金融財務研究会

㈱金融財務研究会
㈱経営調査研究会

確率・統計学入門

午後の部（エクセル編）

ZOOM ミーティング

2022 年 7 月 22 日(木)
9:30－12:30
講師：森谷博之Quasars22

テレビや新聞、雑誌、インターネットには表やグラフが溢れています。自由にダウンロードできるデータも日に日に増えています。また、企業内では豊富なデータが蓄積されています。身の回りにある表やグラフを適切に解釈でき、手元にあるデータを効率よく、効果的で、そして論理的に処理できる人材が求められています。

本セミナーでは、エクセルを用いて、確率・統計の基本的な考え方を身につけます。一般に、記述統計といわれる身の回りの表やグラフの解釈の仕方、得られたデータを散布図、頻度図などに変換して表現する方法などを学びます。また、より論理的にデータを処理する推測統計で必須の知識である確率について学びます。エクセルを用いて問題を表現したり、解いたりすることでデータの処理の仕方の基本を身につけます。

第１部：　記述統計入門 9：30～11：00
- データの種類と収集方法：質的変数、量的変数、全数調査と標本調査
- 基本的な統計グラフ：　度数分布表と散布図
- 基本的な統計量：平均、分散、標準偏差、偏差、相関など
- 相関と因果関係

第２部：　確率の基礎　11：10～12：30
- 確率の基礎：同様に確からしい、事象と確率、確率変数とモデル

■本セミナーに参加して修得できること

発生した問題についてデータをとおして理解する基本的な態度、得られた表やグラフの正しい解釈の仕方、手元のデータを表やグラフを用いて表現する方法がわかります。また、エクセルの活用方法がわかります。演習を多く取り入れ、理解を深めます。

■受講対象者

統計に関して「超初心者」の方を対象としています。

学生時代に確率・統計を学んだが覚えていない、確率・統計学が日々の仕事、研究にどのように役立つのかわからない、統計分析をしたいがどうすれば良いかわからない、部下の統計分析を理解したい、データ分析の本質を理解したい方など。

■使用ソフト：Excel。

■PC には事前に Excel がインストールされている必要があります。

参考文献：

「データの活用」（日本統計学会編）東京図書、
「データの分析」（日本統計学会編）東京図書、
「統計学基礎」（日本統計学会編）東京図書

第1章　記述統計

統計データとはどのようなものでしょうか？政府統計の総合窓口のウェブサイト(https://www.e-stat.go.jp/)では、「統計データを探す」というページがあり、様々な統計データを得ることができます。それらは国税調査、経済センサス、人口推計などに分類されています。データを探す→分野→企業・家計・経済とクリックし小売物価統計調査をみていくと東京都区部のマグロやイワシの平均価格を見ることができます。また、データを探す→分野別→人口・世帯とクリックしファイル→月次→２０２１年１１月のエクセルファイルをみていくと年齢別人口構成を見ることができます。日本の国土の広さ、人口、経済規模を表す国内総生産などは統計データです。これらの統計データは、ある目的をもってデータを集め集計することで出来上がっています。データは過去の記録から得たものや、新たに実験・調査・測定を行ったりして集められています。これらのデータは数値とは限りません。性別、居住地、職業、天気などのデータは文字列です。調査される対象を一般的に、個体またはケース、その項目を変数といいます。本章では、与えられたデータのもつ集団の性質を記述し、要約する方法を学びます。これを記述統計といいます。

1.1.　変数の分類

データはその性質や特性を表す文字列であったり、数値であったりします。私たちの身の回りはデータであふれています。これらのデータを統計的に分析するためには文字列で表された属性が数値に変換されていると便利なことがあります。たとえば、居住地をコンピュータで処理するために数値化することが良くあります。しかし、この数値の平均を求めても何の意味もありません。このようにデータを数値で表すときには、その尺度を理解しておくことが必要です。一般に、このような尺度は４つに分類されます。

- 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：名義尺度のすべての性質に加えて順序(大小関係)が意味をもつ尺度。
  ５段階評価の成績、レストランのランキング、信用評価(AAA, AA, A, , ,)など
- 間隔尺度：順序尺度のすべての性質に加えて、０が相対的な意味をもち、等間隔の大小関係をもち、値の差が意味をもつ尺度。温度、偏差値、西暦など
- 比例尺度：間隔尺度のすべての性質に加えて、単位をもち、ゼロが絶対的な意味をもつ尺度。距離、時間、測度、体重、年齢、身長、収入、絶対温度など。また、乗除の演算が意味をもち、40kg は 20kg の 2 倍ですし、距離を時間で割ると速度という意味をもちます。ほとんどの物理的な量は比例尺度です。

これらの尺度・変数は質的変数と量的変数に分類されます。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。名義尺度と順序尺度は質的変数となり、それらの性質は文字列で表現されます。また、質的変数は２値変数や多値変数で表現できます。一方で、温度や体重などは量的変数です。間隔尺度と比例尺度は量的変数となります。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。

つぎの表は、ポルトガルのミーニョ地方（北西部）ヴィーニョ・ヴェルデのアルコール度数中程度の赤ワインの評価と物理化学的検査の結果です。データは 2004 年 5 月から 2007 年 2 月にかけて収集され公式認証機関（CVRVV）で検査されています。CVRVV は、ヴィーニョ・ヴェルデの品質の向上とマーケティング強化を目的とした専門組織です。ワインのサンプル検査はプロセスを自動的に管理するコンピュータシステムによって記録され

ました。また、評価については，各サンプルを最低 3 人の専門家が評価しています。評価は，0（非常に悪い）から 10（素晴らしい）までのブラインドテイスティングの結果です。これからこのデータベースを活用して、データ分析の手法を学んでいきます。

	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

表 1.1　ワインデータ

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

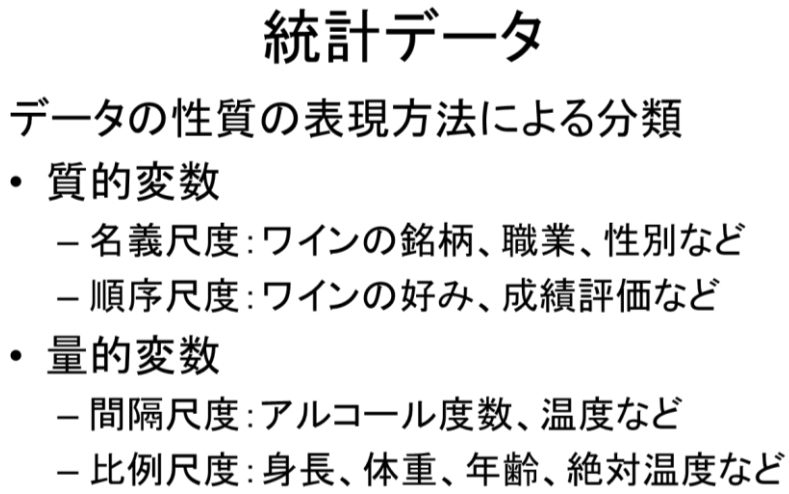


図　1.1 データの性質

厳密なデータの分類は、統計的分析手法の選択の原点です。

1.2. 記述統計

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します。また、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みやすくすることがあります。

1.2.1. 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。

A) 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。

例題 1.1：ワインデータの評価、化学成分 K と B の頻度図を作ってみましょう。

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。図 1.2 は赤ワインの１０段階評価を頻度図にしたものです。

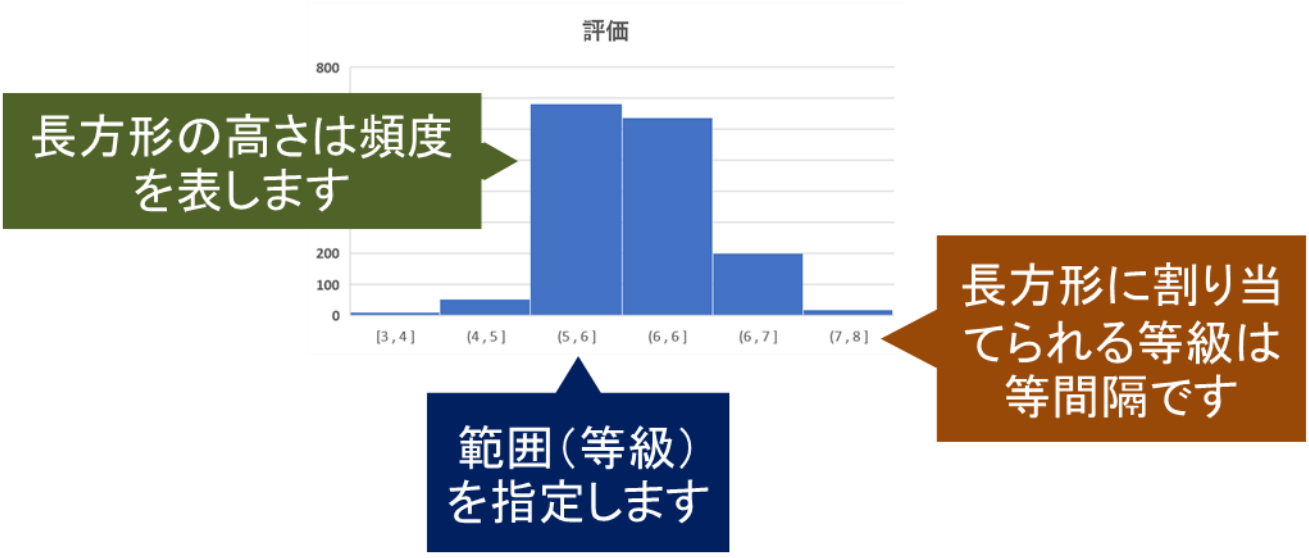


図　1.2　頻度図の製作ヒント

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で最も低い評価は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。頻度の分布はおおよそ左右対称ですので、このような頻度図をベル型と呼びます。

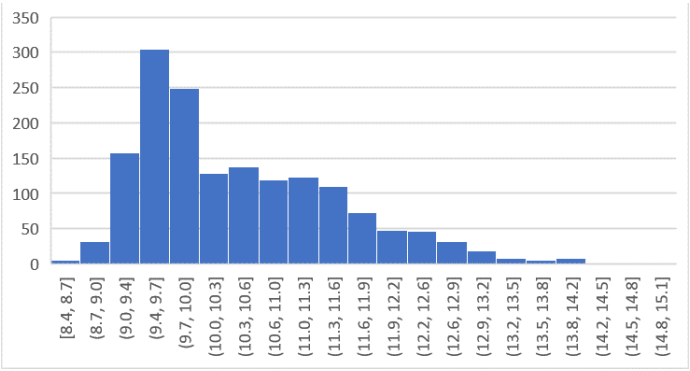


図　1.3　頻度図 化学成分 K

図 1.3 の横軸は化学成分 K です。最も頻度の高い K は 9.5 近辺です。すそ野は右に長くなっています。頻度図の度数は左によっています。これは右にひずんでいます。

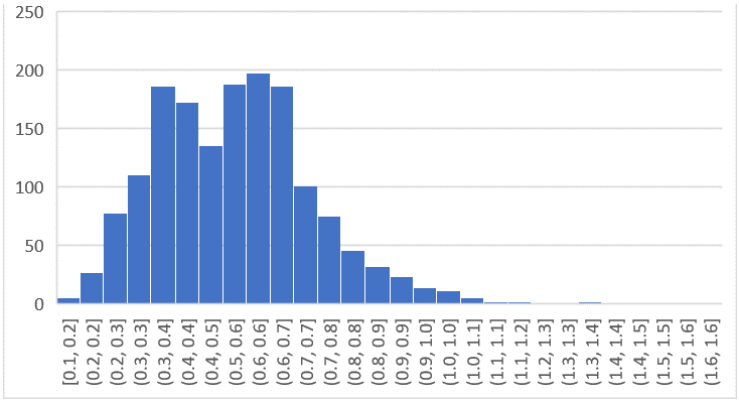


図　1.4　頻度図 化学成分 B

図 1.4 の横軸は化学成分 B です。最も頻度の高い B は 0.6 近辺です。頻度図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つの単峰の頻度図が混じっているようにも見えます。図 1.5 は化学成分 B についてのものです。一番上の図は図 1.4 と同じものです。2 番目は、それよりも等級を細かくしています。一番下は等級を荒くしています。2 こぶが消えてしまっています。幅の大きさによって、頻度図から受けるイメージが変わります。

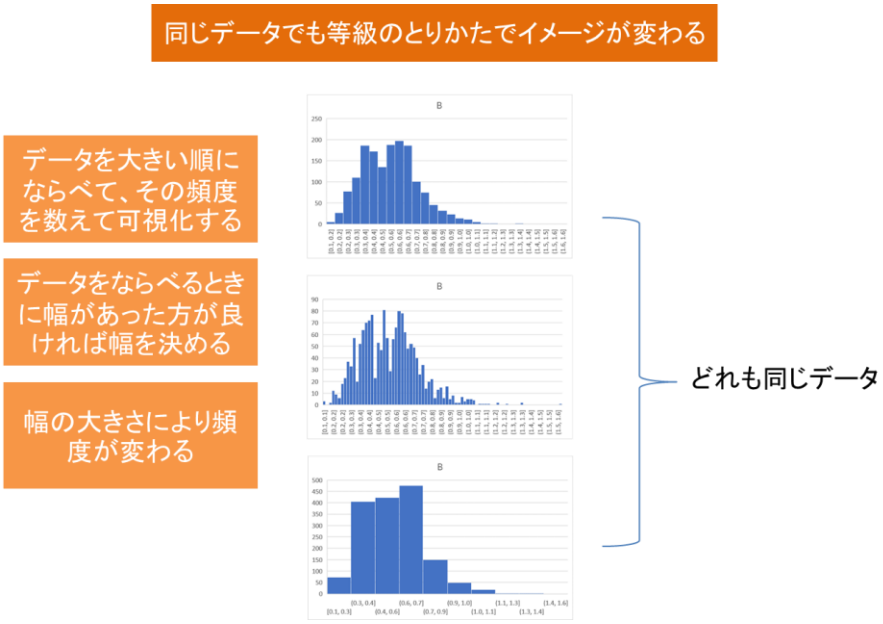


図　1.5　頻度図の等級と与える印象

頻度図の形状は大まかに

- 一様：頻度が横軸の値に対してほぼ均等。

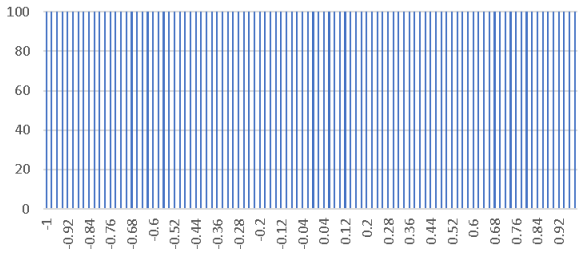


図　1.6　一様な頻度図(例題 2.6)

- ベル型：頻度の高さは横軸に対してベル型。

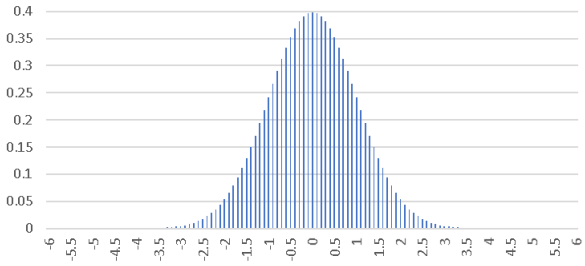


図　1.7　ベル型の頻度図(例題 2.7)

- 右に裾長：右にすそ野が長く、頻度が左寄り。

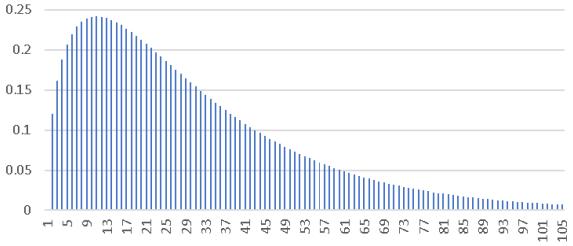


図　1.8 右にすそ長

- 左に裾長：左にすそ野が長く、頻度が右に寄り。

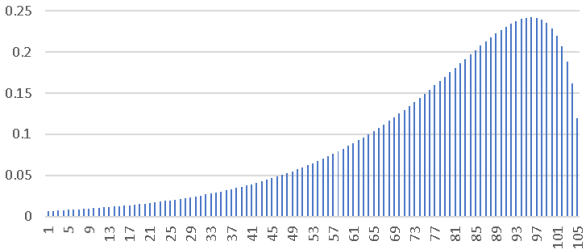


図　1.9　左にすそ長

- 複数のこぶ：複数頻度の山やコブ。

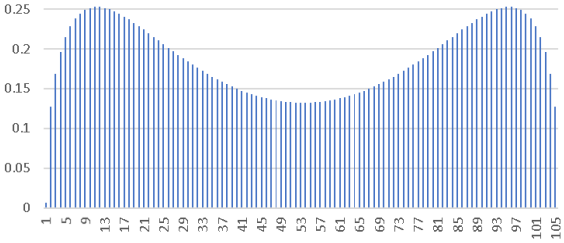


図　1.10　複数頻度の山に分けられます。

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。

B)　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。２つのデータの関係と散らばり具合を大まかにつかむことができます。

例題 1.2：ワインデータの評価と化学成分 K、化学成分 K と化学成分 B、評価と化学成分 B の散布図を作ってみましょう。

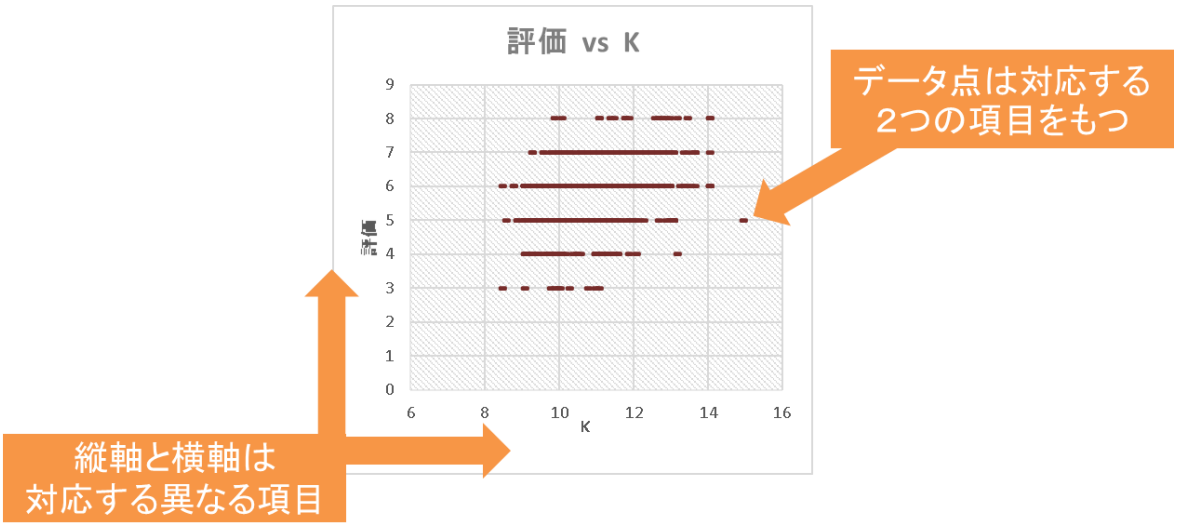


図　1.11　散布図

図 1.11 は、横軸に化学成分 K、縦軸に赤ワインの評価を目盛っています。化学成分 K 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す。また、データ点は横に一列に並んでいる線が平行に 6 本あります。これは評価が離散値であるためです。

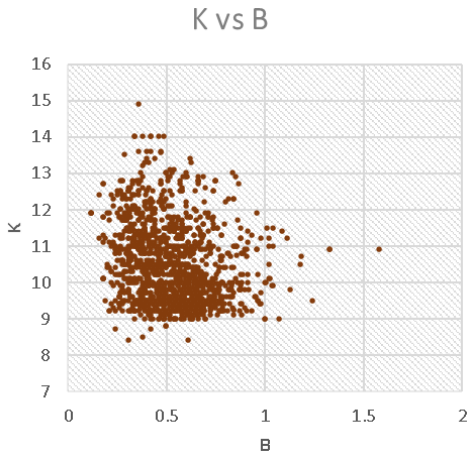


図　1.12　散布図：化学成分 K vs B

図 1.12 は、横軸に化学成分 B、縦軸に化学成分 K を取っています。この散布図から大きな傾向は見られません。化学成分 B が高くなると化学成分 K 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、化学成分 B 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なりこのように見える可能性があります。

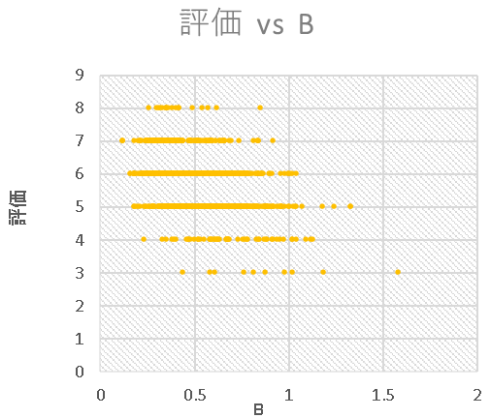


図　1.13　散布図: 評価 vs 化学成分 B

図 1.13 は、横軸に化学成分 B、縦軸に評価を取りました。化学成分 B が上がると評価が下がる傾向がありそうです。しかし、化学成分 B 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。

図　1.14 可視化

1.2.2.　要約統計量

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。記述統計量、基本統計量、代表値ともいいます。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 つ目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を示すものです。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つ目は頻度図(分布)の形状に関するものです。

1.2.2.1.　1 変量要約統計量

まずはなじみの深い平均を見ていきます。

A)　平均(算術平均)

平均は日常生活でもっともなじみの深い基本統計量の１つです。平均にもいろいろな計算方法があるのですが、通常は算術平均のことです。その計算を１から５までの数値を用いて行ってみましょう。

$$\frac{1+2+3+4+5}{5}=3$$

となります。これは何を表しているのでしょうか？平均は与えられた数値のある特定の位置を表す統計量です。まずは元の数値をみてみましょう。

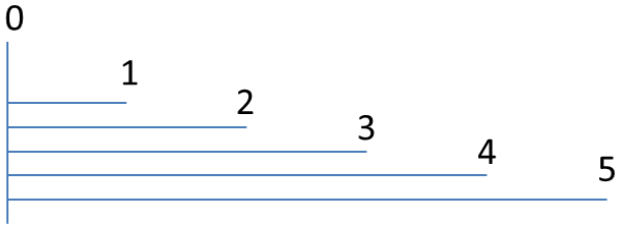


図　1.15　5 つの数値

図 1.15 は 1 から 5 までの数値を、0 を起点に並べてみたものです。数値の大きさを比較するには便利です。つぎに平均の使い方をみてみましょう。 1 から 5 までのそれぞれの値から平均を引いてみましょう。

1−3=−2

2−3=−1

3−3= 0

4−3= 1

5−3= 2

それぞれの計算結果は 1 から 5 までの数と平均との差です。差は距離とも考えられます。つぎにこの計算結果を足し合わせてみましょう。-2-1+0+1+2=0 になります。これは何を意味しているのでしょうか？結果はマイナスのものとプラスのものに分かれました。それらを足し合わせるとゼロになるのですから、平均は与たえられた数値全体の中心の位置を表しています。図 1.16 を見てください。

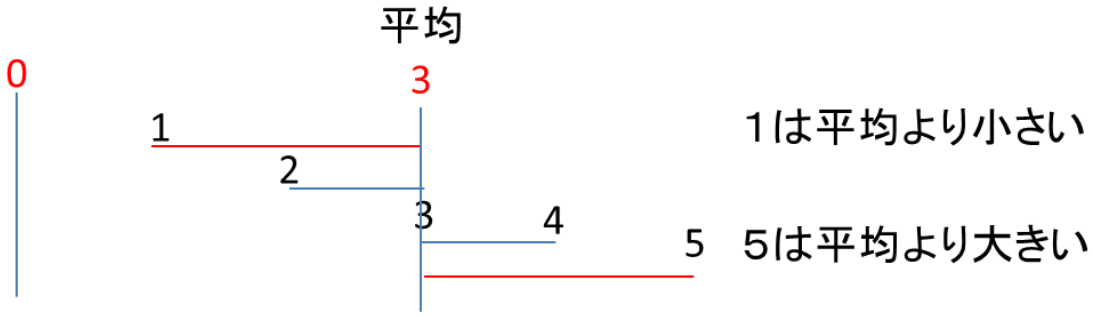


図　1.16 平均の分析

となります。比較の基準を 0 から 3 に変更するだけで見方がだいぶ違ってきます。

n個の数値の平均は、a_iをi番目の数とすると、その計算方法は

$$\frac{a_1 + a_2 +, \ldots, + a_n}{n}$$

となります。これはさらに

$$\bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$$

と書くことができます。$\bar{a}$はaの平均を意味します。

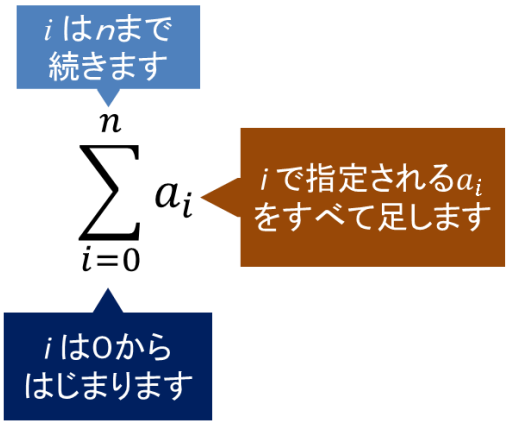


図　1.17 シグマの意味

つぎにデータの散らばりについて見てみましょう。たとえば、１から５の整数をまずならべてみます。

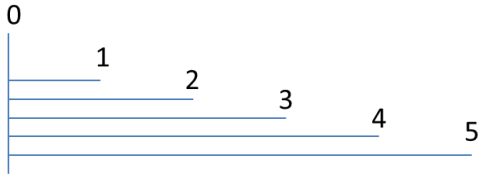


図　1.18　5 つの数値

１から５までの数値のばらつきを考えるときに、それぞれの数値とその平均との差、つまり偏差をとります。これもばらつきの尺度になります。グリーンの数値は偏差を表しています。

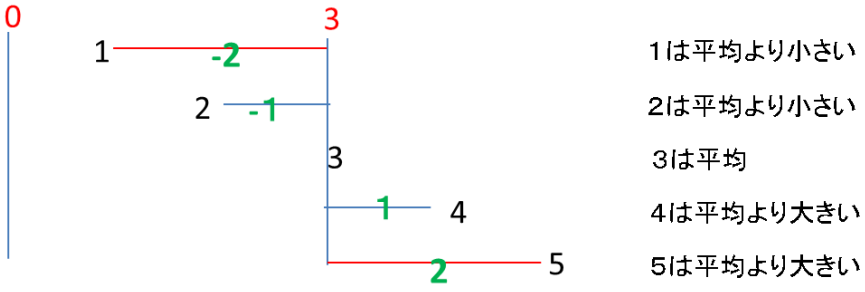


図　1.19　偏差

この偏差を足し合わせるとゼロになってしまうので、ばらつきの尺度としては適当ではありません。

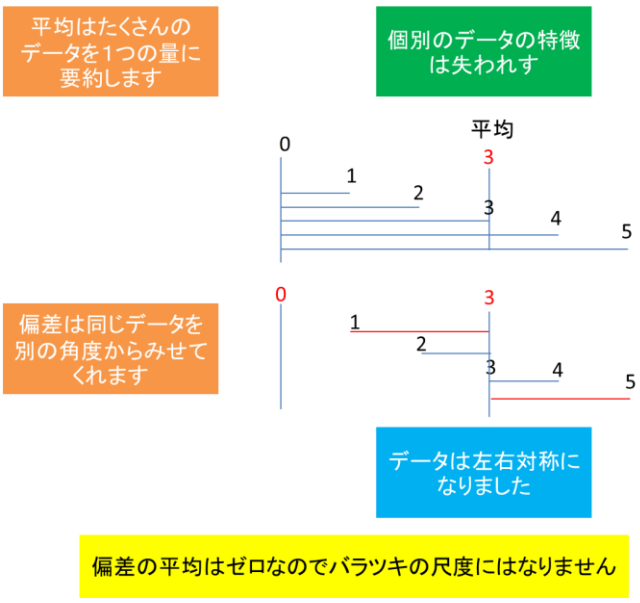


図　1.20　平均と偏差

B） 分散

統計学の分散は、数値の集団の散らばり具合を表します。それぞれの数値と平均との差を取り、それを２乗して総和をとり、数値の数で割ったものです。つぎのように定義されます。

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

ここで、$\bar{x}$は x の平均です。n は数値の数です。つまり、x_i の偏差の２乗の平均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。

偏差を求めて 2 乗して総和を求め、総数で割るという方法が何を意味するのか考えてみましょう。まず、２乗することで負の偏差を正の値に変えることができます。したがって、偏差の２乗を足し合わせてもゼロになることはありません。しかし、２乗して足し合わせただけではデータの数が多くなれば、２乗和はどんどん大きくなってしまいます。そこでその平均を求めて、データの数の影響を排除しているのです。

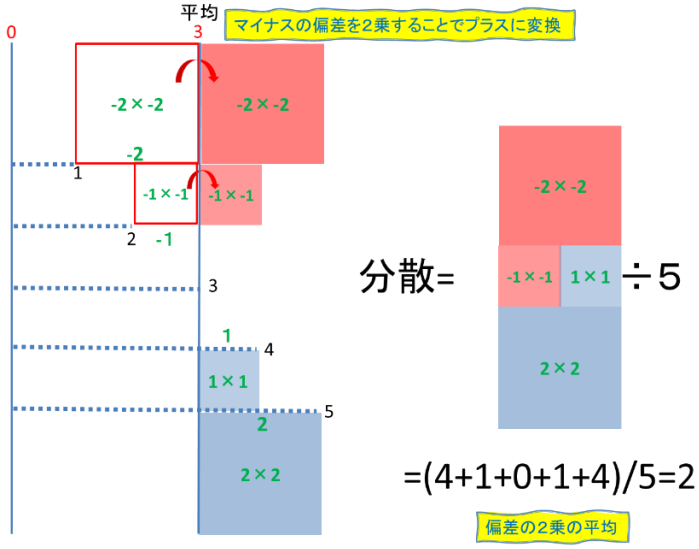


図　1.21　分散の理解

分散は偏差の２乗の平均です。しかし、分散は偏差を２乗しているためにデータの平均とは次元が違うことに注意してください。

図　1.22　平均と偏差と分散の理解

C) 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、数の集団の散らばり具合を表す指標です。

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

標準偏差の意味を考えてみましょう。分散は元のデータの 2 乗を用いて計算しています。したがって、2 次元です。その平方根を取ることで、次元をもとの数値の 1 次元にもどしているのです。標準偏差は分散の弱点を克服しています。図 1.23, 図 1.24 は標準偏差の意味のイメージ図です。

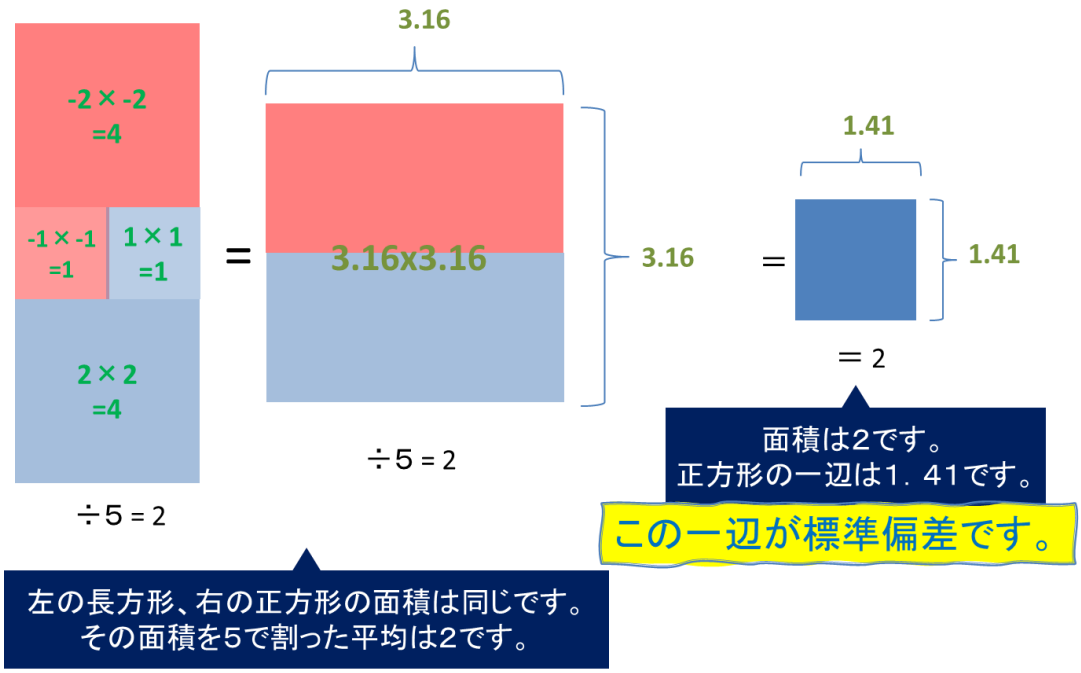


図　1.23 標準偏差のイメージ図

標準偏差はここで求めた面積の一辺だと考えることができます。したがって、分散と違い元のデータの次元と同じです。

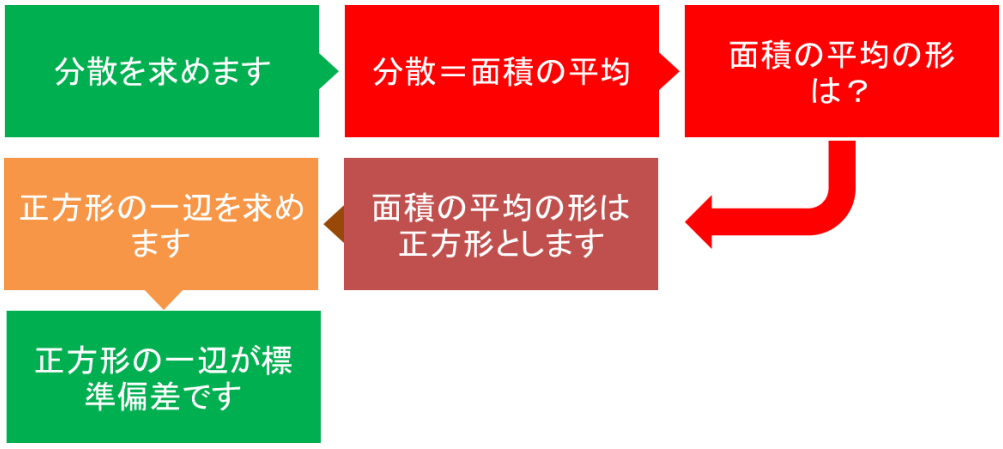


図　1.24 分散と標準偏差

1.2.2.2.　　2 変量要約統計量

平均、中央値、分散、標準偏差は、一変量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。

A) 共分散

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

$$\mathrm{Cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

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す。xとyが同じであると、共分散は分散になります。表 1-2 のように共分散は表(行列)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。

表 1-1　共分散

	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*)

B) 相関

共分散は 2 組のデータ(x, y)のもつ特徴をとらえようとしているのですが、図 1.25 にあるように、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。

$$\rho_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

これが相関です。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。

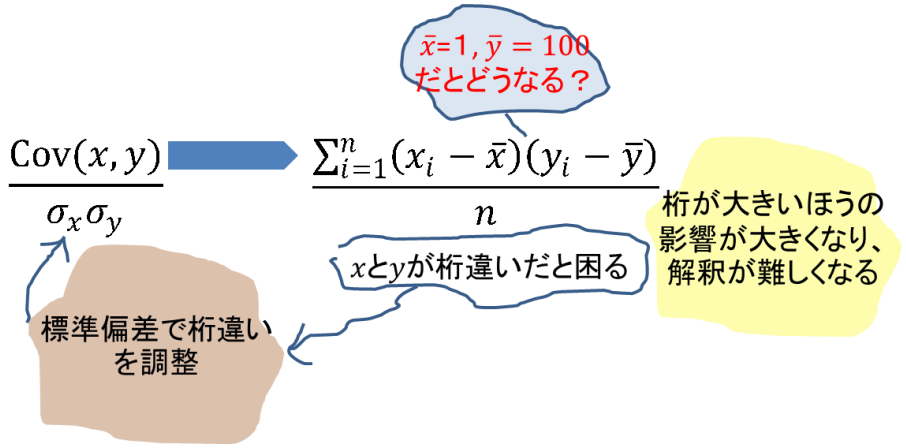


図　1.25 標準偏差と共分散

共分散同様に、相関も行列(マトリックス)を用いて表現すると便利なことがあります。

相関を、散布図を用いて可視化してみましょう。図 1.26~図 1.30 は乱数を用いて確率変数の列を２つ生成し、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。(練習問題 1.9)

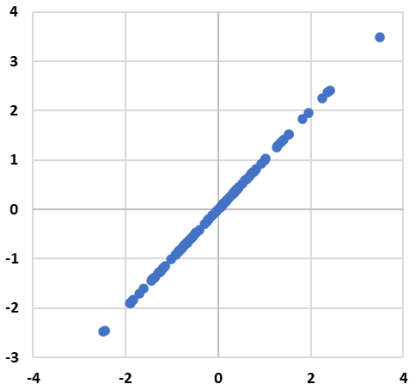


図　1.26　正の完全相関：相関＝ 1

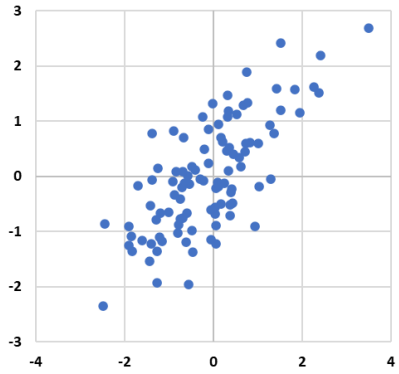


図　1.27 正の相関：相関＝0.7

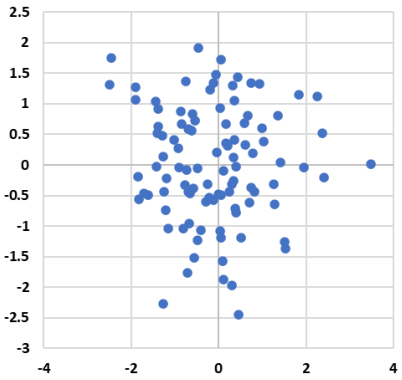


図　1.28　無相関：相関＝0

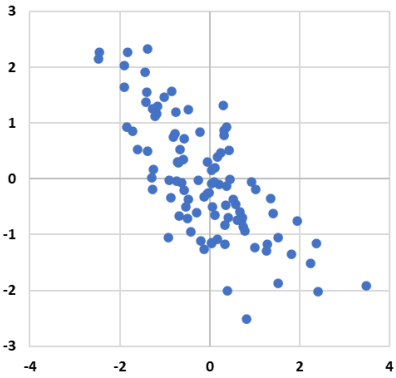


図　1.29　負の相関：相関＝-0.7

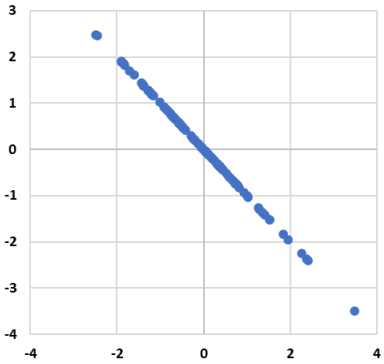


図　1.30 負の完全相関：相関＝-1

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はなりません。この点には注意が必要です。

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。

図 1.31 相関係数

例題 1.3：ワインデータの相関マトリックスを作成してみましょう。

表 1-2 相関行列

	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

エクセルの分析ツールを用いると簡単に相関行列が作れます。

1.2.2.3. 頻度図の形状に関する要約統計量

観測値の頻度の形状を頻度図によりイメージする方法を紹介しましたが、基本統計量でもつかむことができます。

A) 歪度(わいど):skew

頻度図の歪の度合いを表す歪度(skew)は

$$\text{skew} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n} \Big/ \sqrt{\sigma^3}$$

で表すことができます。歪度がゼロであると左右対称の頻度図(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。

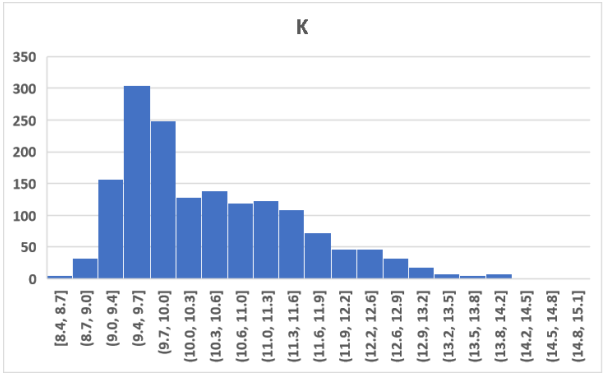


図　1 .28 歪度＝0.86　（練習問題 1.1）

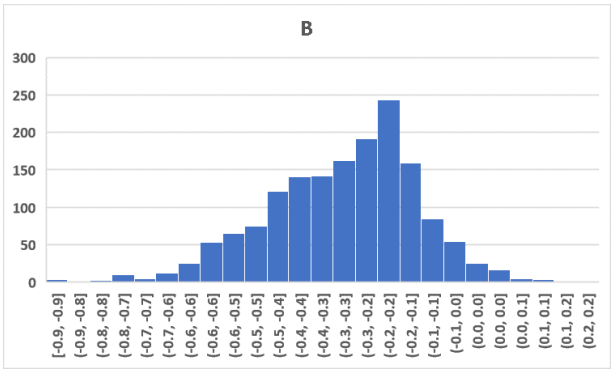


図　1 .29 歪度=- 0.43 (練習問題 4.3)

B)　尖度(せんど): kurt

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。

$$\mathrm{kurt}=\frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n}}{(\sigma^2)^2}-3$$

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。

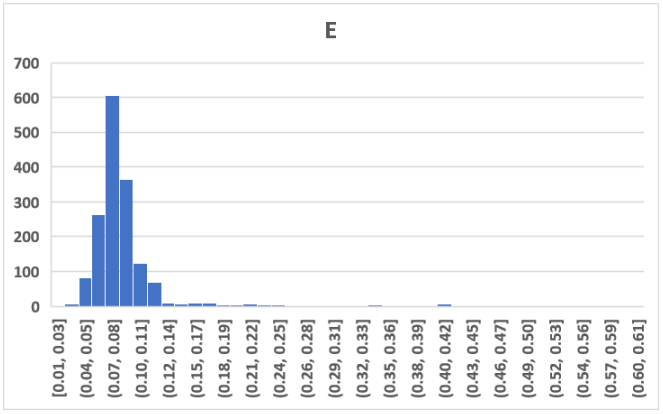


図　1 .30 尖度=1.1 (練習問題 1.1)

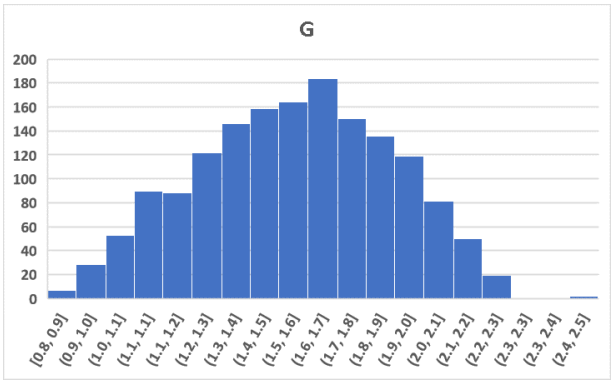


図　1.31　尖度-0.67（練習問題 4.3 データの対数）

例題 1.4：ワインデータの歪度と尖度を計算してみましょう。

表 1-4　歪度と尖度

	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

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。

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。

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割り当てられています。実際の化学成分を使わずに記号を用いている理由は何でしょうか？

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？

練習問題 1.5：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 1.6：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？

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？

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？

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。

第2章 確率変数:確率

本章の主役は確率変数です。確率変数というと予測が不可能は特徴を持った変数というイメージがありませんか？そのイメージを払しょくするために、まず確率と確率分布について学びます。`確率`は通常の会話でもよく使われ、なじみのある単語です。分布も同様で、何となく日常で思い浮かべるイメージがあるのではないでしょうか。本章では、統計学でいう確率と分布、そして確率変数についてサイコロを用いて学んでいきます。これはデータの背後にある集団について理論的に推測するための下準備です。

2.1. 確率

サイコロを投げるとき、その結果は偶然に左右されます。１が出るときもあれば６が出るときもあります。サイコロには６つの面があり、１つ１つの目には１から６までの数字が書き込まれています。この６面に書き込まれた数のように、これ以上分けるこのできない結果を根元事象といいます。サイコロの根元事象は{1},{2},{3},{4},{5},{6}です。サイコロが賭けに使われるときは、目の数よりは、目が偶数であるか奇数であるかに興味があるかもしれません。奇数の目は{1,3,5}で、偶数の目は{2,4,6}です。これは出た目をグループとしてまとめているので、偶数の目、奇数の目は根元事象ではありません。これらは事象です。事象は根元事象で構成されています。サイコロを振って結果を観察することを試行といいます。根元事象とは、試行によって起こる、それ以上に分けられない結果です。事象は、根元事象の特定の集合を指します。標本空間はすべての根元事象の集合です。根元事象全体{1,2,3,4,5,6}を標本空間と呼びます。つまりこれらはサイコロを振る前から確定しています。そして、このような事象の起こりやすさが確率です。サイコロが作られた時点でこの確率も定まっています。サイコロをなんども振っているうちに、角がわずかに欠け、サイコロの目の出方が変わってしまったとします。その際には確率も変わってしまいます。振っているうちに目の出方が変わってしまうようなサイコロは分析の対象にはなりません。

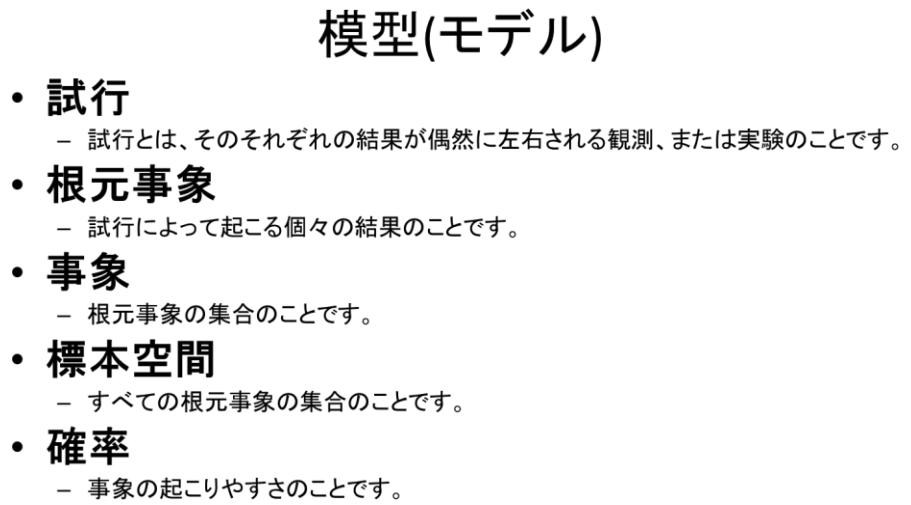


図 2.1 統計モデル

確率には、どれも同じような確からしさで起こるとする古典的な定義、事象の頻度に基づく定義、そして日常的に用いる確率という意味に近い、感覚、主観に基づく定義などがあります。

2.1.1 確率の定義

古典的な確率では、根元事象が生じる確率は等しいと置いて、事象の確率を求めます。この良い例はサイコロの目の出方であるとか、コインの裏表の出方です。根元事象が生じる確率が同様に確からしいとしても、その事象の確率は等しいとは限りません。また、根元事象の生じる確率が等しくない場合もあります。大雨になる確率と小雨の確率は同じであるとは限りません。したがって、発生の頻度に重点を置く考え方もあります。それが頻度確率です。実験や観測により得られた根元事象の相対頻度をもとに確率を求めます。

数学的には、確率は

- 任意の事象Aに対して　$0 \leq P(A) \leq 1$
- 全事象Ωに対して$P(\Omega) = 1$

と定義されます。少し難しく表現しましたが、確率は０から１までの数値であり、何も起こらなければゼロ、全事象の確率を足し合わせると１になることを表現したのです。

2.1.2. 事象と確率

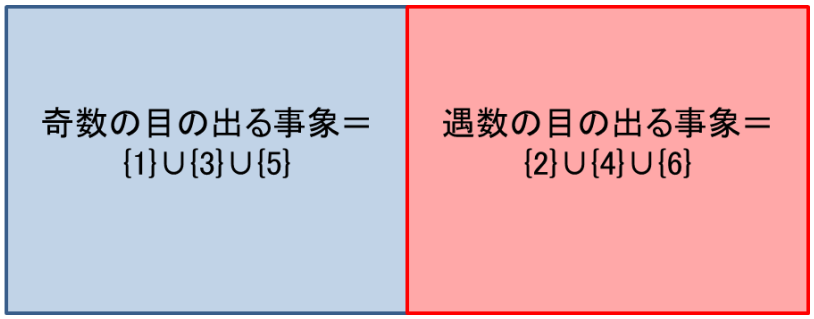
２つの事象AとBの関係について考えてみましょう。典型的な例を２つ紹介します。

A） 和事象

AとBの少なくとも一方が起こる事象を和事象といい、$A \cup B$と書きます。これを「AまたはB」と読みます。

例 2.1：サイコロを一回振って偶数と奇数の目の出る確率を求めましょう。サイコロの目の出方は等確率とします。

偶数の目は$\{2,4,6\}$です。奇数の目は$\{1,3,5\}$です。サイコロの目は全部で６つあるので、$P(1 \cup 3 \cup 5) = 3/6 = 1/2$、$P(2 \cup 4 \cup 6) = 3/6 = 1/2$となります。$P(\cdot)$は確率を表します。奇数の目の出る事象と偶数の目の出る事象は重なり合うものがないため排反事象といいます。排反事象の和事象の確率はそれぞれの事象の和となります。



P(奇数)=P(1)+P(3)+P(5)=1/6+1/6+1/6=1/2　P(遇数)=P(2)+P(4)+P(6)=1/6+1/6+1/6=1/2

図 2.2　奇数の目と偶数の目

B） 積事象

同時に起こる事象を積事象といい、AとBが同時に起こるとき$A \cap B$と書きます。「AかつB」 と読みます。

例題 2.2：Aを奇数の目の出る事象、Bを3の倍数の目の出る事象とするとき$A \cap B$の確率を求めましょう。サ

イコロの目の出方は等確率とします。

$A = (1 \cup 3 \cup 5)$、$B = (3 \cup 6)$となります。$A \cap B$はAとBに含まれる事象ですから{3}となります。$P(A \cap B) = 1/6$です。

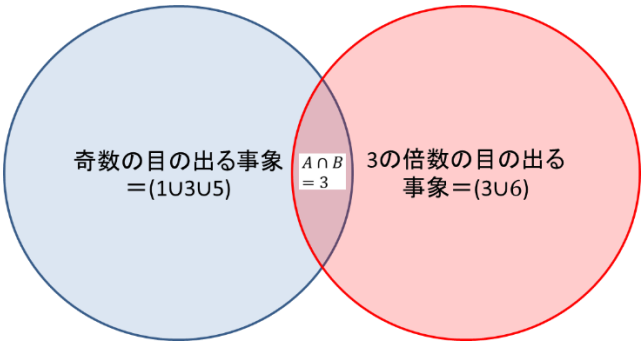


図 2.3　積事象

C）2 つの事象AとBの関係

- 和事象$(A \cup B)$：AとBの少なくとも一方が起こる
- 積事象$(A \cap B)$：AとBが同時に起こる
- 余事象$(^c)$：A^c、Aが起こらない事象；B^c、Bが起こらない事象
- 全事象(Ω)：標本空間全体の事象
- 空事象$(\emptyset)$：何も起こらない事象
- 排反な事象$(A \cap B = \emptyset)$：AとBが同時に起こらない事象

例題 2.3：Aを奇数の目の出る事象、Bを 3 の倍数の目の出る事象とするとき$A \cup B$の確率を求めましょう。サイコロの目の出方は等確率とします。

Aを奇数の目の出る事象：$A = (1 \cup 3 \cup 5)$

Bは 3 の倍数の目の出る事象：$B = (3 \cup 6)$

ですから、重なり合う事象があります。それは{3}です。したがって、排反事象ではありません。排反事象でない事象の和事象の確率をそれぞれの事象の和としてしまうと、重なり合う事象が二重に加算されてしまいます。したがって、その分を差し引く必要があります。一般の和事象は、それぞれの事象の和の確率から、重なり合う積事象の確率を差し引きます。この場合は

$$P(A) = P(1) + P(3) + P(5) = 3/6 = 1/2$$

$$P(B) = P(3) + P(6) = 2/6 = 1/3$$

$$P(A \cap B) = P(3) = 1/6$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 3/6 + 2/6 - 1/6 = 4/6 = 2/3$$

となります。

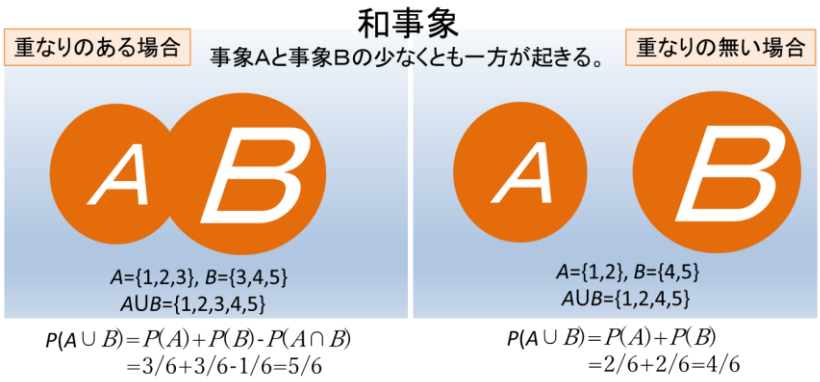


図 2.4　和事象：重なりのある場合とない場合

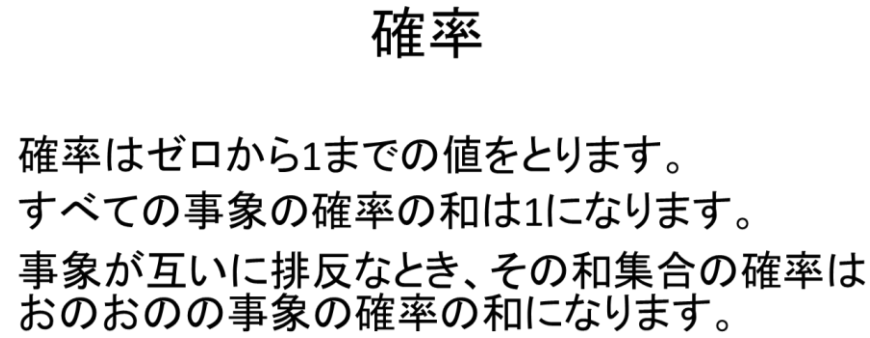


図 2.6　確率とは？

2.1.3. 事象と試行

サイコロの目の出方をいくつかの事象に分類し、その確率を求めてきました。その中で和事象と積事象の確率を扱いました。Aを奇数の目の出る事象、Bを 3 の倍数の目の出る事象とするとき$A \cap B$の確率を求めました。この際の事象Aも事象Bも試行の回数は 1 回です。また、$A \cap B$の意味はサイコロを 1 回振った時に出る目が奇数という性質と 3 の倍数であるという 2 つの性質をあわせもつ結果ということでした。したがって、試行回数はやはり 1 回です。

しかし、事象は必ずしもこのような形であるとは限りません。たとえば、x を赤と青のサイコロの目の和とします。x が 5 になる事象の確率を求めなさいといったときには、実は試行の回数は 2 回です。赤のサイコロを振るという試行と青のサイコロを振るという試行の 2 つから成り立っています。このように事象の確率を求めるときには、事象の意味をよく理解しておく必要があります。

例題 2.4:　2 つのサイコロを同時に振った時に両方とも偶数の目の出る確率を求めてみましょう。サイコロの目の出方は等確率とします。

2 つのサイコロの目の出る組み合わせをすべて書き出してみます。図 2.7 の横軸は 1 つ目のサイコロ、縦軸は 2 つ目のサイコロの目とします。1 つ目のサイコロの目が偶数となるのは出た目が{2,4,6}のときです。{2}が出たとしましょう。その際に 2 つ目のサイコロの出た目が偶数となるのは{2, 4, 6}の目が出たときです。つまり、両方のサイコロの目が偶数であるのは

$$\{2,2\},\{2,4\},\{2,6\},\{4,2\},\{4,4\},\{4,6\},\{6,2\},\{6,4\},\{6,6\}$$

となるときです。クロス表で表現してみましょう。両方のサイコロの目が偶数のものを●としました。

		1番目のサイコロの目					
		1	2	3	4	5	6
2番目のサイコロの目	1						
	2		●		●		●
	3						
	4		●		●		●
	5						
	6		●		●		●
	●2つサイコロとも偶数の目						

図 2.6　2 つのサイコロも目

2つのサイコロの目の組み合わせは全部で36通りあります。この中で両方のサイコロの目が偶数のもの、●の数は9個です。したがって、両方で偶数の目が出る確率は$9/36 = 1/4$です。

例題 2.6：サイコロを振って偶数の目の出る事象をA、奇数の目の出る事象をBとしたとき、青と赤のサイコロを振って起こるすべての事象の確率を求めてみましょう。サイコロの目の出方は等確率とします。

		青のサイコロの目					
		1	2	3	4	5	6
赤のサイコロの目	1	△▲	△●	△▲	△●	△▲	△●
	2	◯▲	◯●	◯▲	◯●	◯▲	◯●
	3	△▲	△●	△▲	△●	△▲	△●
	4	◯▲	◯●	◯▲	◯●	◯▲	◯●
	5	△▲	△●	△▲	△●	△▲	△●
	6	◯▲	◯●	◯▲	◯●	◯▲	◯●
●青の偶数の目；▲青の奇数の目 ◯赤の偶数の目；△赤の奇数の目							
△▲9個；△●9個；◯△9個；◯●9個							

図 2.7　2 つのサイコロの目：奇数と偶数の目

どの事象も確率は$9/36 = 1/4$です。

2.1.4. 事象と独立性

サイコロの目の出方から、いくつかの事象の確率を求めてきました。その中で積事象の確率がそれぞれの事象の確率の積であるものがありました。たとえば、例題 2.2 では、Aを奇数の目の出る事象、Bを 3 の倍数の目の出る事象とするとき$A \cap B$の確率をもとめました。$P(A \cap B)$は$1/6$です。これは$P(A) = 1/2$、$P(B) = 1/3$の積としても求められます。

$$P(A \cap B) = P(A)P(B)$$

が成り立つとき、２つの事象AとBは独立であるといいます。事象Aと事象Bはお互いに影響することなく生起することによります。奇数の目の出る事象は、３の倍数の目の出る事象とは無縁です。同じことが例題 2.4、2.5、2.6 でもいえます。では、どのようなときに

$$P(A \cap B) \neq P(A)P(B)$$

となるのでしょうか。

例題 2.7:青と赤の色の２つのサイコロがあります。青いサイコロの目が奇数であるときそれを事象Aとします。また、赤いサイコロと青いサイコロの目の積が奇数であるときそれを事象 B とします。$A \cap B$の確率をもとめてみましょう。サイコロの目の出方は等確率とします。

事象Aの起こる確率は 1/2 です。事象Bは２つの試行から構成されています。赤いサイコロの目を横軸、青いサイコロの目を縦軸とします。

		赤のサイコロの目					
	積	1	2	3	4	5	6
青のサイコロの目	1	1	2	3	4	5	6
	2	2	4	6	8	10	12
	3	3	6	9	12	15	18
	4	4	8	12	16	20	24
	5	5	10	15	20	25	30
	6	6	12	18	24	30	36
	赤文字:赤と青のサイコロの目の積が奇数						

図 2.8　２つのサイコロの目：独立ではない例

青のサイコロと赤のサイコロの目の積が奇数であるためには、両方の目が奇数である必要があります。$P(A \cap B) = 9/36 = 1/4$となります。Bとなる条件にAが含まれています。このような場合、$P(A \cap B) = P(A)P(B)$とはなりません。$P(A)P(B) = 1/2 \cdot 9/36 = 1/8$となってしまいます。