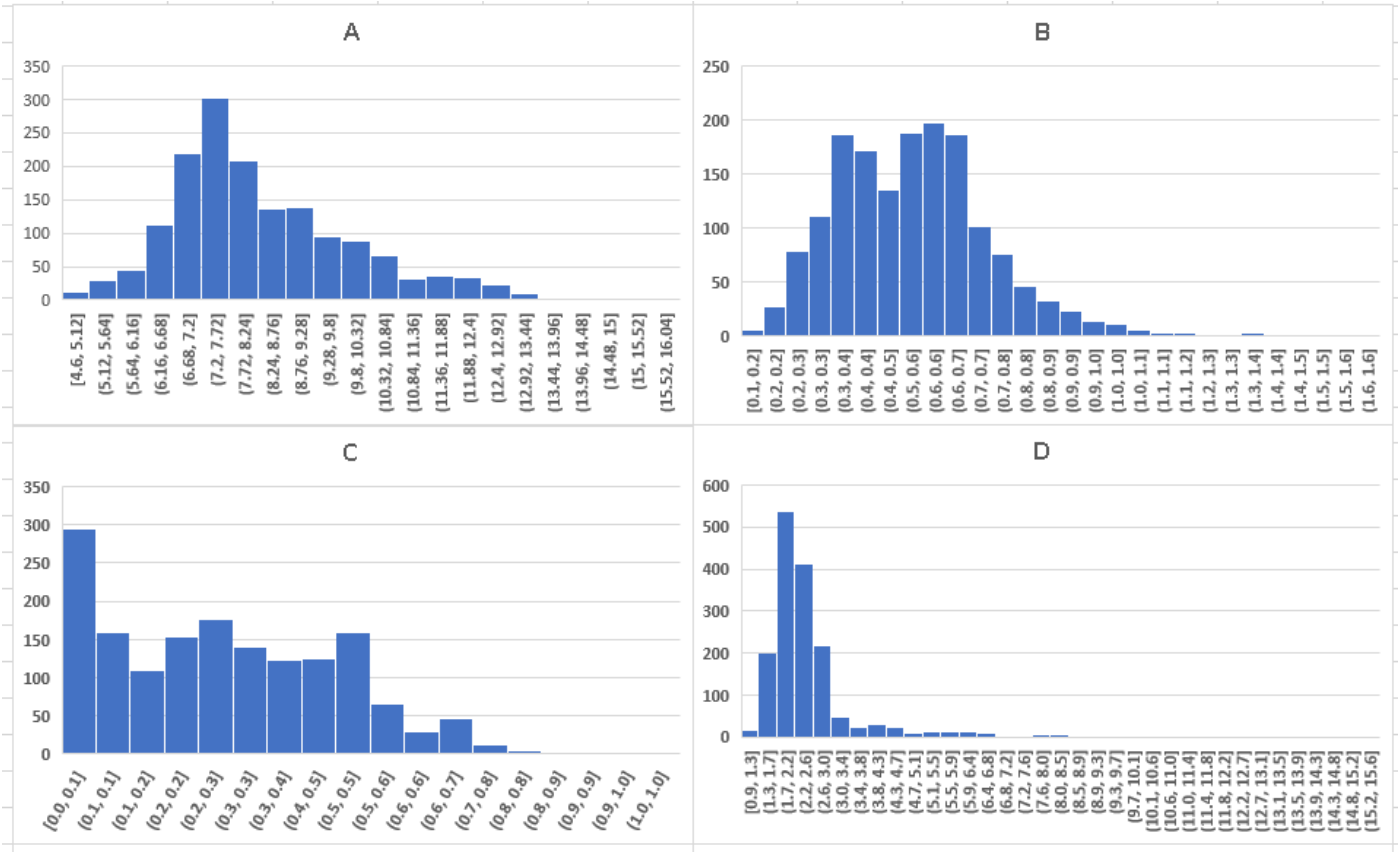
第1章

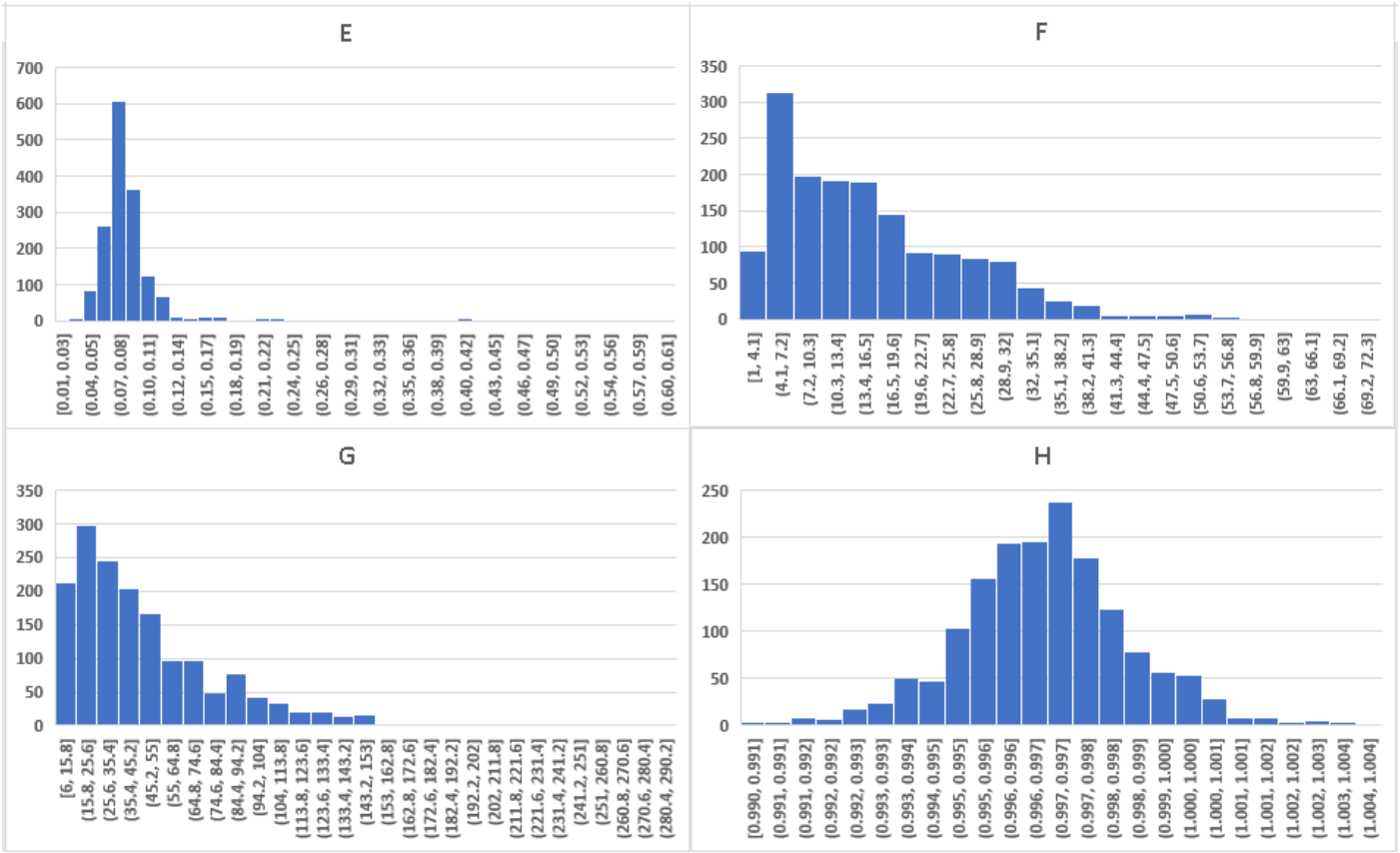
第１章ではデータの要約の方法を主に学びました。頻度図と散布図という２つのグラフについて学びました。また、平均、分散、標準偏差などの基本統計量についても学びました。ポルトガルの生産者のために作られた赤ワインの評価と化学成分のデータ分析を用いて可視化と基本統計量について学びました。基本統計量の中でも’平均’に多くの人はなじみが深いと思われます。しかし、統計学として’平均’が用いられるようになったのはごく最近のことです。１８世紀と１９世紀のはざまでケトレーが平均の概念を明記しています。それ以前は最大値、最小値、最頻値、中央値、範囲などが用いられていたようです。’平均’の概念は紀元前２８０年にはピタゴ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る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。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。

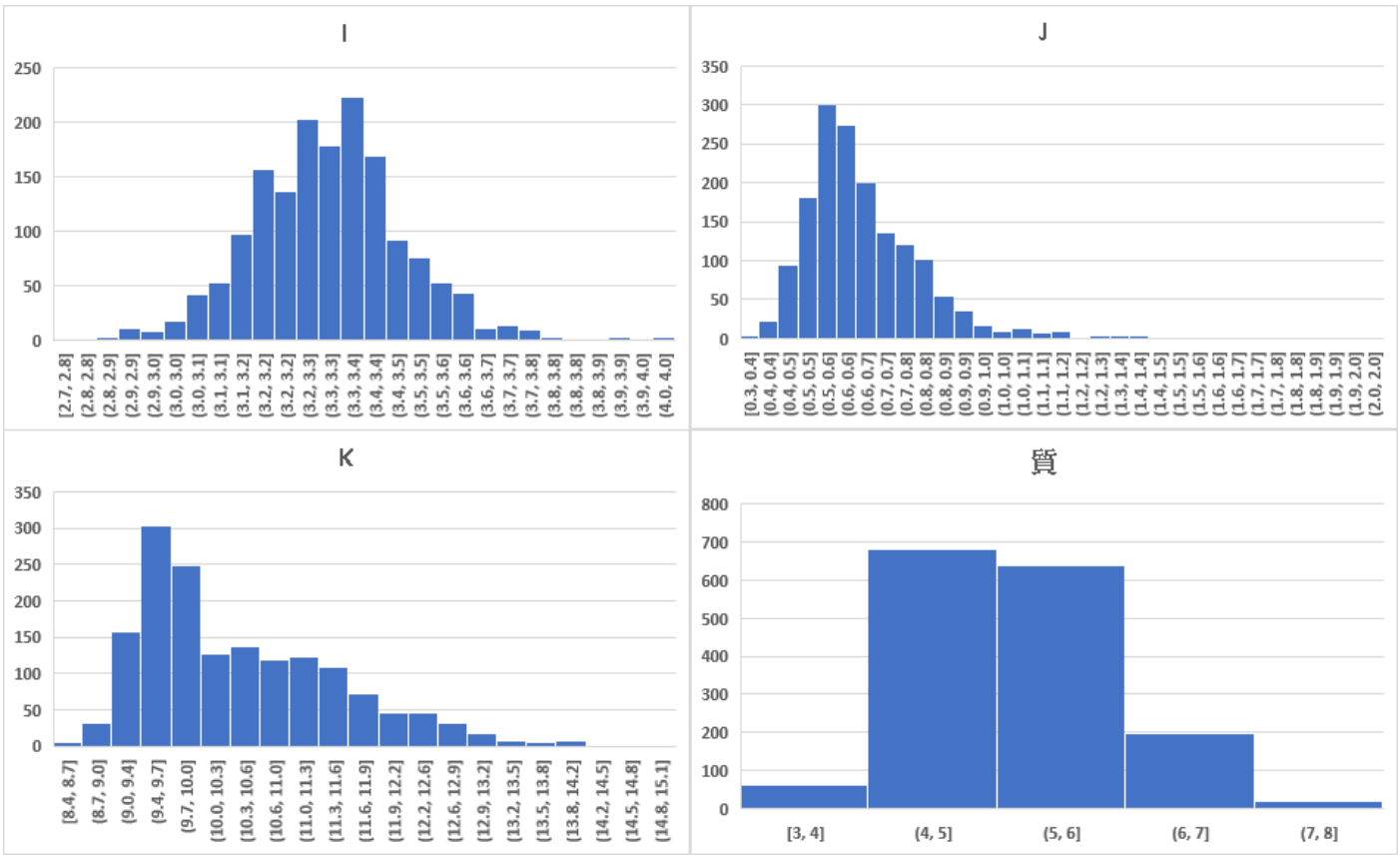
では、ワインデータに関する練習問題をとおして最近のデータ分析の方法を学んでいきましょう。

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。


A
B
C
D

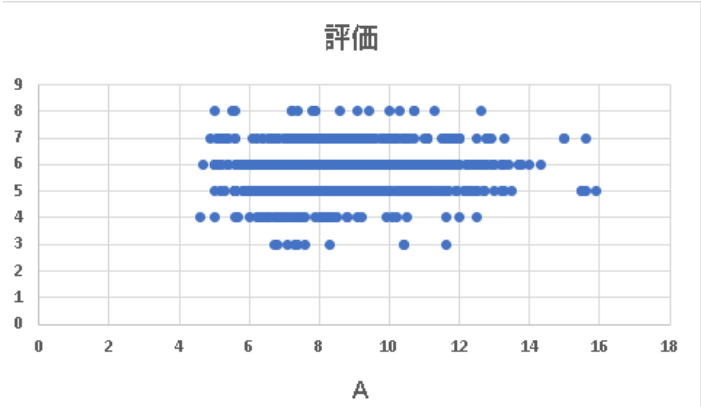

E
F
G
H

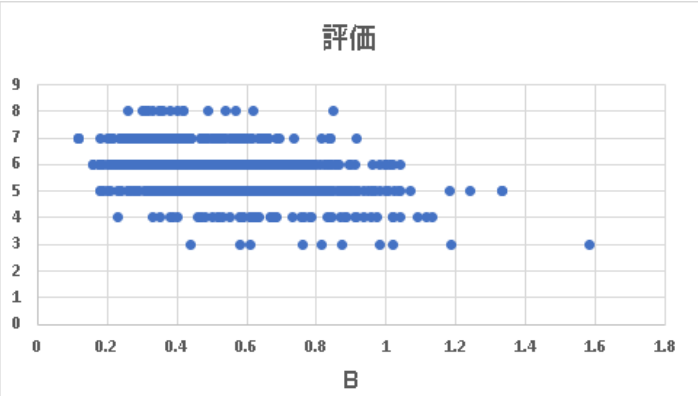


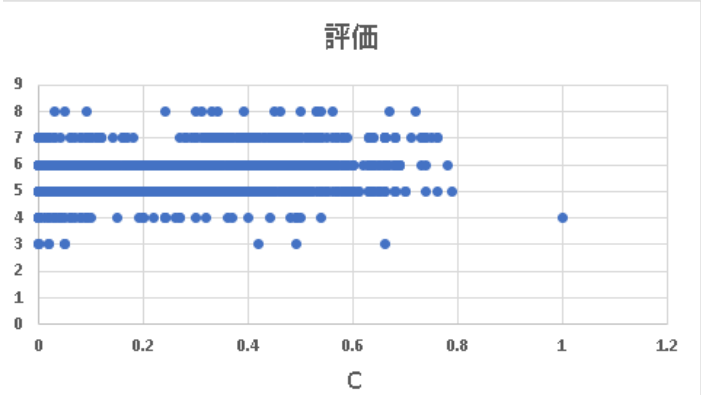
実際の分布はベル型になることが少ないことが分かります。しかし、大まかな傾向はつかめるので分析の出発点を探るには有効な手段です。エクセルシートは練習問題 1.1、エクセルシートの作り方は第３部参照

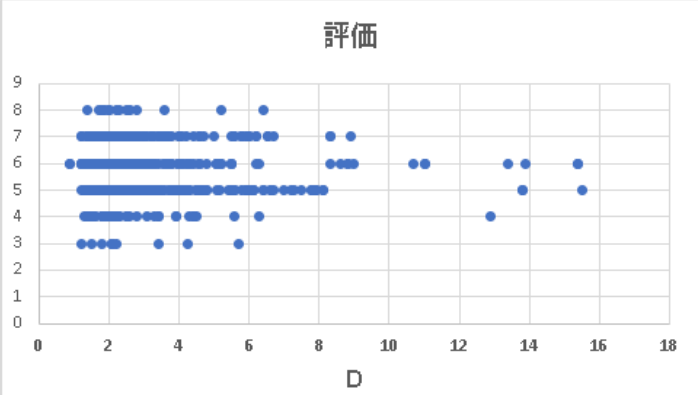
練習問題 1.2:　ワインデータから適当に化学成分を選び、その評価との関係を散布図として描いてみましょう。

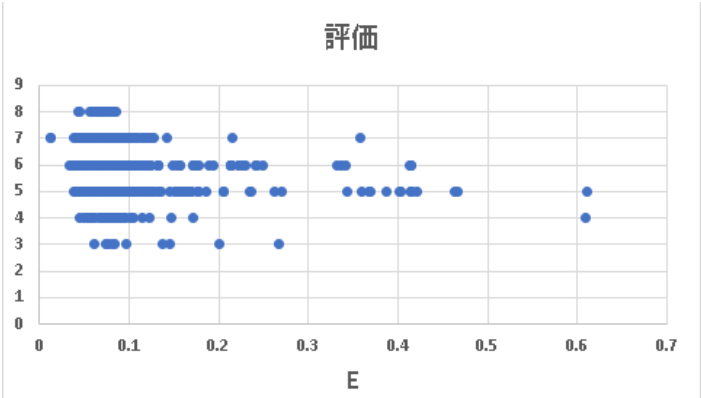
ポルトガルのワイナリーではより評価の高い赤ワインを作ることが目的なので、評価を高くするワインの化学成分が何で、何をどのように調整したらよいのかを知ることが分析の目的です。そうするためには評価と化学成分との間の明確な傾向を見つけることが目的となります。散布図の右上が高く、左下が低いか ⋰ 、または逆の傾向 ⋱ があればよいことになります。見た感じでそのような傾向にあるのは、B, D, I, J, K です。これは人びとの経験と知識に大きく影響される部分であり、正解はありません。散布図を見るときに注意することは頻度の低いデータはばらつきの幅も狭くなるということです。

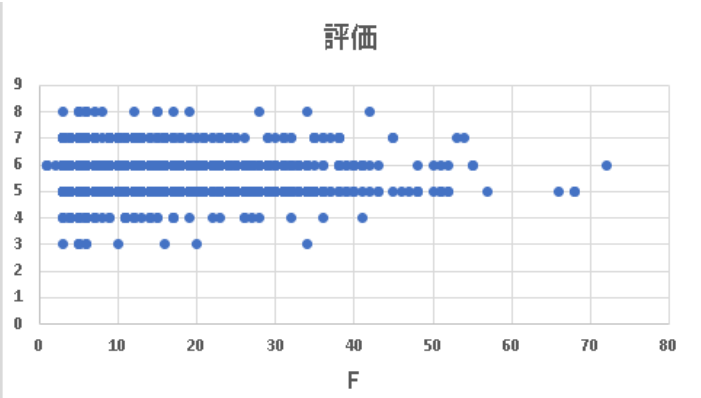

評価
A

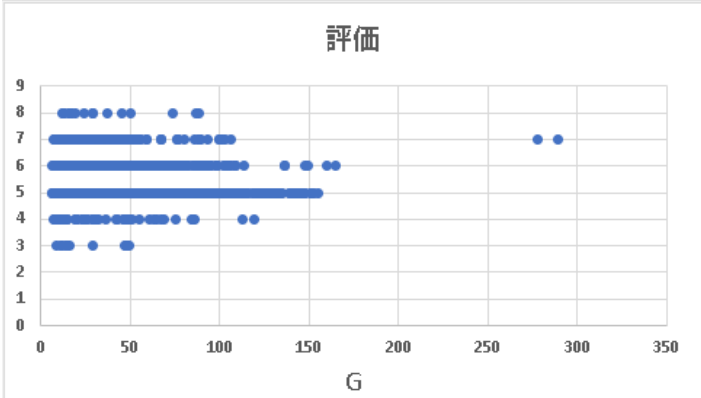

評価
B

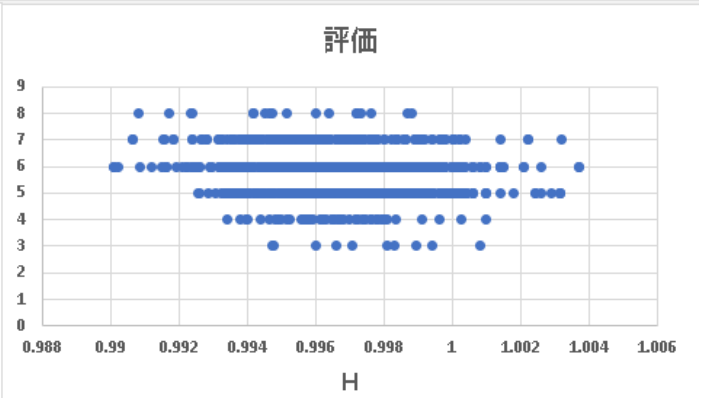

評価
C

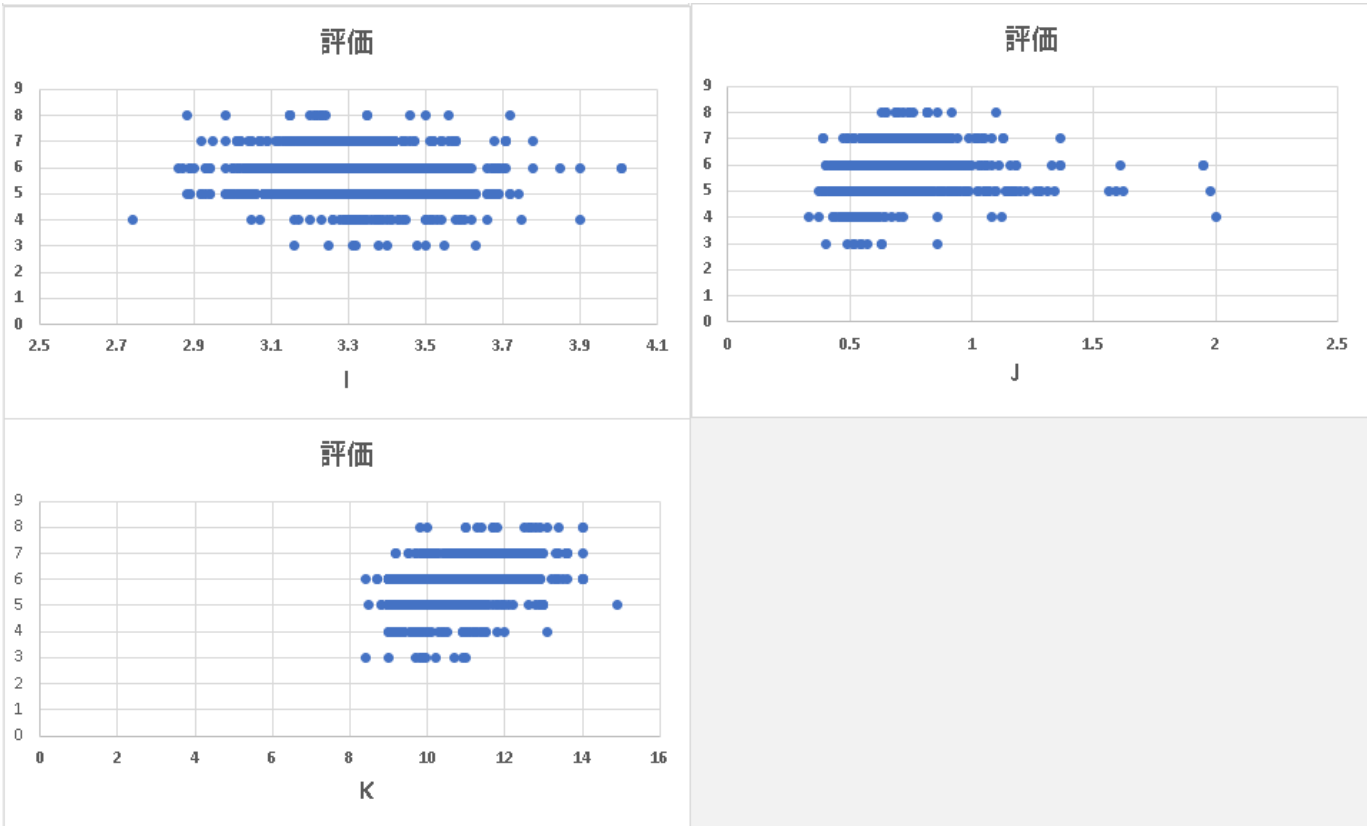

評価
D


評価
E


評価
F


評価
G


評価
H



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、Var(X+Y)=Var(X)+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もちます。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

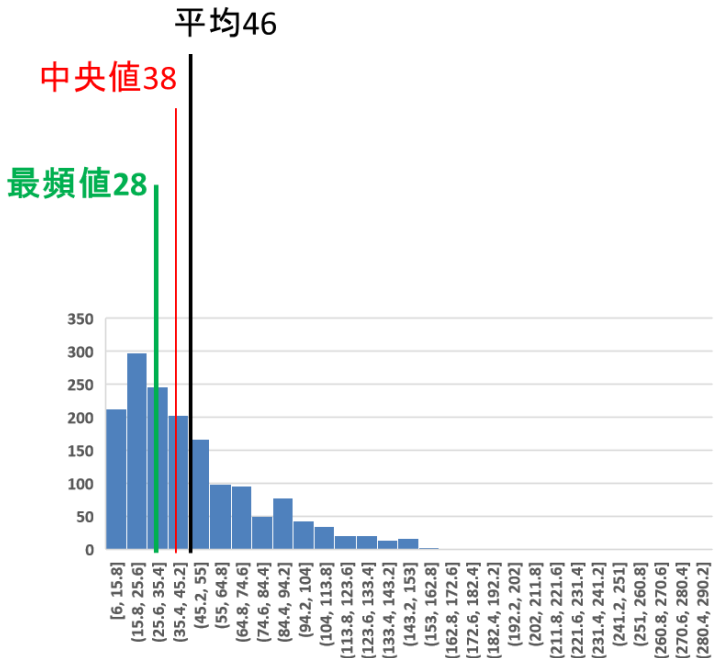
中央値(メディアン)

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。



(練習問題 1?)

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

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

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。

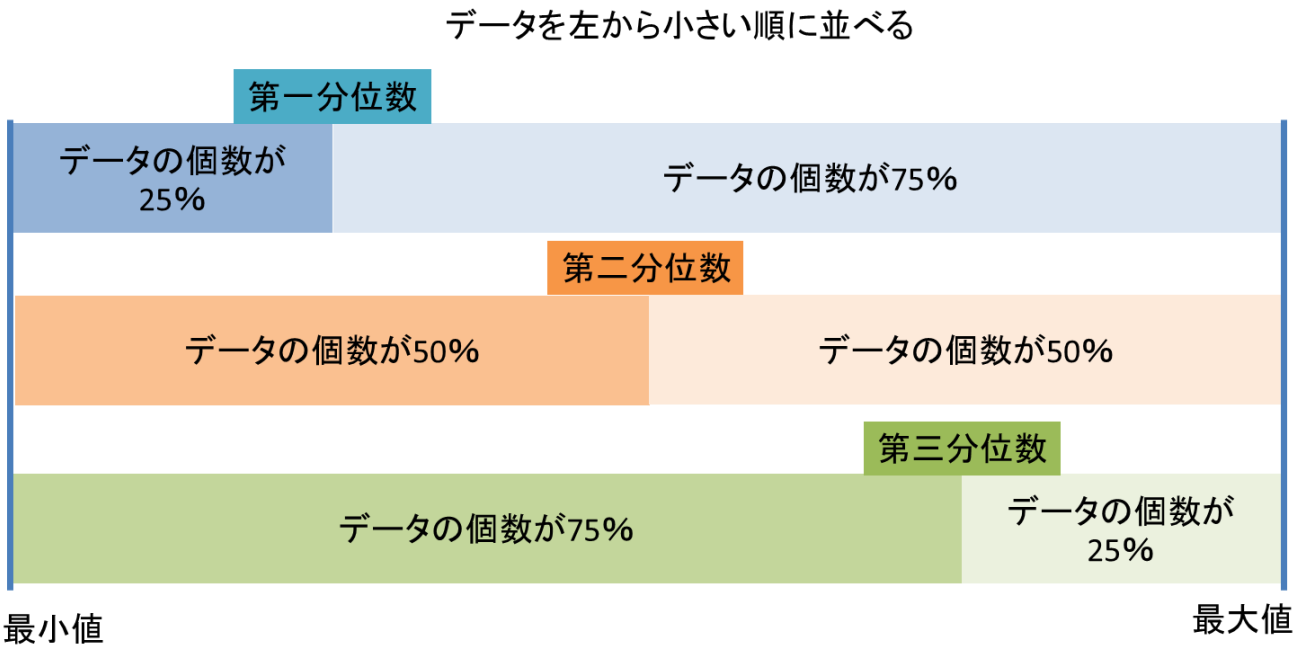
最大値・最小値

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$x_1 \leq x_2 \leq \cdots x_n$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 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。



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

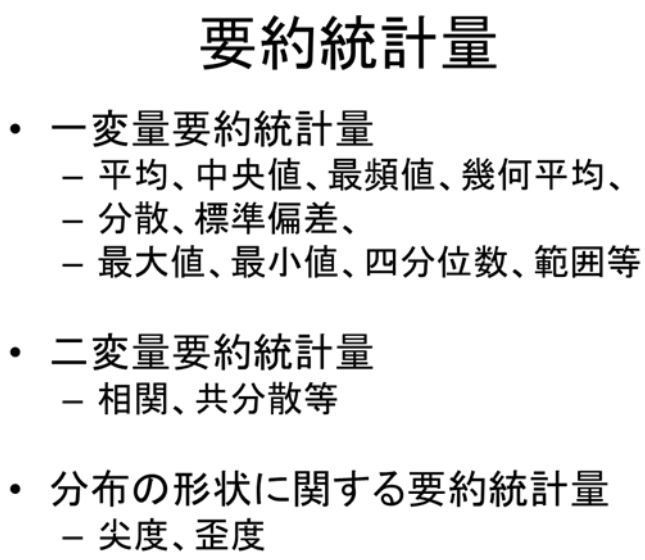
変数の分類と要約統計量

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。



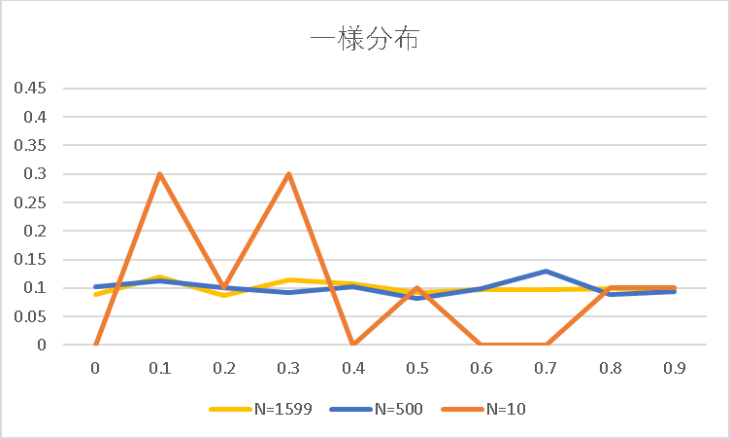
第２章

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。

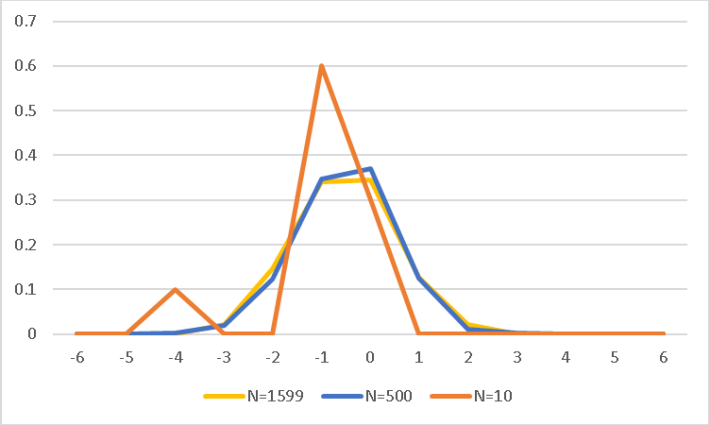
- 一様分布

	累積分布関数			確率密度関数		
	N=1599	N=500	N=10	N=1599	N=500	N=10
0	0	0	0	0.08818	0.102	0
0.1	141	51	0	0.120075	0.112	0.3
0.2	333	107	3	0.086929	0.1	0.1
0.3	472	157	4	0.113821	0.092	0.3
0.4	654	203	7	0.106942	0.102	0
0.5	825	254	7	0.092558	0.082	0.1
0.6	973	295	8	0.097561	0.098	0
0.7	1129	344	8	0.09631	0.13	0
0.8	1283	409	8	0.099437	0.088	0.1
0.9	1442	453	9	0.098186	0.094	0.1
1	1599	500	10			
1.1	1599	500	10			



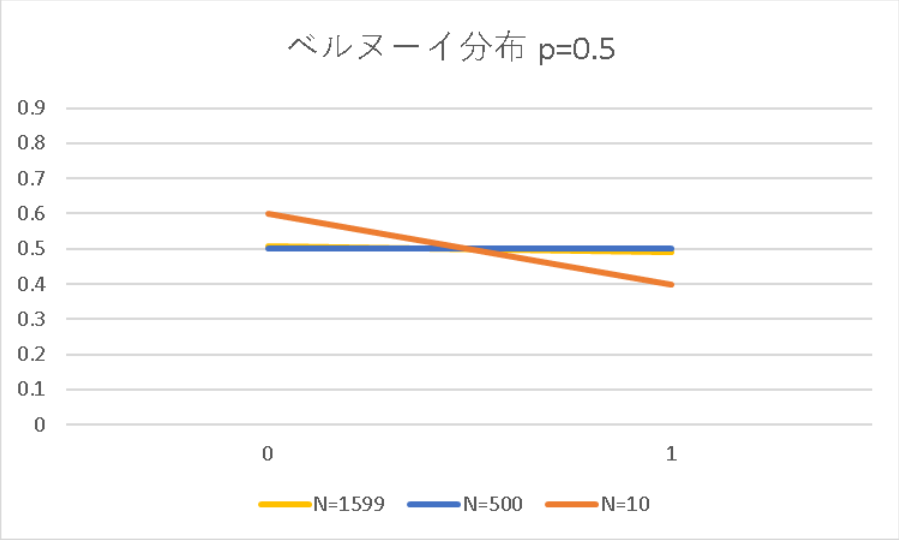
- **正規分布**

	累積分布関数			確率密度関数		
	N=1599	N=500	N=10	N=1599	N=500	N=10
−6	0	0	0	0	0	0
−5	0	0	0	0	0	0
−4	0	0	0	0.000625	0.002	0.1
−3	1	1	1	0.020638	0.02	0
−2	34	11	1	0.145716	0.122	0
−1	267	72	1	0.340838	0.348	0.6
0	812	246	7	0.343965	0.37	0.3
1	1362	431	10	0.126329	0.126	0
2	1564	494	10	0.021263	0.01	0
3	1598	499	10	0.000625	0.002	0
4	1599	500	10	0	0	0
5	1599	500	10	0	0	0
6	1599	500	10	0	0	0
7	1599	500	10			



- ベルヌーイ分布

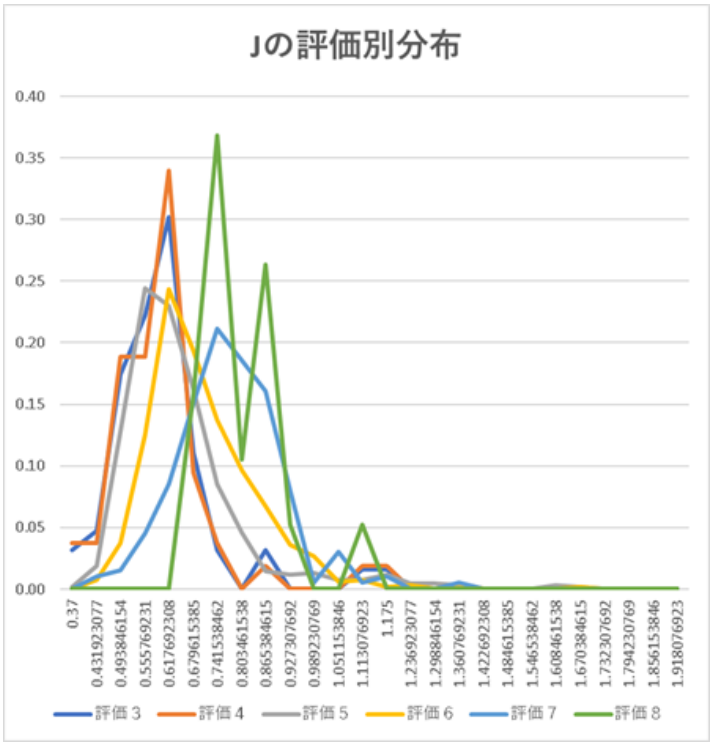
	累積分布関数			確率質量関数		
	N=1599	N=500	N=10	N=1599	N=500	N=10
0	813	250	6	0.508443	0.5	0.6
1	1599	500	10	0.491557	0.5	0.4



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。

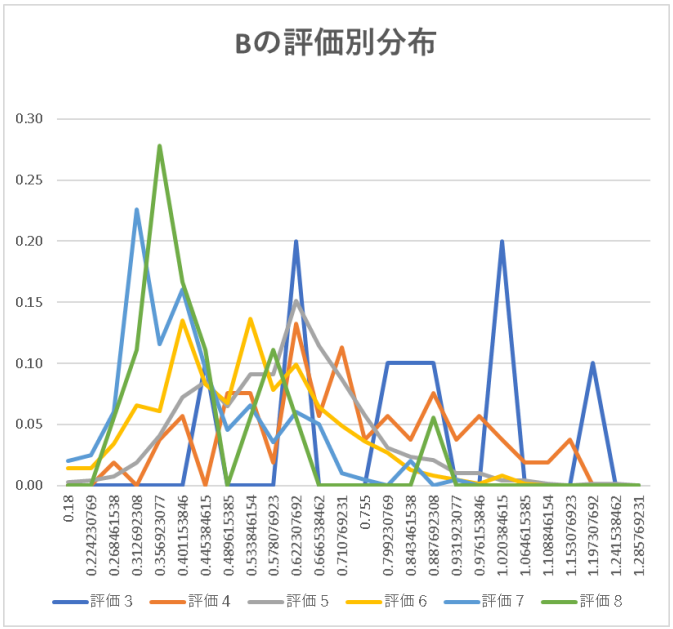
練習問題 2.2: 化学成分 J について評価別分布を作成してみましょう。

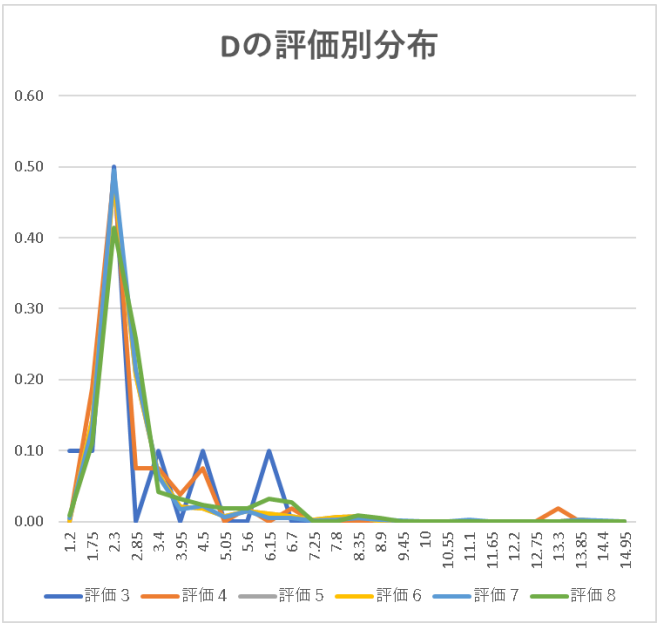
散布図から評価と化学成分 J の間には特別な関係がありそうなので、評価別の頻度図を作成してみました。分布の山が２つありそうなことが見えてきます。

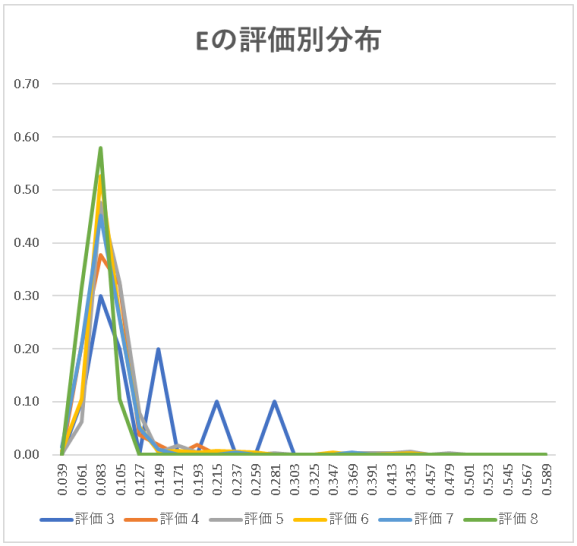


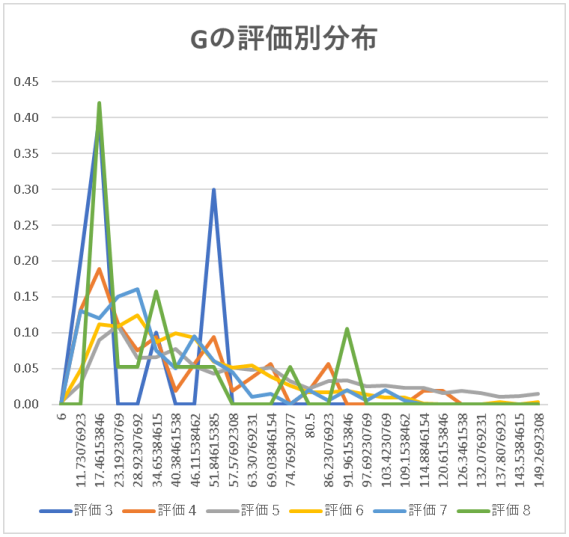
まず、図から得られるイメージは評価３，４，５，６と７，８は同じ分布ではないのかという疑問です。この疑問には統計学は適切な答えを与えてくれます。第６章で分散分析を扱います。

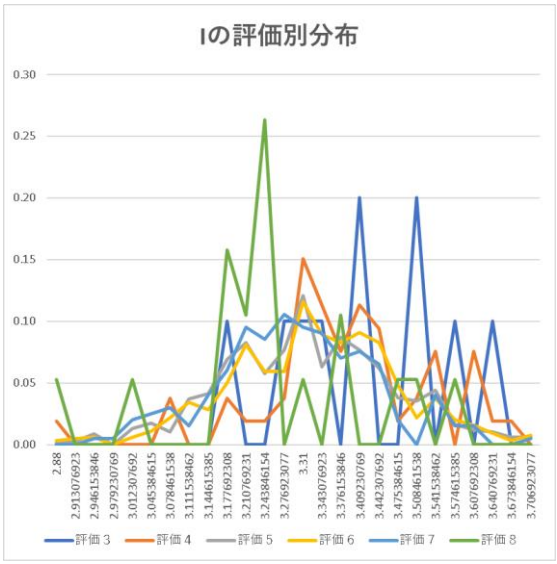
その他の主な化学成分についての評価別分布をつぎに示します。

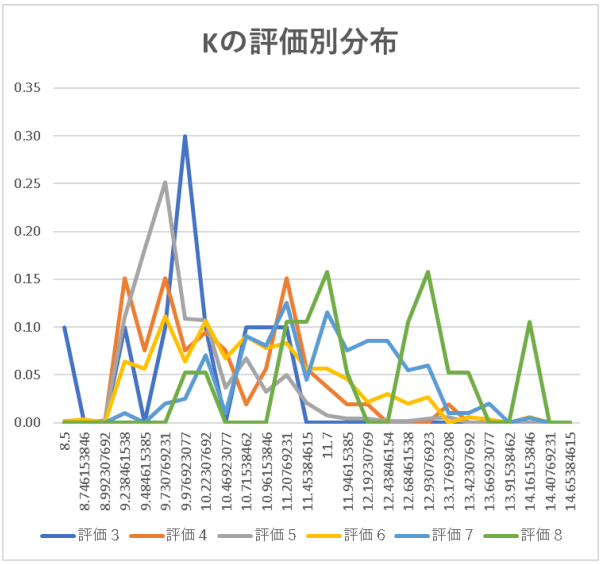












練習問題 2.3: 赤ワインデータの１０段階評価の標本空間と根元事象を示してみましょう。また、その違いを説明して見ましょう。標本空間と根元事象の概念を使って統計分析ができる条件は何でしょうか？

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: トランプの標本空間はなんでしょうか？

標準的なトランプのカードは５２枚で構成されています。そこから１枚のカードを引くとき数値に注目するか、スートに注目するかで答えは違ってきます。数値(番号)に注目すると 1-13 です。またスートに注目するとハート、スペード、ダイヤ、クラブの４つになります。もちろん両方を合わせて考えることもできます。

練習問題 2.5: 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。

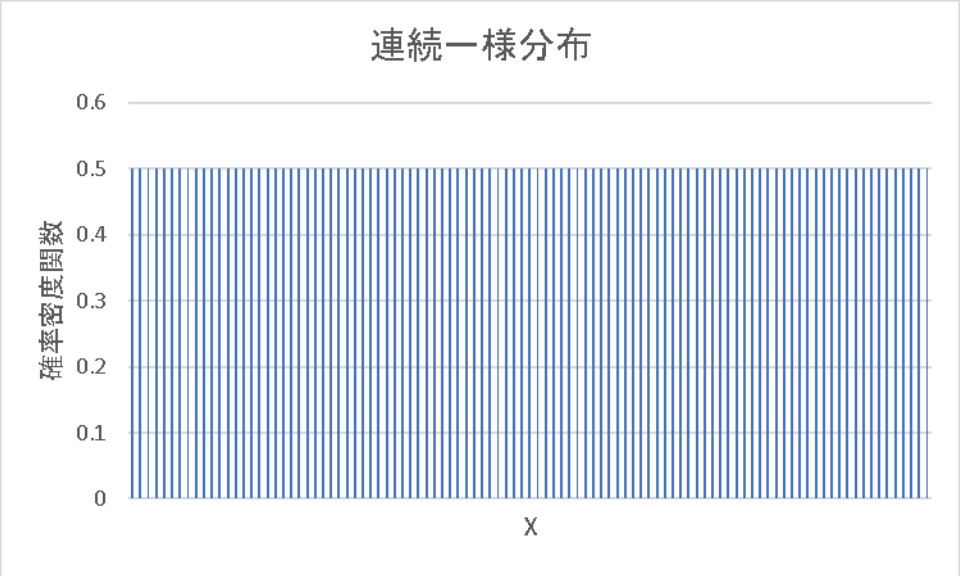
確率変数であるのはそれぞれの化学成分の値ですが、観測された値は観測値となり確定値です。

練習問題 2.6: A と B という事象があって、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)×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 2.7: 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 2.8: 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$(b-a)^2/16$で与えられます。b は上限, a は下限です。

練習問題 2.9: 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 2.10: 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。

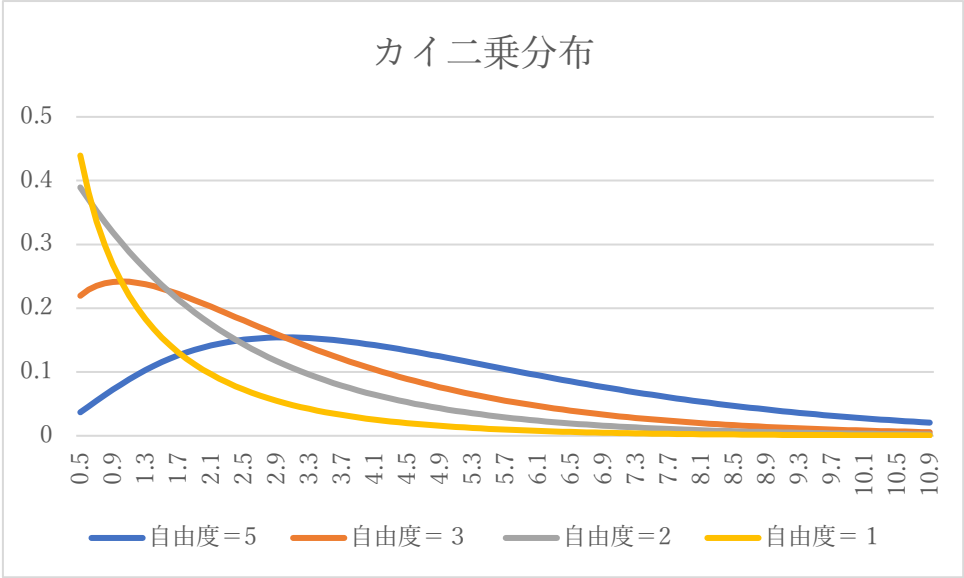
第３章

３章では母数と標本を扱いました。例題 3.1 では統計データを用いた身近な例について母集団と標本を記述してみました。”選挙の当選予想”では”全有効票数”を母集団として”出口調査で得られた票数”を標本としました。母集団全体を対象とする調査もあります。それを全数調査といいます。全数調査の代表例は国税調査です。しかし、このような調査は時間と費用がともないます。したがって、母集団の一部を調査の対象とする標本調査が主流です。標本に含まれるデータの数を標本の大きさ(標本サイズ)といいます。母集団から標本の大きさ n の標本を得ることを標本抽出といいます。標本データから母集団について推測するためには、標本が母集団を適切に代表している必要があります。それを偏りのない標本と呼びます。偏りのない標本を得るためには、母集団の特性をよく知る人の経験に頼る方法と、人の知識や経験に頼らない、確率的、機械的に標本を得る方法があります。前者を有意抽出用、後者を無作為抽出法と呼びます。

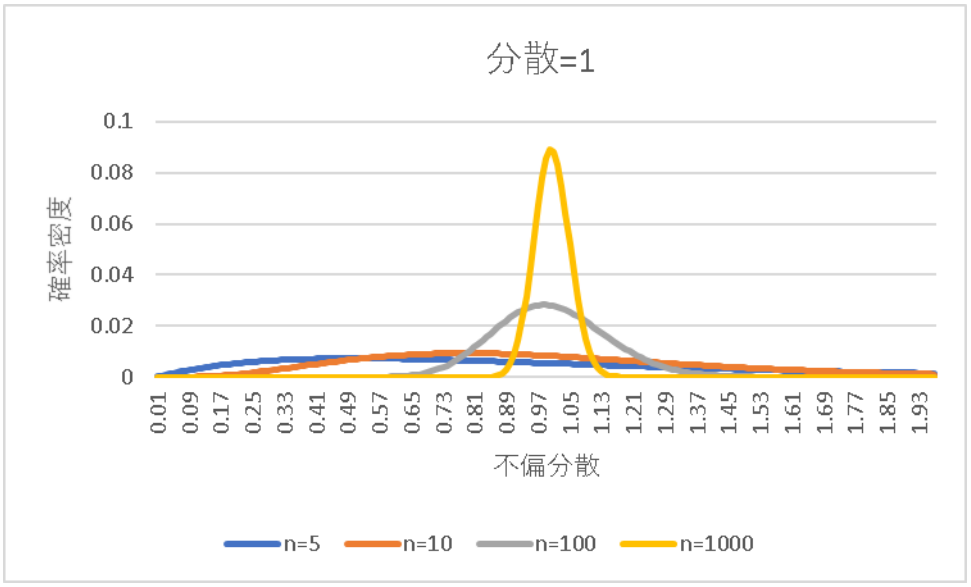
練習問題 3.1: 平均と期待値の違いを説明してみましょう。

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。

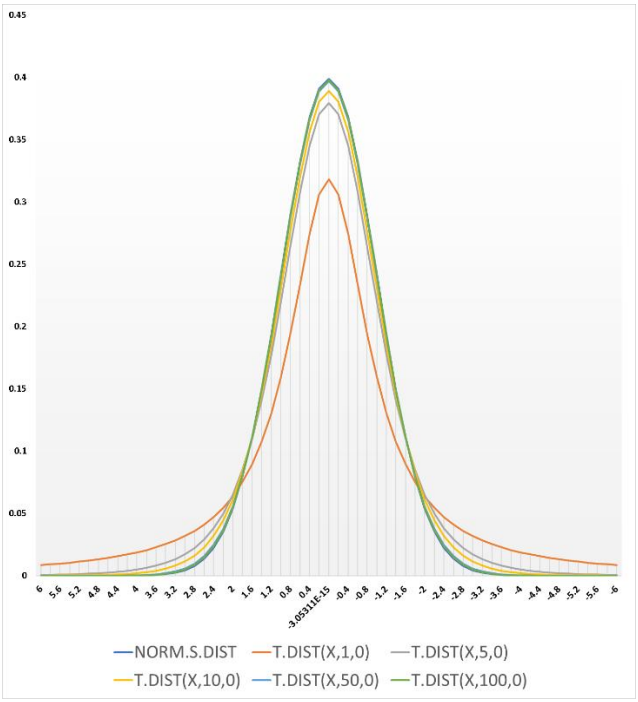


練習問題 3.3: カイ二乗分布と標本分散の関係についてエクセルで表示してみましょう。



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

得られたデータから平均値を計算したり、頻度図として表現して、手元のデータから得られる情報を明らかにすることは、記述統計の分野に入ります。推測統計では、得られたデータを全体の一部と考え、適切な仮説をたてて、一部のデータから見えていない部分を含む全体を推測します。ときには、経験分布を説明する適切な理論分布が見当たらない場合には、経験分布に加工を施すこともあります。

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	-1.1581	-0.3513	-0.2789	-0.8513					
3	-0.5759	-0.5264	0.2313	0.4335	平均	0.30	0.07	-0.04	-0.00
4	0.9886	-0.8316	0.0300	0.4108	標準誤差	0.30	0.10	0.03	0.01
5	-0.5308	0.0291	-1.3474	1.4242	中央値（メジアン）	0.01	0.06	-0.09	-0.00
6	0.0011	0.9192	-0.6797	-2.0640	最頻値（モード）	#N/A	#N/A	#N/A	#N/A
7	-0.0725	-1.1206	-1.5177	-2.3878	標準偏差	1.11	0.88	1.02	1.00
8	1.3521	-1.2713	1.6957	-1.9152	分散	0.92	1.03	1.03	1.01
9	1.2720	-1.2790	1.0740	-1.0014	歪度	0.09	0.04	0.02	0.03
10	0.0227	-0.4571	1.1195	0.9842	尖度	-1.39	-0.20	-0.22	-0.01
11	1.6575	1.3443	0.6147	-0.3746	範囲	2.82	5.02	6.22	8.05
12		1.7856	-0.2841	-0.5142	最小	-1.16	-2.32	-2.97	-3.70
13		-0.3411	-1.1428	-0.5801	最大	1.66	2.69	3.25	4.36
14		0.4412	0.3660	1.0387	合計	1.93	-4.42	-2.98	-18.69
15		1.1251	-0.5674	-0.5778	データの個数	10	100	1000	10000
16		-0.4149	-0.2383	-0.2197	信頼度(95.0%)(95.0%	0.80	0.17	0.06	0.01969

n の数が増えるつれ、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な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$n=1$、$\alpha=\frac{1}{1599}=0.00625$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

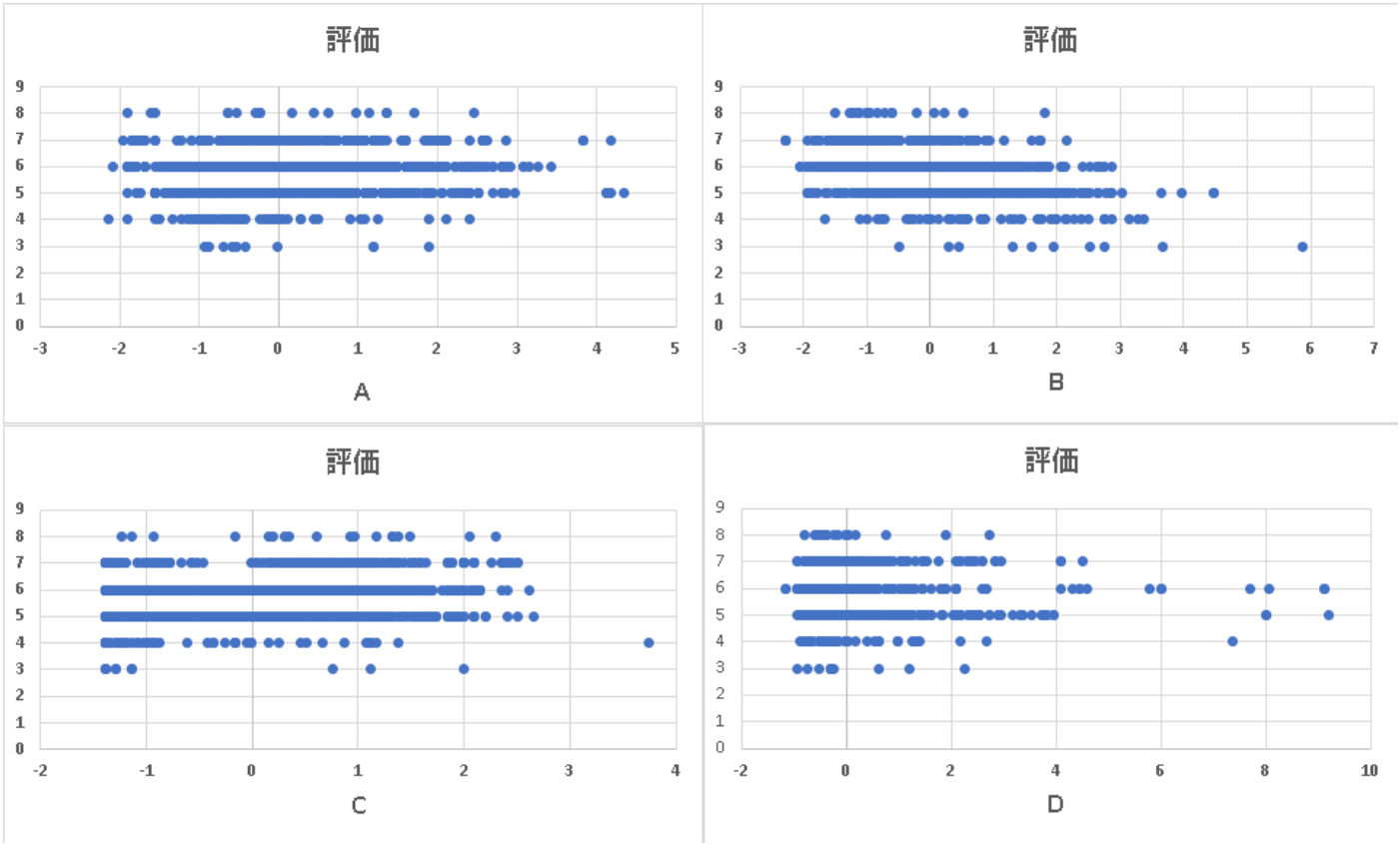
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	平均	8.32	0.53	0.27	2.54	0.09	15.88	46.47	1.00	3.31	0.66	10.42	5.64
2	分散	3.03	0.03	0.04	1.99	0.00	109.42	1082.14	0.00	0.02	0.03	1.14	0.65
3	標準偏差	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
4	最大値	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00
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

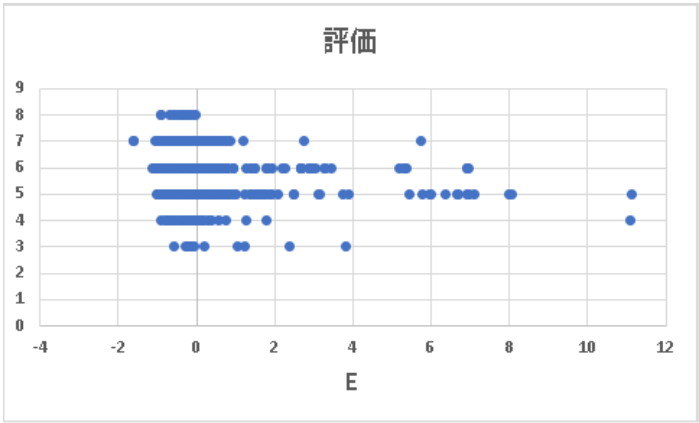
となり、結果はわずかに改善しますが、実際は混合分布の区間推定となるので問題はより複雑です。本書の範囲を超えます。

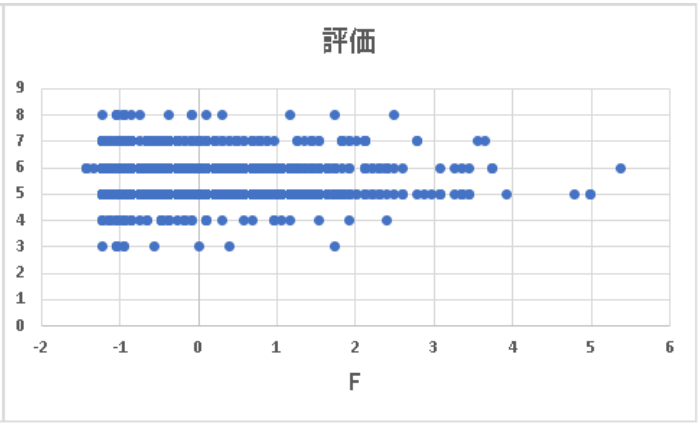
練習問題 4.2:　赤ワインデータベースの母平均の上限と下限を推測してみましょう。

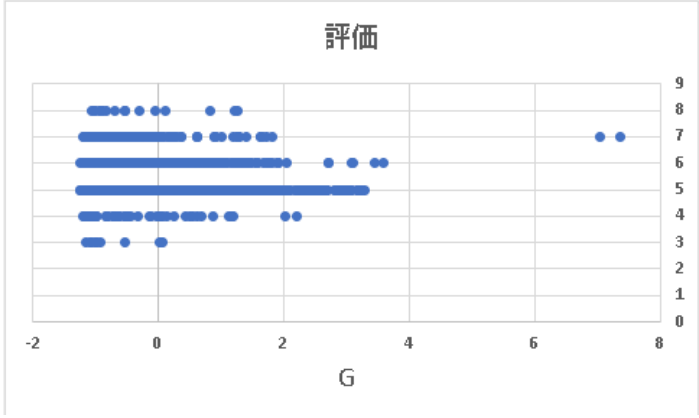
練習問題 4.3: ひずんだ分布を修正する方法があるかどうかを試してみましょう。

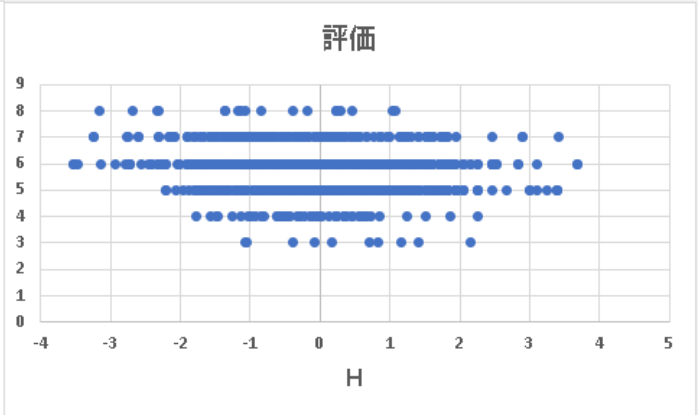
平均と分散を用いて標準化してみました。

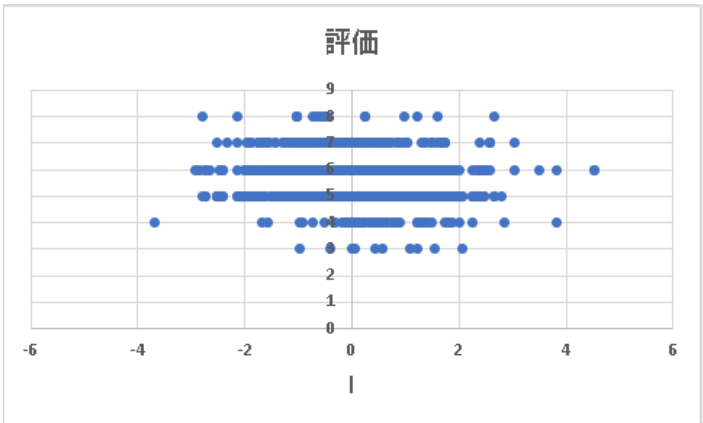


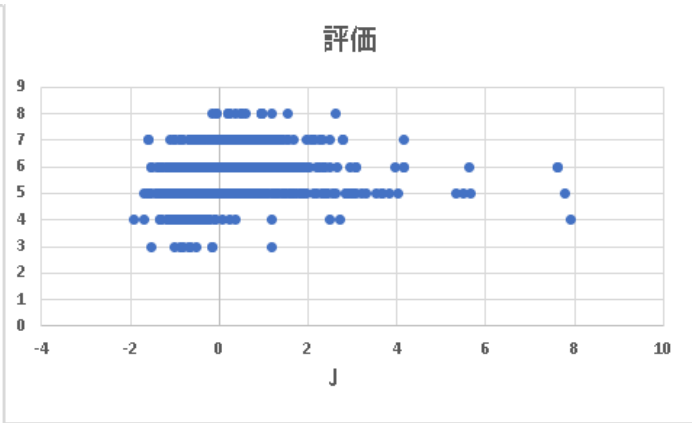


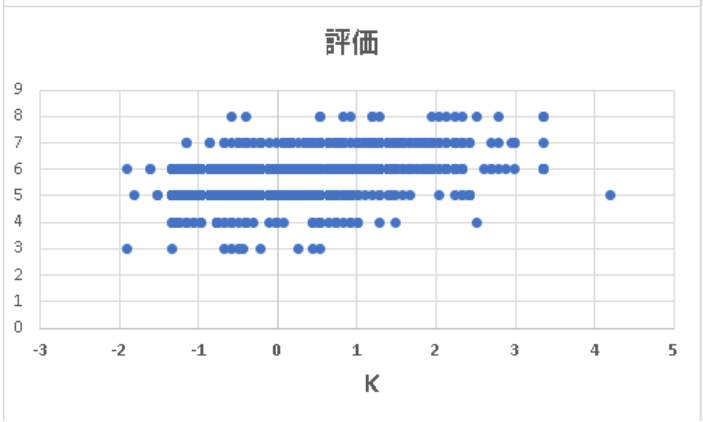




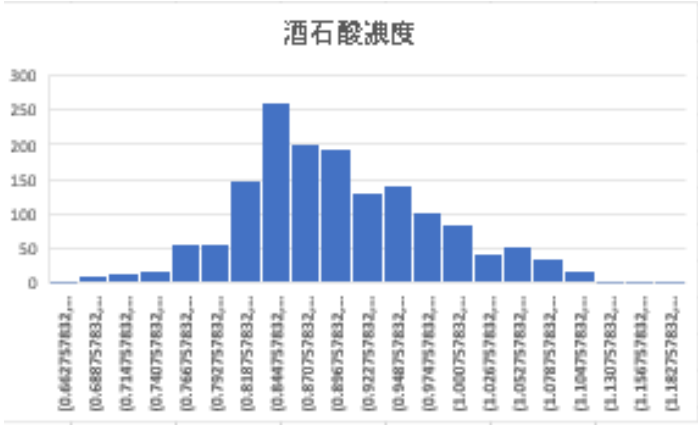


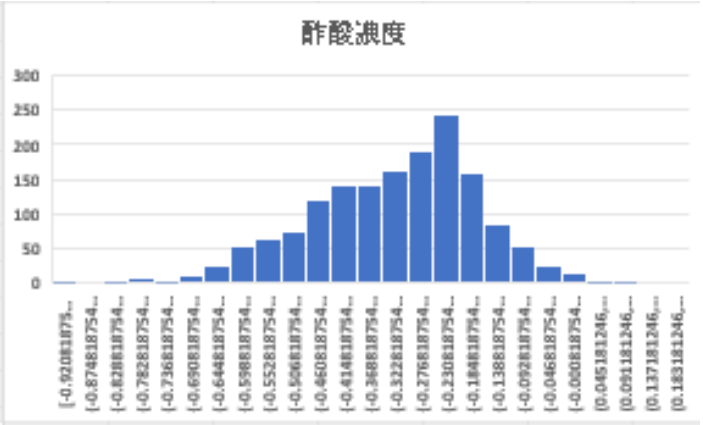


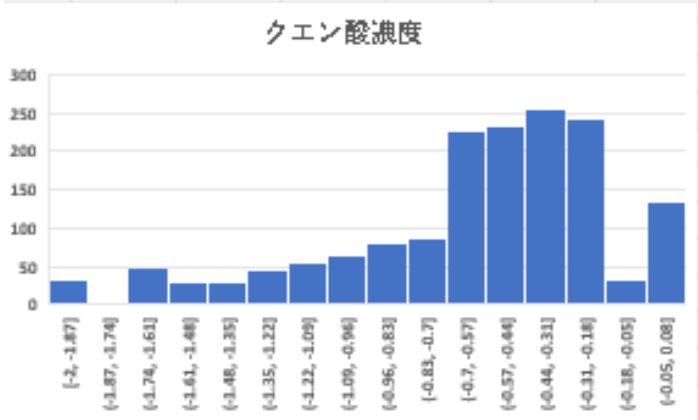


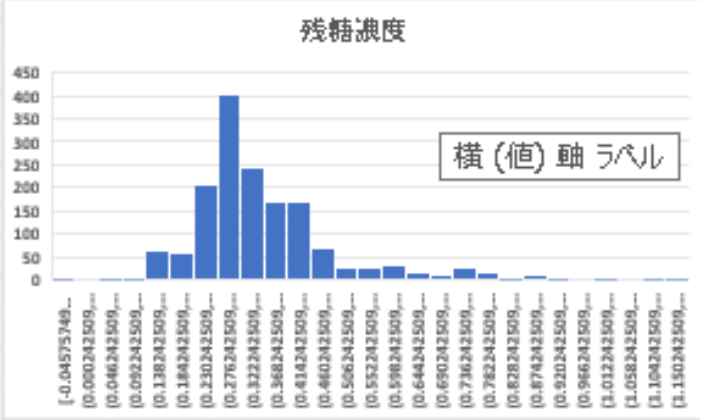


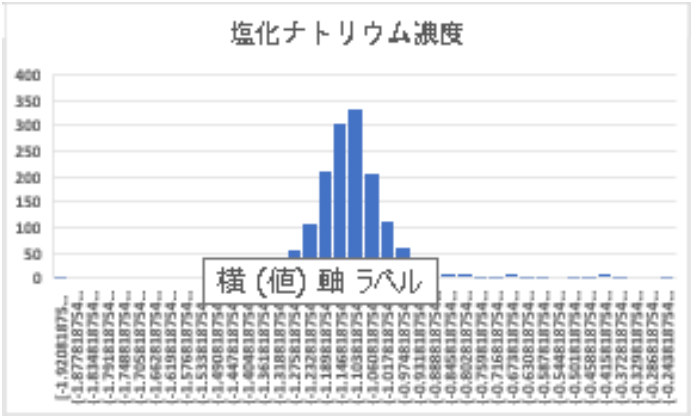
対数を取るとクエン酸、密度、評価を除いて正規性を改善することができます。

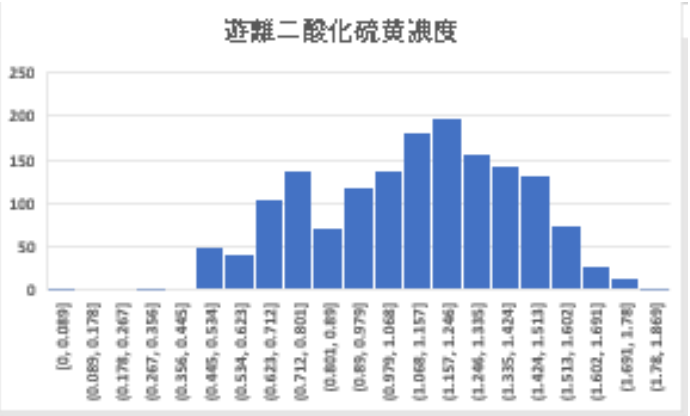

酒石酸濃度

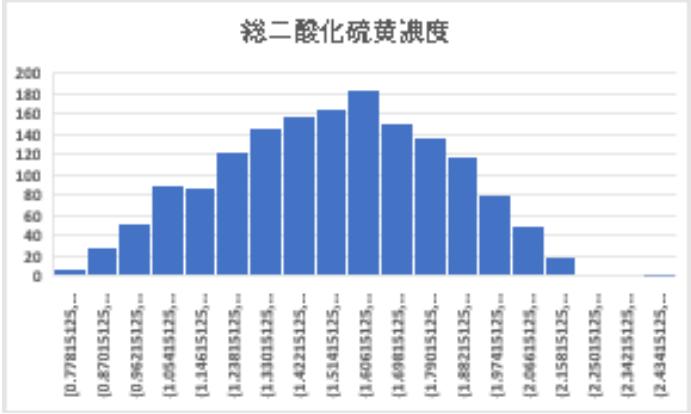

酢酸濃度

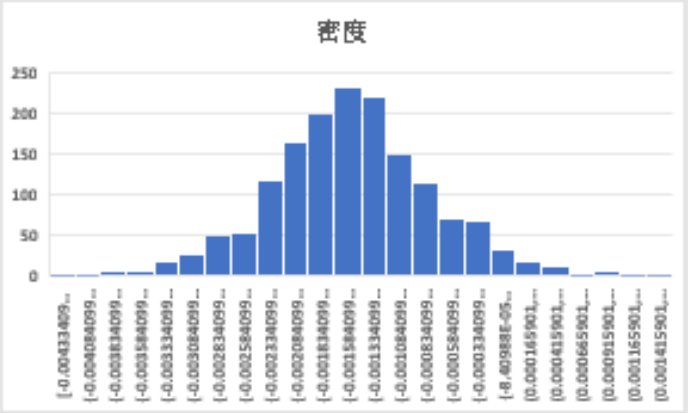

クエン酸濃度

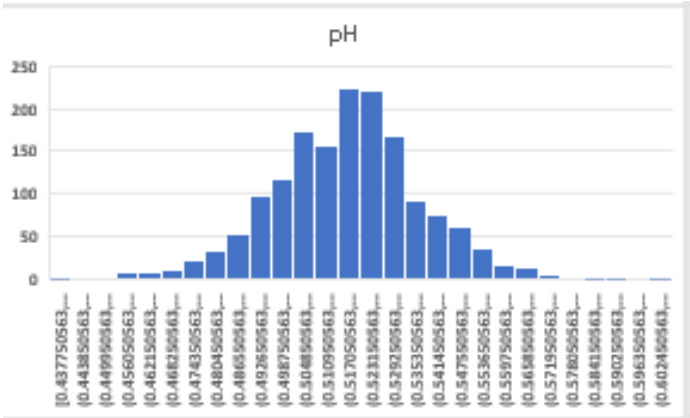

残糖濃度
横 (値) 軸 ラベル

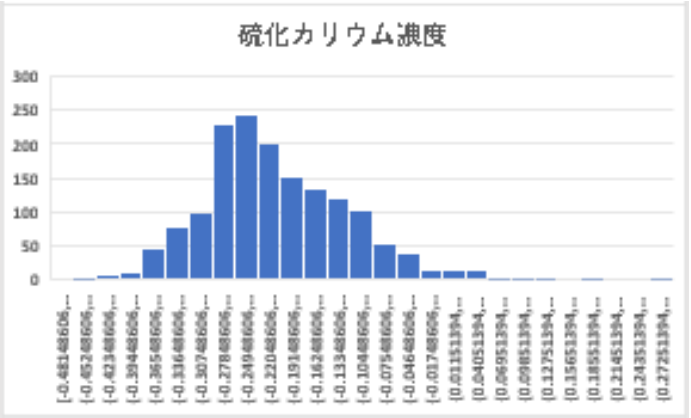

塩化ナトリウム濃度
横 (値) 軸 ラベル

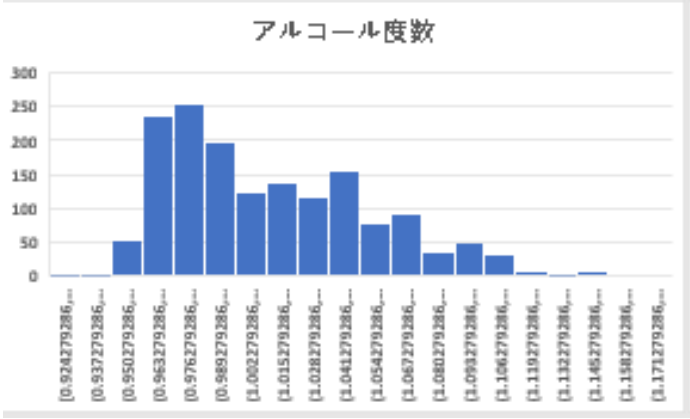

遊離二酸化硫黄濃度

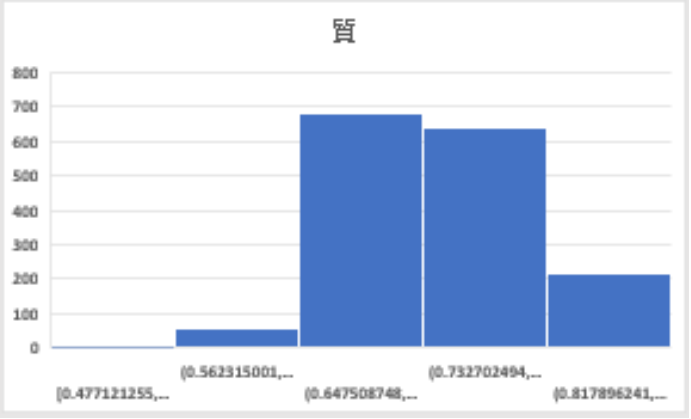

総二酸化硫黄濃度


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

第５章

練習問題 5.1　帰無仮説を立てるときには$\mu = \mu_0$というように＝を使って仮説を立てます。$\mu > \mu_0$とか$\mu < \mu_0$とか$\mu \neq \mu_0$と立てないのはなぜでしょうか？

仮説検定でその仮説をすてる、すてないを判断する際には、その仮説の母集団を１つに絞る必要があります。

練習問題 5.2 例題 5.2 でニキビの患者に薬を投与して、その平均が薬を投与する前の患者の平均よりも低いかどうかを検定しました。もしその結果が低いと判断されるとこの薬に効果があるといえるでしょうか？

たとえば、薬を投与したある患者のニキビの数が極端に減ったとします。かつ、他の多くの患者にニキビの数の増加が見られたとします。この場合でも平均のニキビの数が減る可能性があります。したがって、母平均と標本平均を比べただけでは、薬の効果の判断はできません。このような場合にはそれぞれの患者のニキビの増減を調べる必要があります。

ニキビの治療で患者が病院を訪れた場合、薬を服用する前の患者 A, B, C, D, E のニキビの数を記録しておく必要があります。A は１１、B は９、C は１２、D は８、E は１０とします。この観測値から得られる標本平均は１０です。薬を投与後の被験者のニキビの数は A は 9、B は９、C は１0、D は 9、E は 8 とします。

被験者	薬を投与前	薬を投与後	差
A	11	9	−2
B	9	9	0
C	12	10	−2
D	8	9	+1
E	10	8	−2

差の検定には”独立な場合の平均の差の検定”と”対応のある場合の平均の差の検定”があり、これらの検定の臨海値(棄却限界)を求めるためには t 分布を使います。この場合には$H_0: \bar{X} = \mu_0, H_1: \bar{X} < \mu_0$

$$t = \frac{\hat{d} - \mu}{\frac{s_d}{\sqrt{n}}}$$

上式を用いて、左側片側検定の式を得ると

$$\bar{d} - t_{(\alpha, n-1)} \frac{s_d}{\sqrt{n}} < \mu$$

となります。この式を用いて t 値を計算すると−1.58 になります。この値は有意水準 5%の t 統計量よりも大きいので帰無仮説を棄却することはできません。

対応のある差の検定	
投与前ごの差A	-2
投与前ごの差B	0
投与前ごの差C	-2
投与前ごの差D	1
投与前ごの差E	-2
差の平均	-1
標準偏差	1.414214
t	-1.58114
t(0.05,8)	-2.13185

つぎに薬を投与後の結果を

被験者	薬を投与前	薬を投与後	差
A	11	9	-2
B	9	9	0
C	12	10	-2
D	8	8	0
E	10	8	-2

この場合の対応のある差の検定を行うと t は-2.44 となるので、帰無仮説を棄却できます。

対応のある差の検定		
投与前ごの差A	-2	-2
投与前ごの差B	0	0
投与前ごの差C	-2	-2
投与前ごの差D	1	0
投与前ごの差E	-2	-2
差の平均	-1	-1.2
標準偏差	1.414214	1.095445
t	-1.58114	-2.44949
t(0.05,8)	-2.13185	-2.13185

クラゲの予測が当たるか当たらないかを統計的に判断するために、予測が当たる確率を p として帰無仮説を p=0.5 としました。一方で、ニキビの薬が効くか効かないかを判断するために、薬を投与する前の被験者のニキビの平均値を基準として、帰無仮説をμ＝１０としました。この２つの帰無仮説の根本的な違いは何でしょうか？

クラゲの予測が当たらないという仮説を帰無仮説としています。この場合、予測が当たる当たらないの 2 値で考えています。これは当たる確率が 0.5 よりも大きければ当たることになます。当たる確率が 0.5 よりも小さければ予測は当たらにことになります。しかし、当たらないことの逆は当たることなので、予測があたらないためには当たる確率が 0.5 である必要があります。なのでこの場合の帰無仮説は原理として決まる当たる確率＝0.5 となるのです。一方で、ニキビの場合には帰無仮説のμ＝１０は観測値から得らたものです。つまり、観測値の平均なので、薬を投与する前の被験者のニキビの標本平均であり、母平均ではありません。クラゲの予測の p=0.5 は母数です。つまり、帰無仮説を設定するときの値が、母数であるか、統計量であるかの違いがこの２つの帰無仮説の根本的な違いです。

帰無仮説が棄却されるという意味

仮説検定では標本にもとづいて帰無仮説と対立仮説のどちらかを統計的な判断で選択しています。対立仮説が選択されたというときには、この選択が誤りであるという確率は α 以下であると保証されています。つまり、対立仮説が強く成り立っていると主張することができます。

もちろん、確率変数が、帰無仮説が仮定する条件を満たしていないために棄却された場合には、この限りではありません。

つぎに、帰無仮説が棄却されなかったからといって、それを積極的に支持する理由にはなりません。それが誤りである確率が低いということはどこにも言及されていないのです。つまり、帰無仮説が棄却されないからといって、強く支持することはできないのです。帰無仮説が単に棄却される理由が不十分であったに過ぎないのです。ですので、帰無仮説を受容するとは言わずに、”棄却するには十分ではない”とか”すてることはできない“と表現するのです。

ここで注意が必要なのは確率変数が独立でないとか、一様でないとかという理由で棄却されてしまう場合があります。確率変数 X がある確率分布にしたがうとしたときは、確率変数は条件を満たしていなければなりません。確率変数 X がそれを満たさなければ、X はその確率分布にしたがいません。そうすると統計的検定には意味がなくなります。

第６章

練習問題 6.1：身長・体重の単回帰分析の例の有意 F について、エクセルシートを使って自分で求めてみましょう。

- “データ分析”の”回帰分析”の設定メニューで残差にチェックを入れると観測値と予測値の差である残差が出力されます。また、予測値が出力されているので、観測値と予測値の相関の 2 乗を計算することで決定係数が得られます。その平方根を取ると重相関 R が得られます。補正 R2 はデータの総数と回帰の自由を用いて調整することで得ることができます。
- 標準誤差は残差平方和をデータの自由度と回帰の自由度を用いて調整することで得ることができます。
- 回帰の分散(変動)は被説明変数の分散に自由度の合計を掛けたのもから残差平方和を引くことで得られます.
- 残差の分散は残差平方和を残差の自由度で割ることで得られます。
- 観察された分散比は回帰の分散を残差の分散で割ることで得られます。
- ++++++++
- F 値は観測された分散比を回帰の自由度と残差の自由度で調整した F 値を 1 から引いたものです。

	A	B	C	D	E	F	G	H	I	J
1					残差出力				確率	
2	身長	体重		観測値	予測値: Y	残差	標準残差		百分位数	Y
3	113	20		1	116.68	-3.68	-1.43		2.63	111
4	111	19		2	115.57	-4.57	-1.78		7.89	112
5	118	20		3	116.68	1.32	0.52		13.16	113
6	112	16		4	112.25	-0.25	-0.10		18.42	114
7	114	19		5	115.57	-1.57	-0.61		23.68	114
8	118	23		6	120.00	-2.00	-0.78		28.95	116
9	124	22		7	118.89	5.11	1.99		34.21	117
10	118	20		8	116.68	1.32	0.52		39.47	118
11	122	23		9	120.00	2.00	0.78		44.74	118
12	119	19		10	115.57	3.43	1.33		50.00	118
13	118	22		11	118.89	-0.89	-0.35		55.26	118
14	116	22		12	118.89	-2.89	-1.12		60.53	118
15	118	21		13	117.78	0.22	0.08		65.79	119
16	114	19		14	115.57	-1.57	-0.61		71.05	119
17	119	22		15	118.89	0.11	0.04		76.32	119
18	117	20		16	116.68	0.32	0.13		81.58	119
19	119	20		17	116.68	2.32	0.90		86.84	120
20	119	24		18	121.10	-2.10	-0.82		92.11	122
21	120	20		19	116.68	3.32	1.29		97.37	124
22	11.005848	=VAR(A3:A21)				118.9487	=SUMSQ(f3:f21)			
23	198.10526	=A22*B37				6.996982	=F22/B36 -> 残差->分散			
24						79.15657	=+A23-F22 ->回帰->変動、分散			
25	概要									
26	回帰統計									
27	重相関 R	0.6321	=SQRT(L6)							
28	重決定 R2	0.3996	=CORREL(A3:A21,E3:E21)^2							
29	補正 R2	0.3642	=1-(1-CORREL(A3:A21,E3:E21)^2)*(B31-1)/(B31-B35-1)							
30	標準誤差	2.6452	=+SQRT(F22)/SQRT(B31-B35-1)							
31	観測数	19								
32										
33	分散分析表									
34		自由度	変動	分散	観測された分散比	有意 F				
35	回帰	1	79.157	79.157	11.31295808	0.003688	=1-F.DIST(E35,B35,B36,1)			
36	残差	17	118.95	6.997	=D35/D36					
37	合計	18	198.11							

線形回帰.xlsx=>身長_体重_検証

練習問題 6.2: 赤ワインデータについて重回帰分析により評価を予測できるどうか確かめて見よ。

- 回帰統計はモデルの適性を示しています。
- 有意 F はゼロですべての回帰係数がゼロであるという帰無仮説を棄却しています。
- P-値ではじかれたデータは酒石酸濃度、クエン酸濃度、残糖濃度です。
- 最も高い調整済み係数はアルコール濃度です。

	N	O	P	Q	R	S	T	U
1	概要							
3	回帰統計							
4	重相関 R	0.993547						
5	重決定 R2	0.987135						
6	補正 R2	0.986424						
7	標準誤差	0.648018						
8	観測数	1599						
9								
10	分散分析表							
11		自由度	変動	分散	観測された分散比	有意 F		
12	回帰	11	51167.16	4651.56	11077.07007	0		
13	残差	1588	666.8439	0.419927				
14	合計	1599	51834					
15								
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	調整済み係数
17	切片	0.0000						
18	酒石酸濃度	0.0042	0.0165	0.2540	0.7995	-0.02809	0.036448	0.01
19	酢酸濃度	-1.0995	0.1201	-9.1554	0.0000	-1.33509	-0.86396	-0.24
20	クエン酸濃度	-0.1839	0.1472	-1.2497	0.2116	-0.47258	0.104749	-0.04
21	残糖濃度	0.0071	0.0121	0.5854	0.5583	-0.01658	0.030694	0.01
22	塩化ナトリウム濃度	-1.9117	0.4178	-4.5763	0.0000	-2.73115	-1.09234	-0.11
23	遊離二酸化硫黄濃度	0.0046	0.0022	2.1115	0.0349	0.000325	0.008817	0.06
24	総二酸化硫黄濃度	-0.0033	0.0007	-4.5715	0.0000	-0.00475	-0.0019	-0.14
25	密度	4.5299	0.6253	7.2440	0.0000	3.30337	5.756499	0.01
26	pH	-0.5231	0.1600	-3.2695	0.0011	-0.83693	-0.20928	-0.10
27	硫化カリウム濃度	0.8871	0.1108	8.0061	0.0000	0.669737	1.104386	0.19
28	アルコール度数	0.2970	0.0173	17.2163	0.0000	0.263166	0.330841	0.39

赤ワインデータ.xlsx=>重回帰分析

練習問題 6.3：AIC と BIC を用いてスペクターとマッテーオのデータのモデルを再度検討してみてください。

説明変数を変えて、2 つのタイプについて評価しました。その結果はつぎのようになっています。

成績の改善/ 経済の評価点/ 参加・不参加	AIC	-1.93915
	BIC	-59.0077
成績の改善/ 生徒の評価点/ 参加・不参加	AIC	6.374183
	BIC	-50.6943
成績の改善/ 経済の評価点/ 生徒の評価点/ 参加・不参加	AIC	-2.32019
	BIC	-59.3887

結果として経済の成績、生徒の評価点、参加・不参加のデータを用いたものが一番良いという結果になりました。

重回帰分析.xlsx=>重回帰分析、重回帰分析（2）、重回帰分析（3）

概要	原系列

回帰統計	
重相関 R	0.68
重決定 R2	0.46
補正 R2	0.40
標準誤差	0.37
観測数	32

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	3.32	1.11	7.93	0.00
残差	28	3.90	0.14		
合計	31	7.22			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-1.51	0.50	-2.99	0.01	-2.54	-0.47
PCA1	0.03	0.02	1.89	0.07	0.00	0.07
PCA2	0.46	0.16	2.95	0.01	0.14	0.78
参加・不参	0.43	0.13	3.20	0.00	0.16	0.71
	残差標準偏差	0.35				
	AIC	-2.32				
	BIC	-59.39				

概要	標準化

回帰統計	
重相関 R	0.68
重決定 R2	0.46
補正 R2	0.40
標準誤差	0.37
観測数	32

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	3.32	1.11	7.93	0.00
残差	28	3.90	0.14		
合計	31	7.22			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	0.17	0.09	1.96	0.06	-0.01	0.34
PCA1	0.18	0.06	3.18	0.00	0.06	0.30
PCA2	0.17	0.12	1.43	0.16	-0.08	0.42
参加・不参加	0.43	0.13	3.20	0.00	0.16	0.71
	残差標準偏差	0.35				
	AIC	-2.32				
	BIC	-59.39				

主成分分析.xlsx => 重回帰原系列、重回帰標準化

データに原系列を使ったものと、標準化したものと２つについて試してみました。両者のパフォーマンスはAIC,BIC 基準、決定係数を基準とした場合には変わりません。

第７章

練習問題 7.1：主成分分析を用いて重回帰の際の多重共線性が解消できるかどうかをスペクターとマッテーオのデータで試してみましょう。

概要	原系列					
回帰統計						
重相関 R	0.68					
重決定 R2	0.46					
補正 R2	0.40					
標準誤差	0.37					
観測数	32					
分散分析表						
	自由度	変動	分散	観測された分散比	有意 F	
回帰	3	3.32	1.11	7.93	0.00	
残差	28	3.90	0.14			
合計	31	7.22				
	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-1.51	0.50	-2.99	0.01	-2.54	-0.47
PCA1	0.03	0.02	1.89	0.07	0.00	0.07
PCA2	0.46	0.16	2.95	0.01	0.14	0.78
参加・不参	0.43	0.13	3.20	0.00	0.16	0.71
	残差標準偏差	0.35				
	AIC	-2.32				
	BIC	-59.39				

概要	標準化					
回帰統計						
重相関 R	0.68					
重決定 R2	0.46					
補正 R2	0.40					
標準誤差	0.37					
観測数	32					
分散分析表						
	自由度	変動	分散	観測された分散比	有意 F	
回帰	3	3.32	1.11	7.93	0.00	
残差	28	3.90	0.14			
合計	31	7.22				
	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	0.17	0.09	1.96	0.06	-0.01	0.34
PCA1	0.18	0.06	3.18	0.00	0.06	0.30
PCA2	0.17	0.12	1.43	0.16	-0.08	0.42
参加・不参加	0.43	0.13	3.20	0.00	0.16	0.71
	残差標準偏差	0.35				
	AIC	-2.32				
	BIC	-59.39				

主成分分析.xlsx => 重回帰原系列、重回帰標準化

データに原系列を使ったものと、標準化したものと２つについて試してみました。両者のパフォーマンスはAIC,BIC 基準、決定係数を基準とした場合には変わりません。

第８章

練習問題 8.1：因子分析を用いて重回帰の際の多重共線性が解消できるかどうかをスペクターとマッテーオのデータで試してみましょう。

概要

回帰統計	
重相関 R	0.60
重決定 R2	0.36
補正 R2	0.29
標準誤差	0.41
観測数	32

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	2.58	0.86	5.18	0.01
残差	28	4.64	0.17		
合計	31	7.22			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	0.19	0.10	1.90	0.07	-0.02	0.40
因子1	-0.22	0.46	-0.48	0.64	-1.16	0.72
因子2	0.55	0.51	1.10	0.28	-0.48	1.59
参加・不参加	0.47	0.15	3.15	0.00	0.16	0.77
	残差標準偏差	0.39				
	AIC	3.24				
	BIC	-53.83				

因子分析.xlsx => 重回帰分析

結果は主成分分析、重回帰分析の結果に若干劣ります。