Principia I

統計講座

データを整理し
客観的な指標へ

数字は嘘をつかないが
解釈は人を惑わす

分散と標準偏差と相関係数

神奈川県立横須賀高等学校
SSH推進委員会　作成

SSH
Super Science High school

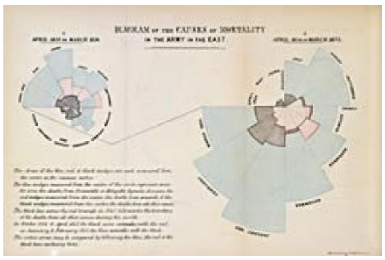
【統計学は、「真実をあぶり出す」ための道具】

フロレンス＝ナイチンゲール（1820-1910）

ナイチンゲールは、上流階級の家庭に生まれ、歴史・語学・音楽など高いレベルの教育を受けました。また、若いころから「近代統計学の父」ベルギー人アドルフ＝ケトレー（1796-1874）を尊敬し、数学や統計に強い興味を持ち、すぐれた家庭教師について勉強したと言われています。

ナイチンゲールは、1854年、イギリス政府によって看護師団のリーダーとしてクリミア戦争（ロシアとトルコの間の戦争で、イギリスはフランスとともにトルコに味方してロシアと戦った）に派遣されると野戦病院で骨身をけずって看護活動にはげみ、病院内の衛生状態を改善することで傷病兵の死亡率を劇的に引き下げました。

統計に関する知識を存分に使ってイギリス軍の戦死者・傷病者に関する非常にたくさんのデータを分析したナイチンゲールは、かれらの多くが戦いで受けた傷そのものではなく、傷を負った後の治療や病院の衛生状態が十分でないことが原因で死亡したことを明らかにしたのです。



クリミア戦争における死因分析を表したグラフ

ナイチンゲールが取りまとめた報告は、統計になじみのうすい国会議員や役人にも分かりやすいように、当時としてはめずらしかったグラフを用いて、目で見てわかりやすいプレゼンテーションを工夫しました。今も「＿＿＿＿＿＿＿＿」と呼ばれる円グラフの一種はこの過程で彼女によって考え出されたものです。

1860 年には、ケトレーが立ち上げた国際統計会議のロンドン大会に出席し、衛生統計の統一した基準を提案しました。統計のとり方がバラバラであっては、分析に支障を来し、医療技術の向上にもつながらないと考えたのです。提案は会議の分科会で討議され、各国政府に送付する決議が採りあげられました。

国をまたいで統計調査の形式や集計方法をそろえることは、今日でも簡単なことではありません。ナイチンゲールには現場の経験と統計の知識に裏付けられたゆるぎない信念があったのでしょう。

ナイチンゲールは、「英国陸軍の衛生改革への統計的手法の活用」という業績が認められて、1858年、女性として初めて王立統計学会のフェローに推薦され、そして1874年には、米国統計学会の名誉会員にも選ばれています。

「白衣の天使」ナイチンゲールは、祖国イギリスでは統計学の草分けとして今も人々の記おくに刻まれています。

【参考文献】
丸山健夫（2008）
『ナイチンゲールは統計学者だった！―統計の人物と歴史の物語―』日科技連出版社
多尾清子（1991）
『統計学者としてのナイチンゲール』医学書院

宇宙統計ステーションNARUHODO　HP参照（一部改）

【統計学の全体像】

【１】　　・・・・数学Ⅰ「データの分析」

記述統計学（Descriptive Statistics）とは、一言で言えば「手元にあるデータの性質を、分かりやすく整理・要約して伝えるための手法」のことです。数学Ⅰで学ぶ「データの分析」の内容は、ほぼすべてこの「記述統計学」に該当します。

記述統計学には、大きく分けて3つの表現方法があります。

１．数値でまとめる（代表値と散らばり）

データ全体の特徴を、一つの数値で代表させたり、データの「バラツキ」を数値化する。

①＿＿＿＿＿＿＿＿＿＿

データの「中心」がどこにあるかを示す。

②＿＿＿＿＿＿＿＿＿＿　　（数学Ⅰの「データの分析」で学びます）

データが平均値の周りにどのくらい「散らばって」いるかを示す。

③＿＿＿＿＿＿＿＿＿＿

データを大きさ順に並べて4等分し、データの分布の偏りを見る。

２．グラフや表で可視化する

数値の羅列では読み取れない「データの形」を、図解して一目で分かるようにする。

①＿＿＿＿＿＿＿＿＿＿

どの範囲にどれくらいのデータが集まっているか、視覚化する。

②＿＿＿＿＿＿＿＿＿＿

最小値、最大値、中央値などを一つの図にまとめ、複数グループの比較に適する。

③＿＿＿＿＿＿＿＿＿＿

２つのデータの間の関係性（相関）を視覚化する。

３．記述統計学の目的

記述統計学の最大の目的は、「膨大なデータを、人間が理解できる情報に変換すること」です。例えば、学年全体のテスト結果を一人ずつ眺めても、学年全体の傾向は分かりません。しかし、「平均点（代表値）」を出し、「点数の分布（ヒストグラム）」を描くことで、「平均点付近が多い」、「二極化している」といったデータが持つ意味が見えてきます。

【２】＿＿＿＿＿＿＿＿＿＿　・・・数学B　「統計的な推測」

推測統計学（Inferential Statistics）とは、「抽出されたデータ（標本）を使って、まだ見ぬ全体像（母集団）を推測する手法」のことです。

数学Ⅰ「データの分析（記述統計学）」が「＿＿＿＿＿＿」ものだとすれば、数学B「統計的な推測（推測統計学）」は「＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿」ための学問です。

「未知に挑む」をテーマに課題研究を進める皆さんにとって、推測統計学は最大の武器となるはずです。１年生では、推測統計学の１つである「＿＿＿＿」を少しだけ学びます。本格的な推測統計学は２年生で学びます。

推測統計学には、以下の４つの柱があります。

１．データを確率で捉える

目の前のデータが「どのような確率のルール（分布）に従っているか」を考える。

① 確率変数

サイコロの目やテストの点数のように、値が確率的に決まる変数

② 二項分布

「結果が２通りしかないこと」を繰り返した時に、ある結果が何回起こるかを表す確率分布

２．データを「型」に当てはめる

①＿＿＿＿＿＿

平均値を中心に左右対称な鐘型をした確率分布のことです。平均値・中央値・最頻値がすべて一致するという特徴があります。身長・体重・テストの点数・製品のサイズの誤差など、世の中の多くのデータが自然にこの形に近づくため、統計学において最も重要な分布です。

② 標準化

データの種類が違っても、比較できるように変換すること。

※皆さんご存じの「偏差値」は「標準化」の考えを使っています。標準化によって、そのデータが全体の中でどのくらいの位置にいるのかを判断できるようになります。

３．一部から全体を予想する

① 標本平均

母集団から、ランダム抽出された一部のデータ（標本）の平均値のこと。

② 母平均の推定

標本平均を使って、母集団の平均値（母平均）を予測すること。

③ 95%の信頼区間

95%の確率で母平均を含む区間のこと。（母平均が動くのではなく、区間が動く）

※統計学では標本平均が「170cm」だからといって、母平均も「ぴったり170cm」とは言いません。「母平均は169cm～171cmにあるはず」というように、＿＿＿＿＿＿＿＿＿＿で予想します。

４．仮説検定

「新薬を飲んだら熱が下がった」という結果が、偶然か薬の効果かを判断するのが仮説検定です。「薬に効果はない」という仮説を立てて、それが起きる確率を計算します。その確率が極めて低ければ（５%以下など）「これは偶然ではない、意味がある差だ」と判断します。

① 仮説検定

実験や調査で得られた結果が、単なる「たまたま起きた偶然（誤差）」なのか、それとも「意味のある変化（必然）」なのかを、確率を使って科学的に判断する手法のことです。

② 帰無仮説と対立仮説

帰無仮説：示したい仮説と反対の仮説

対立仮説：示したい仮説

③ ＿＿＿＿＿

どのくらい珍しいことが起きたら偶然ではないと認めるか、というボーダーラインのこと。一般的には５％を使うことが多い。

【データの整理と可視化】

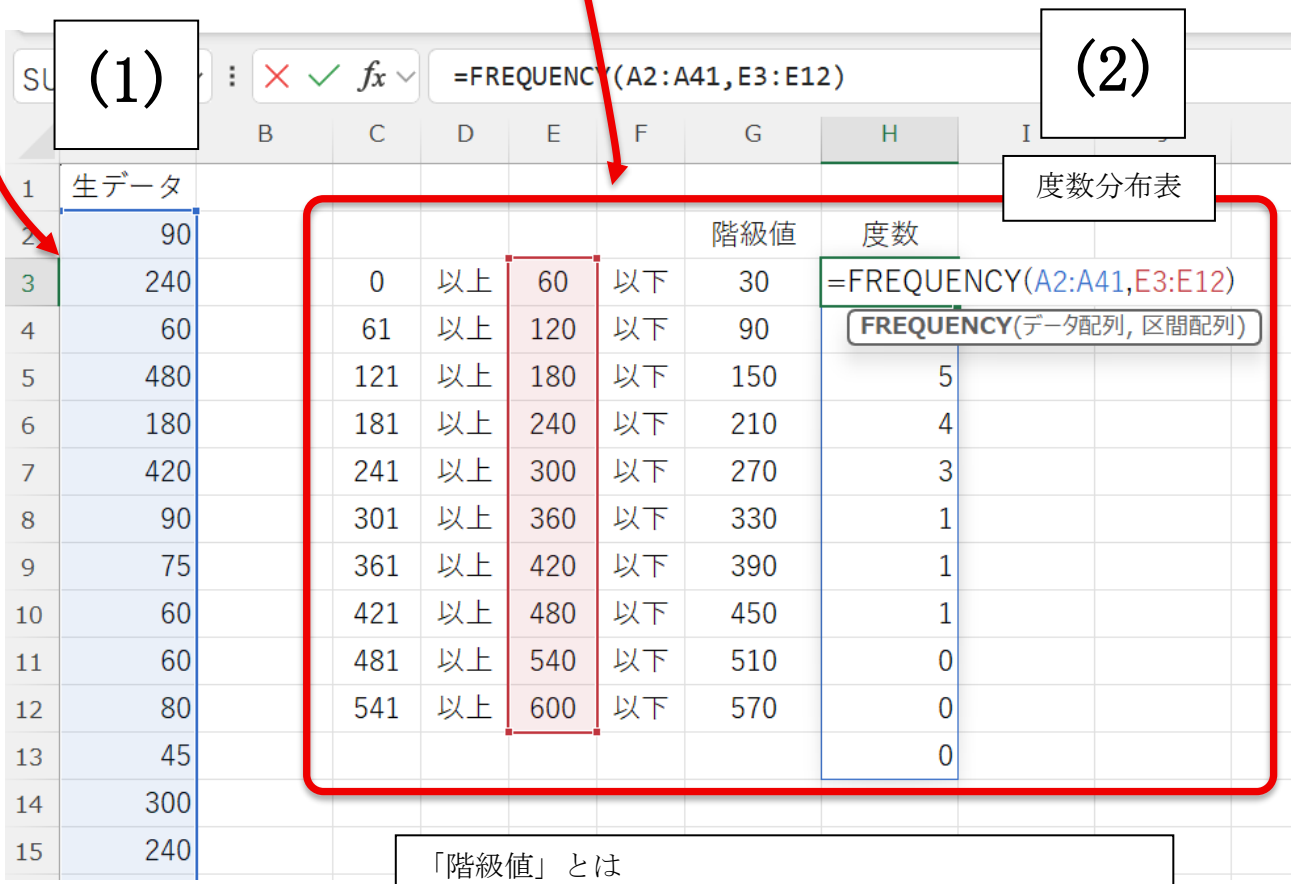
研究を進める過程で得られる加工前の数値を「　　」と呼びます。この生データを詳しく分析することで、一見バラバラな数字の背後に隠された特徴を読み取り、深い考察を行うことが可能になります。今回は、生データから**「データの整理」**を行い、**「度数分布表」**や**「グラフ」**を用いて**「データの可視化」**しましょう。統計の手法を使ってデータの個性を浮き彫りにする、**「データサイエンス」**の基礎を学びます。

【実習１】

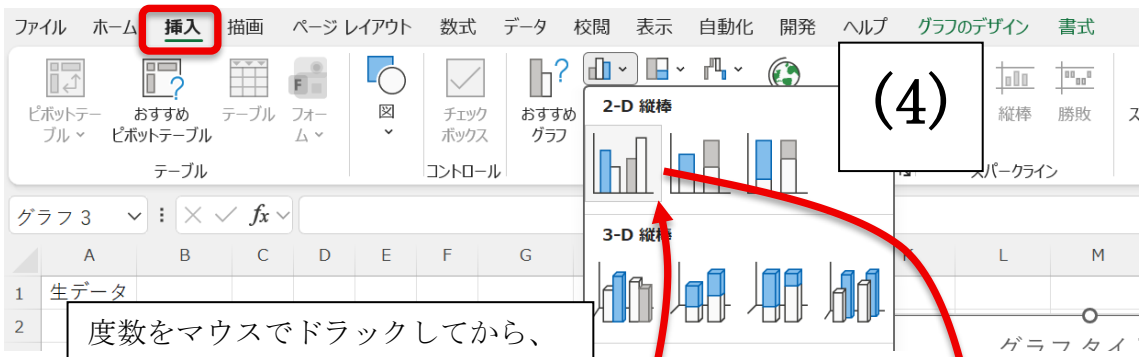
以下は、あるクラスの生徒の「昨日の家庭学習の時間」40人分です。（時間：分）

90	180	60	300	45	120	210	80	300	180
240	420	60	240	150	90	0	360	90	150
60	90	80	100	75	120	80	120	180	60
480	75	45	120	30	200	60	30	270	90

（１）このクラスの学習状況を説明してください。
（２）Excelに、上記の表の値を入力してください。　・**「資料の整理」**
（３）度数分布表を作成してください。　・・・・・・**「データの可視化」**
（４）ヒストグラムを作成してください。　・・・・・・**「データの可視化」**
（５）箱ひげ図を作成してください。　・・・・・・**「データの可視化」**
（６）平均値、最頻値、中央値を求めてください。・・**「データの可視化」**



「階級値」とは
その区間を代表する値（一般的にその階級の中央値）



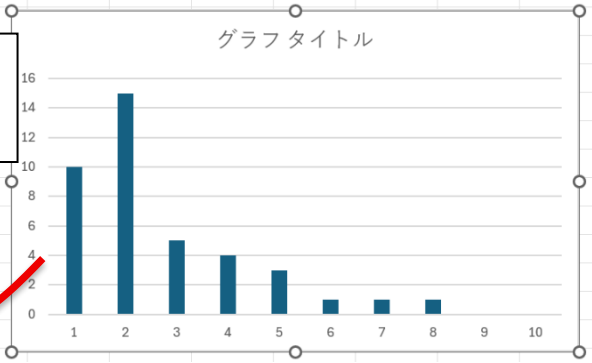
(4)

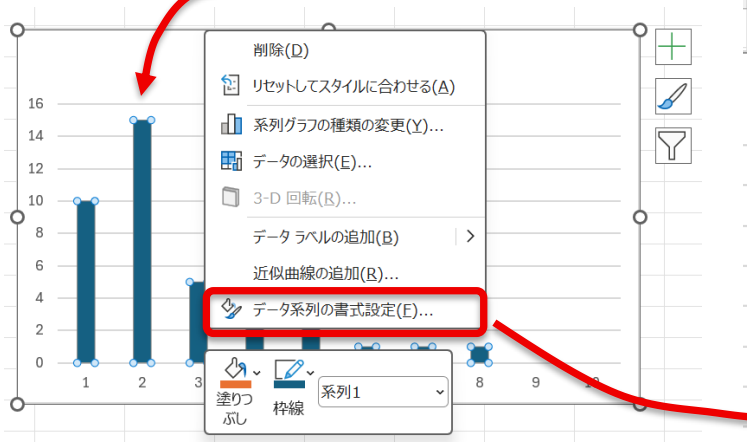
度数をマウスでドラックしてから、
棒グラフ（2-D縦棒）を選ぶ。
（ヒストグラムは使いにくいから）

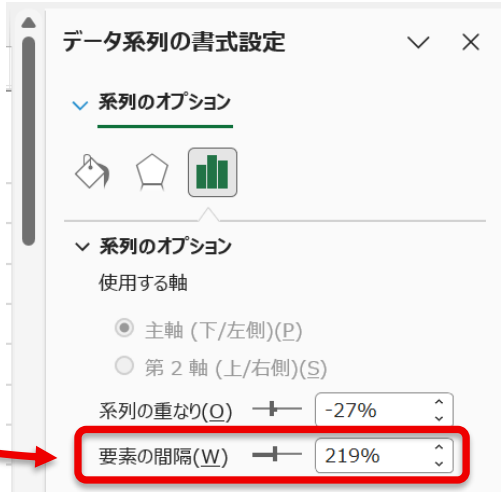
こんな感じのグラフになる。
ヒストグラムにするには、棒と棒の
隙間を無くさないといけないので

				階級値	度数
0	以上	60	以下	30	10
61	以上	120	以下	90	15
121	以上	180	以下	150	5
181	以上	240	以下	210	4
241	以上	300	以下	270	3
301	以上	360	以下	330	1
361	以上	420	以下	390	1
421	以上	480	以下	450	1
481	以上	540	以下	510	0
541	以上	600	以下	570	0

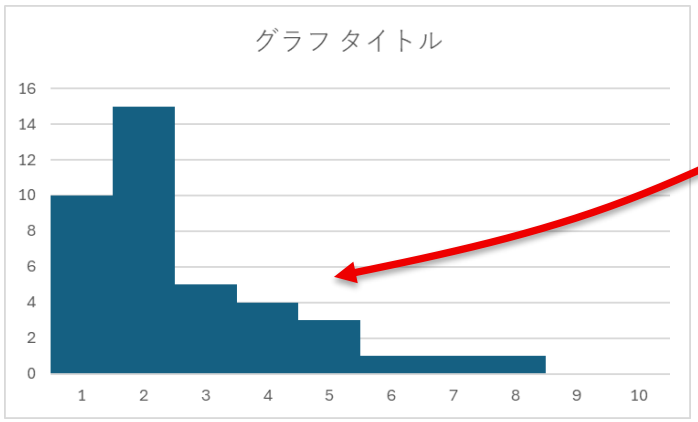
(3)

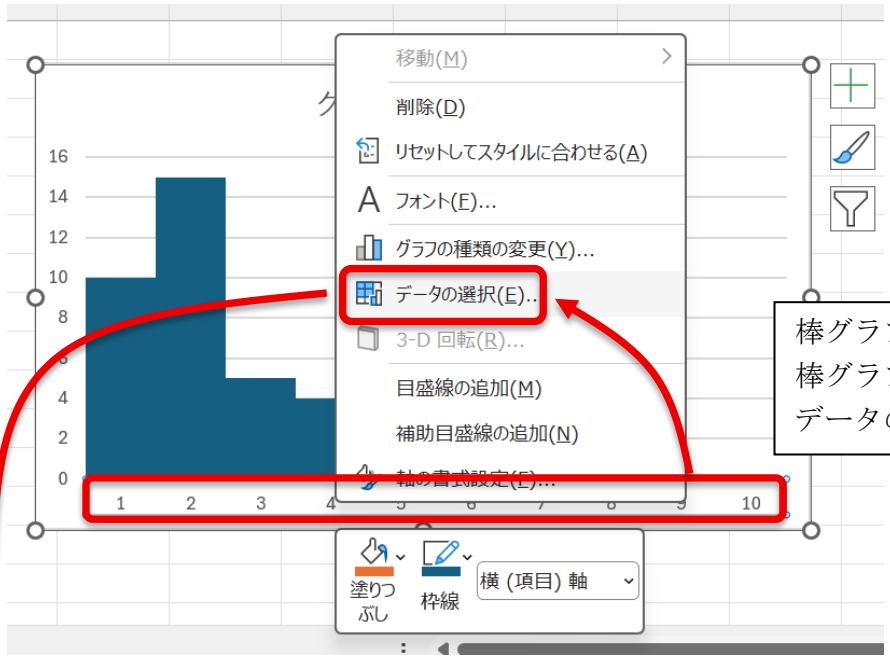




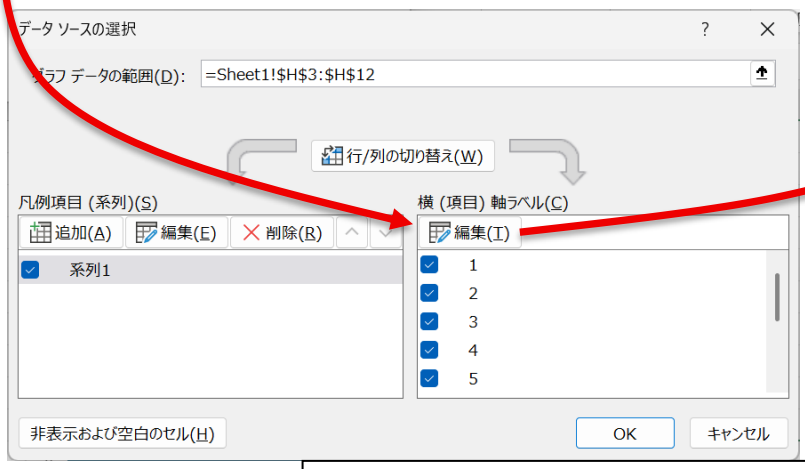


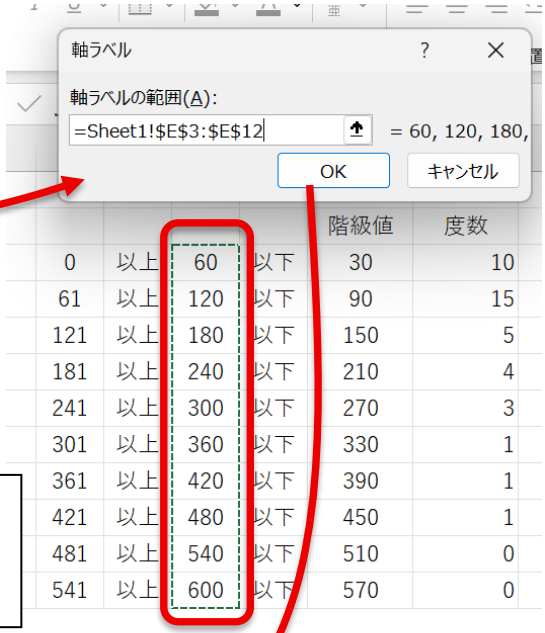
要素の間隔を０％にすると、棒グラフの隙間がなくなります。





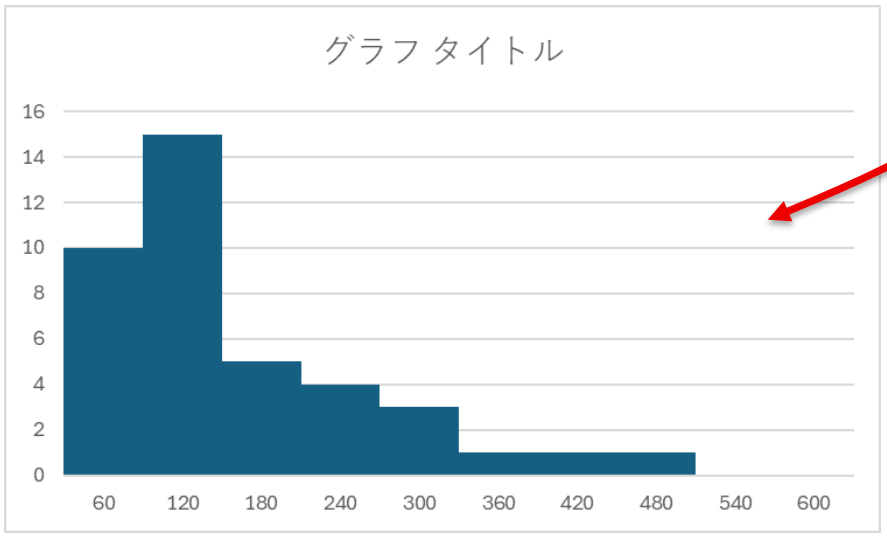
棒グラフの横軸を変えます。
棒グラフの横軸を右クリックして
データの選択をクリックする。

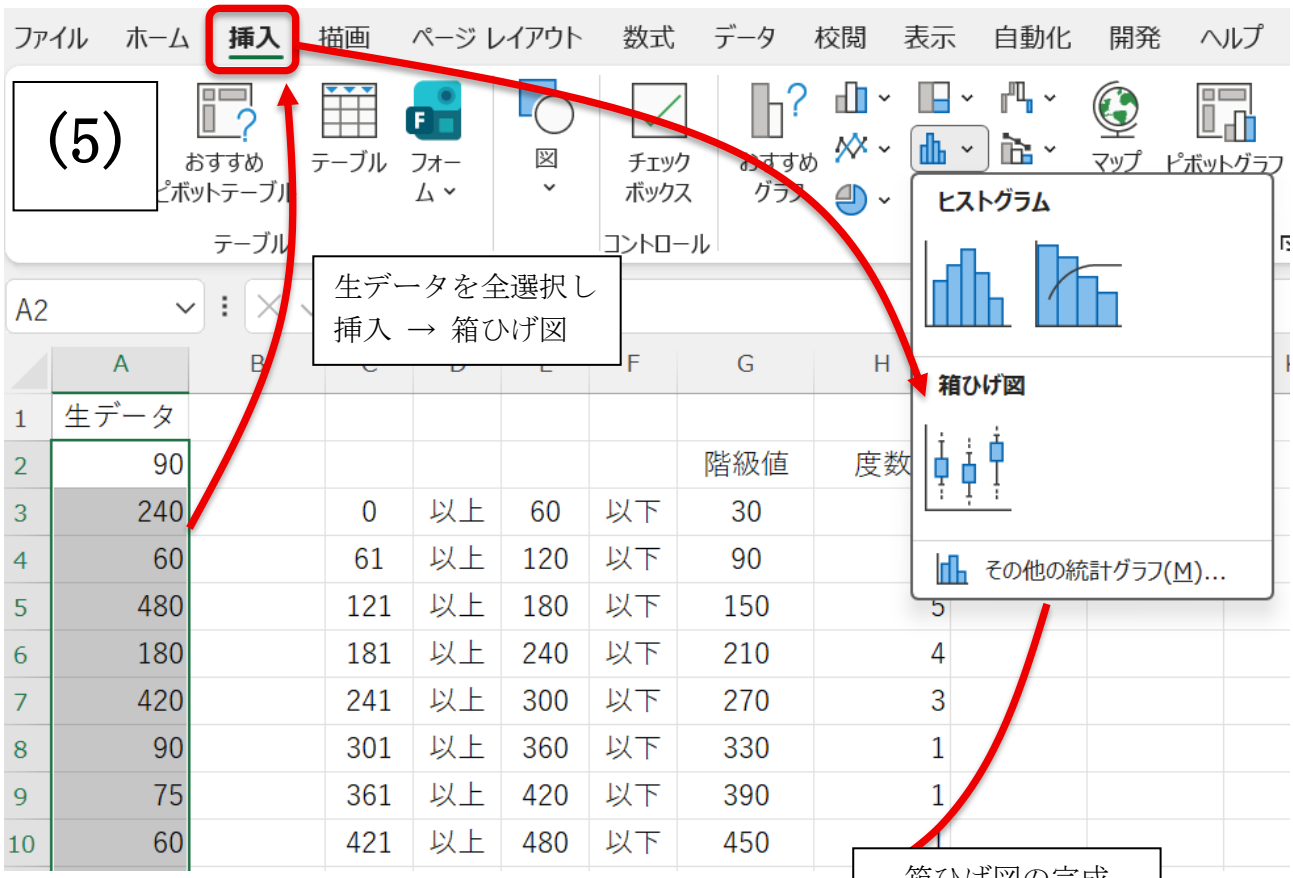




				階級値	度数
0	以上	60	以下	30	10
61	以上	120	以下	90	15
121	以上	180	以下	150	5
181	以上	240	以下	210	4
241	以上	300	以下	270	3
301	以上	360	以下	330	1
361	以上	420	以下	390	1
421	以上	480	以下	450	1
481	以上	540	以下	510	0
541	以上	600	以下	570	0

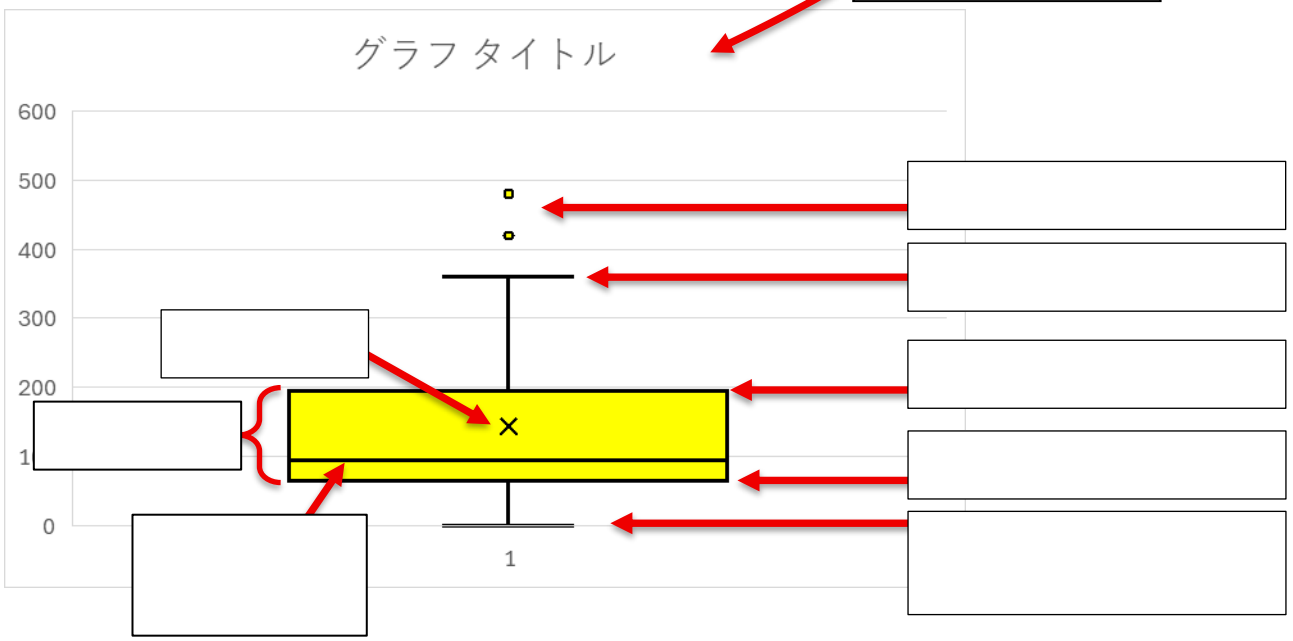
横（項目）軸ラベルを選択し
軸ラベルの範囲を選択し
軸にしたい所を選択する。





	A	F	G	H
1	生データ			
2	90		階級値	度数
3	240	0 以上 60 以下	30	
4	60	61 以上 120 以下	90	
5	480	121 以上 180 以下	150	5
6	180	181 以上 240 以下	210	4
7	420	241 以上 300 以下	270	3
8	90	301 以上 360 以下	330	1
9	75	361 以上 420 以下	390	1
10	60	421 以上 480 以下	450	1

箱ひげ図の完成



『はずれ値』とは

　他の値から極端にかけ離れたデータがあることがあります。そのような値をはずれ値と呼びます。はずれ値は除外すべき値とは限りません。その背景を探ることが大切です。測定ミスや入力ミスでなければ、そこに問題発見や解決の手がかりがあることもあります。なお、はずれ値は、　　とされます。Excelの場合はこのルールに従います。

※標準偏差σを用いて　　の場合もある。（有意水準５%の棄却域）

	A	B	C	D	E	F	G	H
1	生データ							
2	90						階級値	度数
3	240		0	以上	60	以下	30	10
4	60		61	以上	120	以下	90	15
5	480		121	以上	180	以下	150	5
6	180		181	以上	240	以下	210	4
7	420		241	以上	300	以下	270	3
8	90		301	以上	360	以下	330	1
9	75		361	以上	420	以下	390	1
10	60		421	以上	480	以下	450	1
11	60		481	以上	540	以下	510	0
12	80		541	以上	600	以下	570	0
13	45							0
14	300		平均値	143	"=AVERAGE(A2:A41)"			
15	240		最頻値	90	"=MODE.SNGL(A2:A41)"			
16	100		中央値	95	"=MEDIAN(A2:A41)"			
17	120							

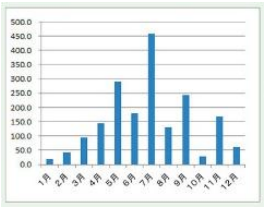
(6)

平均値、最頻値、中央値を求める。関数の前後についている「 ” ” 」は削除してください。

【情報の視覚化】

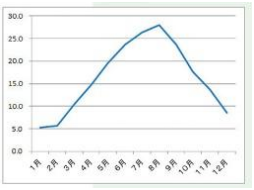
データにはそれぞれ「伝えたいメッセージ」があり、目的に応じて最適なグラフを選択することが重要です。グラフには量の比較に適した棒グラフ、推移を示す折れ線グラフ、内訳を表す円グラフなど、各々に得意分野があります。各グラフの特徴を正しく理解し、数値の背後にあるストーリーを最も効果的に伝える手法を選ぶことが重要です。

（1）



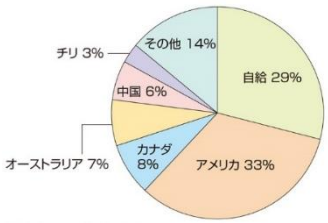
このグラフは、数値の大きさを「棒の高さ」で表し、並べて比較するために用いるグラフです。**種類の異なるカテゴリーを並べて「どれが一番多いか」「どれくらい差があるか」を一目でハッキリさせたいときに適しています。**縦軸は「０から始める」のが鉄則です。変化の勢いではなく、「純粋な量の比較」をしたい場合に最も有効な手段となります。

（2）



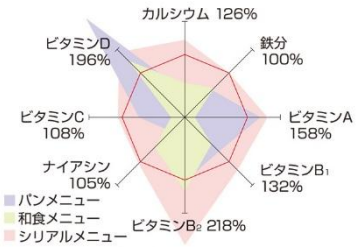
数値の地点を線で結ぶことでデータの「変化」や「推移」を視覚的に捉えるためのグラフです。**時間の経過に伴う数値の上下や、増減の傾向を一目で把握したいときに適しています。**複数の線を重ねれば異なる項目の変化の仕方を比較することも可能です。気温の変化、株価の推移といった継続的な記録の動向を伝えたい場合に最も威力を発揮します。

（3）



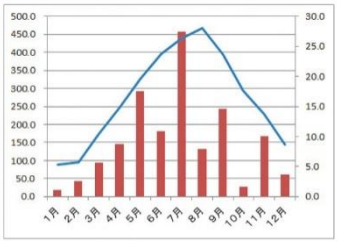
丸い図を扇形に区切ることで全体に対する各項目の「割合」や「内訳」を視覚的に表すグラフです。**全体を100%としたとき、特定の項目がどの程度のシェアを占めているかを一目で把握するのに適しています。**アンケートの回答比率や予算の配分などを説明する際によく用いられ、構成比の大きさを直感的に伝えたい場合に有効です。なお、項目数が多すぎると見づらくなるため、重要な数項目に絞って表示するのがコツです。

（4）



中心から放射状に伸びる複数の数値軸をつなぎ、データの「バランス」や「特性」をクモの巣のような形で表すグラフです。**複数の項目（能力や評価など）を同時に並べることで、突出した強みや不足している弱点を一目で把握するのに適しています。**ゲームキャラのステータスや多角的な満足度調査などでよく用いられ、全体のシルエットが正多角形に近いほどバランスが良いと判断できるのが最大の特徴です。

（5）



このグラフは、棒グラフと折れ線グラフなど、異なる種類のグラフを一つの図に重ねて表示するグラフです。「降水量（棒）」と「気温（折れ線）」のように、**単位や性質が異なる二つのデータを同時に並べることで、それらの相関関係を一目で把握するのに適しています。**左右に異なる目盛りを持つことが多く、複数の視点から一つの事象を分析したい場合に非常に有効です。情報の密度を高め、多角的なストーリーを伝えることができます。

問　以下データにはどのグラフが最も適しているか。

【選択肢】

円グラフ / 棒グラフ / レーダーチャート / 複合グラフ（棒＋折れ線） / 折れ線グラフ

（1）クラスの生徒に「一番好きなスポーツ」を1つ選んでもらった。各スポーツを選んだ人の割合を表現したいときは？

（2）1学期～3学期までの、自分の「計算問題の得点推移」をわかりやすく示したいときは？

（3）A班～E班まで、それぞれの班が「1週間に読んだ本の合計冊数」を単純比較したいときは？

（4）各教科の偏差値（国数英理社）のバランスをひと目で確認したいときは？

（5）月ごとの「平均気温」の変化と、その月の「降水量」を1つのグラフで表したいときは？

（6）理科の実験で、熱した水の温度の変化を1分おきに記録した結果をまとめるときは？

（7）都道府県別の「お米の収穫量」を、多い順に並べて大きさを比べたいときは？

（8）家庭の支出のうち「食費」や「住居費」が何%を占めているか、シェアを見たいときは？

（9）店の「接客」「味」「価格」「清潔さ」「メニューの多さ」の5項目を評価したいときは？

(10) 月の「売上額」を棒グラフで表し、「売上目標」の推移をも重ねて表示したいときは？

【分散、標準偏差】

平均値などの代表値だけでは、データの真の姿は見えません。同じ平均でも、値が中央に固まった集団と、上下に激しく散らばった集団では性質が異なるからです。そこで分散や標準偏差を用い、データの「バラツキ」を数値化することで、実態を客観的に把握できます。さらに相関係数を使えば、二つの事象間の「関係性の強さ」を測定でき、データの個性を正確に読み解く鍵となります。

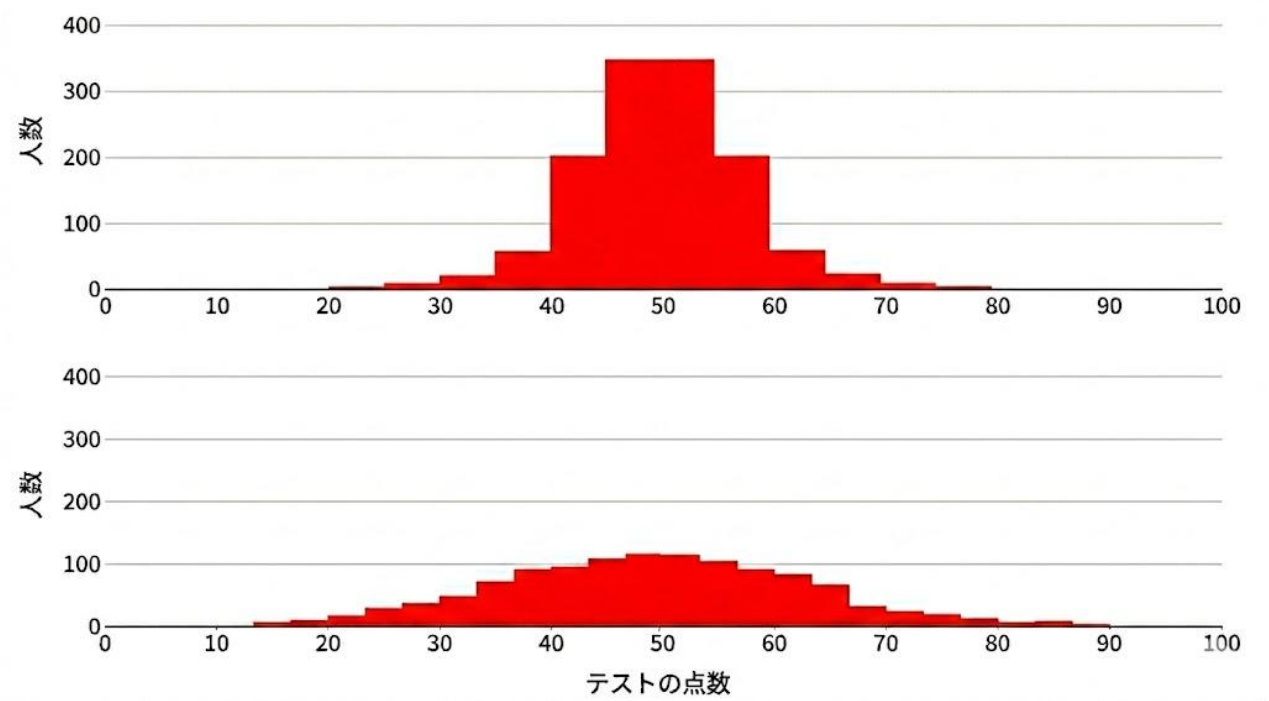
1. _____

データのバラツキ具合を数値化したものです。各データが平均値からどれくらい離れているかを計算し、その差を2乗して平均した値です。

2. _____

分散の平方根をとったもので、最もよく使われるバラツキの指標です。

※以下の２つのヒストグラムは、「平均値、中央値、最頻値」がともに50点で同じですが、データの散らばり方が異なります。このデータの散らばり方を数値化したものが、分散・標準偏差です。ちなみに、あなたの得点が60点の場合、上のグラフだと偏差値70ぐらい、下のグラフだと偏差値55ぐらいになります。



※偏差値

みなさんお馴染みの「偏差値」とは、集団の中での自分の位置を客観的に示す数値です。平均点を「50」とし、点数のバラつき（標準偏差）を基準に算出されます。テストの難易度や平均点が異なっても、偏差値を見れば「上位何%にいるか」が分かります。一般的に50より高ければ平均以上、低ければ平均以下を意味し、受験だけでなくデータの比較に広く活用される指標です。

偏差値の数式

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

問　以下の場合における偏差値を計算せよ。

(1)　あなたの得点70点、平均点50点　標準偏差10点

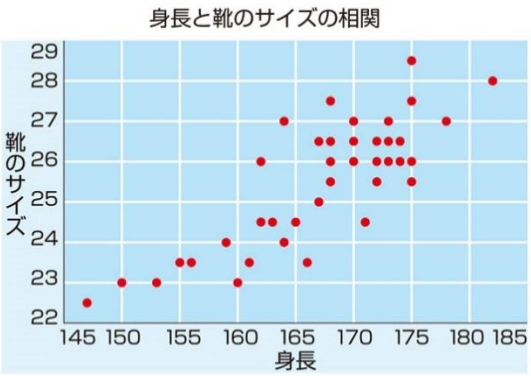
(2)　あなたの得点70点、平均点50点　標準偏差20点

※ちなみに、標準偏差＝$\sqrt{分散}$ なので、(1)は分散100、(2)は分散400となります。

※なぜ、分散ではなく、標準偏差を使うのか

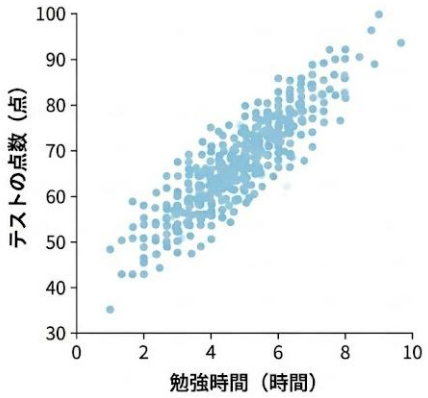
元のデータの単位と揃えるためです。分散を求めるには、各データの値から平均値を引き、その差を2乗します。この「2乗する」というステップがあるため、単位も一緒に2乗されてしまいます。しかし、標準偏差は、分散の正の平方根を取ることで求められます。2乗したものをルートで元に戻すため、単位も元のデータと同じに戻ります。

【散布図】

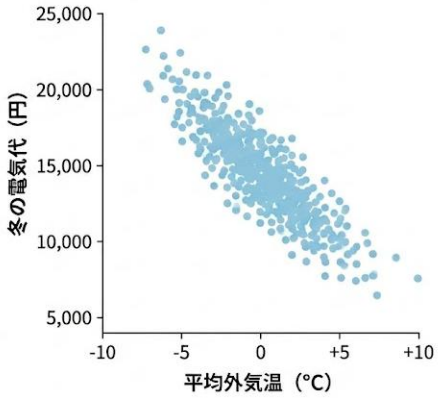


散布図とは、＿＿＿＿＿＿＿＿＿＿＿＿を視覚的に把握するためのグラフです。横軸と縦軸にそれぞれ異なる項目（身長と靴のサイズなど）を割り当て、各データを点としてプロットします。点の広がり方を見ることで、一方が増えるともう一方も増える「正の相関」や、逆に減る「負の相関」といったデータの傾向を直感的に確認できます。

正の相関の散布図



負の相関の散布図



【相関係数】

１．相関係数

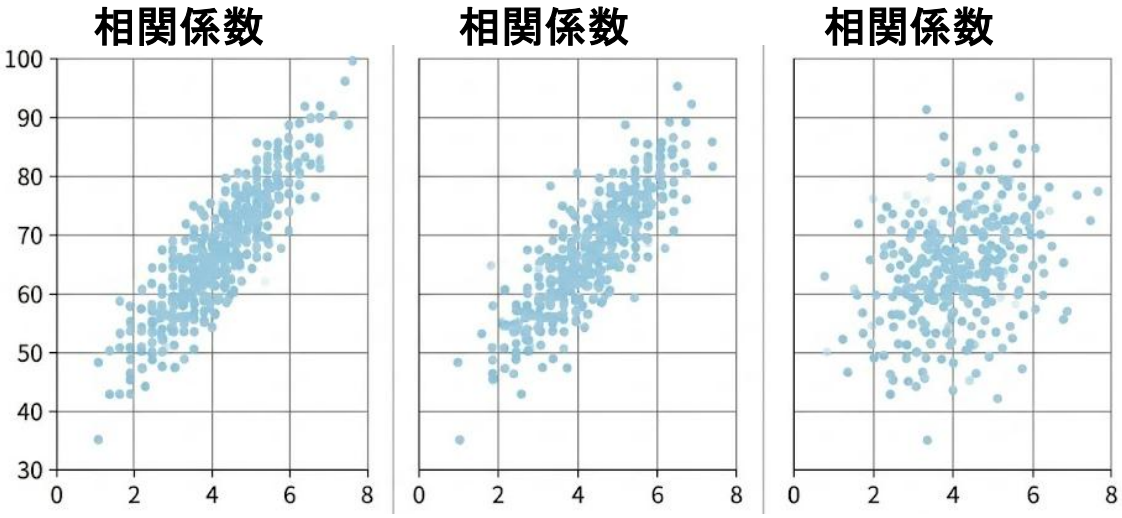
２つのデータの間の「関係の強さ」を＿＿＿＿＿の範囲で表した数値です。

①　　１に近い（正の相関）：

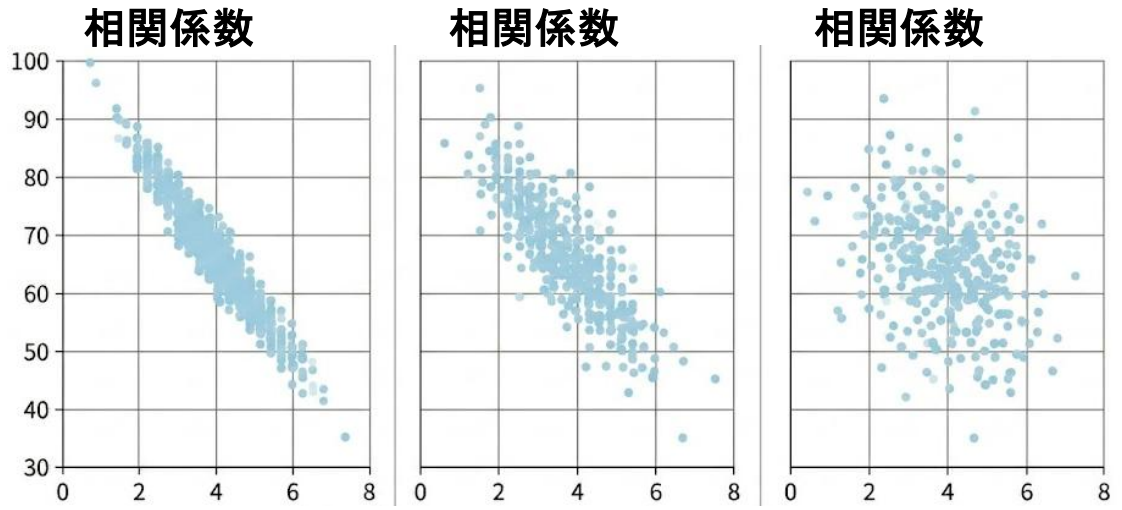
②　－１に近い（負の相関）：

③　　０に近い　　：

正の相関



負の相関



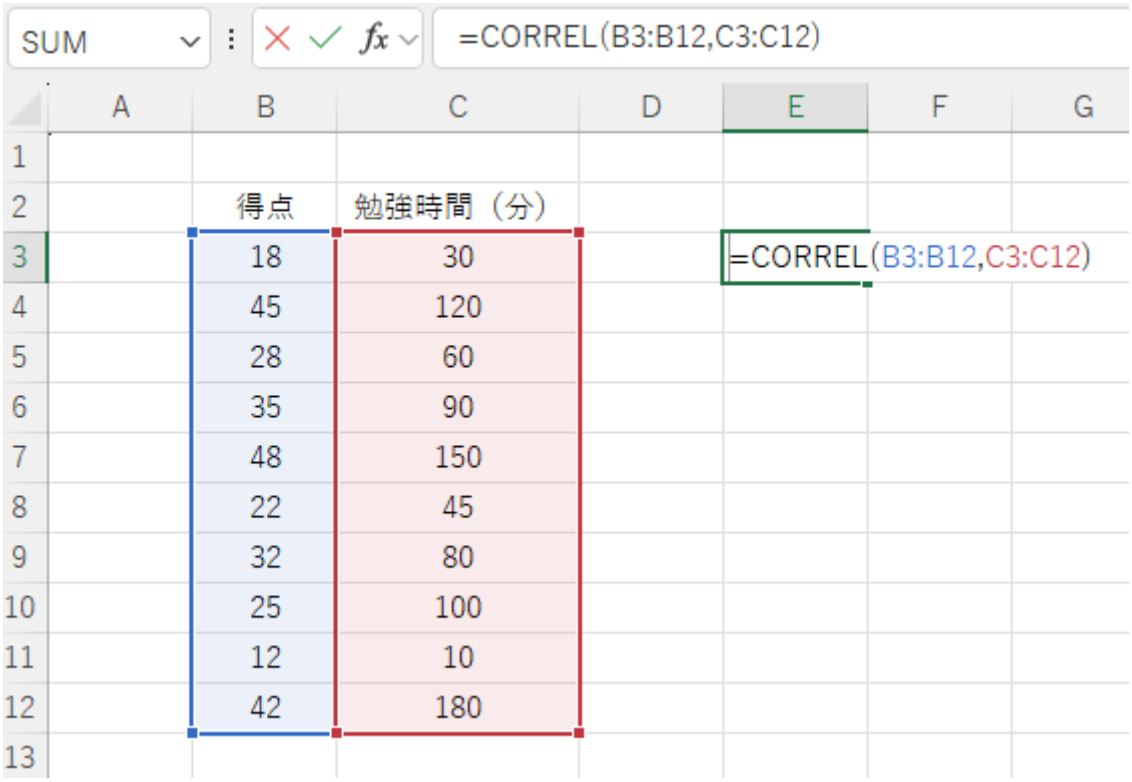
相関係数 r は右のように求める。

$$r = \frac{x,yの共分散}{(xの標準偏差) \times (yの標準偏差)}$$

※＿＿＿＿＿とは、２つの変数の関係性を示す指標です。一方が増えるときにもう一方も増えれば「正」、減れば「負」の値をとります。**データの散らばり具合を掛け合わせて平均したもの**で、相関係数の計算の基礎となります。

※相関係数を求める最大のメリットは、「２つのデータの関連性の強さ」を＿＿＿＿＿＿で把握できることです。共分散ではデータの単位によって値が大きく変わってしまいますが、**相関係数は単位に依存しない**ため、異なる項目同士でも比較が可能です。これにより、経験や勘に頼らず「気温が上がるとビールの売上は本当に増えるのか？」といった仮説をデータに基づき論理的に判断できるようになります。

Excelを使った相関係数の求め方



【実習２】

今から、簡単な計算問題を解いてもらいます。その計算問題の「得点」と、平日の「勉強時間（分）」、自宅から学校までの「通学時間（分）」、昨日の「睡眠時間（分）」「スマホの１日の使用時間（分）」、「数学の好き嫌い（１～５の５段階評価）」をフォームに入力してください。その入力結果を用いてそれぞれの相関係数を求めてみよう。

【相関関係と因果関係は違う】

１．

２つの事象が「連動して変化する」状態です。一方が増えるともう一方も増える（または減る）関係ですが、どちらが原因かは問いません。

（例）アイスの売上が上がると、水難事故が増える。

これは「気温上昇」という共通原因によるもので、アイスが事故を起こすわけではありません。

２．

「Aが原因でBが起きた」という、時間的な前後と直接的な結びつきがある状態です。

（例）気温が上がったから、アイスが売れた。

見極めのポイント

相関関係があっても、実は**「第３の要因」が隠れていたり、単なる偶然だったりする**ことが多々あります。「連動しているだけか、一方がもう一方を生んでいるか」を疑うことが、データに騙されないコツです。

問　以下の(1)～(5)のケースについて、「因果関係（原因と結果の関係）」といえるか、それとも単なる「相関関係（連動しているだけ）」なのか判断してください。

(1)読書量と年収
　『統計的に、本をたくさん読む人ほど年収が高い傾向にある。』

(2)降雨と傘の売上
　『雨が降る量が増えると、傘の売上も増加する。』

(3)交番の数と犯罪件数
　『ある地域の交番の数が多いほど、その地域での犯罪認知件数も多い。』

(4)広告費と商品の売上
　『企業がテレビCMなどの広告費を増やすと、その商品の売上が伸びた。』

(5)朝食の摂取とテストの点数
　『毎日朝食を食べる子供は、食べない子供よりもテストの平均点が高い。』

【実習３】

実習２で取得したデータを用いて、以下の内容に取り組もう。

(1)　第３の要因を見つけるトレーニング

例えば「数学の好き嫌い」と「計算の得点」に強い相関が出た場合、そこには共通の要因が潜んでいるはずです。その共有要因を予測してみよう。

(2)　次に取るべきデータのアイディアを出す。

分析の精度を上げるために、次に取るべきデータのアイディアを出し、実際にデータを取得してみよう。