

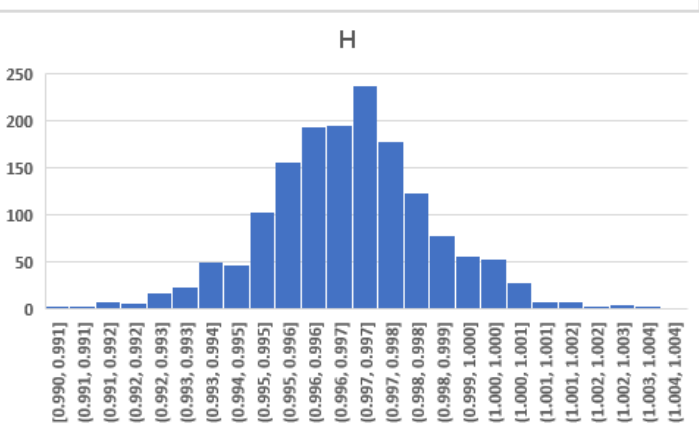
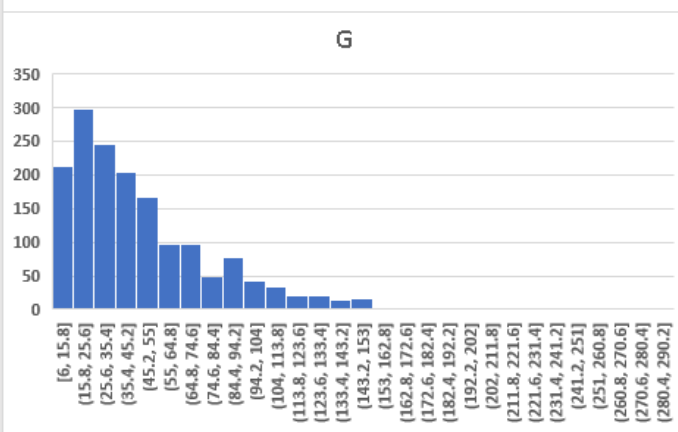
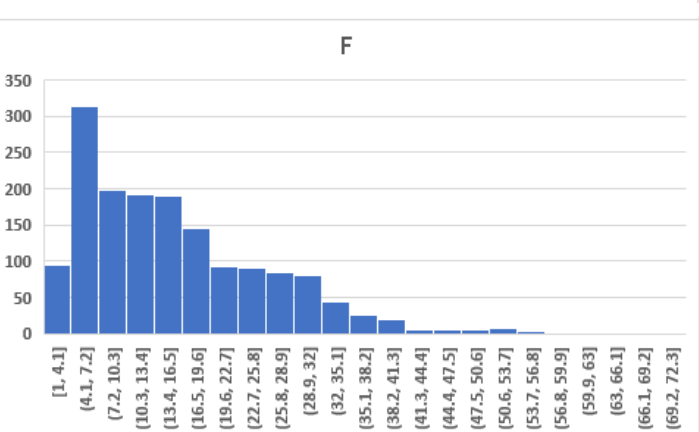
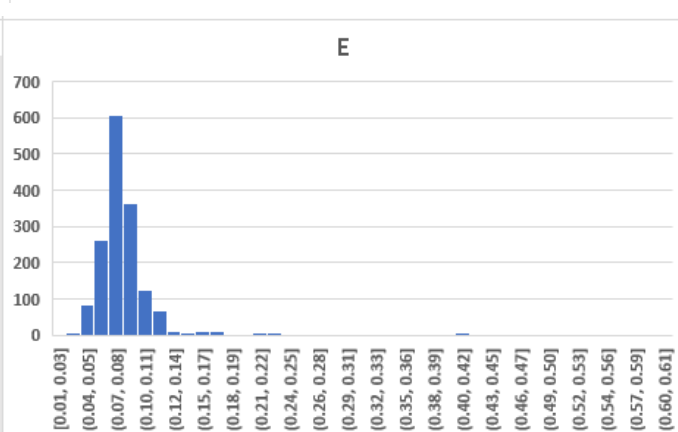
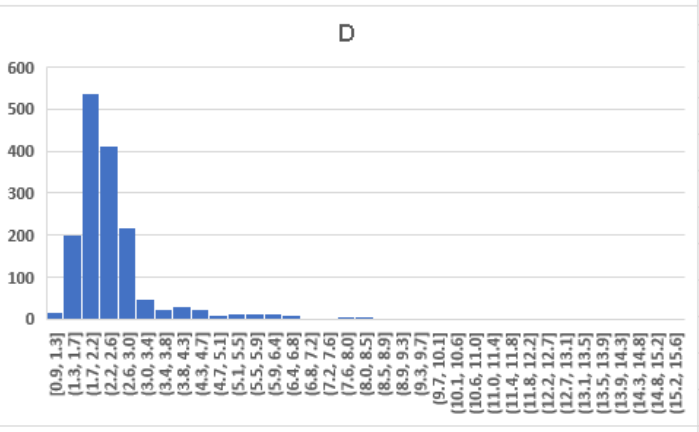
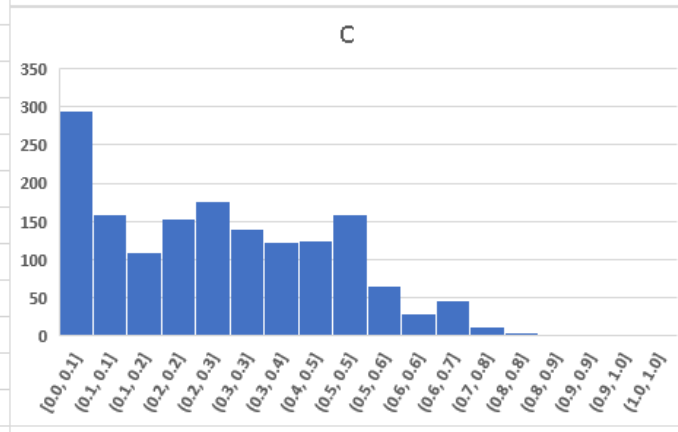
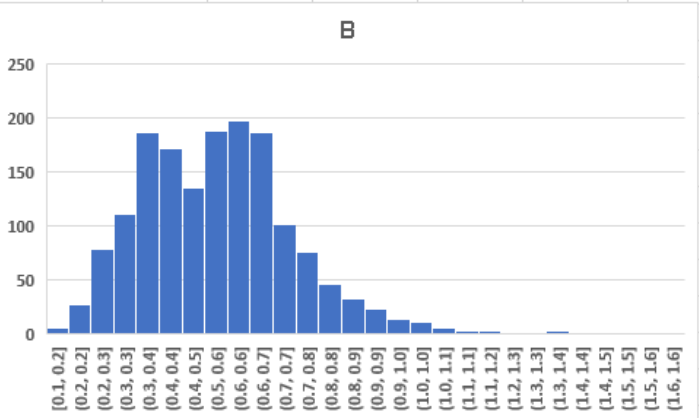
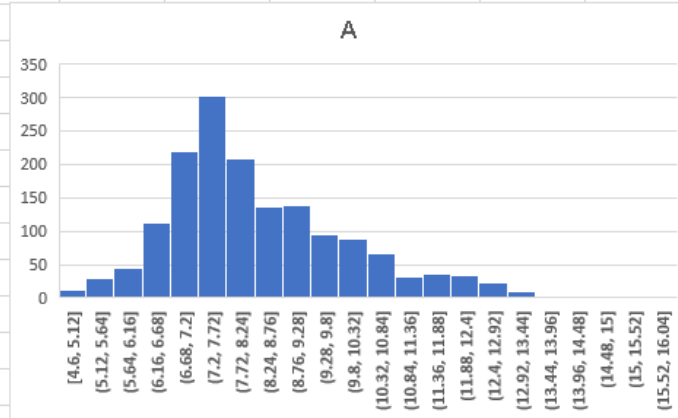
## 第1章

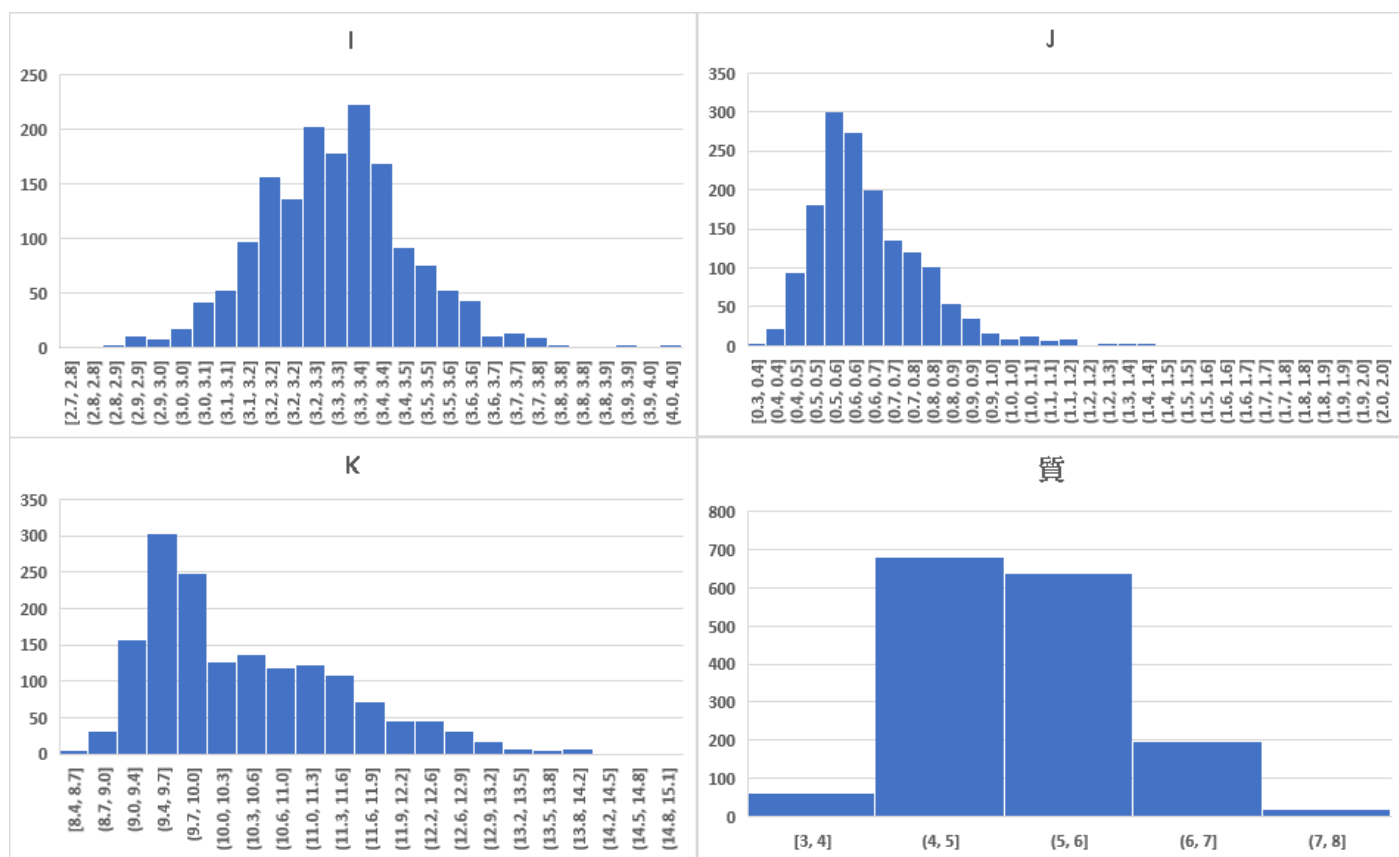
第1章ではデータの要約の方法を主に学びました。頻度図と散布図という2つのグラフについて学びました。また、平均、分散、標準偏差などの基本統計量についても学びました。ポルトガルの生産者のために作られた赤ワインの評価と化学成分のデータ分析を用いて可視化と基本統計量について学びました。基本統計量の中でも‘平均’に多くの人はなじみが深いと思われます。しかし、統計学として‘平均’が用いられるようになったのはごく最近のことです。18世紀と19世紀のはざまでケトレーが平均の概念を明記しています。それ以前は最大値、最小値、最頻値、中央値、範囲などが用いられていたようです。‘平均’の概念は紀元前280年にはピタゴラス学派により議論されていました。当時すでに算術平均、幾何平均、調和平均が知られていましたが、それは統計学としてではなく、哲学的な意味においてでした。最頻値が最初に登場するのは紀元前428年です。物語の中で敵の城壁に梯子を掛けるときにその梯子の長さを予測の際に最頻値が用いられています。近代統計学が登場するまでの間、人びとは最大値のような極端な量を統計量として使っていたようです。大きい数値を提示することで、人びとの注目を集めようとしたのかもしれませんが。これはデータの中の一点を選んで、それ以外のデータを捨てていたことに相当します。そうすることでデータから最大の情報を引き出そうとしていたのです。ちなみに1フィートという長さの単位が法律で定められた時の様子がコベールの銅版画として残されています。ここでは1フィートが16人の市民の足の長さを測定して決められています。16人の足の長さの合計を16フィートと定め、これを1ロッドと定めたのです。

では、ワインデータに関する練習問題をとおして最近のデータ分析の方法を学んでいきましょう。

**練習問題 1.1:** ワインデータから適当に化学成分を選び、頻度図を描いてみましょう。

A, B, …といったローマ字は化学成分を表しています。縦軸は頻度を表しています。赤ワインの銘柄数は1599です。横軸は化学成分の濃度や密度です。



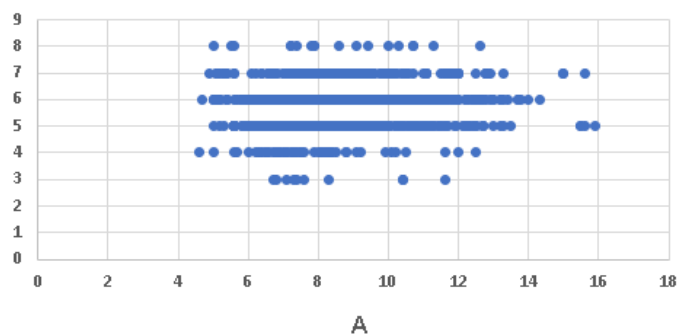


実際の分布はベル型になることが少ないことが分かります。しかし、大まかな傾向はつかめるので分析の出発点を探るには有効な手段です。エクセルシートは練習問題 1. 1、エクセルシートの作り方は第 3 部参照

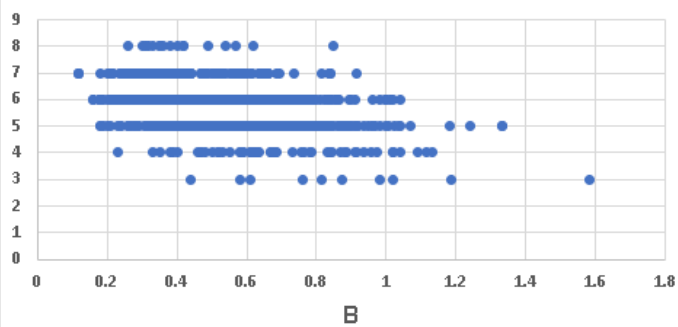
**練習問題 1. 2:** ワインデータから適当に化学成分を選び、その評価との関係を散布図として描いてみましょう。

ポルトガルのワイナリーではより評価の高い赤ワインを作ることが目的なので、評価を高くするワインの化学成分が何で、何をどのように調整したらよいのかを知ることが分析の目的です。そうするためには評価と化学成分との間の明確な傾向を見つけることが目的となります。散布図の右上が高く、左下が低いか  $\nearrow$ 、またはその逆の傾向  $\nwarrow$  があればよいことになります。見た感じでそのような傾向にあるのは、B, D, I, J, K です。これは人びとの経験と知識に大きく影響される部分であり、正解はありません。散布図を見るときに注意することは頻度の低いデータはばらつきの幅も狭くなるということです。

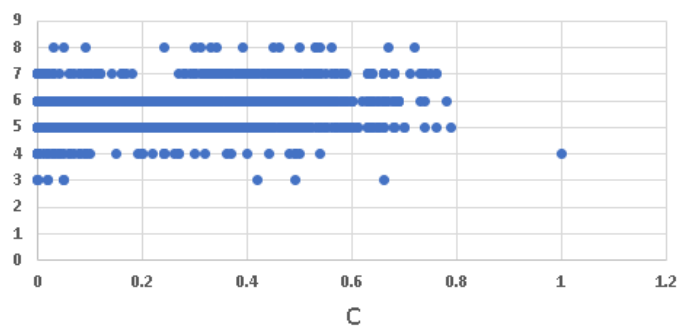
評価



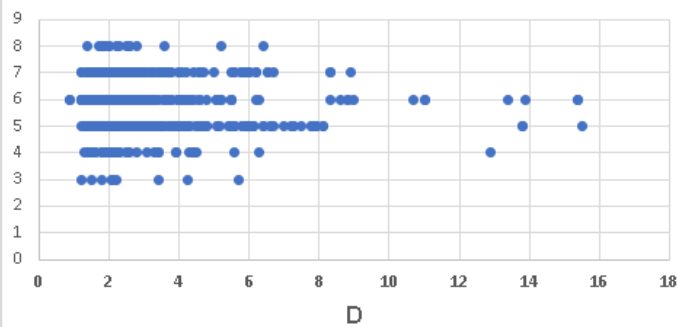
評価



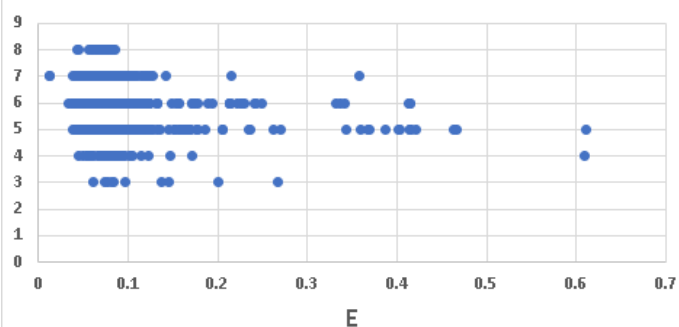
評価



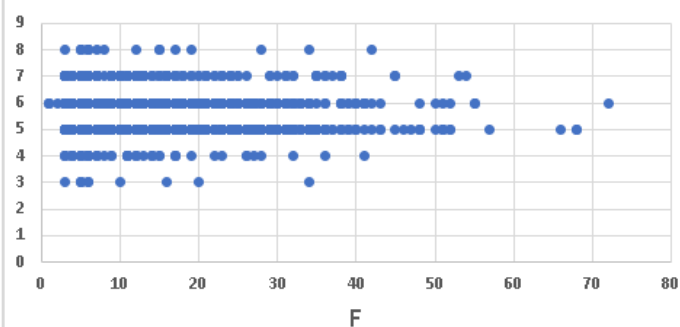
評価



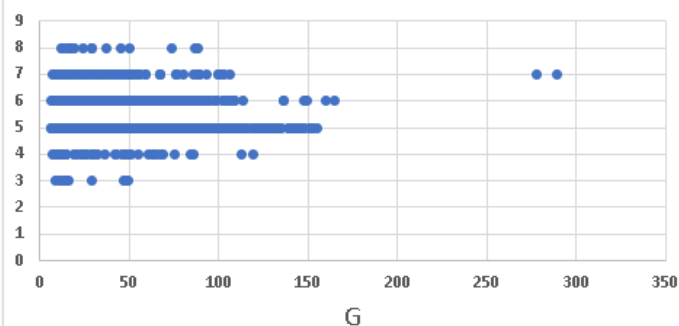
評価



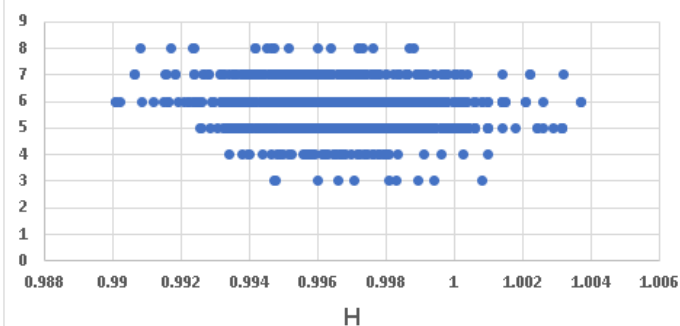
評価

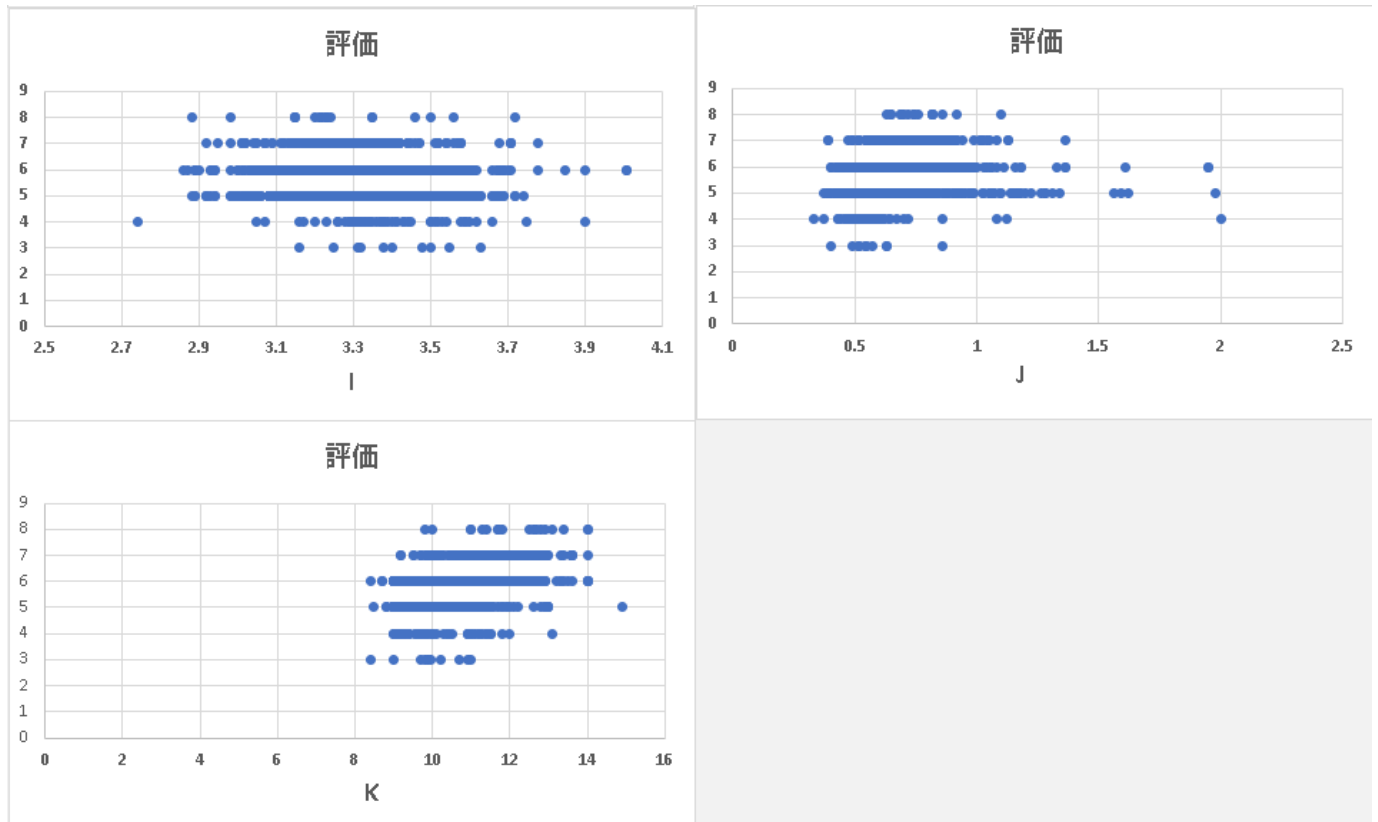


評価



評価





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。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。頻度図と散布図の両方を見比べることも傾向を見極めるヒントになります。エクセルシートは練習問題 1.2、エクセルシートの作り方は第 3 部参照

**練習問題 1.3:** ワインデータのそれぞれの化学成分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、ポルトガルのワイナリーのためにワインの高い評価を得るための指針を提供するために作成されました。ワインの特定の化学成分とワインの評価で構成されていますが、その評価方法と特定の成分は、長年による経験と試行錯誤により決められたものです。また、その運営組織は様々な議論を得て設立されています。用いられている化学成分は十分な検討の後に選ばれたものであり、また、評価は公平性が十分に考慮されていると考えられます。ですので化学成分に関する事前の知識と経験に惑わされることなく、客観的、統計的に選ばれたデータを分析するために、化学成分を伏せました。

**練習問題 1.4:** 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 1.5:** 分散と標準偏差を比べてそれぞれを用いる利点は何でしょうか？

分散は偏差の二乗和をそのデータ数で割って平均を取ったものです。そして標準偏差はその平方根を取ったものです。分散は二乗和の平均であるためにその単位は変数の単位の二乗になります。変数が長さの単位をもつと

すると、それを二乗するので分散は面積に相当します。したがって分散と平均の2つ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。分散は理論計算に積極的に用いられます。変数が独立であれば、 $\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$  が成り立ちます。これを分散の加法性といいます。

#### 練習問題 1.6 : 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 1.5 が関連しています。共分散は、解釈が難しい統計量の1つです。特に分散の値に桁違いがあるときにはその解釈が難しくなります。一方で、相関は共分散を2つの標準偏差で割ってあるので、指数化されています。共分散は、経済、金融の分野、特に投資に関するポートフォリオの構成を決めるときに大きな役割を果たしています。現代ポートフォリオ理論、資本資産価格モデルで共分散が用いられています。また、統計学、画像解析では、6章で扱う主成分分析に用いられています。

#### 練習問題 1.7 : 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乗の平均が分散、偏差の3乗の平均と標準偏差の3乗の比が歪度、偏差の4乗の平均と標準偏差の4乗の比が尖度です。偏差を偶数乗するとその値は対称性を持ちます。一方で奇数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乗を用います。また、 $n$ 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 1.8 : 要約統計量を用いる利点と欠点は何でしょうか？

ひとつひとつの観測値を見たのではどのように解釈するべきかは、つかみにくいものです。要約統計量は、データのもつ性質を要約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。

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、観測の時刻などの影響を受けます。データにはこのような情報を記載しておくことが重要です。紀元前3000年にはシュメール人によって、統計分析に相当する記録があることが分かっています。粘土板には統計分析に用いたデータの出所まで記述されています。当時の人はデータが誰によって測定されたものかをひとつの情報としてとらえていたことが分かります。

#### 練習問題 1.9 : 乱数を用いて正の完全相関、正の相関、無相関、負の相関、負の完全相関を散布図を用いて可視化してみましょう。

図は本文参照、エクセルシートは練習問題 1.9、エクセルシートの作り方は第3部参照

### 1章で扱わなかった1変量要約統計量

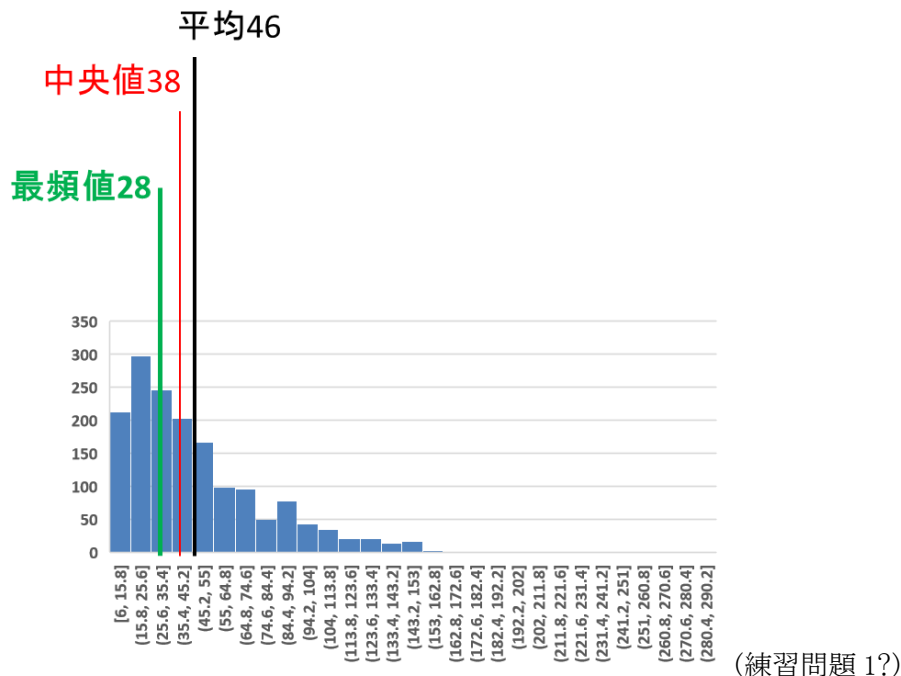
#### 中央値(メディアン)

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。このような基本統計量は、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。頻度図の山の頂点です。

化学成分 B の頻度図に中央値、平均値、最頻値を書き込んでみましょう。



この例のように平均、中央値、最頻値は、同じ値になるとは限りません。

## 幾何平均

広義の平均の 1 つとして、幾何平均があります。データ  $(x_1, \dots, x_n)$  の幾何平均とはつぎの式で定義されます。

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

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。成長率などの性質をもつ変数に使われます。ビジネスの分野では、売り上げの成長率とかに使われます。

## 最大値・最小値

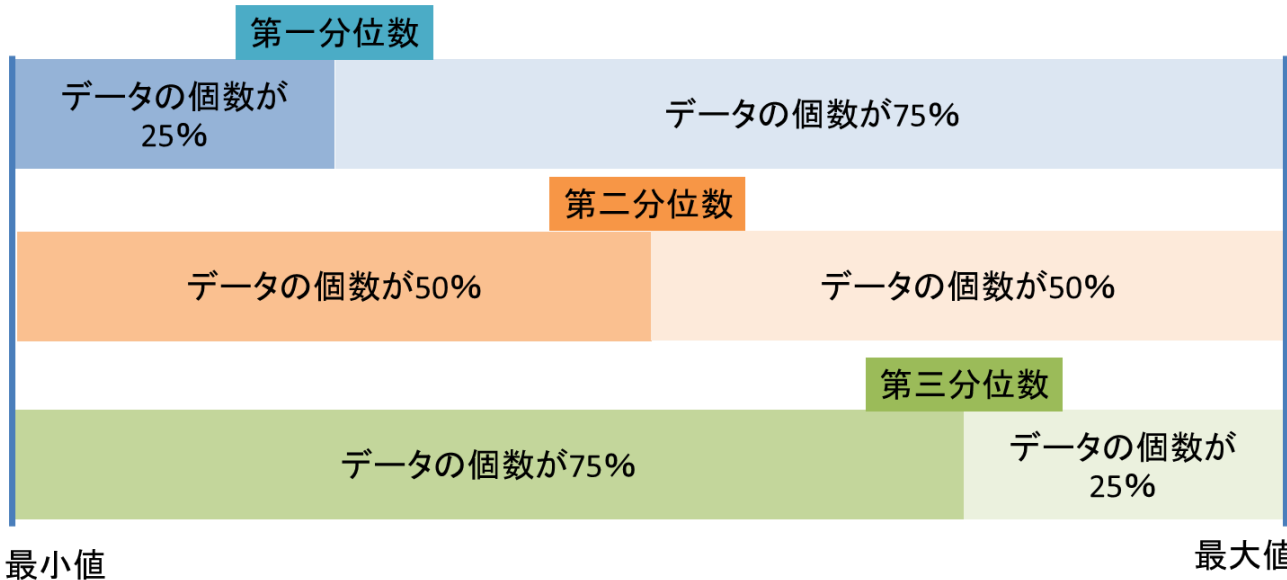
最大値は観測値のうちで最も大きな値、最小値はデータのうちで最も小さな値です。

## 分位数

$n$  個のデータを左から小さい順に並べ  $x_1 \leq x_2 \leq \dots \leq x_n$  とします。  $i/n = \alpha/100$  とすると、  $P_\alpha = (x_i + x_{i+1})/2$  よりも小

さな値をとるデータの割合が  $\alpha$  %、それよりも大きな値をとるデータの割合が  $100-\alpha$  %となります。このような  $P_\alpha$  を第  $\alpha$  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。

データを左から小さい順に並べる



範囲

最大値から最小値を差し引いた値です。

四分位範囲（しぶんいはんい）

データを左から小さ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。

ワインデータの一変量要約統計量を求めてみましょう。

	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

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、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。

## 要約統計量

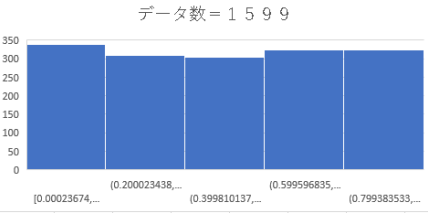
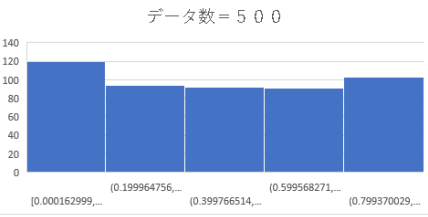
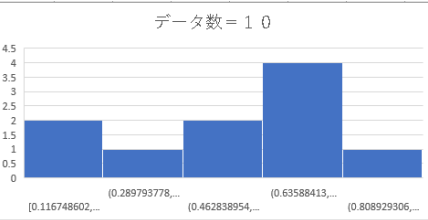
- 一変量要約統計量
  - 平均、中央値、最頻値、幾何平均、
  - 分散、標準偏差、
  - 最大値、最小値、四分位数、範囲等
- 二変量要約統計量
  - 相関、共分散等
- 分布の形状に関する要約統計量
  - 尖度、歪度

## 第2章

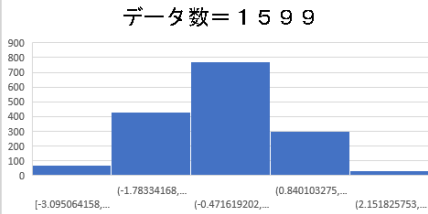
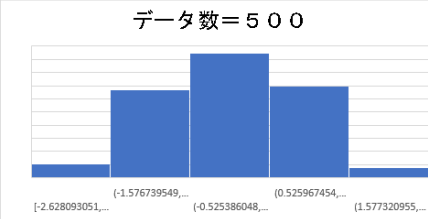
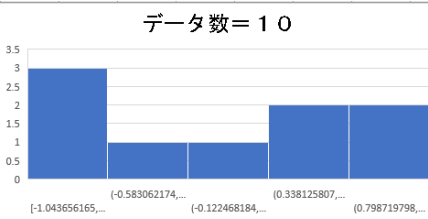
データ分析を学ぶ際に、厄介ないくつかの用語があります。模型、確率、確率変数、実現値(観測値)、乱数などです。模型(モデル)という用語は日常生活でも頻繁に用いられているために、意味が広く、データ分析で何を意味するのかはつかみにくいものです。同様のことが、確率、確率変数、ランダム、観測値などの用語にも起こります。模型というとか何か複雑なものを思い浮かべてしまいませんか？もっとも簡単な模型は第2章で学んだ確率分布です。一様分布、正規分布、t分布は模型です。確率変数は確率分布から抽出された値です。乱数も同様です。実現値は実際に得られたデータです。乱数はデータ分析で重要な役割を果たします。また、データ分析の学習にも便利な道具です。特に模型(モデル)の理解には欠かせない道具です。本書では乱数を積極的に用いてデータ分析を理解していきます。

**練習問題 2.1:** エクセルを用いて乱数を発生させ、頻度図を描きましょう。乱数は一様分布、ベルヌーイ分布、正規分布から発生させてみましょう。その際にデータ数を  $n=10, 500, 1599$  と変化させてみましょう。

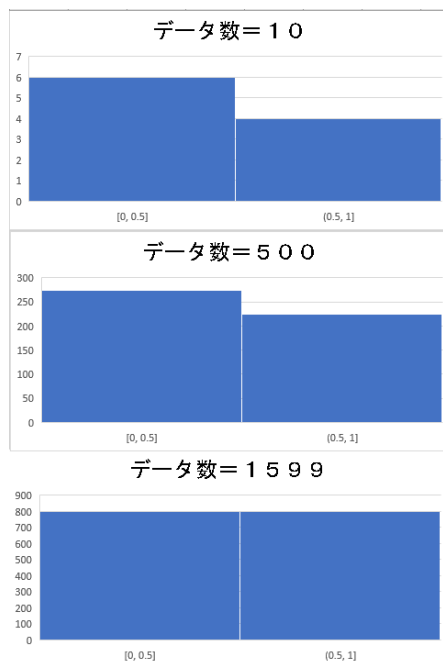
- 一様分布



正規分布



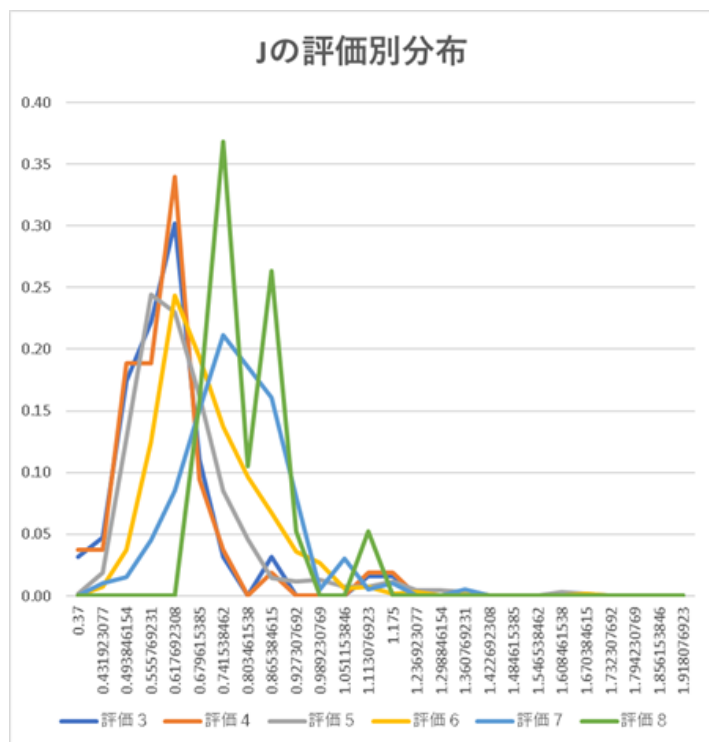
ベルヌーイ分布



F9 を押して乱数を生起すると乱数の数が小さいと分布の形状が安定しないことが分かります。

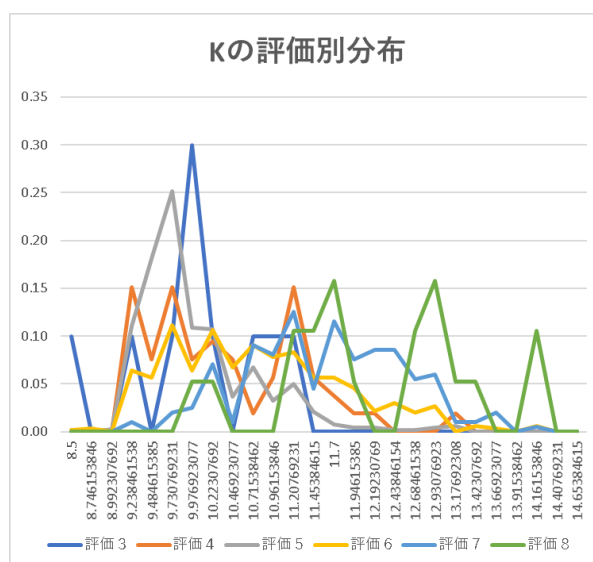
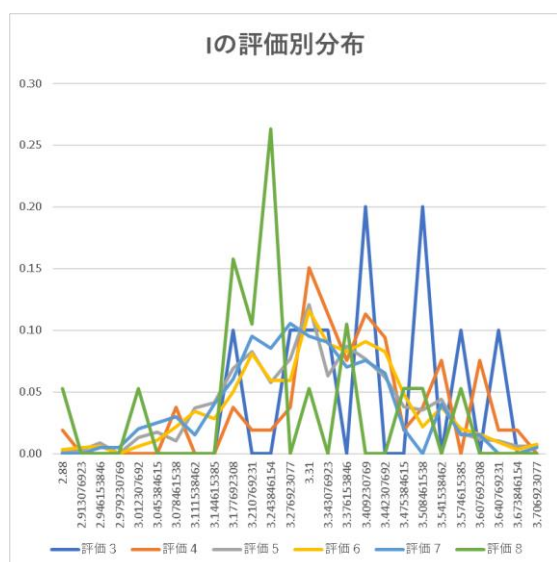
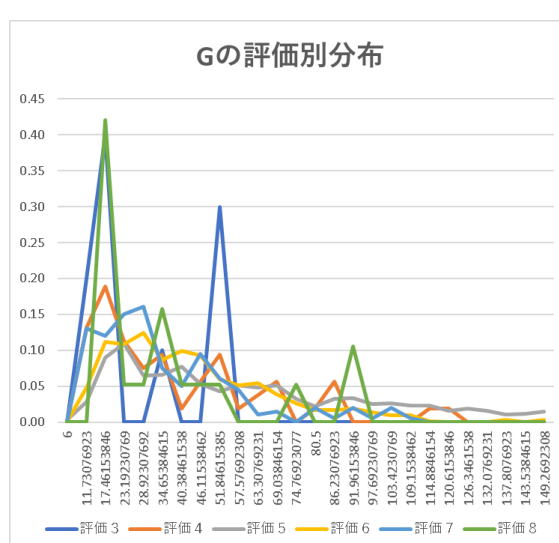
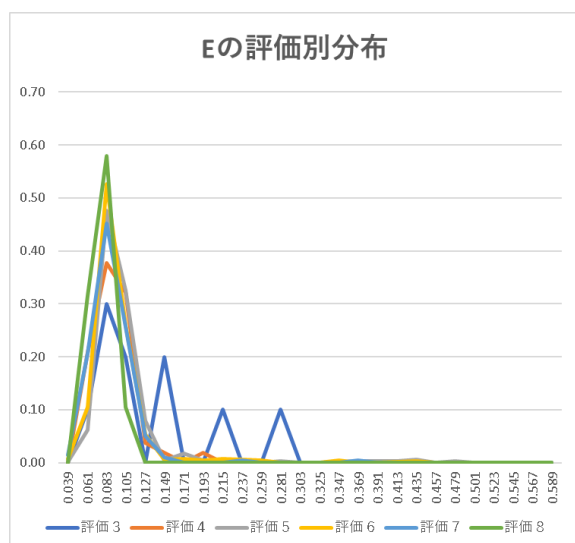
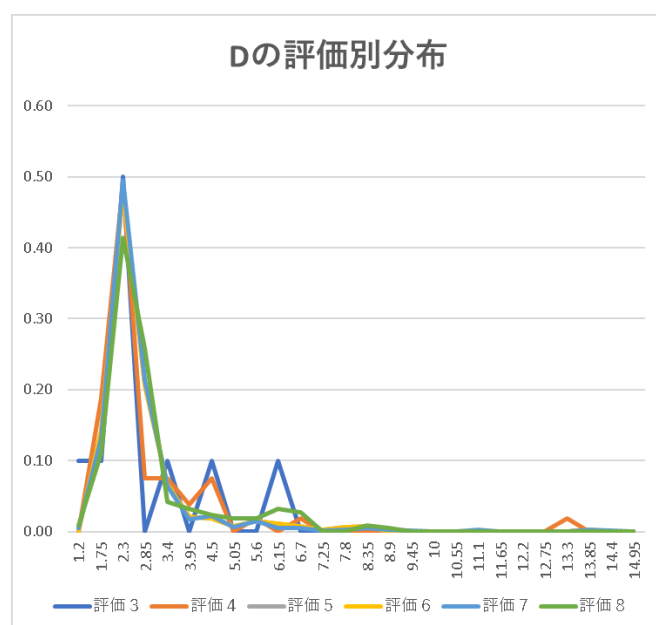
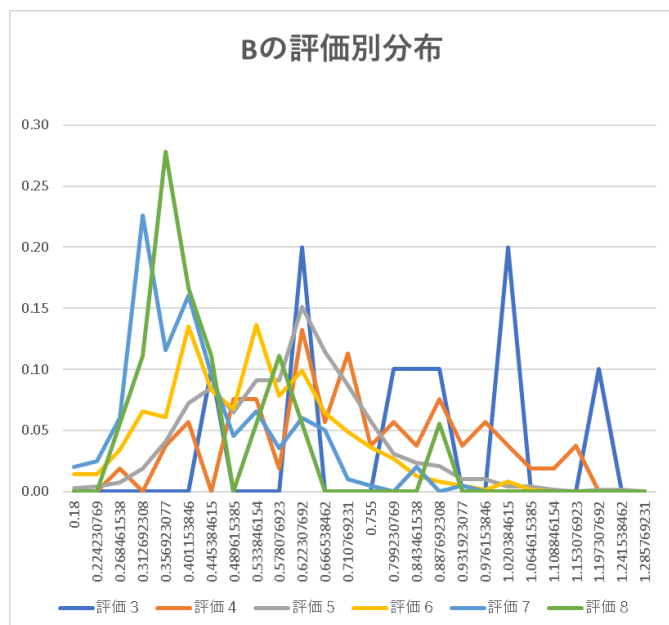
練習問題 2.2: 化学成分 J について評価別分布を作成してみましょう。

散布図から評価と化学成分 J の間には特別な関係がありそうなので、評価別の頻度図を作成してみました。分布の山が 2 つありそうなが見えてきます。



まず、図から得られるイメージは評価 3, 4, 5, 6 と 7, 8 は同じ分布ではないのかという疑問です。この疑問には統計学は適切な答えを与えてくれます。第 6 章で分散分析を扱います。

その他の主な化学成分についての評価別分布をつぎに示します。



練習問題 2.3: 赤ワインデータの10段階評価の標本空間と根元事象を示してみましょう。また、その違いを説明して見ましょう。標本空間と根元事象の概念を使って統計分析ができる条件は何でしょうか？

根元事象は1, 2, 3, 4, 5, 6, 7, 8, 9, 10それぞれの評価です。標本空間はこれらすべてです。

違いは、根元事象が最小単位の事象を表すのに対して、標本空間は全体の空間を表しています。

統計学では標本空間と根元事象が固定されていて変化しないことが前提です。

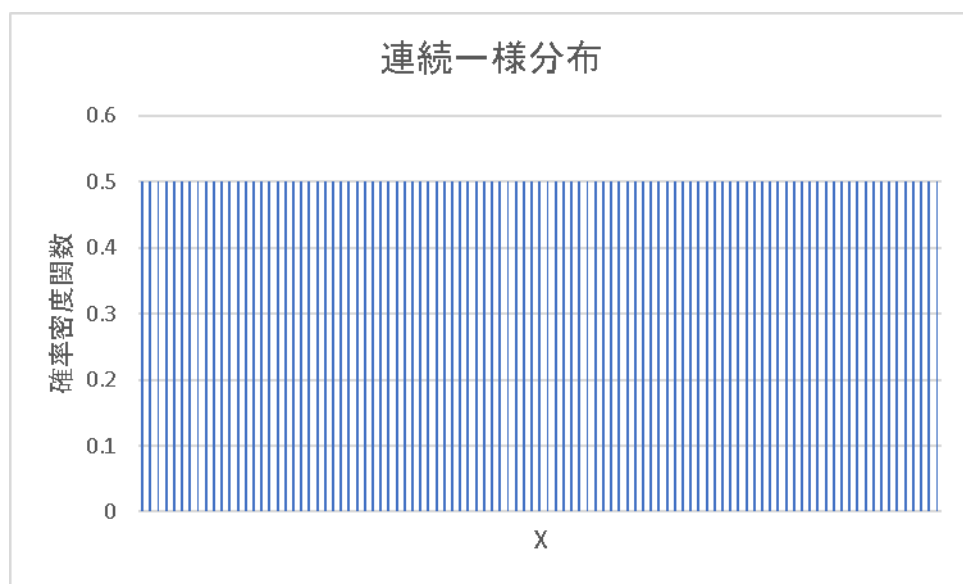
練習問題 2.4: 赤ワインデータについてどれが確率変数であるかを考察してみましょう。

確率変数であるのはそれぞれの化学成分の値ですが、観測された値は観測値となり確定値です。

練習問題 2.5: A と B という事象があつて、それが独立である場合と相関のない場合の違いについて説明してみましょう。

A と B が独立であるとは A と B が同時に生起する確率は  $P(A) \times P(B)$  となる事象です。どちらがどちらかを起こすという因果関係はありません。一方で相関は単なる平均的な関係を示しているだけです。

練習問題 2.6: 上限を 2，下限を  $-2$  として、連続一様分布を図で描いてみましょう。



直線で長方形として描く方法もありますが、連続変数であっても範囲で考える必要があるために、頻度図で描いてみました。

練習問題 2.7: 一様分布、正規分布についてバラツキとは何かについて考察してみましょう。

一様分布、または正規分布から得られた確率変数のばらつきを表すには、分散を用います。一様分布の分散は  $(b - a)^2/12$  で与えられます。b は上限, a は下限です。

練習問題 2.8: ベルヌーイ分布の例をあげてみましょう。

コイン投げの表裏の結果、サイコロを振った時の 6 が出るときとそうでないときの結果、野球の勝敗など

練習問題 2.9: 離散確率データの確率については理解しやすいです。(頻度 ÷ 頻度の総数) で得られます。連続確率変数の場合には分母の頻度の総数は無限になってしまいます。そう考えると確率はゼロになってしまいま

す。正しいでしょうか？

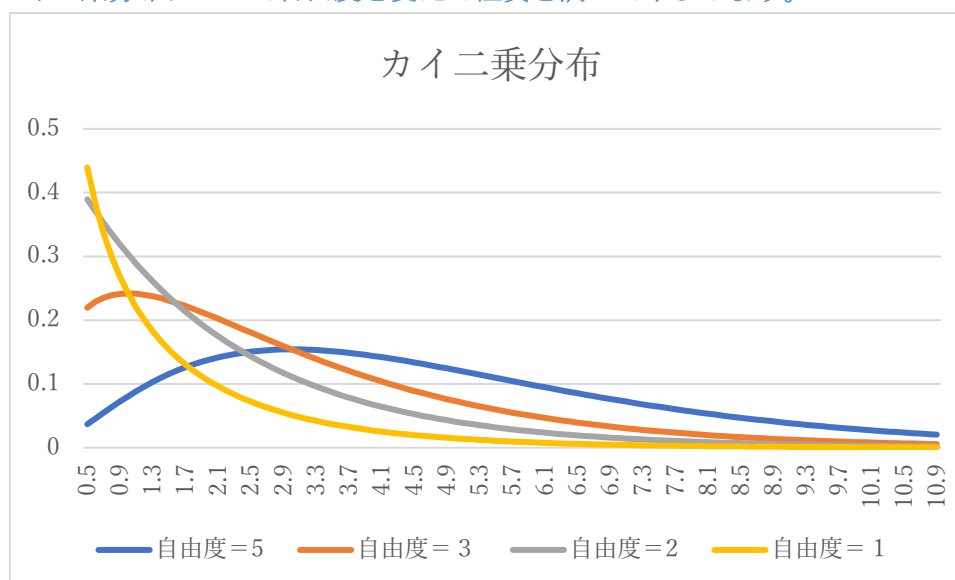
公平な6面体のサイコロの出る目は6分の1と明確です。サイコロに無限面あるとすると1面が出る確率はゼロになってしまいます。しかし、面の数を有限として1面が出る確率を考えれば確率は有限です。これは面を時間に変えても長さに変えても同じです。わずかでも有限の範囲が指定されればその範囲内で事象が生じる確率があります。連続確率変数の分布関数を確率密度関数と呼ぶ所以です。

## 第3章

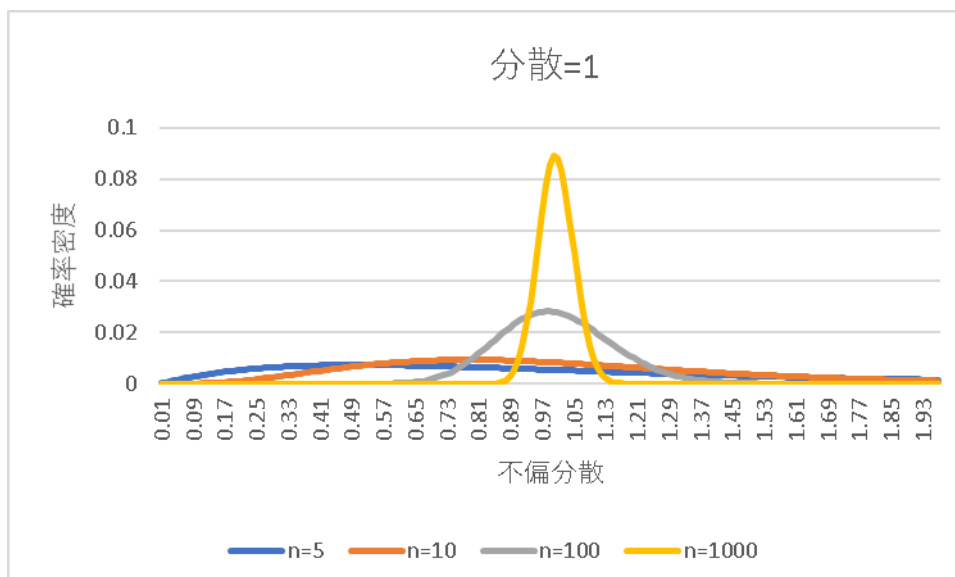
練習問題 3.1: 平均と期待値の違いを説明してみましょう。

平均は手ものに得られたデータ、観測値、実験結果などから得られた平均的な結果でしか過ぎませんが、期待値は確率分布をもとに計算されています。

練習問題 3.2: カイ二乗分布について自由度を変えて性質を調べてみましょう。

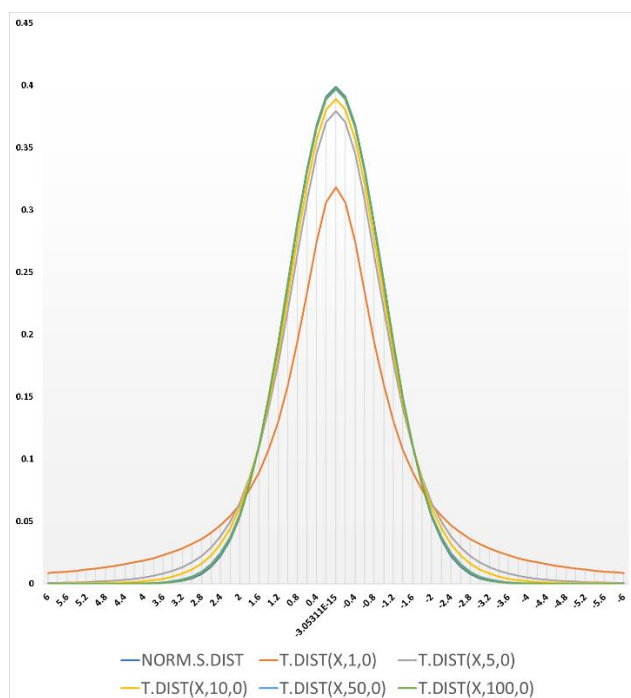


練習問題 3.3: カイ二乗分布と標本分散の関係についてエクセルで表示してみましょう。



n は標本の大きさ

練習問題 3.4:  $t$  分布について、 $n$  を 1, 5, 10, 50, 100 と変化させグラフに描いてみましょう。



練習問題 3.5: 練習問題 3.4 の結果から  $t$  分布の性質を記述してみましょう。

$t$  分布は標本の大きさが小さいとすそ野の厚い分布になります。標本の大きさが大きくなるにつれて標準正規分布に近づいていきます。 $t$  分布は標本の大きさによらずに歪度はゼロを維持します。つまり、いつでも左右対称です。尖度は標本の大きさが大きくなるにしたがいゼロに近づきます。

第4章

練習問題 4.1: excel で正規乱数を発生させ基本統計量をとってみましょう。乱数の数を 10、100、1000、10000 といろいろと変えてやってみましょう。

	A	B	C	D	E	F	G	H	I
1	n=10	n=100	n=1000	n=10000		n=10	n=100	n=1000	n=10000
2	0.0256	-0.4084	-0.5663	0.5908					
3	0.3024	0.3715	-0.1164	1.5459	平均	0.19	-0.04	-0.00	-0.01
4	1.3627	0.2776	-0.6109	-0.6595	標準誤差	0.35	0.09	0.03	0.01
5	-1.3102	-0.8951	-0.2415	-1.0657	中央値(メジアン)	0.05	-0.03	-0.04	0.00
6	-1.4631	-0.6161	-1.2640	-0.5843	最頻値(モード)	#N/A	0.68	0.55	-0.20
7	1.4241	-0.1953	0.3619	-1.8044	標準偏差	1.11	0.88	1.02	1.01
8	-0.0226	0.6138	0.9161	-1.5252	分散	1.24	0.77	1.04	1.01
9	0.0736	-1.5966	-0.4703	-0.7742	尖度	-0.77	-0.08	-0.01	-0.02
10	-0.3462	-0.9394	1.8675	-1.0486	歪度	0.02	0.35	0.01	-0.02
11	1.8839	0.0660	-0.4714	-0.4286	範囲	3.35	4.14	6.61	7.07
12		1.1774	0.3906	1.2542	最小	-1.46	-1.87	-3.56	-3.61
13		-0.0678	-0.9795	-0.6785	最大	1.88	2.27	3.05	3.46
14		0.1515	0.3074	-0.3249	合計	1.93	-4.42	-2.98	-91.15
15		0.5689	-0.8693	-1.5284	データの個数	10	100	1000	10000
16		-0.0856	-0.2768	-0.1147	信頼度(95.0%)	0.80	0.17	0.06	0.02

n の数が増えるにつれ、平均がゼロから外れる確率を小さくできます。これは大数の弱法則です。標本の大きさが大きくなるにつて標本平均は母平均に近づきます。標本の大きさを無限大にすると標本平均は確率 1 で母平均になります。これは大数の強法則です。

練習問題 4.2 : エクセルによるワインデータの主要要素の最大値と最小値を推定してみましょう。

標本平均を母平均、標本分散を母分散として

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

を用いて、 $\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ 、また  $n = 1$ 、 $\alpha = \frac{1}{1599} = 0.000625$  とすると  $\mu - 3.23\sigma < X < \mu + 3.23\sigma$  となります。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	平均	8.32	0.53	0.27	2.54	0.09	15.88	46.47	1.00	3.31	0.66	10.42	5.64
2	分散	3.03	0.03	0.04	1.99	0.00	109.42	1082.14	0.00	0.02	0.03	1.14	0.65
3	標準偏差	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
4	最大値	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00
5	最小値	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
11	自由度	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598
12	信頼係数	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937
13	最大値の推定	14.27	1.14	0.94	7.36	0.25	51.65	158.98	1.00	3.84	1.24	14.07	8.40
14	最小値の推定	2.36	-0.08	-0.40	-2.28	-0.07	-19.90	-66.05	0.99	2.78	0.08	6.78	2.87
15	(最大値-最小値)/(上限-下限)	0.95	1.19	0.75	1.51	1.86	0.99	1.26	1.05	1.20	1.44	0.89	0.91
16		A	B	C	D	E	F	G	H	I	J	K	評価
17		7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

半数以上の範囲(最大値-最小値)は推測範囲(上限-下限)よりも大きくなりました。

また、例題 4.2 で求めた母数の区間推定を用いると



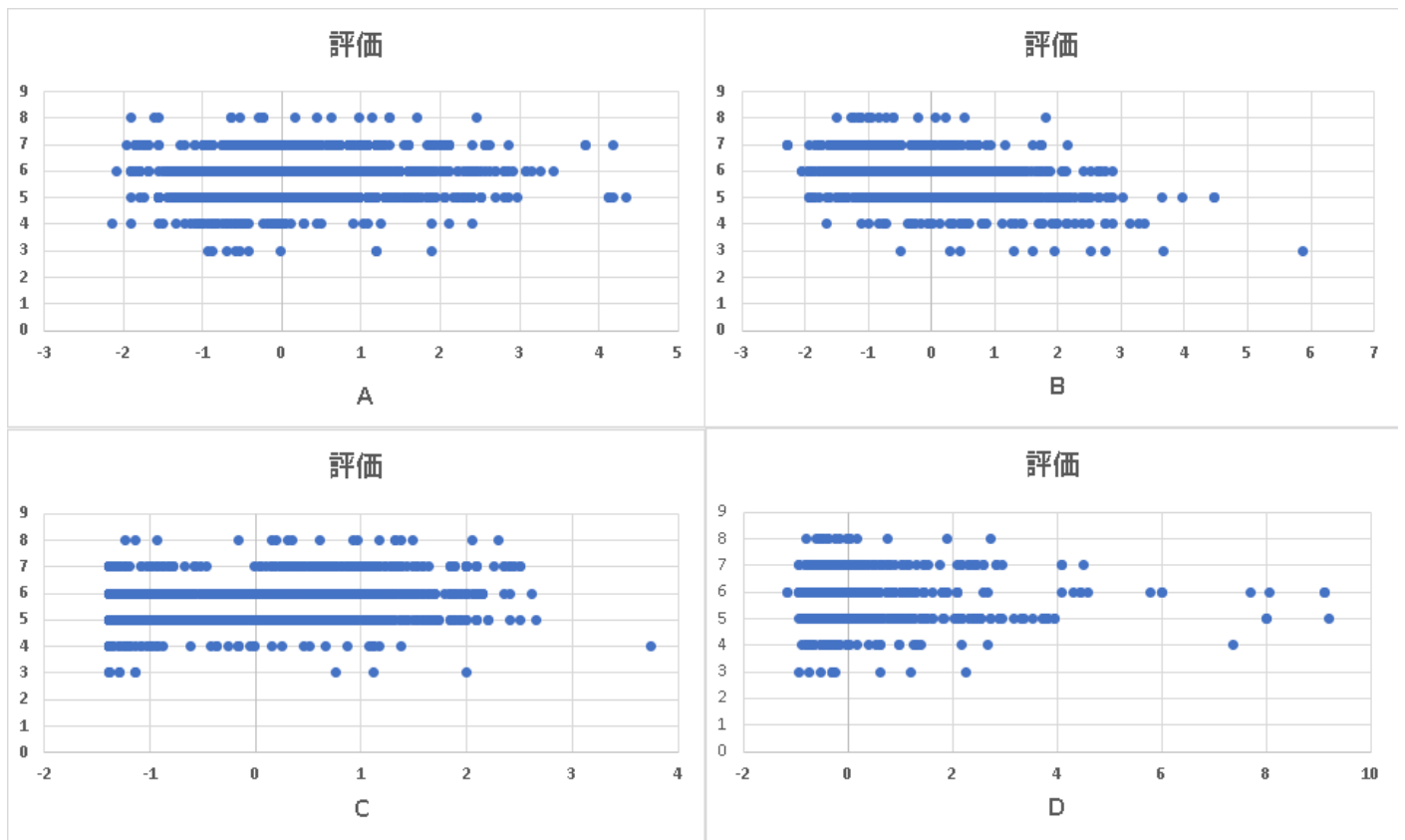
	A	B	C	D	E	F	G	H	I	J	K	L	M
1 平均		8.32	0.53	0.27	2.54	0.09	15.88	46.47	1.00	3.31	0.66	10.42	5.64
2 分散		3.03	0.03	0.04	1.99	0.00	109.42	1082.14	0.00	0.02	0.03	1.14	0.65
3 標準偏差		1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
4 最大値		15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00
5 最小値		4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
11 自由度		1598	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598	1598
12 信頼係数		0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937	0.99937
13 上限		8.47	0.54	0.29	2.66	0.09	16.77	49.29	1.00	3.32	0.67	10.51	5.71
14 下限		8.17	0.51	0.25	2.42	0.08	14.98	43.65	1.00	3.30	0.64	10.33	5.57
15 最大値の推定		14.42	1.16	0.95	7.48	0.25	52.55	161.80	1.00	3.85	1.25	14.16	8.47
16 最小値の推定		2.22	-0.10	-0.41	-2.40	-0.08	-20.80	-68.87	0.99	2.77	0.06	6.69	2.80
17 (最大値-最小値)/(上限-下限)		0.93	1.16	0.73	1.48	1.81	0.97	1.23	1.03	1.17	1.40	0.87	0.88

となり、結果はわずかに改善しますが、実際は混合分布の区間推定となるので問題はより複雑です。本書の範囲を超えます。

練習問題 4.2: 赤ワインデータベースの母平均の上限と下限を推測してみましょう。

練習問題 4.3: ひずんだ分布を修正する方法があるかどうかを試してみましょう。

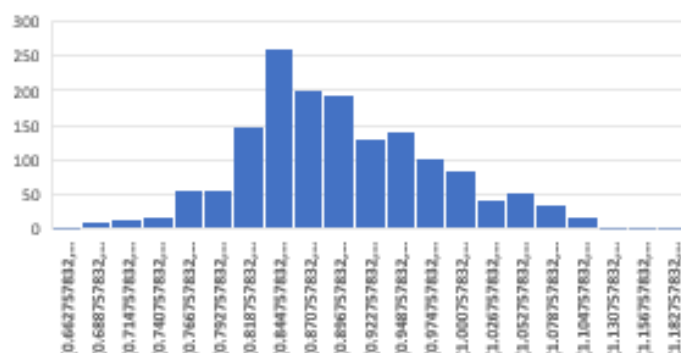
平均と分散を用いて標準化してみました。



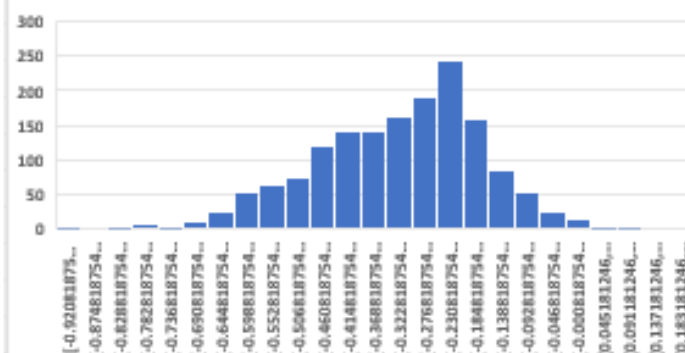


対数を取るとクエン酸、密度、評価を除いて正規性を改善することができます。

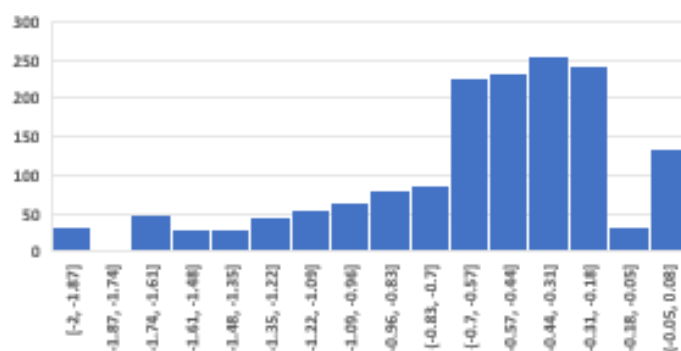
酒石酸濃度



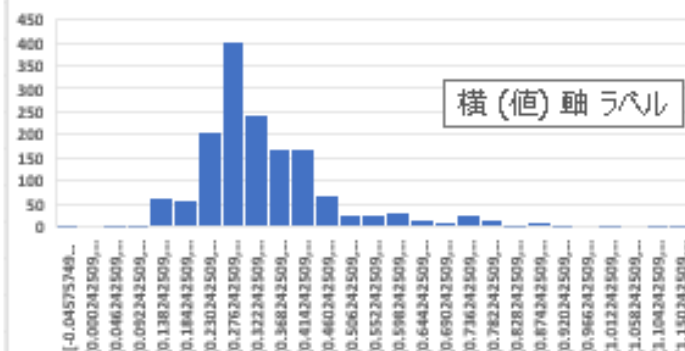
酢酸濃度



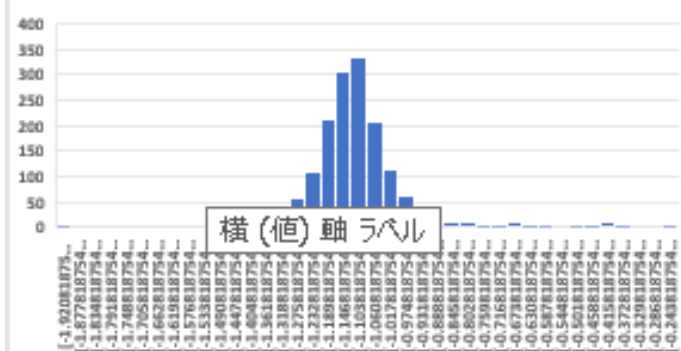
クエン酸濃度



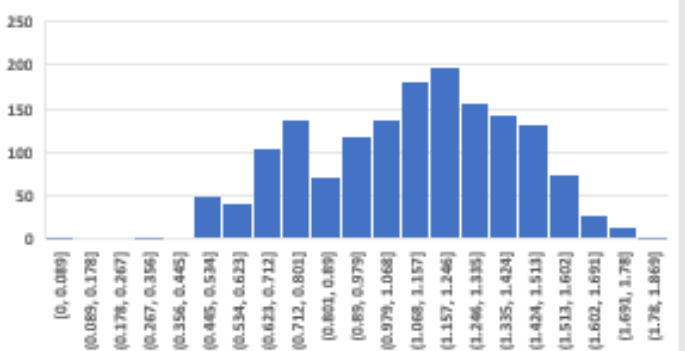
残糖濃度



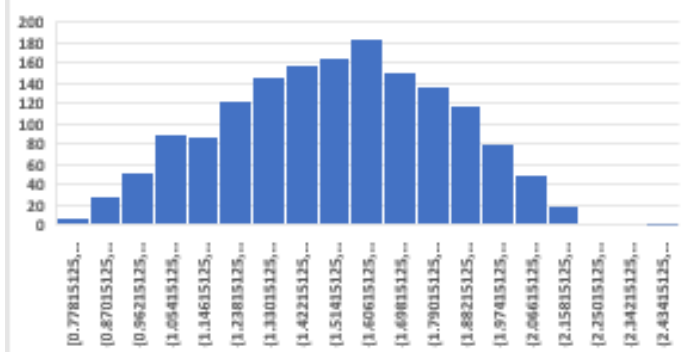
塩化ナトリウム濃度



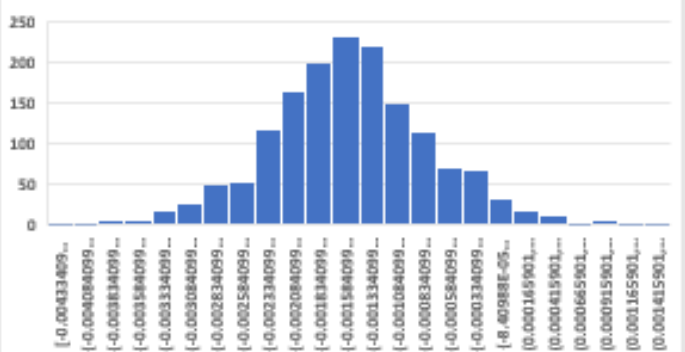
遊離二酸化硫黄濃度

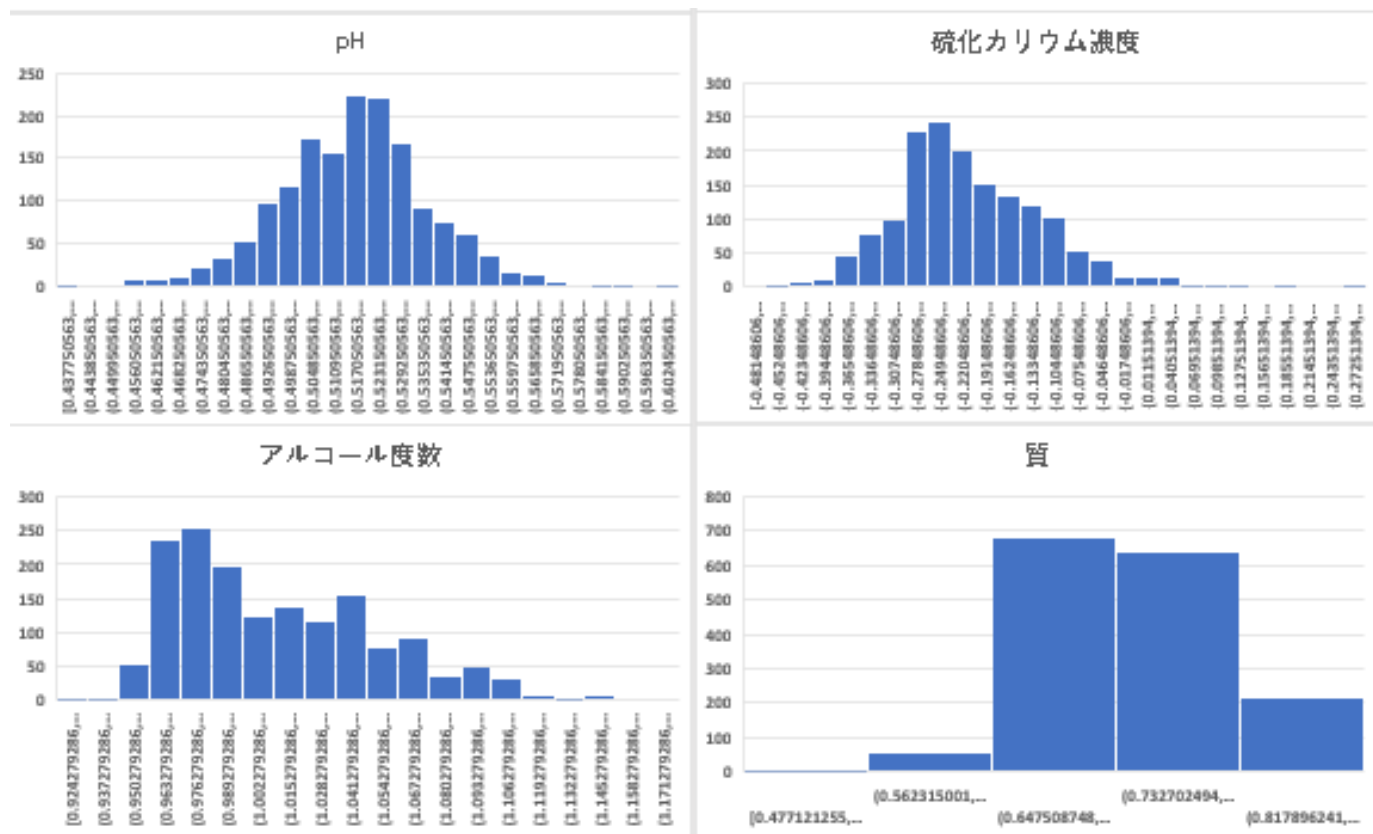


総二酸化硫黄濃度



密度





	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	質
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
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
調整後												
歪度	0.39	-0.43	#NUM!	1.81	1.75	-0.23	-0.08	0.06	0.01	0.92	0.66	-0.37
尖度	0.10	0.31	#NUM!	4.95	9.40	-0.46	-0.67	0.93	0.67	2.12	-0.29	1.30