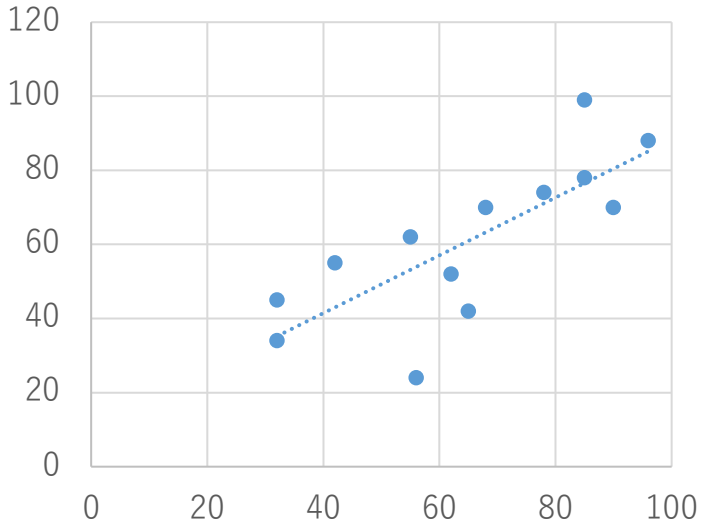
実際のデータ分析では、
相関係数のみを見て
関係性を判断せずに、
散布図を用いて目視で確認
することが非常に重要。



正の相関・相関なし・負の相関

正の相関

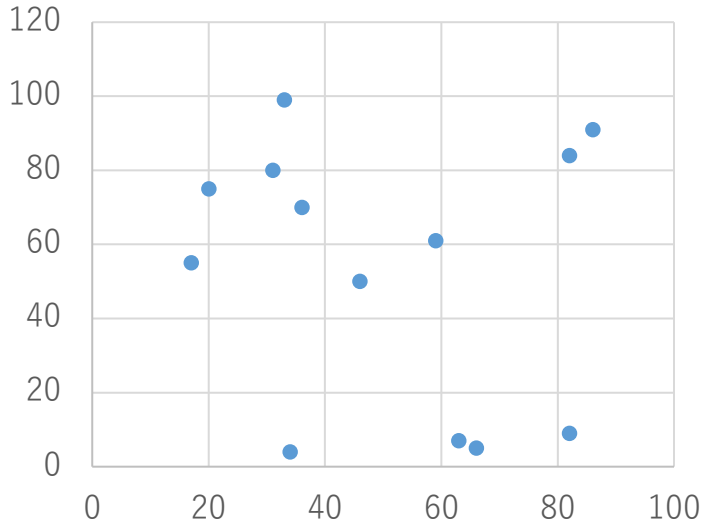
横軸の値（x）が増加すると縦軸の値（y）も増加するという関係のこと。



相関係数：0.769

相関関係なし

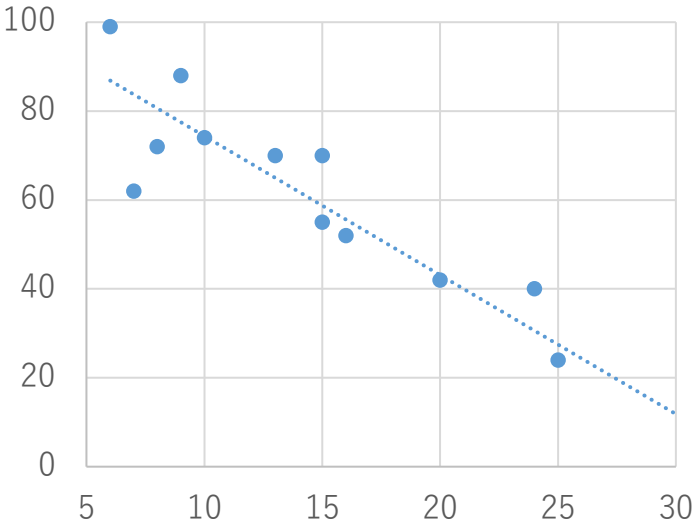
xが増加してもyに増減の傾向が見られない関係のこと。



相関係数：0.178

負の相関

xが増加するとyが減少するという関係のこと。

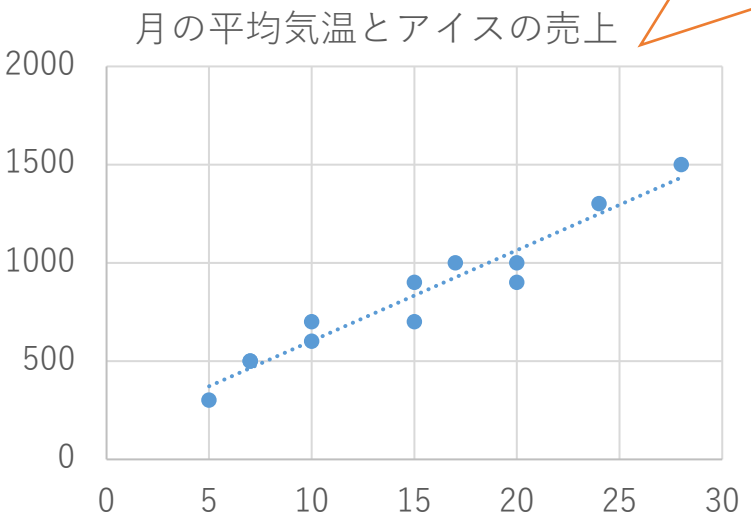


相関係数：-0.932

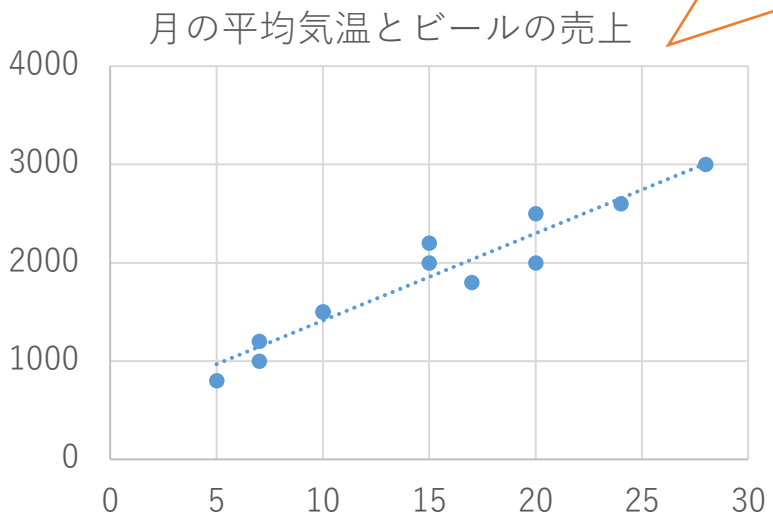
因果関係

2つの事象のうち一方が原因となって他方の結果があるという関係のこと。

月の平均気温が高い
↓
アイスの売上が高くなる

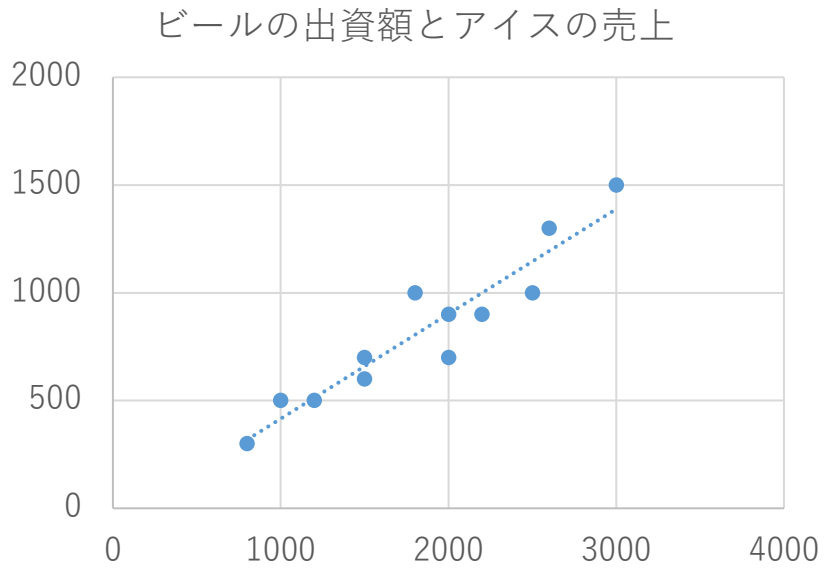


月の平均気温が高い
↓
ビールの売上が高くなる



疑似相関

因果関係がないのにあたかも因果関係があるように見えること。



ビールの売上が高い

↓

アイスの売上が高くなる

ビールの売上もアイスの売上も

気温の高さが原因で変化しており

どちらかが原因というわけではない。

相関関係があっても
因果関係があるとは限らない。

混同行列

混同行列

あるデータを分類したときに、その正解・不正解の数を
整理しておく表のこと。

例）PCR検査を受けた際の検査結果と、実際にコロナウイルスに罹患していたかどうか。

	PCR検査で陽性反応	PCR検査で陰性反応
実際に コロナウイルスに 罹患している	**真陽性(True Positive)** ・陽性と判定したものが正解（真）である	**偽陰性(False Negative)** ・陰性と判定したものが不正解（偽）である
実際は コロナウイルスに 罹患していない	**偽陽性(False Positive)** ・陽性と判定したものが不正解（偽）である	**真陰性(True Negative)** ・陰性と判定したものが正解（真）である

評価指標

AIの性能を測定するための指標のこと。
どの評価指標が最善なのかは目的によって異なる。

下図は、PCR検査とコロナの罹患状況をもとに、混同行列の各項目の人数をまとめたもの。
この結果をもとに各評価指標を算出していく。

例）100人がPCR検査を受けた際の検査結果と、実際にコロナに罹患していたかどうか。

	PCR検査で陽性反応	PCR検査で陰性反応	
コロナウイルスに罹患している	**真陽性(TP)** 30人	**偽陰性(FN)** 15人	合計45人
コロナウイルスに罹患していない	**偽陽性(FP)** 5人	**真陰性(TN)** 50人	合計55人
	合計35人	合計65人	全体100人

正解率

全データのうち、正解したデータ数の割合。

正解した人数(TP30+FP50)÷全体の人数(100)＝**正解率80%**

真陽性(TP) 30人	偽陰性(FN) 15人
偽陽性(FP) 5人	**真陰性(TN)** 50人

正解した人数80人

全体の人数 100人

真陽性(TP) 30人	**偽陰性(FN)** 15人
偽陽性(FP) 5人	**真陰性(TN)** 50人

メリット
最もシンプルで分かりやすい。

デメリット
クラスごとの評価データ数が著しく異なると不適切。

100人中1人だけコロナ患者がいて検出したい場合に、
実際には検査をせず「全員陰性」とするだけでも
正解率は99%になってしまう。

再現率

実際のコロナ患者のうち、検査で陽性となった人の割合。

検査で陽性となったコロナ患者数(TP30)÷実際のコロナ患者数 (TP30+FN15)＝**再現率66.67%**

真陽性(TP) 30人	偽陰性(FN) 15人
偽陽性(FP) 5人	

検査で陽性となったコロナ患者数30人

実際のコロナ患者数 45人

真陽性(TP) 30人	**偽陰性(FN)** 15人
偽陽性(FP) 5人	

メリット
取りこぼし（FN）を減らすことが目的の場合、
目的に合った学習ができているかを確認できる。

FNが減れば減るほど、再現率TP/(TP+FN)の式の値は大きくなるため。

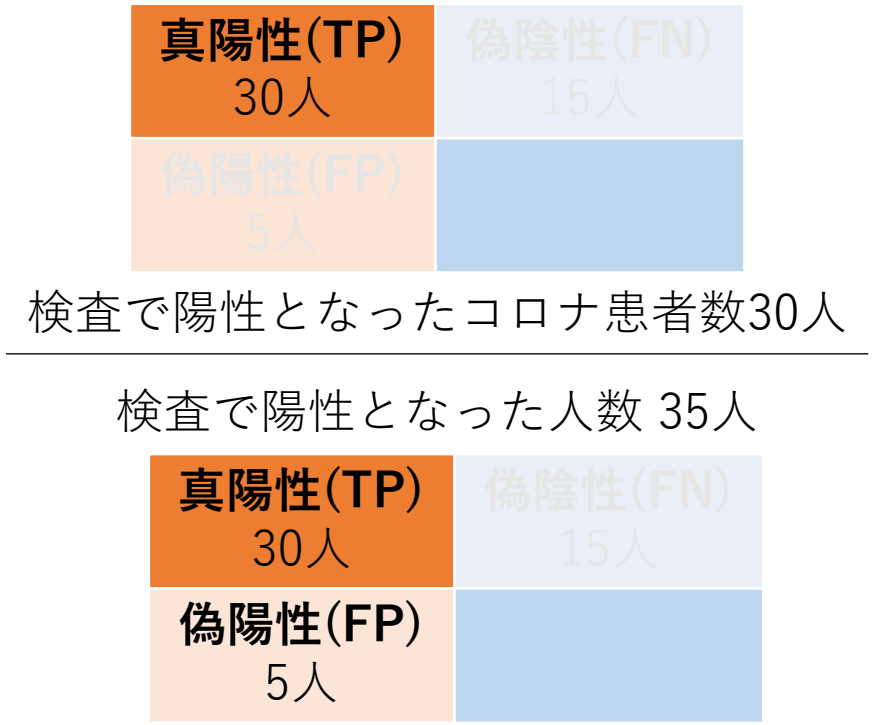
デメリット
誤検知（FP）の善し悪しを確認できない。

再現率 TP/(TP+FN) の式にはFPが含まれないので、誤検知がいくら大きくなっても、再現率には影響しないため。

適合率

検査で陽性と判別した人のうち、実際のコロナ患者の割合。

検査で陽性となったコロナ患者数(TP30)÷検査で陽性となった人数 (TP30+FP5)＝**適合率85.71%**



メリット
誤検知（FP）を減らすことが目的の場合、
目的に合った学習ができているかを確認できる。

FPが減れば減るほど、適合率TP/(TP+FP)の式の値は大きくなるため。

デメリット
取りこぼし（FN）の善し悪しを確認できない。

適合率 TP/(TP+FP) の式にはFNが含まれないので、取りこぼしがいくら大きくなっても、適合率には影響しないため。

F値

再現率と適合率の調和平均(逆数の平均の逆数)。

再現率$\frac{30}{45}$、適合率$\frac{30}{35}$のため　　$1 \div \{ (\frac{45}{30}+\frac{35}{30}) \div 2\}$ ＝**F値75%**

メリット
取りこぼし（FN）、誤検知（FP）を均等に判断できる。

再現率と適合率はトレードオフの関係のため、(再現率を上げると適合率が下がり適合率を上げると再現率が下がる)再現率と適合率を一方に偏らせずに均等に評価したい場合に使われる。

デメリット
数値の解釈が難しくなる。

以上で「予測に必要な統計の基礎知識」は終了です。

お疲れ様でした！

関連情報の紹介

総務省統計局　なるほど統計学園
https://www.stat.go.jp/naruhodo/index.html

統計WEB　統計学の時間
https://bellcurve.jp/statistics/course/