1.1 統計データ

ある事象がもつ特性と関連すると注目される要素、またはその集まりから数値上の性質、規則性の有無を見出すとき、その要素やその集合のことをデータ、または統計データといいます。これらのデータは過去の記録から得たり、新たに実験・調査を行ったりして集めます。この統計データのうち質的な要素であるものを属性データ、量的な要素を計量データといいます。また、1 種類の要素に注目するとき、その統計データは 1 変量といい、複数を同時に得るとき、多変量であるといいます。データを特性に応じて分類することは、データ分析の出発点です。

1.2 変数の分類

変数は質的変数と量的変数に区別されます。性別、血液型、レストランのランキングなどは質的変数です。一方で、温度や体重などは量的変数です。質的変数はさらに名義尺度と順序尺度に分類されます。

- 名義尺度：同じ値のときだけに意味をもち、それ以外では意味をもたない尺度
  名字、名前、血液型、性別、好きな株式銘柄など
- 順序尺度：名義尺度の性質に加えて順序(大小関係)が意味をもちます。
  成績の５段階評価、レストランのランキング、クレジット・レーティング(AAA,AA,A,,,)など

量的変数には間隔尺度と比例尺度の２つがあります。

- 間隔尺度：０が相対的な意味をもち、等間隔の大小関係をもち、値の差が意味をもちます。
  温度、偏差値、西暦など
- 比例尺度：間隔尺度の性質に加えて、ゼロが絶対的な意味をもちます。
  体重、年齢、身長、収入、絶対温度

さらに質的変数は２値変数と多値変数、量的変数は離散変数と連続変数に分けることができます。

統計データ

データの性質の意味合いからの分類

- 質的変数
  - 名義尺度：ワインの銘柄、職業、性別など
  - 順序尺度：ワインの好み、成績評価など
- 量的変数
  - 間隔尺度：アルコール度数、温度など
  - 比例尺度：身長、体重、年齢、絶対温度など

厳密なデータの分類は、統計的分析手法の選択の原点になることがあります。

例：つぎの表はポルトガル産のワインの成分とその評価を表したものです。12 項目からなり、ワインの評価は 10 段階で 1 から 10 までの整数を用いて行われています。１０が最も良く、１が最も悪いことを表します。ここでのワインの評価は順序尺度です。

	A	B	C	D	E	F	G	H	I	J	K	L
1	比例尺度											順序尺度
2	A	B	C	D	E	F	G	H	I	J	K	評価
3	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
4	7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5
5	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
6	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
8	7.4	0.66	0.00	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
9	7.9	0.60	0.06	1.6	0.069	15	59	0.9964	3.30	0.46	9.4	5
10	7.3	0.65	0.00	1.2	0.065	15	21	0.9946	3.39	0.47	10	7

データの出所：http://archive.ics.uci.edu/ml/datasets/Wine+Quality

1.3 要約統計量

実際の調査や観測で得られたデータを観測値といいます。実験や観測では複数のデータを集めます。しかし、大量のデータ、１つ１つの観測値を見ても、なかなかそのデータそのもののもつ特徴はとらえられません。グラフを用いると直感的に特徴をとらえられたり、その特徴を１つの数値で表すとデータのもつイメージがつかめたりすることがあります。

1.3.1 データの可視化

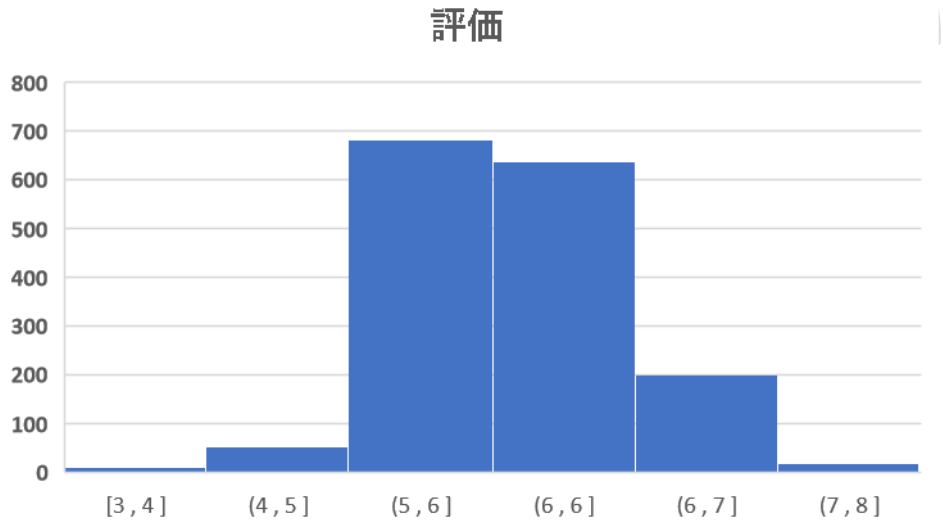
データをグラフとして視覚的に要約することで、全体の特徴をとらえることができます。

- ヒストグラム(頻度図)の作成

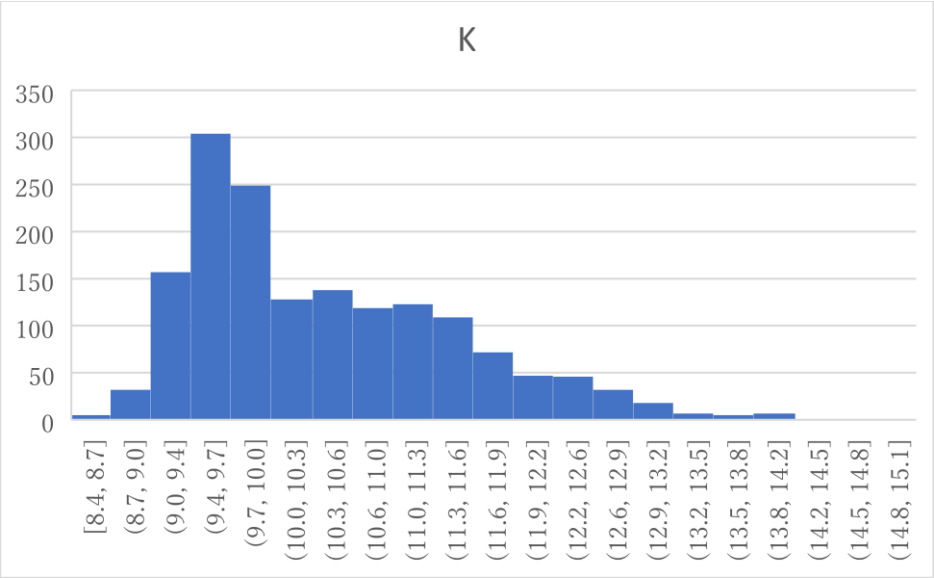
頻度図は、横軸に変数をその大きさ、または階級などに応じて並べ、縦軸にそれらの頻度を表したグラフです。

例：ワインデータの主な要素の頻度図を得ます。

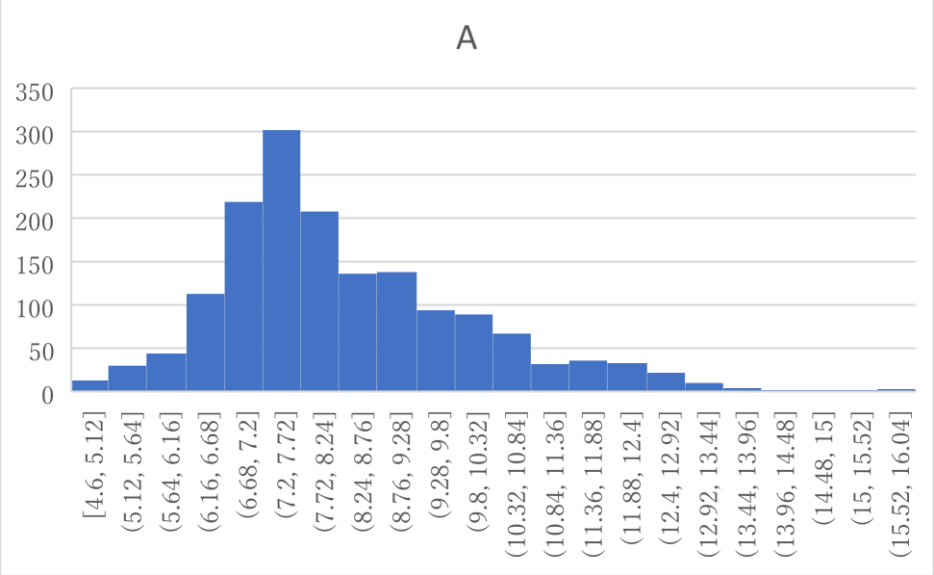
このデータベースは赤ワインを１０段階の評価結果とワインの特徴の科学的分析結果を集めたものです。つぎの頻度図は赤ワインの１０段階評価を頻度図にしたものです。



横軸は１０段階評価、縦軸はその頻度です。最も頻度の多い評価は５です。つぎが６です。最も高い評価は８で低い評価は３です。頻度が評価の中央に位置していて単峰の山のようであるのが分かります。頻度の分布は左右対称ですので、このような分布の形をベル型の分布と呼びます。



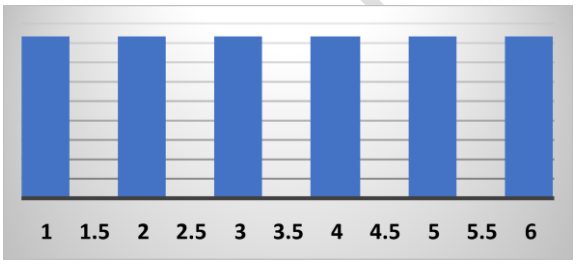
横軸はアルコール度数です。最も頻度の高いアルコール度数は9.5近辺です。すそ野は右に長くなっています。分布の度数は左によっています。これを右にひずんだ分布といいます。



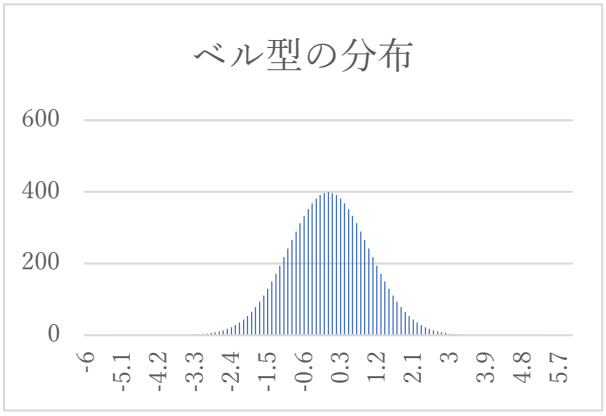
横軸は酢酸濃度です。最も頻度の高い酢酸濃度は0.6近辺です。分布の形状は天井が平らで、左右のすそ野はなだらかに減少している台形にも見えますし、２つの単峰の分布が混じっているようにも見えます。

頻度図の形状は大まかに

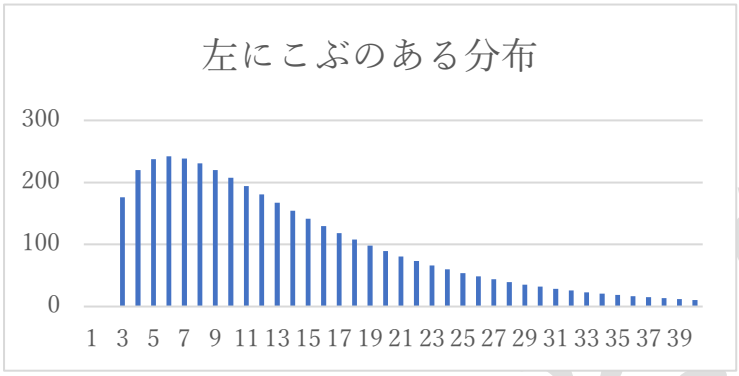
- 一様な分布：頻度が横軸の値に対してほぼ均等。



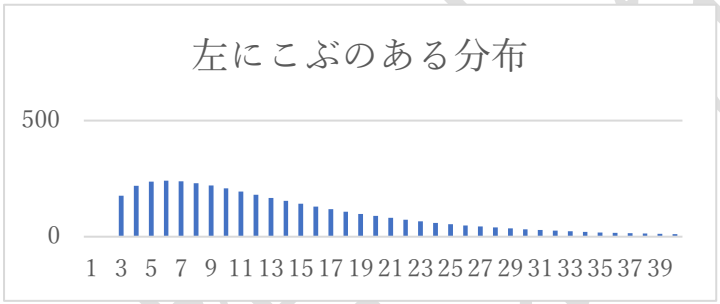
- ベル型の分布：頻度の高さは横軸に対してベル型。



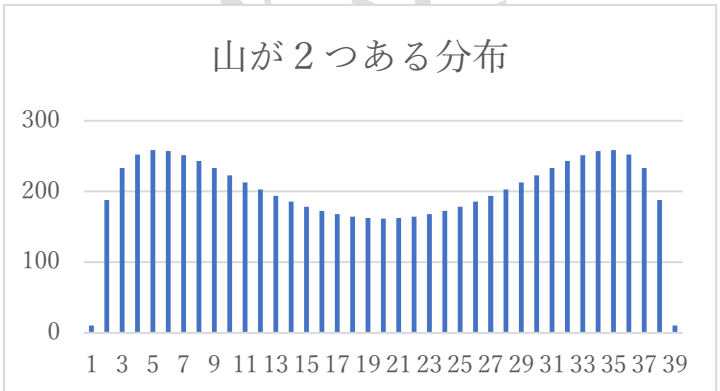
- 右にひずんだ分布：右にすそ野が長く、頻度が左寄った分布。



- 左にひずんだ分布：左にすそ野が長く、頻度が右寄った分布。



- 複数の山をもつ分布：いくつもの分布が混じった分布。

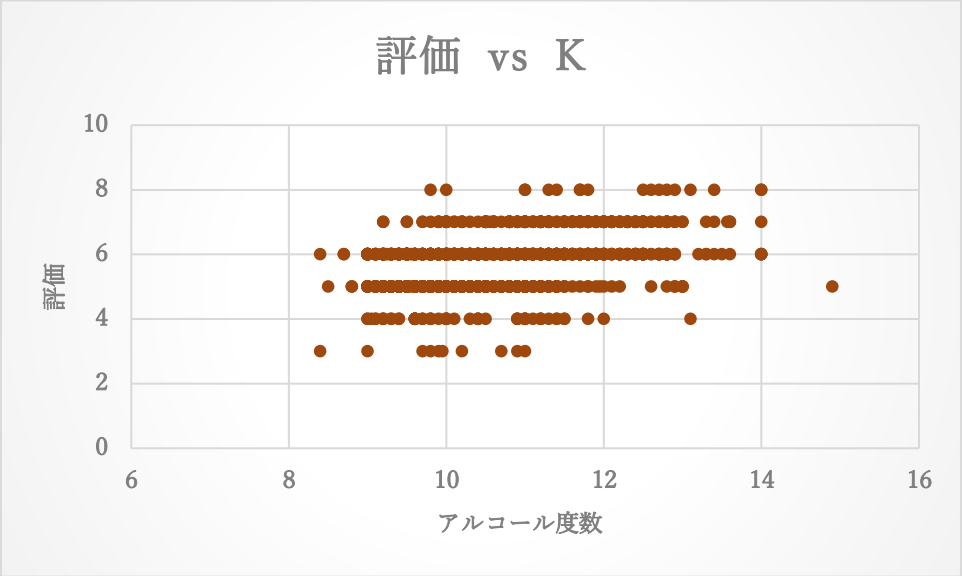


に分けられます。頻度図により変数の幅、ばらつき具合、頻度の高低などの大まかな傾向が一目でつかめます。

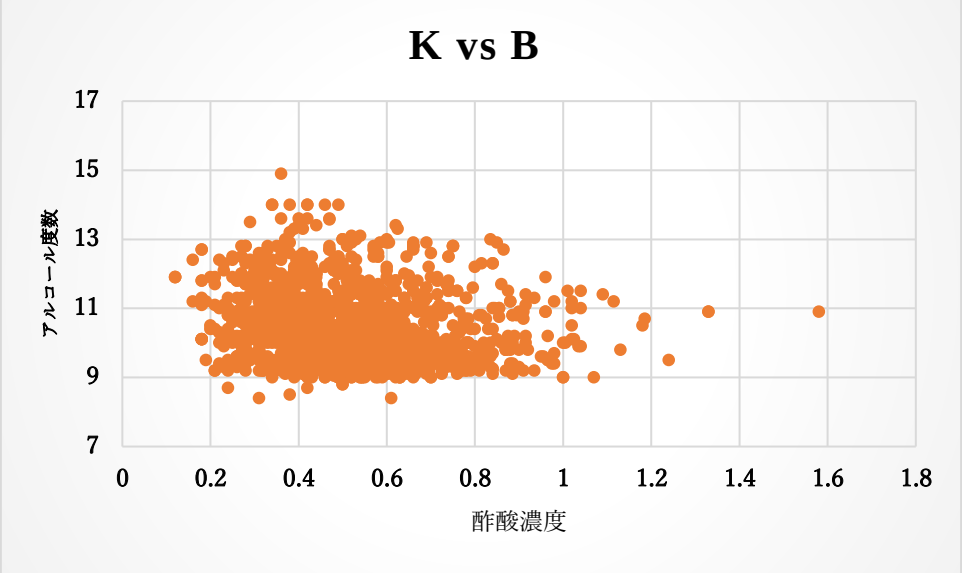
- 散布図の作成

散布図は横軸と縦軸に二つの異なるデータを割り当て，観測値を打点して作るグラフです。

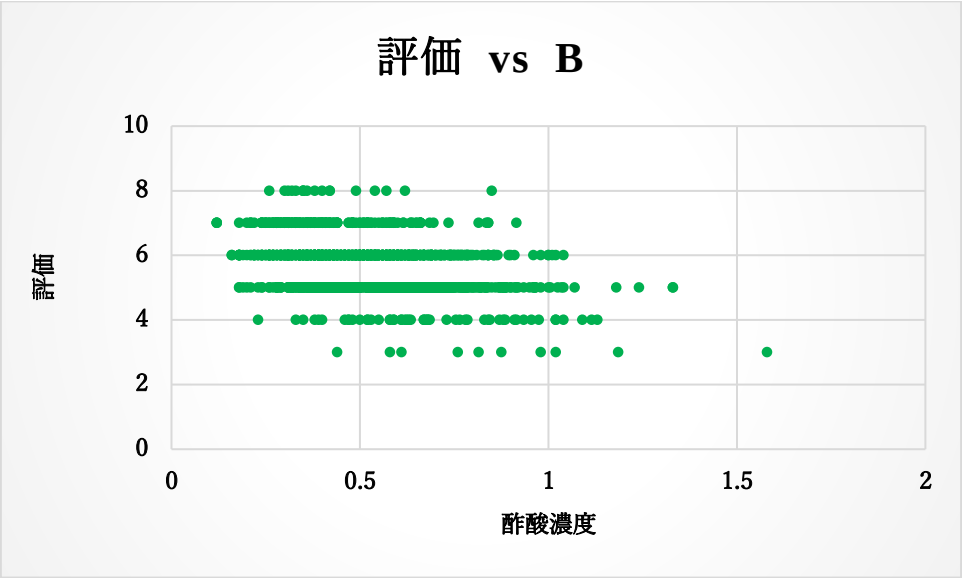
例：ワインデータの主な要素の散布図



横軸にアルコール濃度、縦軸に赤ワインの評価を目盛っています。アルコール濃度が増えると評価が高くなる傾向がありそうです。しかし、それはかなり大まかな傾向であることが分かります。



横軸に酢酸濃度、縦軸にアルコール濃度を取っています。この散布図から大きな傾向が見られません。酢酸濃度が高くなるとアルコール濃度の幅が狭まり、9 から 12 の中に納まっているように見えます。しかし、酢酸濃度は高くなると頻度が低くなるので、ただ単にデータ点の数が少なくこのように見える可能性があります。



横軸に酢酸濃度、縦軸に評価を取りました。酢酸濃度が上がると評価が下がる傾向がありそうです。しかし、酢酸濃度の頻度は両端に行くほど低くなっているので、その影響を考慮する必要があります。

３つの散布図を見ましたが、このような可視化は２つの変数の大まかな傾向をとらえるときに有効です。

データの特徴を１つの数値として表現すると便利なときもあります。たとえば、小さいことが良い指標となるときには最小値を使うでしょう。大きいことを誇りたいときには最大値が便利でしょう。このように１つの数値でデータの特徴を表現するととても便利なときがあります。記述統計量、基本統計量、代表値ともいわれます。要約統計量ですが、４つのタイプに大きく分けることができます。１つはどの辺にデータが集中しているか、２つ目はどの程度のばらつきがあるのか、そして３つ目はデータ間の関係をとらえる指標です。最後の４つ目は分布の形状に関するものです。

1.3.2 1 変量要約統計量

データの集中の度合いについて見てみましょう。さまざまな尺度があり、それぞれに特徴があります。

平均(算術平均)

平均とはデータ $(x_1,\ldots,x_n)$ の位置を示す代表値の 1 つです。x_i は i 番目の観測値を示します。n は観測値の数です。最もなじみのある計算方法は n 個のデータの総和をその数で割った値です。これを相和平均とか単純平均と呼びます。それは

$$\frac{x_1 + x_2 + x_3 \cdots + x_n}{n}$$

と書くことができます。一般に平均という場合には算術平均のことです。

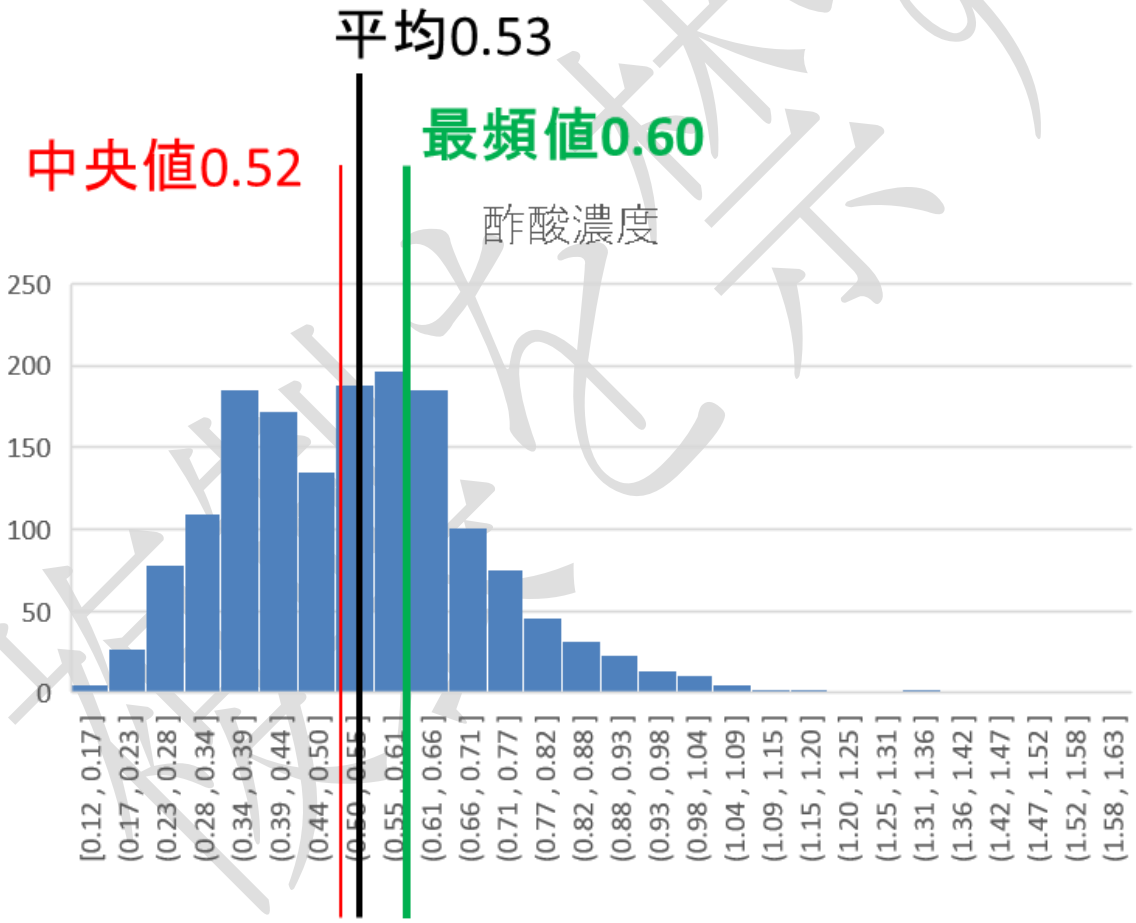
中央値(メディアン)

中央値は、データを大きい順、または小さい順に並べたときにその中央に来る値です。もし中央に位置する値が２つある場合には、その２つの値の算術平均を取ります。このような要約量は、データを２つに分割するときの目安になります。また、データの分布が左右対称とならずにどちらかに偏っている場合、異常に大きな値、または小さな値があるときの代表値として、平均よりも適しているときがあります。

最頻値(モード)

データの中で最も頻度の高い値です。

つぎの図は酢酸濃度の頻度図に中央値、平均値、最頻値を重ねたものです。



平均、中央値、最頻値は、酢酸濃度のように常に同じ値になるとは限りません。

幾何平均

データ $(x_1,\ldots,x_n)$ の幾何平均とはつぎの式で定義されます。

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

それぞれの数値の積を求め、その n 乗根を取って得られます。成長率などの指数関数的な性質をもつ変数に使われます。

つぎにデータのばらつきの度合いについて見てみましょう。

分散

分散はデータが平均値からどれくらいばらついているかを表す指標でつぎにように定義されます。

$$\mathrm{var}(x) = \frac{\sum (x_i - \bar{x})^2}{n}$$

ここで、$\bar{x}$は x の平均です。n はデータの数です。つまり、x_i の平均からの偏差の 2 乗の和として定義されます。分散がゼロであれば、ばらつきはありません。分散が大きくなるとばらつきも大きくなります。

データのばらつきを求める際に、それぞれのデータの値とその平均との差を求めて、その総和を求めるという方法もあります。しかし、そのような方法ではそれらの差の正と負の値がほぼ同数となると、多くが相殺されてしまいます。したがって、平均からのばらつきの指標にはなりません。そこで、それぞれのデータと平均との差を 2 乗して総和を求め、総数で割るという方法が取られているのです。

標準偏差

分散の正の平方根を標準偏差と呼びます。分散同様に、データの散らばり具合を表す指標です。

$$\sigma_x = \sqrt{\mathrm{var}(x)}$$

分散は元のデータの 2 乗を用いて計算しているので、元のデータとは次元が異なるために直接比較することはできません。平方根を取ることで、比較ができるようになり、　標準偏差は分散の弱点を克服しています。

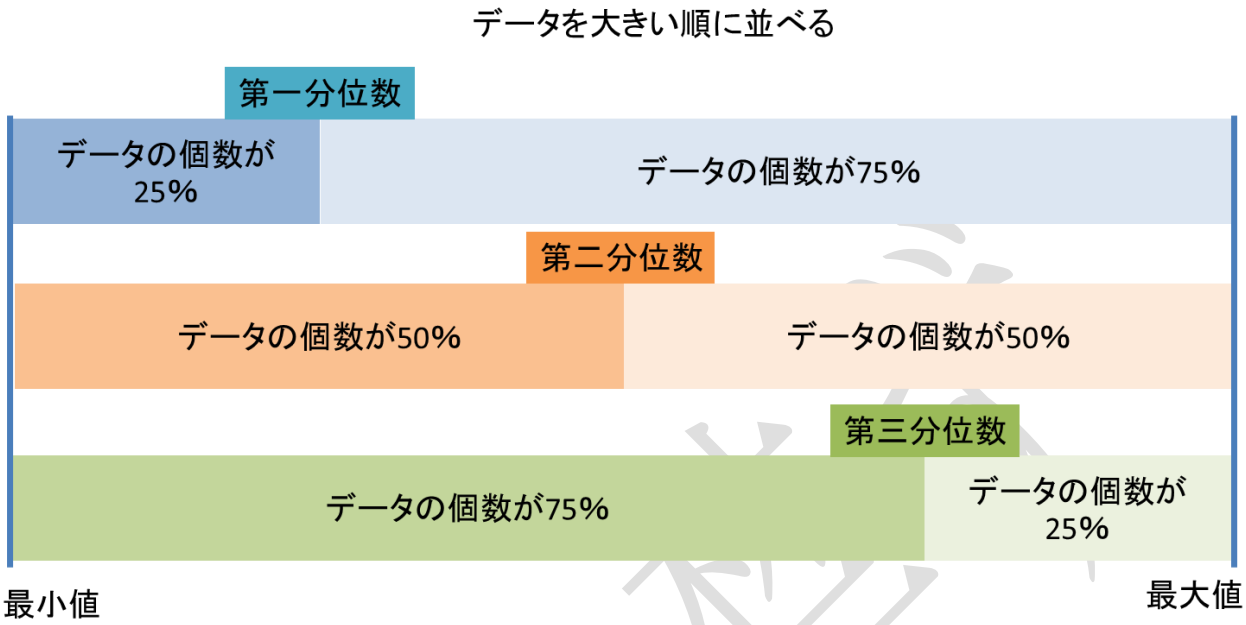
最大値・最小値

最大値は観測値のうちで最も大きな値、最小値はデータのうちで最も小さな値です。

分位数

n 個のデータを大きい順に並べ$x_1 \leq x_2 \leq \cdots x_n$とします。$i/n = \alpha/100$ とすると、P_α $=(x_i+x_{i+1})/2$ よりも小さな値

をとるデータの割合が α%、それよりも大きな値をとるデータの割合が 100−α%となります。このような P_α を第 α 百分位数といいます。第 25 百分位数、第 50 百分位数、第 75 百分位数をそれぞれ第 1 四分位数、第 2 四分位数、第 3 四分位数といいます。四分位範囲とはこの第 1 四分位数と第 3 四分位数の間の範囲のことです



範囲

最大値から最小値を差し引いた値です。

四分位範囲

データを大きい順に並べたときにそのデータの最初の 25%から 75%までの範囲のことです。

例：エクセルによる得られたワインデータの主な要約統計量

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	平均	8.3	0.53	0.27	2.5	0.087	16	46	0.9967	3.31	0.66	10.4	6
2	分散	3.0	0.03	0.04	2.0	0.002	109	1082	0.0000	0.02	0.03	1.1	1
3	標準偏差	1.7	0.18	0.19	1.4	0.047	10	33	0.0019	0.15	0.17	1.1	1
4	最大値	15.9	1.58	1.00	15.5	0.611	72	289	1.0037	4.01	2.00	14.9	8
5	最小値	4.6	0.12	0.00	0.9	0.012	1	6	0.9901	2.74	0.33	8.4	3
6	第1四分位範囲	7.1	0.39	0.09	1.9	0.070	7	22	0.9956	3.21	0.55	9.5	5
7	第3四分位範囲	9.2	0.64	0.42	2.6	0.090	21	62	0.9978	3.40	0.73	11.1	6
8	範囲	11.3	1.46	1.00	14.6	0.599	71	283	0.0136	1.27	1.67	6.5	5
9													
10		A	B	C	D	E	F	G	H	I	J	K	評価
11		7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

主な 1 変量要約統計量のエクセル関数

	A	B
1	平均	=AVERAGE(B$11:B$1609)
2	分散	=VAR.P(B$11:B$1609)
3	標準偏差	=STDEV.S(B$11:B$1609)
4	最大値	=MAX(B$11:B$1609)
5	最小値	=MIN(B$11:B$1609)
6	第1四分位範囲	=QUARTILE.EXC(B$11:B$1609,1)
7	第3四分位範囲	=QUARTILE.EXC(B$11:B$1609,3)
8	範囲	=+B4-B5

1.3.3 変数の分類と要約統計量

上述の 4 つの尺度にどの統計量が利用できるのか？主なものとして

- 名義尺度：度数、最頻値
- 順序尺度：度数、最頻値、中央値、四分位数
- 間隔尺度：度数、最頻値、中央値、四分位数、平均、標準偏差
- 比例尺度：度数、最頻値、中央値、四分位数、平均、標準偏差、変動変数、幾何平均

があります。名義尺度は単に区別のために用いられています。度数とか最頻値を計算して、他と比較することができます。順序尺度は、順序や大小関係を表現するために用いられます。そのために中央値、四分位数が計算できます。間隔尺度はデータの差に意味をもたせ、間隔や距離を測るために用いることができます。したがって、順序尺度に加えて、平均、標準偏差が計算できます。比例尺度は間隔尺度に加えて比率にも意味があるものをいいます。したがって、幾何平均などの計算ができます。

1.3.4 2 変量要約統計量

平均、中央値、分散、標準偏差は、それぞれの要素の統計的な性質を説明しています。つぎは対となる 2 組(または、それ以上の組)のデータの間の特徴をとらえる要約統計量を説明します。

共分散

2 組のデータ (x_1,y_1)，(x_2,y_2)，…(x_n,y_n)，の共分散は、つぎにように定義されます。

$$\mathrm{Cov}(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

ここで、$\bar{x}$，$\bar{y}$はそれぞれ x, y の平均を表します。共分散は 2 組のデータの平均からの偏差の積の単純平均であ

ると考えられます。x と y が同じであると、共分散は分散になります。要素の数が 2 つ以上になるとマトリックスとして表現されます。a, b, c, d は要素を表しています。対角線上の Cov(a,a), Cov(b,b), Cov(c,c), Cov(d,d)は分散を表しています。対角線を境に対称で同じ色のセルの共分散は同じものです。

	a	b	c	d
a	Cov(a,a)	Cov(b,a)	Cov(c,a)	Cov(d,a)
b	Cov(a,b)	Cov(b,b)	Cov(c,b)	Cov(d,b)
c	Cov(a,c)	Cov(b,c)	Cov(c,c)	Cov(d,c)
d	Cov(a,d)	Cov(b,d)	Cov(c,d)	Cov(d,d)

相関

共分散は 2 組のデータ(a, b)のもつ特徴をとらえようとしているのですが、その計算結果は対となるデータのそれぞれの平均からの偏差の大きさ(標準偏差)に大きな影響を受けます。何らかの判断の材料にするためには経験を要します。そこで、共分散を各標準偏差で割ることで、-1 から+1 までの数値に収まるようにします。

$$\frac{\mathrm{Cov(a,b)}}{\sigma_a \sigma_b}$$

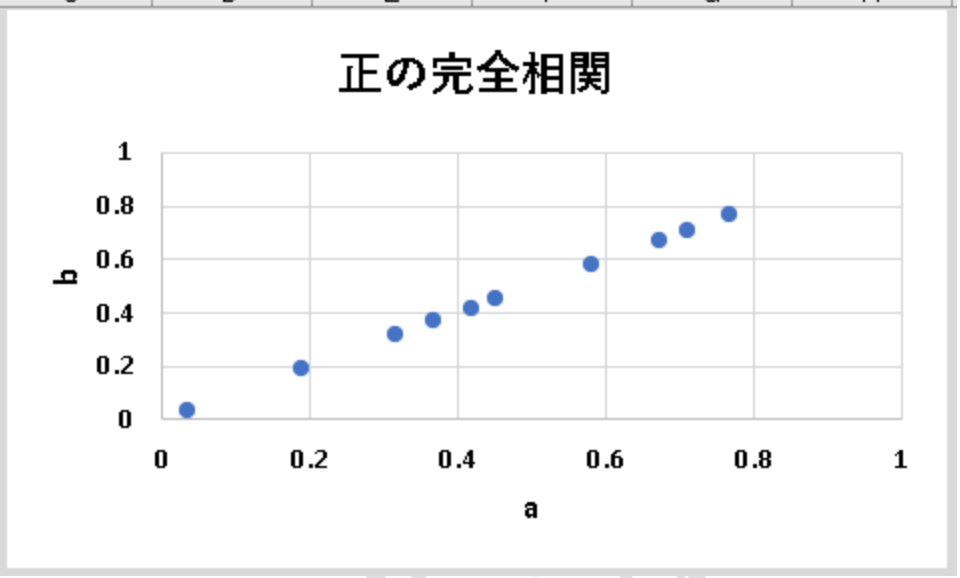
このようにすることで、相関が 1 に近ければ 2 組のデータは同じような動きになり、ゼロに近ければ、関係がなく、-1 に近ければ逆の動きをしていることになります。相関が 1 のときを正の完全相関、-1 のときを負の完全相関といいます。共分散同様に、要素が 2 つ以上ある場合には相関もマトリックスを用いて表現します。

相関は便利で使いやすいのですが、使い方に注意が必要です。相関は単なる平均的な関係を示すだけで、たとえば A と B の相関が高いからといって、それが，A が B の原因であるとか、B が A の原因であるとか、事象の因果関係を示すことになりません。この点には注意が必要です。

例：相関を可視化してみましょう。

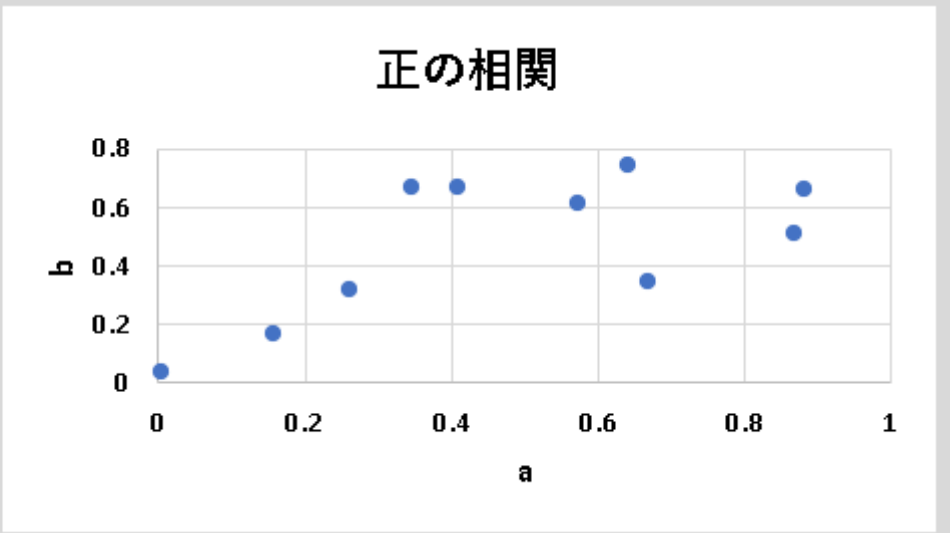
相関を、散布図を用いて理解してみましょう。つぎの図は乱数(RAND 関数)を用いて確率変数 a, b を生成し、正の完全相関、正の相関、無相関。負の相関、負の完全相関を散布図として表現したものです。正の完全相関、負の完全相関を生成するのは簡単ですが、それ以外についてはそれぞれの実現値 a, b の相関を示してあります。RAND 関数を用いて生成しているために、エクセルスプレッドシート上の値は再計算のたびに変化します。

	A	B
1	a	b
2	0.418147	0.418147
3	0.672999	0.672999
4	0.03329	0.03329
5	0.766086	0.766086
6	0.187438	0.187438
7	0.449648	0.449648
8	0.367128	0.367128
9	0.581226	0.581226
10	0.710097	0.710097
11	0.31577	0.31577
12		
13	相関	1

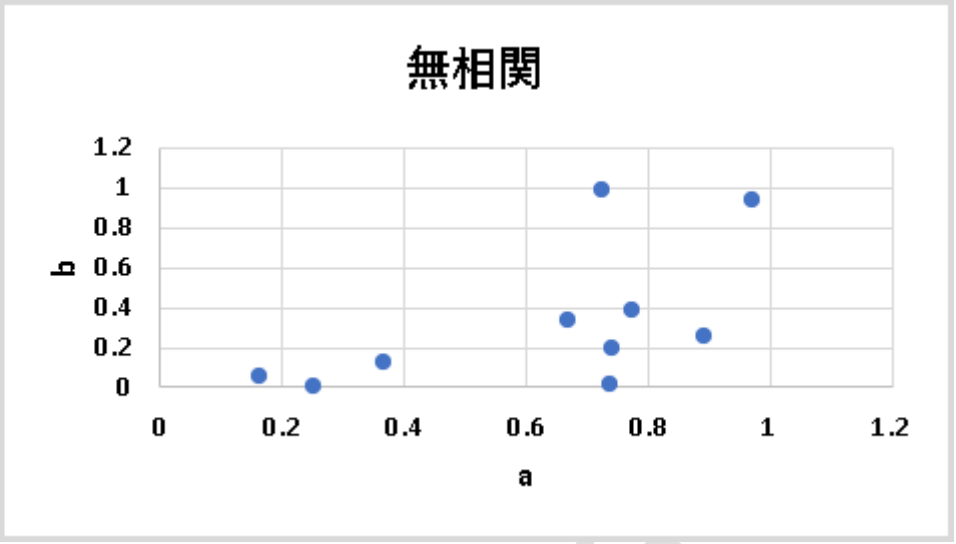


	A	B
1	a	b
2	=RAND()	=+A2
3	=RAND()	=+A3
4	=RAND()	=+A4
5	=RAND()	=+A5
6	=RAND()	=+A6
7	=RAND()	=+A7
8	=RAND()	=+A8
9	=RAND()	=+A9
10	=RAND()	=+A10
11	=RAND()	=+A11
12		
13	相関	=CORREL(A2:A11,B2:

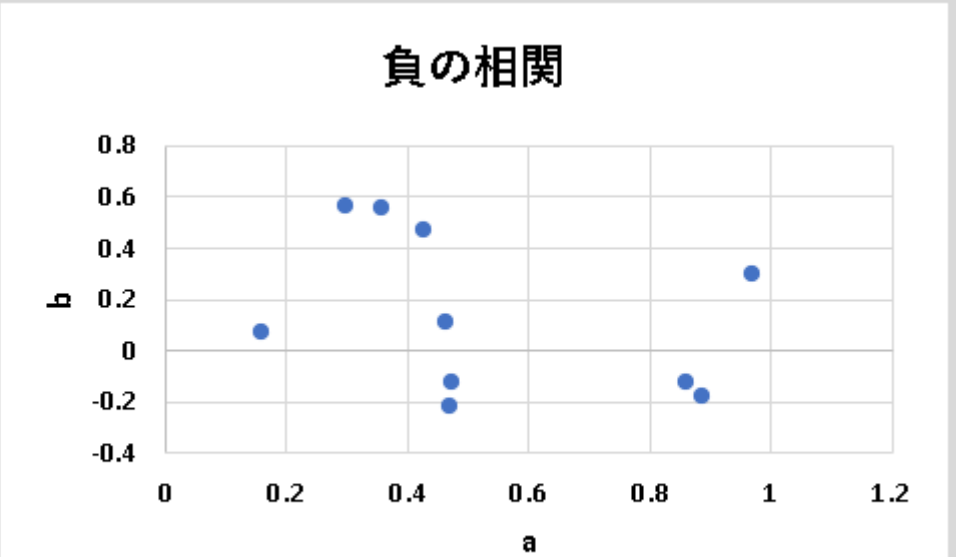
a	b
0.261179	0.319787
0.867516	0.510086
0.410221	0.666242
0.669737	0.344558
0.882037	0.665307
0.34592	0.666036
0.640114	0.74127
0.156747	0.166402
0.572299	0.613479
0.004114	0.038647
相関	0.65836



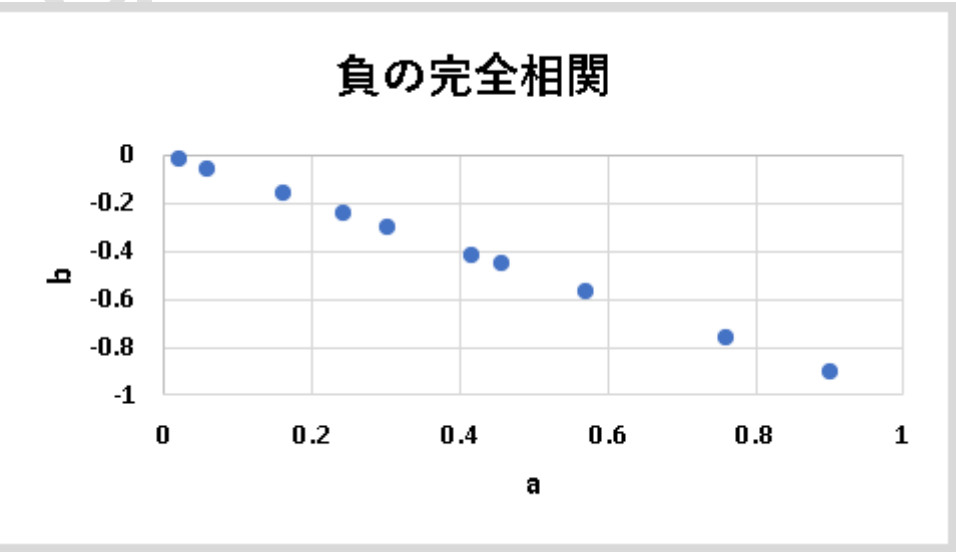
a	b
0.97075	0.939692
0.726239	0.988143
0.739592	0.011938
0.252276	0.009686
0.367643	0.129518
0.164134	0.057461
0.775865	0.386807
0.74139	0.191862
0.894188	0.25479
0.668387	0.340191
相関	0.580575



a	b
0.475074	−0.12594
0.885917	−0.17567
0.158306	0.074127
0.470641	−0.21922
0.298975	0.564101
0.970142	0.295994
0.860583	−0.12552
0.464784	0.107955
0.357405	0.556642
0.427084	0.474639
相関	−0.35873



a	b
0.41648	−0.41648
0.303659	−0.30366
0.020853	−0.02085
0.569892	−0.56989
0.455995	−0.456
0.162883	−0.16288
0.760633	−0.76063
0.242157	−0.24216
0.057514	−0.05751
0.902755	−0.90275
相関	−1



散布図は相関を明確に表現してくれますが、正の相関、負の相関、無相関の境界はそれぞれの状況で判断する

必要があります。

相関係数

- ペアーとなる2つの確率変数の間の関係の強さを
- 1から1までの数値で表します。
- 相関は便利で使いやすいのですが、使い方には注意が必要です。
- 相関は平均的な関係の強さを示しているだけです。
- AとBの相関が高いからといって、
  AがBの原因であるとか
  BがAの原因であるとか

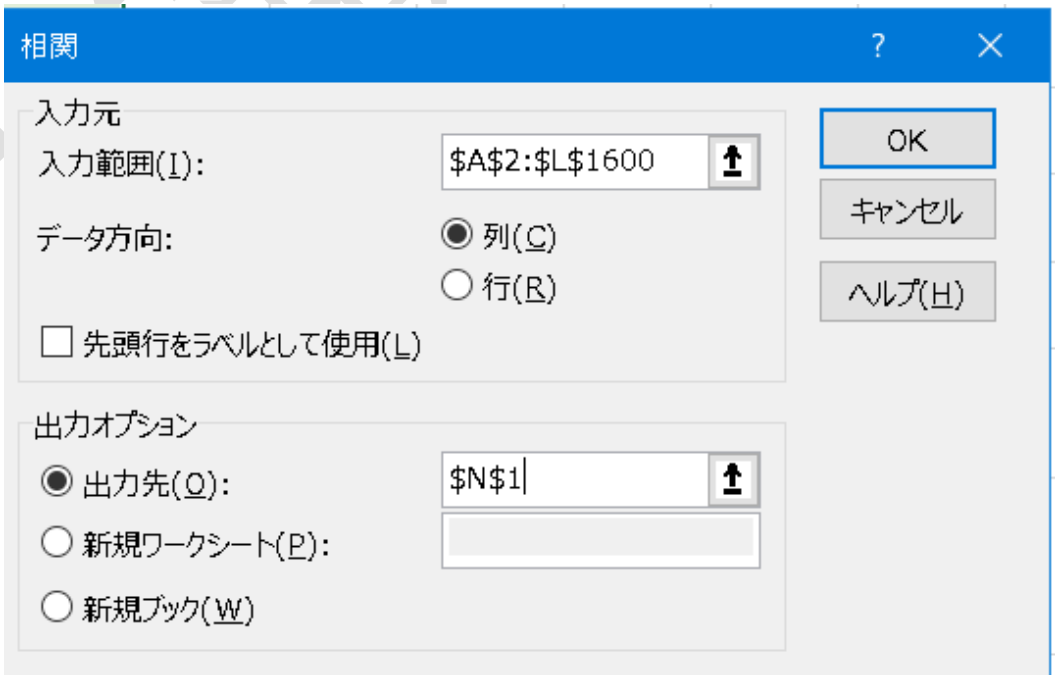
という因果関係を示しているものではありません。

例：ワインデータの相関を計算してみましょう。

CORREL 関数を用いて、相関を得ることができます。

	A	B	M	N	O	P
1	A	B			A	B
2	7.4	0.7		A	=CORREL($A$2:$A$1600,A2:A1600)	
3	7.8	0.88		B	=CORREL($A$2:$A$1600,B2:B1600)	=CORREL(B2:B1600,B2:B1600)

また、相関マトリックスを作成するのは変数の数が多いと手間がかかります。よって、メニューバー ＞ データ ＞ データ分析 ＞ 相関(共分散)を用いて、簡単に相関マトリックスを作ることができます。



	A	B	C	D	E	F	G	H	I	J	K	評価
A	1.00											
B	-0.25	1.00										
C	0.67	-0.55	1.00									
D	0.11	0.00	0.14	1.00								
E	0.09	0.06	0.20	0.05	1.00							
F	-0.15	-0.01	-0.06	0.19	0.01	1.00						
G	-0.11	0.08	0.03	0.20	0.05	0.67	1.00					
H	0.67	0.02	0.37	0.36	0.20	-0.02	0.07	1.00				
I	-0.68	0.23	-0.54	-0.08	-0.26	0.07	-0.06	-0.34	1.00			
J	0.18	-0.26	0.31	0.00	0.37	0.05	0.04	0.15	-0.20	1.00		
K	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	
評価	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.18	-0.17	-0.06	0.25	0.48	1.00

1.3.5 分布の形状に関する要約統計量

分布の形状の度合いについての要約統計量を見てみましょう。

歪度

分布の歪の度合いを表す歪度(skew)は

$$skew = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}$$

で表すことができます。歪度がゼロであると左右対称の分布となります。歪度が正の値ですと、右にすそ野が長くなります。これはx_iの平均との差の３乗が正となることから平均よりも大きいほうに偏りがあることが分かります。負の値ですと平均よりも小さいほうに偏りがあります。

尖度

尖度(kurt)は分布の中心の尖り具合、すそ野の厚さを表します。

$$kurt = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}$$

正規分布の尖度は３です。これは発案者であるカール・ピアソンの提案にしたがっています。また、エクセルなどではゼロになります。注意をしましょう。尖度が正の値になると分布は正規分布よりも、中心の尖り具合が強く、すそ野が厚くなります。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	歪度	0.98	0.67	0.32	4.54	5.68	1.25	1.52	0.07	0.19	2.43	0.86	0.22
2	尖度	1.13	1.23	-0.79	28.62	41.72	2.02	3.81	0.93	0.81	11.72	0.20	0.30
3		A	B	C	D	E	F	G	H	I	J	K	評価
4		7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

分布形状についてのエクセル関数

	A	B
1	歪度	=SKEW(B4:B1602)
2	尖度	=KURT(B4:B1602)
3		A
4		7.4

要約統計量

- 一変量要約統計量
  - 平均、中央値、最頻値、幾何平均、
  - 分散、標準偏差、
  - 最大値、最小値、四分位数、範囲等
- 二変量要約統計量
  - 相関、共分散等
- 分布の形状に関する要約統計量
  - 尖度、歪度

練習問題 1.1:　ワインデータから適当に要素を選び、頻度図を描いてみましょう。

練習問題 1.2:　ワインデータから評価とその他の要素の散布図を描いてみましょう。

練習問題 1.3:　ワインデータについて、エクセルのデータ分析/分析ツールを用いて各種要約統計量を計算してみましょう。

練習問題 1.4:　ワインデータのそれぞれの要素にはローマ字が用いられています。実際の要素の名称を用いずに記号が用いられている理由は何でしょうか？

練習問題 1.5:　分散は要約統計量、基本統計量の１つだと紹介しました。それは量なのでしょうか？割合なのでしょうか？それとも何か別のものなのでしょうか？

練習問題 1.6：分散と標準偏差を比べて分散を用いる利点は何でしょうか？

練習問題 1.7：共分散と相関を比べて共分散を用いる利点は何でしょうか？

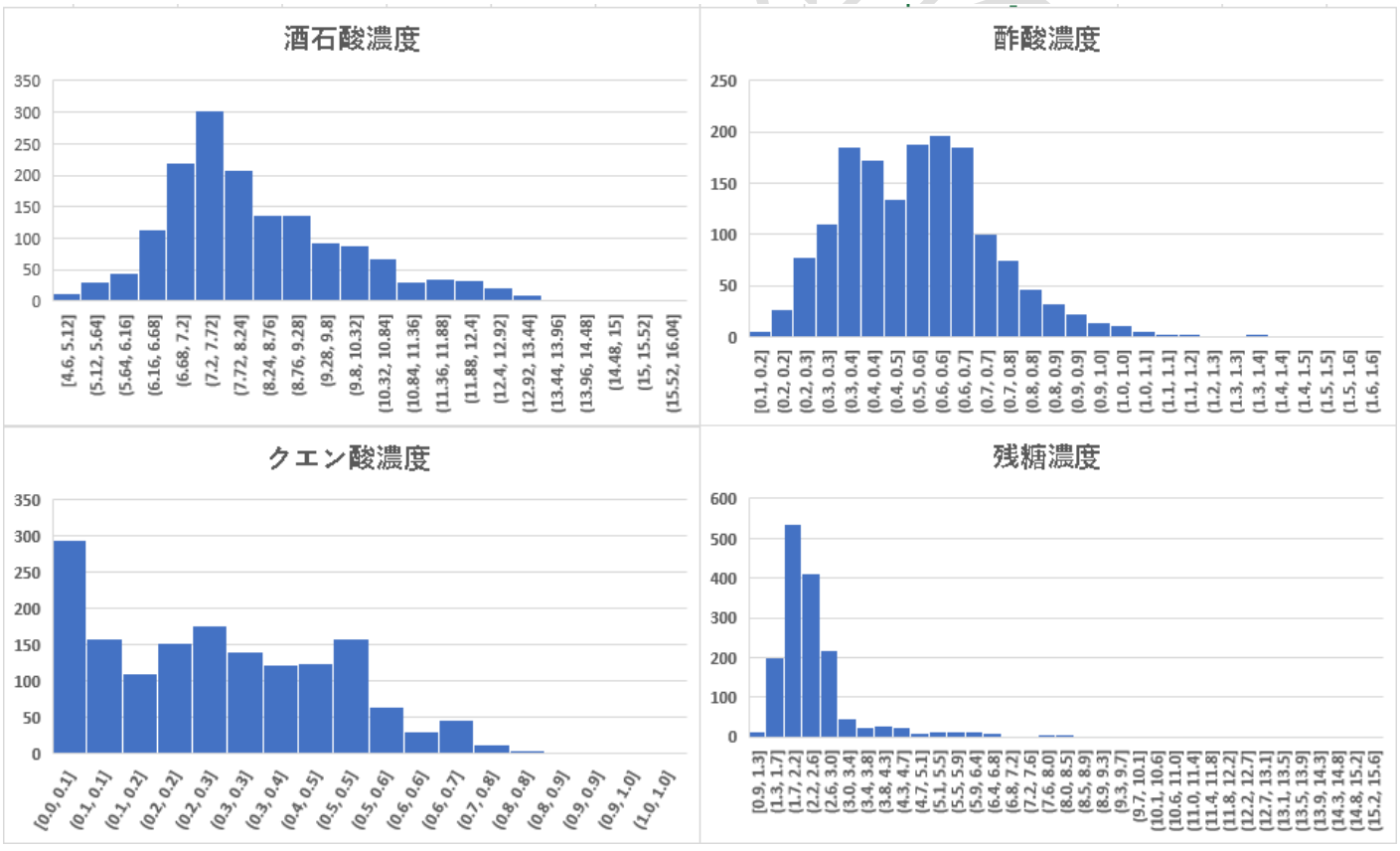
練習問題 1.8：歪度は偏差の３乗、尖度は偏差の４乗を用いています。それはなぜでしょうか？また、これは平均、分散と何か共通点があるのでしょうか？

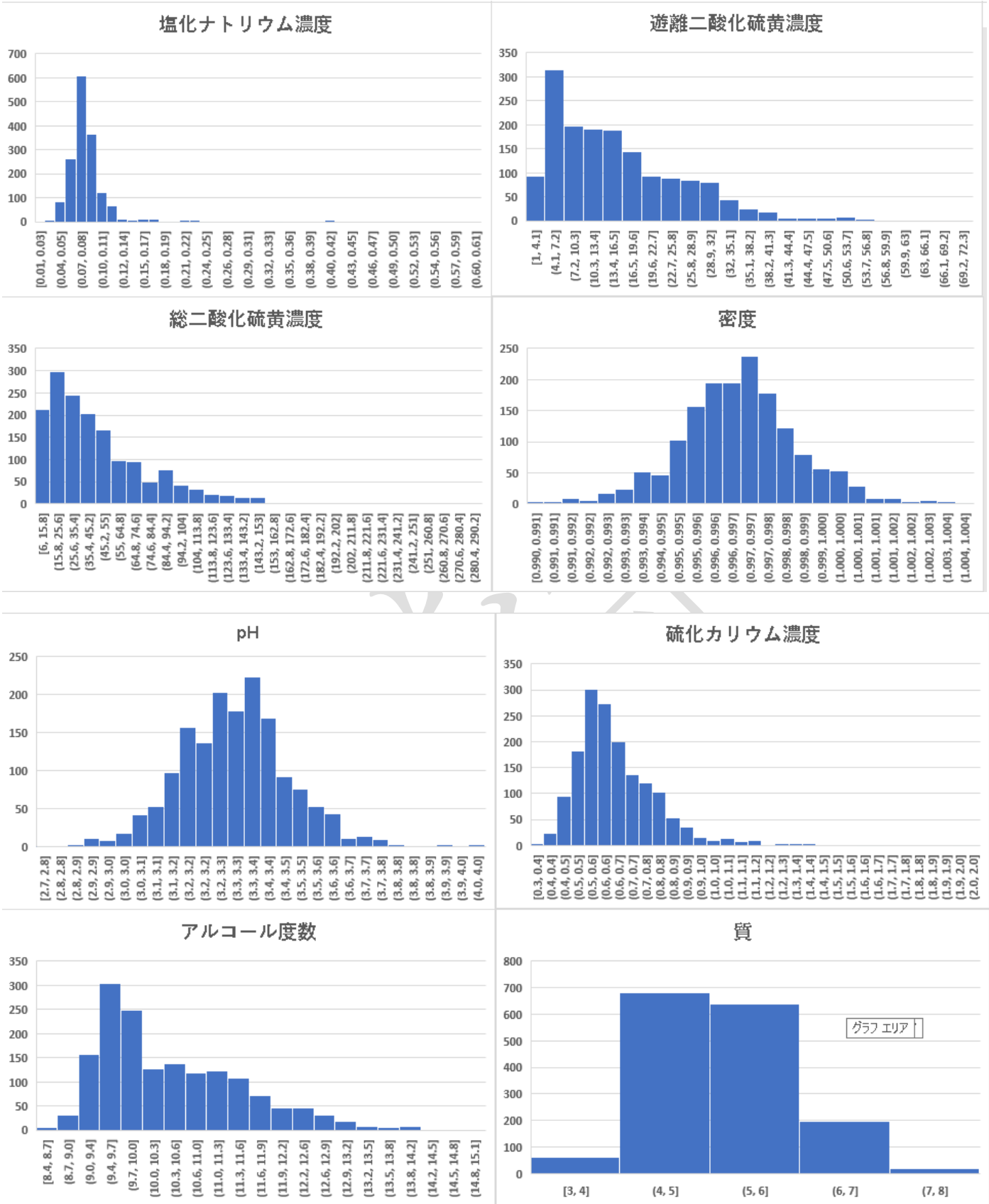
練習問題 1.9：要約統計量を用いる利点と欠点は何ですか？

第９章　練習問題の解

第1章

練習問題 1.1　ワインデータから適当に要素を選び、頻度図を描いてみましょう。

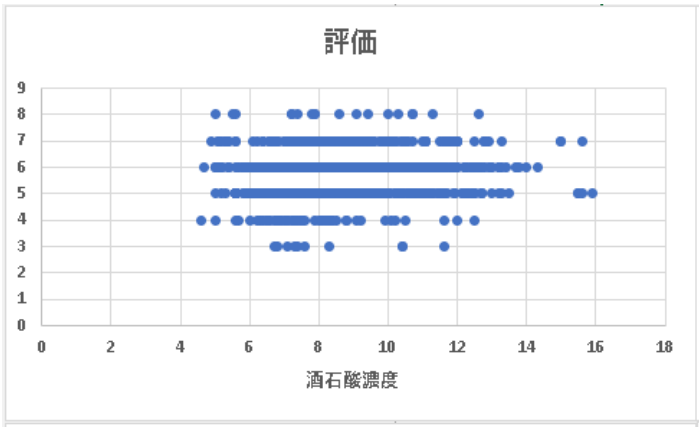


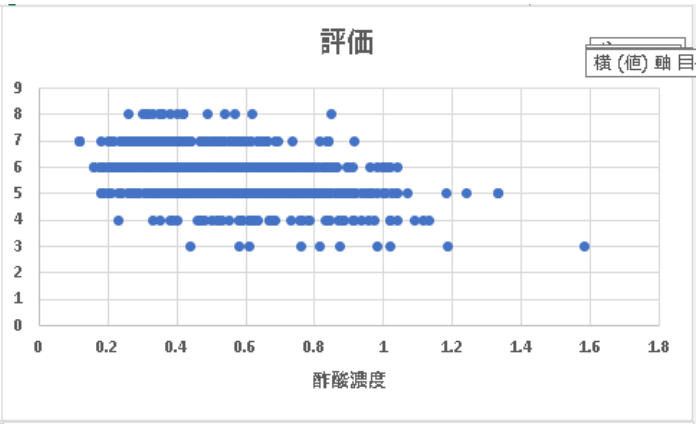


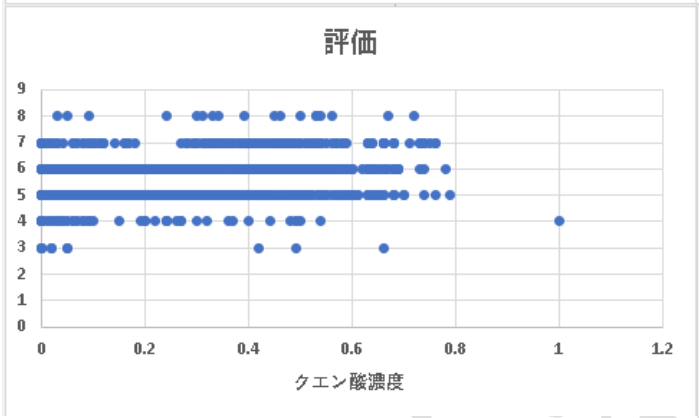
winequality_red>ヒストグラム

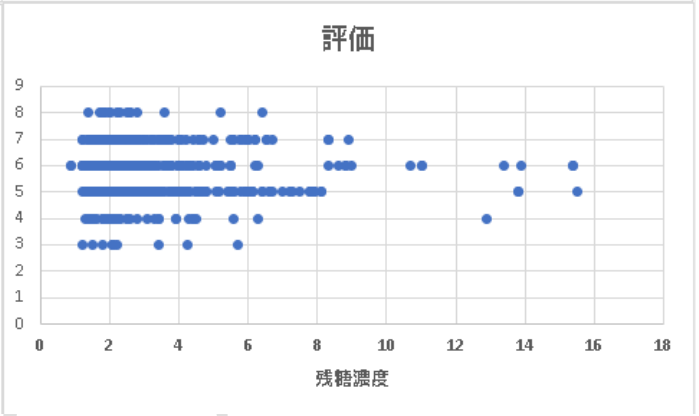
実際の分布は正規分布のようにベル型になる分布が少ないことが分かります。しかし、大まかな傾向はつかめるので分析の出発点を探るには有効な手段です。

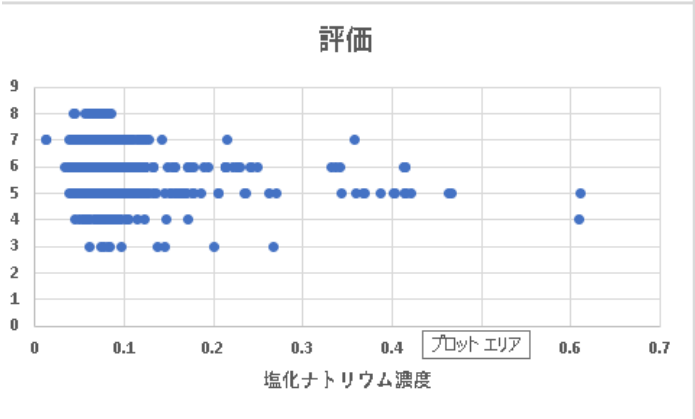
練習問題 1.2　ワインデータから評価とその他の要素の散布図を描いてみましょう。

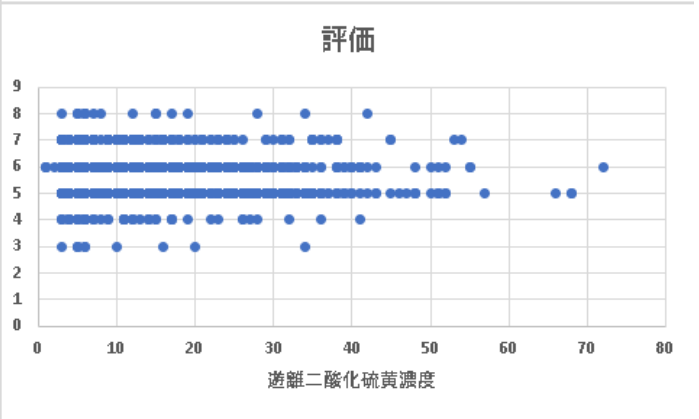


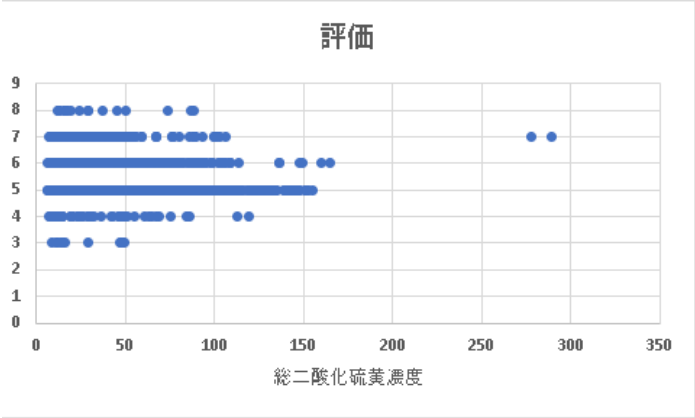


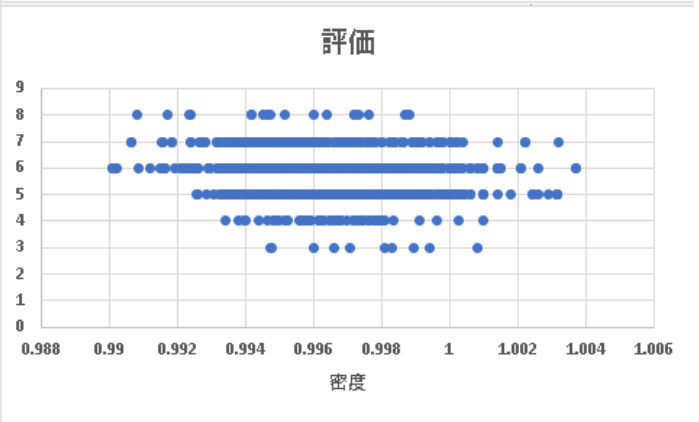


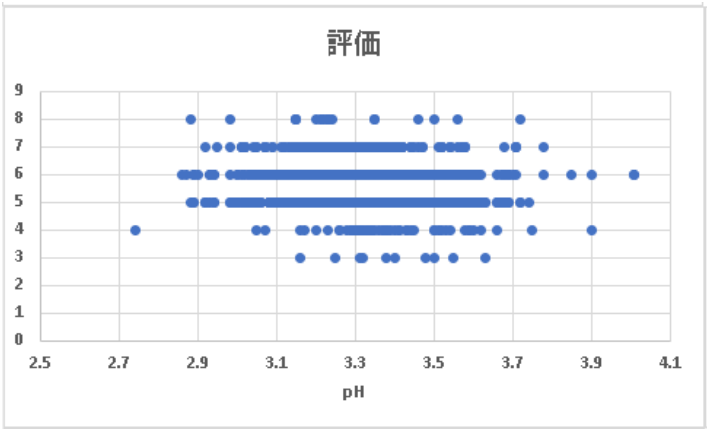


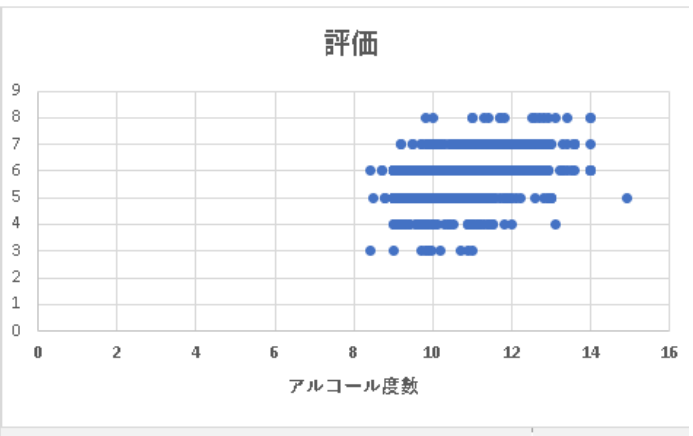


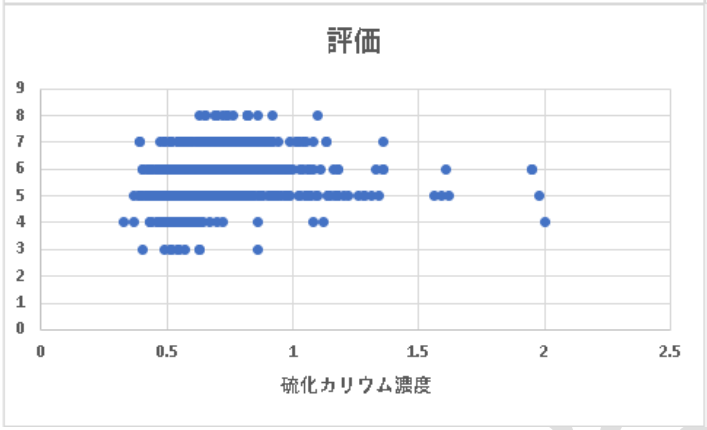








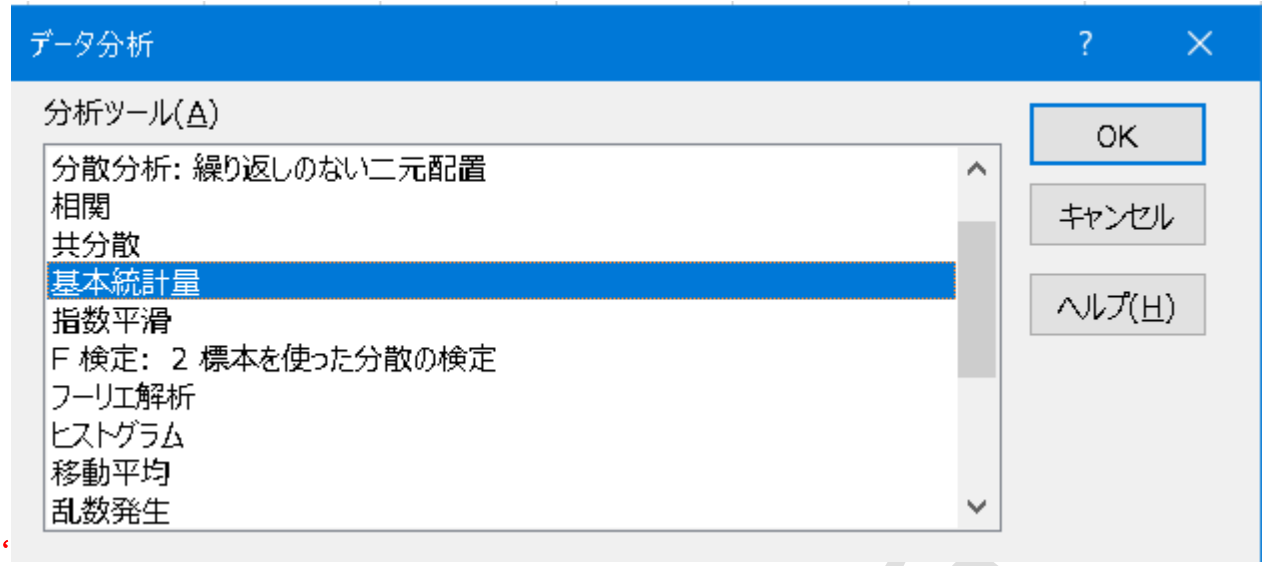




散布図はデータ分析で最も使う頻度の多い可視化ツールですが、明確な傾向が出ることは少ないことが分かります。大まかな傾向はつかみやすいので、分析の出発点となります。

練習問題 1.3　ワインデータについて、エクセルのデータ分析/分析ツールを用いて各種要約統計量を計算してみましょう。

エクセルの”データ分析”ツールは、スプレッドシート上のエクセル関数を用いるよりも効率が良いものが多数あります。



入力範囲にデータのレンジを入力して OK を押すと直ちに結果を出力してくれます。また、出力先も指定できます。

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1		酒石酸濃度	酢酸濃度	クエン酸濃度	残糖濃度	塩化ナトリウム濃度	遊離二酸化硫黄濃度	総二酸化硫黄濃度	密度	pH	硫化カリウム濃度	アルコール度数	評価
2													
3	平均	8.32	0.528	0.271	2.539	0.087	15.88	46.47	0.997	3.311	0.658	10.42	5.636
4	標準誤差	0.044	0.004	0.005	0.035	0.001	0.262	0.823	5E-05	0.004	0.004	0.027	0.02
5	中央値 （メジアン）	7.9	0.52	0.26	2.2	0.079	14	38	0.997	3.31	0.62	10.2	6
6	最頻値 （モード）	7.2	0.6	0	2	0.08	6	28	0.997	3.3	0.6	9.5	5
7	標準偏差	1.741	0.179	0.195	1.41	0.047	10.46	32.9	0.002	0.154	0.17	1.066	0.808
8	分散	3.031	0.032	0.038	1.988	0.002	109.4	1082	4E-06	0.024	0.029	1.136	0.652
9	尖度	1.132	1.226	-0.789	28.62	41.72	2.023	3.809	0.934	0.807	11.72	0.2	0.297
10	歪度	0.983	0.672	0.318	4.541	5.68	1.251	1.515	0.071	0.194	2.429	0.861	0.218
11	範囲	11.3	1.46	1	14.6	0.599	71	283	0.014	1.27	1.67	6.5	5
12	最小	4.6	0.12	0	0.9	0.012	1	6	0.99	2.74	0.33	8.4	3
13	最大	15.9	1.58	1	15.5	0.611	72	289	1.004	4.01	2	14.9	8
14	合計	13303	844	433.3	4060	139.9	25385	74303	1594	5294	1052	16666	9012
15	データの個数	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599
16	信頼度(95.0%)	0.085	0.009	0.01	0.069	0.002	0.513	1.614	9E-05	0.008	0.008	0.052	0.04

winequality_red>一変量要約統計量

練習問題 1.4:　ワインデータのそれぞれの要素にはローマ字が用いられています。実際の要素の名称を用いずに記号が用いられている理由は何でしょうか？

ワインデータセットはイタリアのワイン製造現場での質の改善を目的に、収集されたワインの化学分析の結果です。データ収集・分析の過程で様々な考察と議論を得てすでに取捨選択されたいるデータセットです。ですので後は客観的な分析をするのみなので、むしろ成分に惑わされることなく、統計的に分析されるべきです。

練習問題 1.5:　分散は要約統計量、基本統計量の１つだと紹介しました。それは量なのでしょうか？割合なのでしょうか？それとも何か別のものなのでしょうか？

分散の計算の仕方を見るとまず偏差を求めてそれを２乗して、その総和を求めて、データの数で割っています。したがって、分散は二乗偏差の平均値です。

練習問題 1.6：分散と標準偏差を比べて分散を用いる利点は何でしょうか？

幾つかの棒を手に入れてその長さを測り平均を求めるとそれは長さの平均値です。それを二乗するとそれは面積に相当します。したがってその２つの数値を直接比べることはできません。そこで分散の平方根を取って次元を長さの単位に戻します。そうすると直接比べることができるようになります。しかし、そうすることによる損失はどのようなものでしょうか？１つは計算時間の問題です。計算の量か増えてしまいます。これは大量のデータ処理をする際には問題になることがあります。つぎに計算精度の問題です。平方根を取ることである程度の丸目の誤差が出ます。３つ目は、わずかなデータの変化を取ら得たい場合には、分散の方がその差が２乗されてい

るので、大きく反映されます。

練習問題 1.7：共分散と相関を比べて共分散を用いる利点は何でしょうか？

共分散と相関についてのこの議論にも練習問題 1.6 が関連しています。また、２つのデータの大きさに桁違いがある場合などには、共分散の変化がどちらの変数が原因であるかが、つかめる場合があります。

練習問題 1.8：歪度は偏差の３乗、尖度は偏差の４乗を用いています。それはなぜでしょうか？また、これは平均、分散と何か共通点があるのでしょうか？

偏差は平均との差として求められます。そして、その偏差の２乗和の平均が分散、３乗和の平均が歪度、４乗和の平均が尖度です。偏差を偶数で累乗すると平均について対称性をもちます。一方奇数での累乗では対称性を持ちません。したがって、非対称の度合いを見たい場合は奇数の累乗を用います。累乗のｎが大きくなると、大きな偏差がより強調されるようになります。

練習問題 1.9：要約統計量を用いる利点と欠点は何ですか？

何かを観測してデータを得る場合には、観測した人、観測に用いた機器、観測の時刻などを記載しておくことが重要です。得られたデータを要約する際にはこのような要素からの影響が排除されている必要があります。このようなデータでもひとつひとつの観測値を見たのではどのような意味がかるかはつかみにくいものです。それを要約統計量として示すとデータのもつ性質を要約統計量として一つの数値として表すことができます。しかし、同時に多くの特徴が捨て去られてしまいます。捨てるデータの特徴よりも、要約統計量とした方が便利である場合にだけそれは用いられます。