1.1 統計データ

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、人口推計などにまず分類されています。データを探す→分野別→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。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ります。データは過去の記録から得たり、新たに実験・調査・測定を行ったりして集めます。これらのデータは数値とは限りません。性別、居住地、職業、天気などの場合は文字列です。

また、1 種類のデータ、または変数に注目するとき、そのデータは 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。たとえば、ワインの銘柄リストがあり、その化学成分が分かっているときに、どのような成分構成のワインの評価が高いかなどの分析は多変量を対象にした分析になります。ある銘柄の株価の特性を過去の価格データを利用して明確にするとき、それは１変量分析です。

1.2 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します。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統計処理の基本になります。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。名義尺度と順序尺度は質的変数です。これらの性質は文字列で表現できます。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です。また、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。

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）で検査されました。CVRVV は、ヴィーニョ・ヴェルデの品質とマーケティングを向上させることを目的

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

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

統計データ

データの性質の表現方法による分類

- 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です。

1.3 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

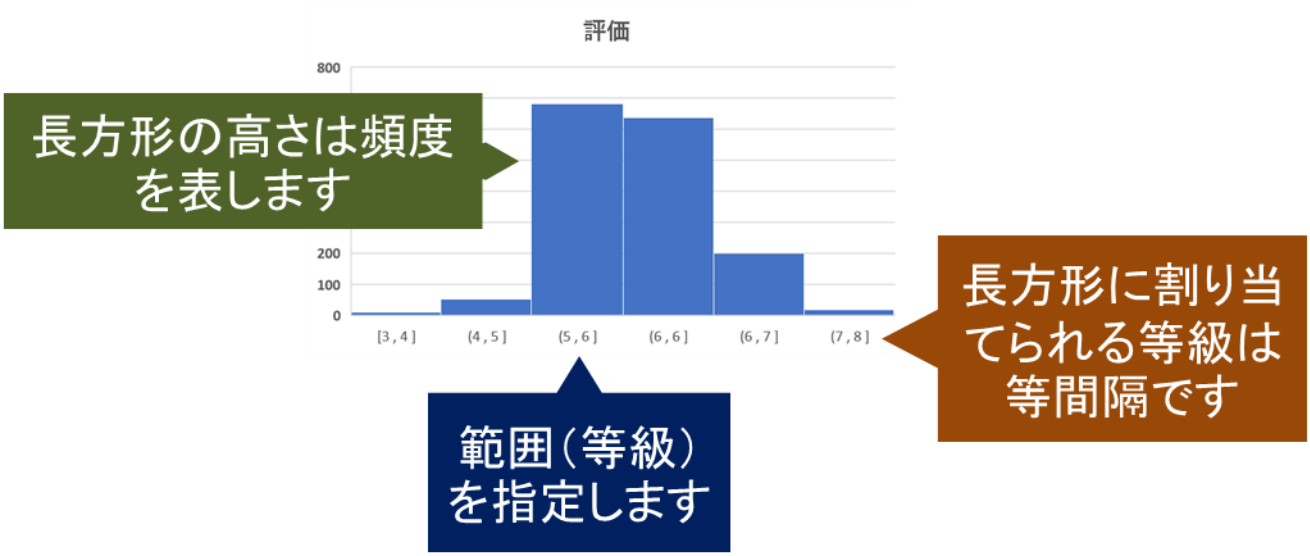
データをグラフとして視覚的に要約することで、全体の特徴をとらえることができます。

- ヒストグラム(頻度図)の作成

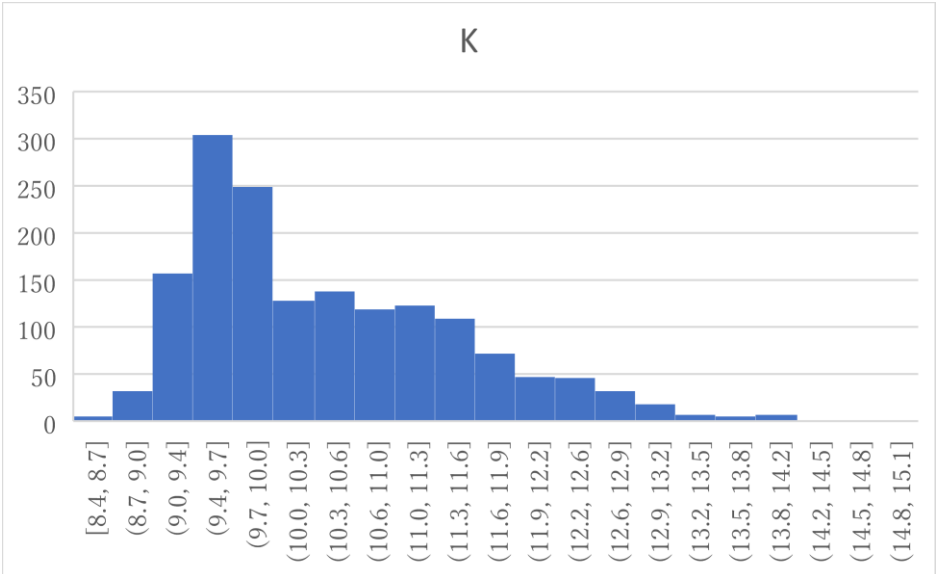
頻度図は、横軸に変数をその大きさ、または階級などに応じて並べ、縦軸にそれらの頻度を表したグラフです。

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。

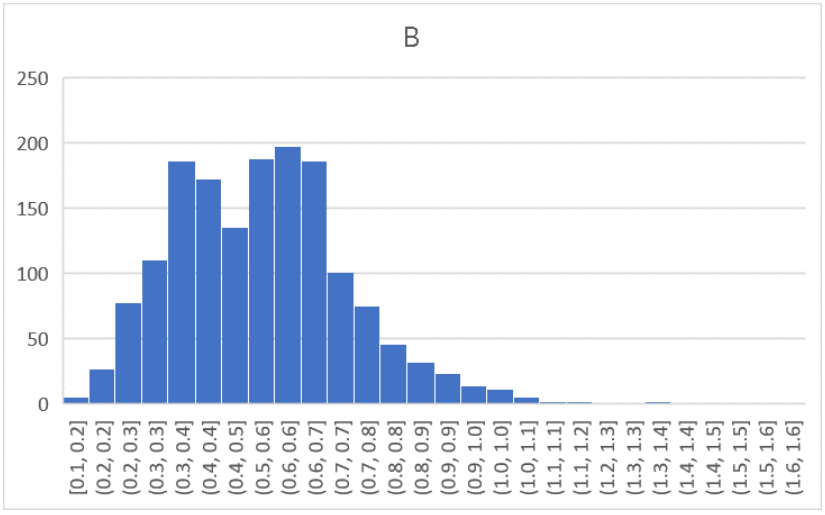
このデータベースは、赤ワインを１０段階の評価結果とワインの特徴の科学的分析結果を集めたものです。つぎは赤ワインの１０段階評価を頻度図にしたものです。



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。頻度の分布はおおよそ左右対称ですので、このような分布をベル型の分布と呼びます。



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。分布の度数は左によっています。これは右にひずんだ分布です。



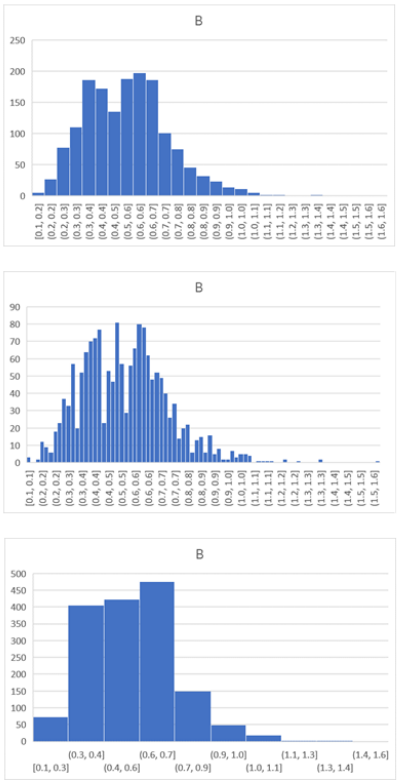
横軸は化学成分 B です。最も頻度の高い B は 0.6 近辺です。分布の形状は天井が平らで、左右のすそ野はなだらかに減少している台形にも見えますし、２つの単峰の分布が混じっているようにも見えます。

同じデータでも等級のとりかたでイメージが変わる

データを大きい順にならべて、その頻度を数えて可視化する

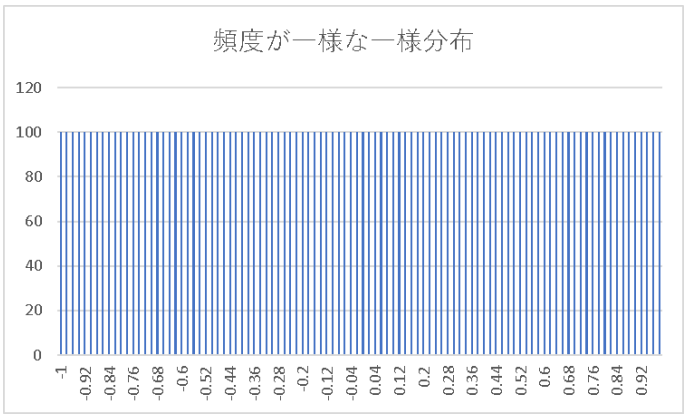
データをならべるときに幅があった方が良ければ幅を決める

幅の大きさにより頻度が変わる



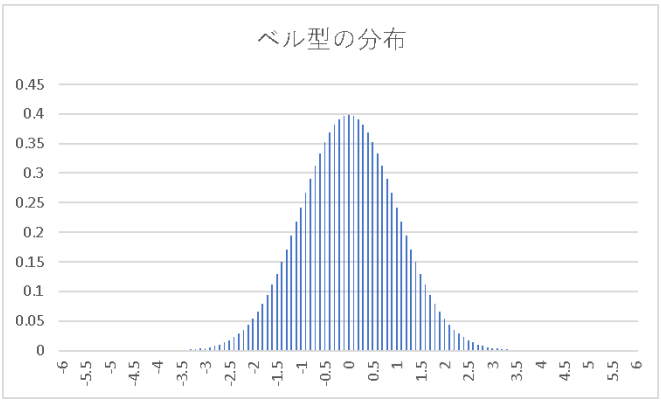
頻度図の形状は大まかに

- 一様な分布：頻度が横軸の値に対してほぼ均等。



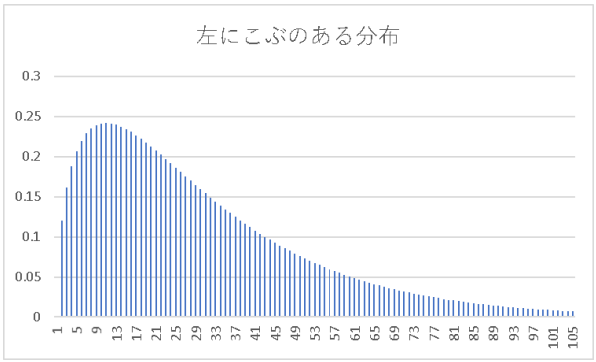
(例題 2.6)

- ベル型の分布：頻度の高さは横軸に対してベル型。



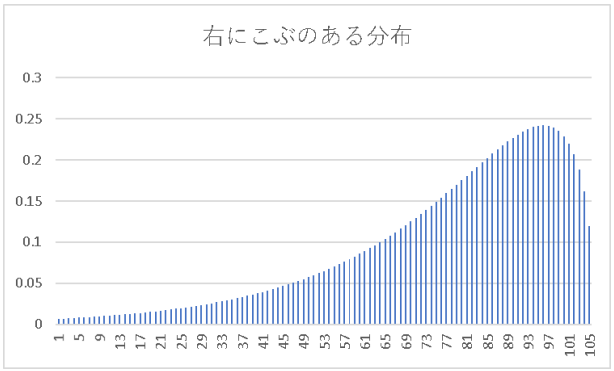
(例題 2.7)

- 右にひずんだ分布：右にすそ野が長く、頻度が左に寄った分布。



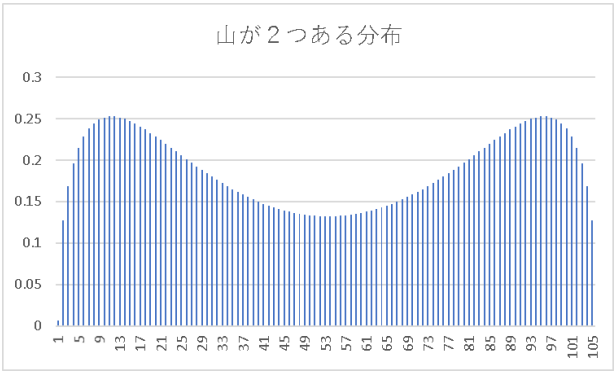
(例題３.?)

- 左にひずんだ分布：左にすそ野が長く、頻度が右に寄った分布。



(例題３.?)

- 複数の山をもつ分布：いくつもの分布が混じった分布。



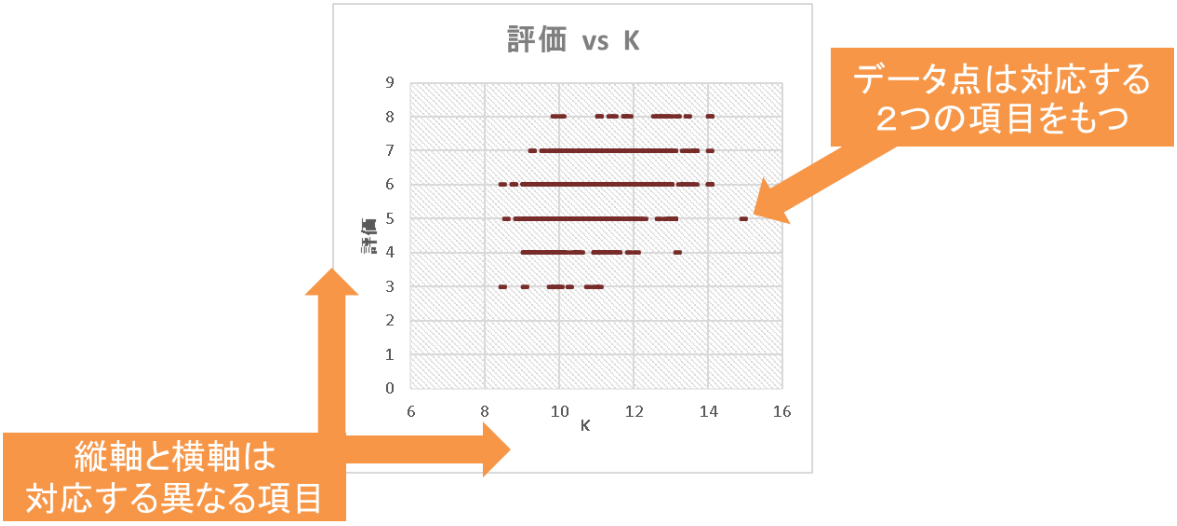
(例題３.?)

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。

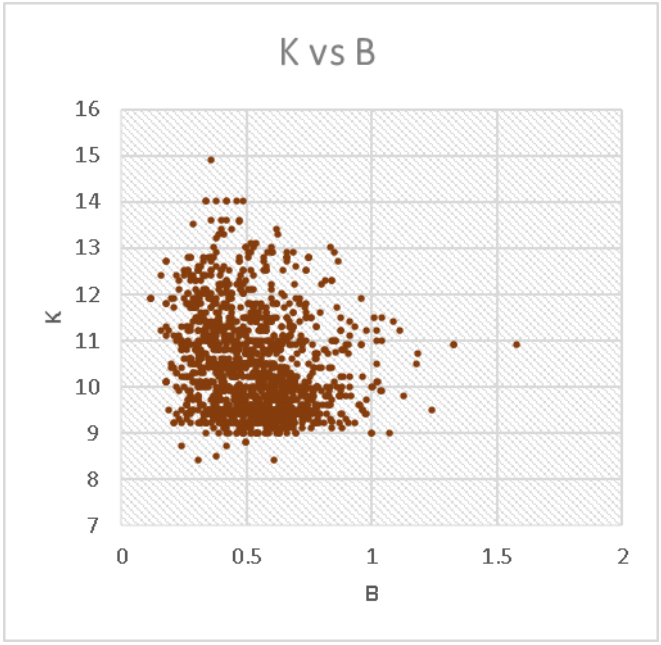
- 散布図をつくる目的

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。

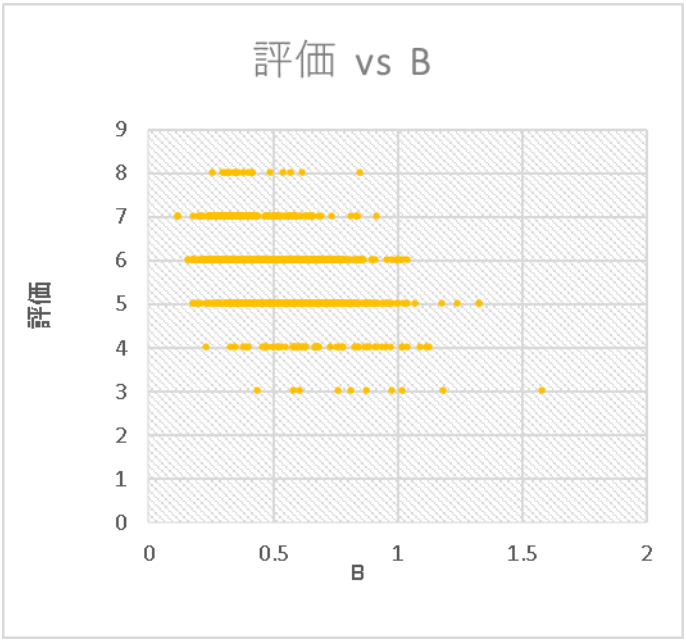
例題 1.2：ワインデータの評価と化学成分 K、化学成分 K と化学成分 B、評価と化学成分 B の散布図を作ってみましょう。



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あることが分かります。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影響です。



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このように見える可能性があります。



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。

可視化の際にはグラフの数は多ければ多いほど良い

いろいろなグラフからイメージを得る

1.3.2 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 一変量要約統計量

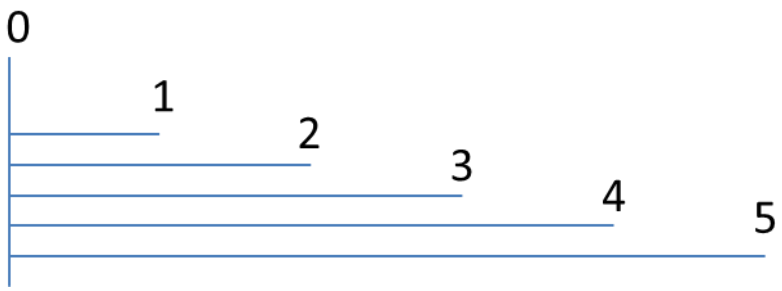
まずはなじみの深い平均を見ていきます。

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。



これは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

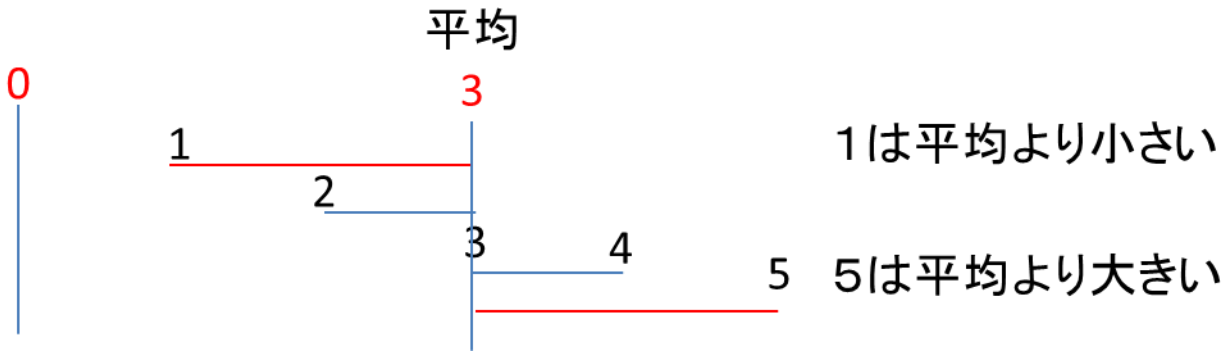
2-3=-1

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。図で表すと



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。

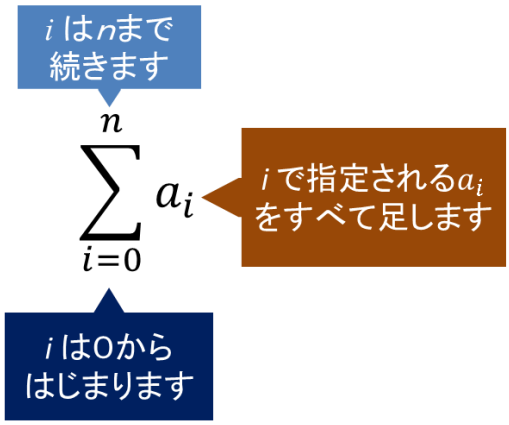
n 個の数値の平均は、a_i を i 番目の数とすると、その計算方法は

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

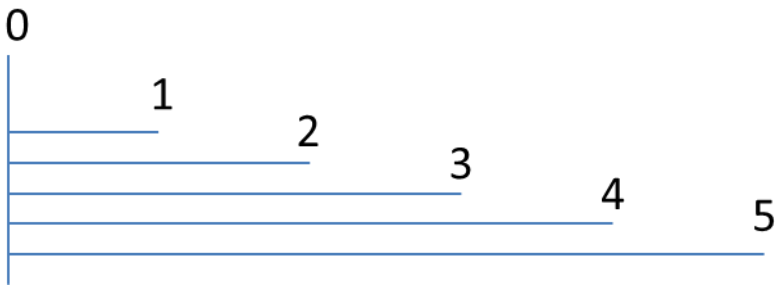
となります。これはさらに

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

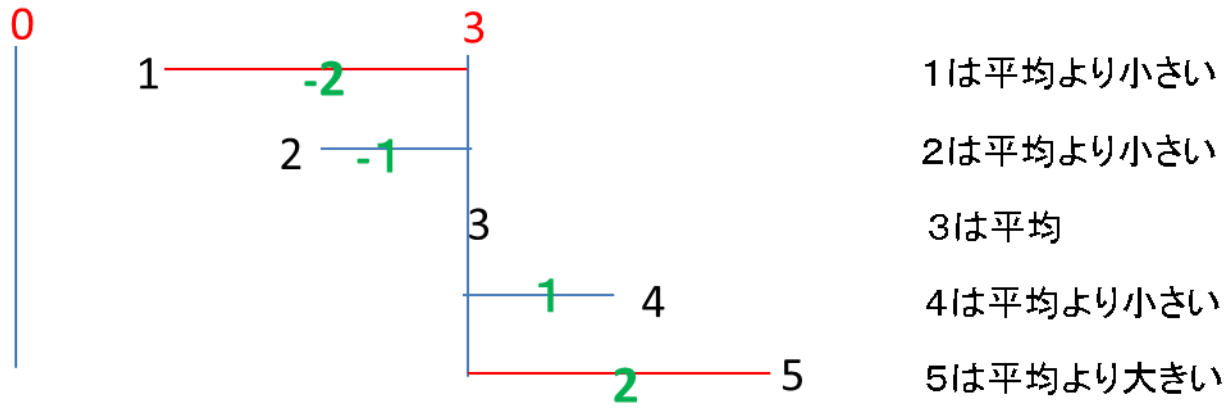
と書くことができます。$\bar{a}$はaの平均を意味します。



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。



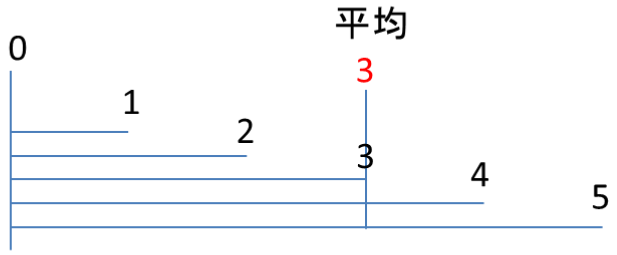
１から５までの数値のばらつきを考えるときに、それぞれの数値とその平均との差、つまり偏差をとります。これもばらつきの尺度になります。グリーンの数値は偏差を表しています。



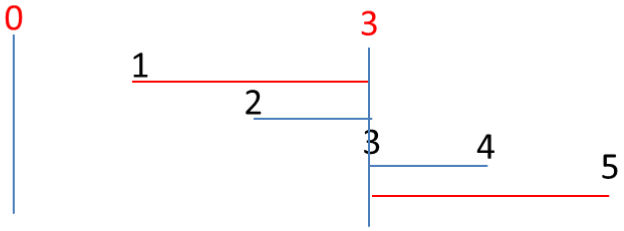
この偏差を足し合わせるとゼロになってしまうので、ばらつきの尺度としては適当ではありません。

平均はたくさんのデータを1つの量に要約します

個別のデータの特徴は失われました



偏差は同じデータを別の角度からみせてくれます



データは左右対称になりました

偏差の平均はゼロなのでバラツキの尺度にはなりません

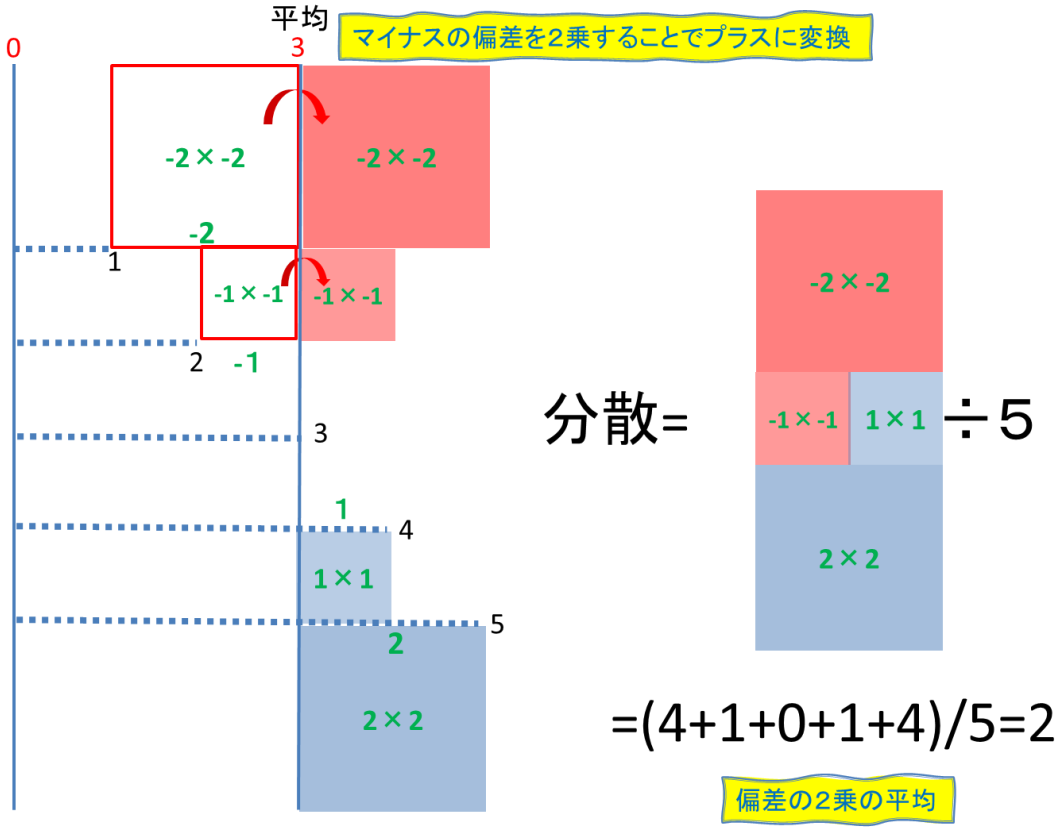
分散

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。

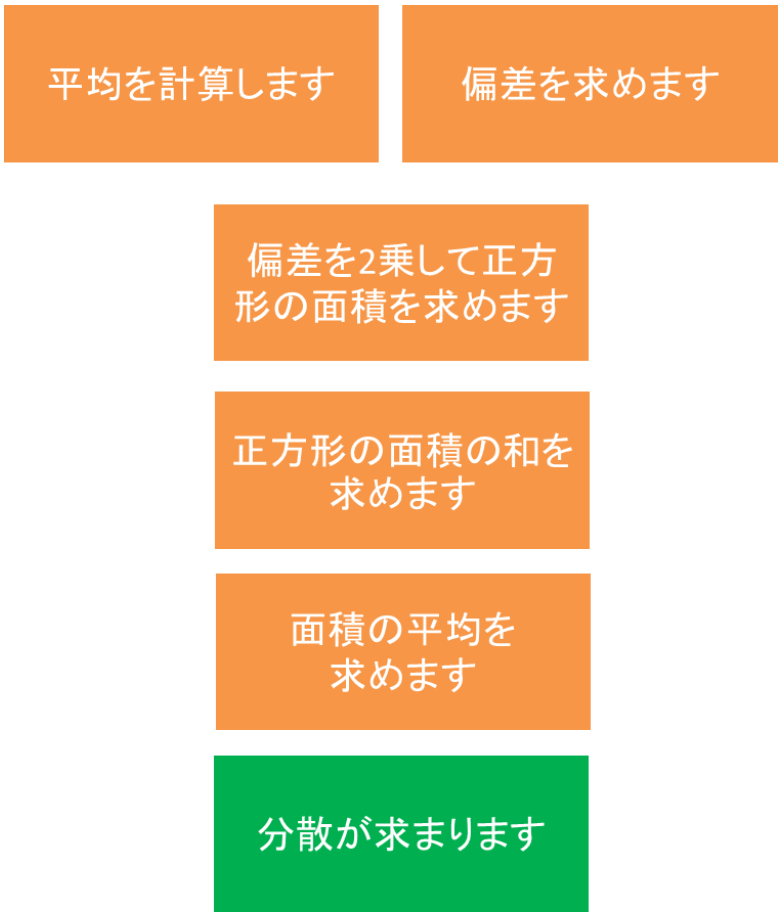
$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

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。



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。

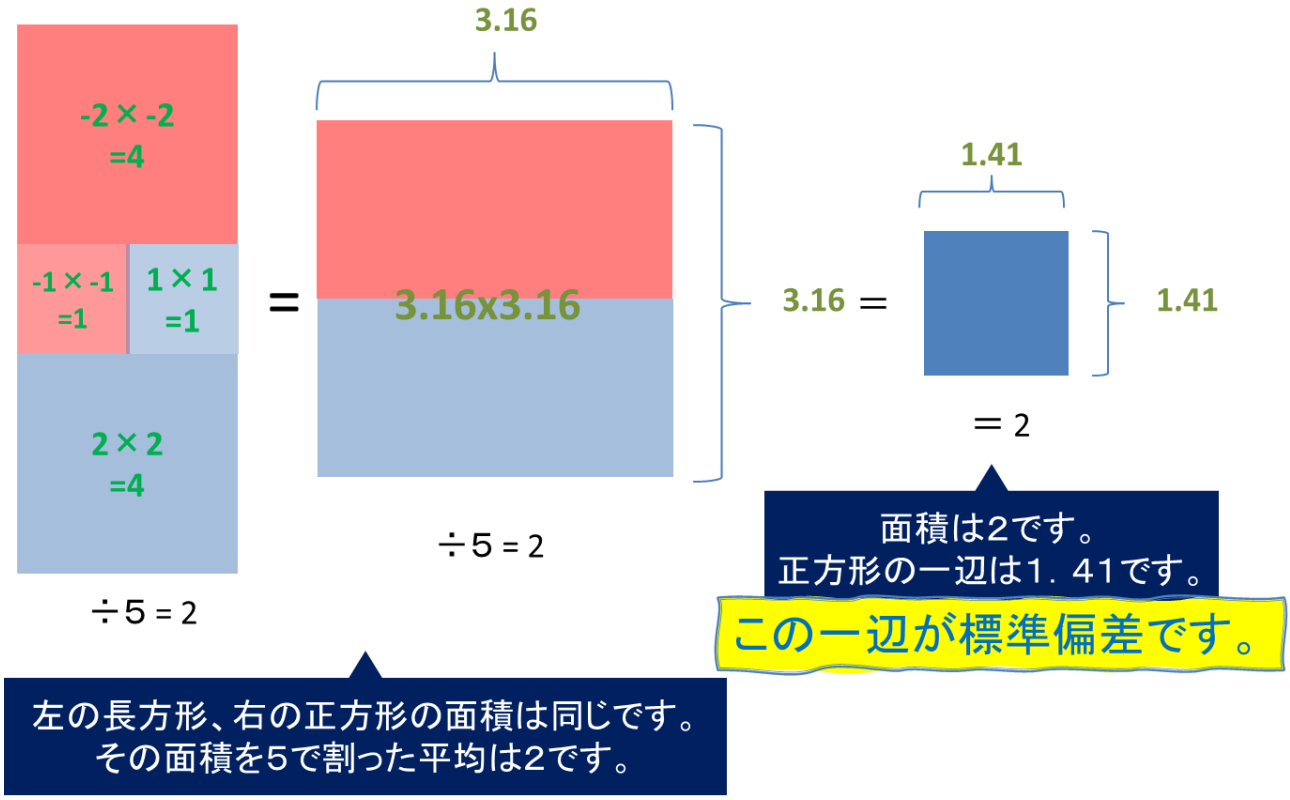


標準偏差

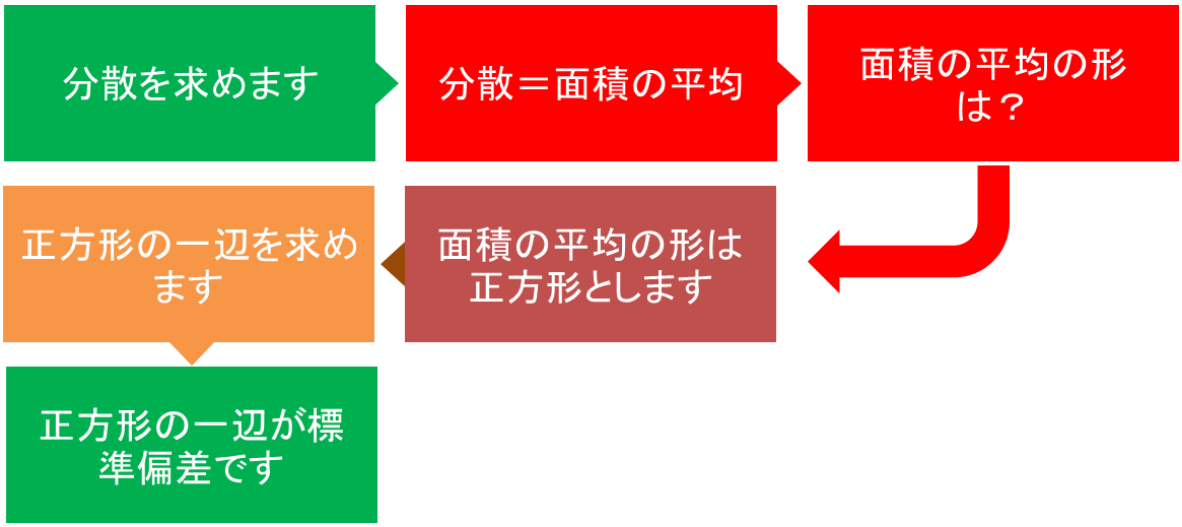
分散の正の平方根を標準偏差と呼びます。分散同様に、数の集団の散らばり具合を表す指標です。

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

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。つぎの図は標準偏差の意味のイメージ図です。



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。



1.3.3　　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。

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。

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

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。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。

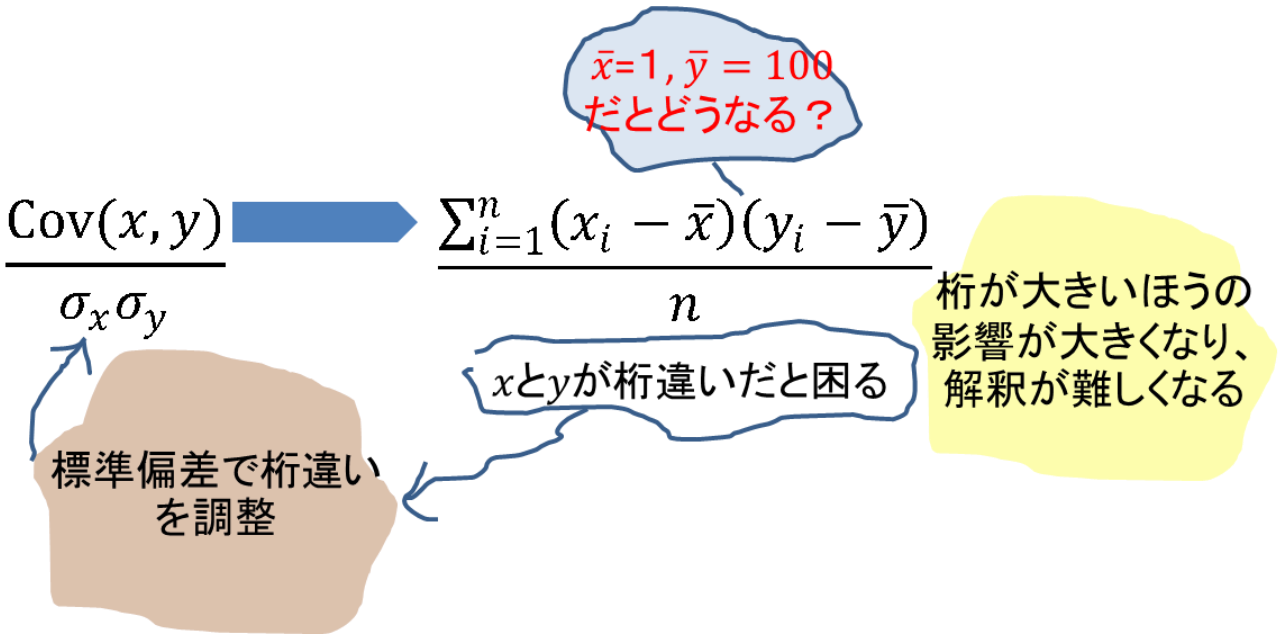
	a	b	c	d
a	Cov(a,a)	Cov(b,a)	Cov(c,a)	Cov(d,a)
b	Cov(a,b)	Cov(b,b)	Cov(c,b)	Cov(d,b)
c	Cov(a,c)	Cov(b,c)	Cov(c,c)	Cov(d,c)
d	Cov(a,d)	Cov(b,d)	Cov(c,d)	Cov(d,d)

相関

共分散は 2 組のデータ(x,y)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。

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

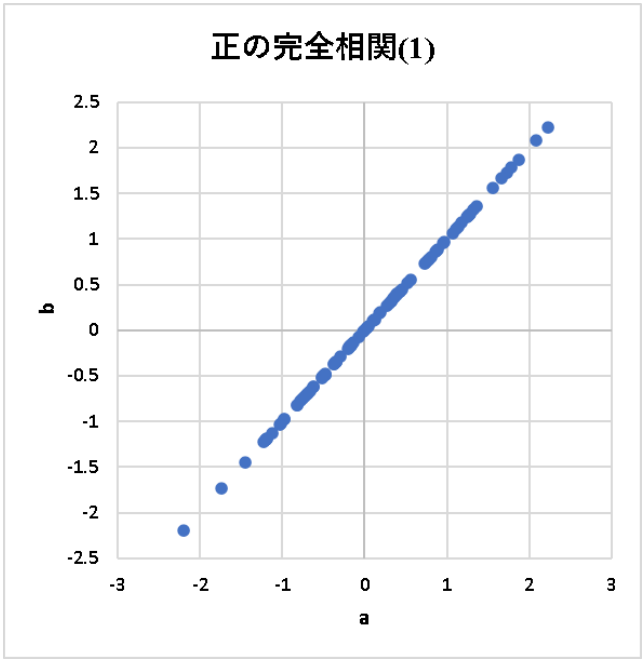
これが相関です。このようにすることで、相関が 1 に近ければ 2 組のデータは同じような動きになり、ゼロに近ければ、関係がなく、−1 に近ければ逆の動きをしていることになります。相関が 1 のときを正の完全相関、−1 のときを負の完全相関といいます。

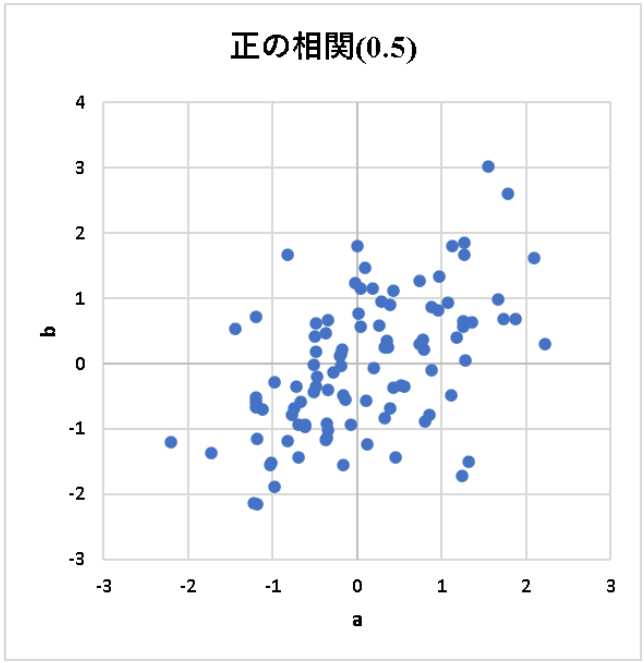


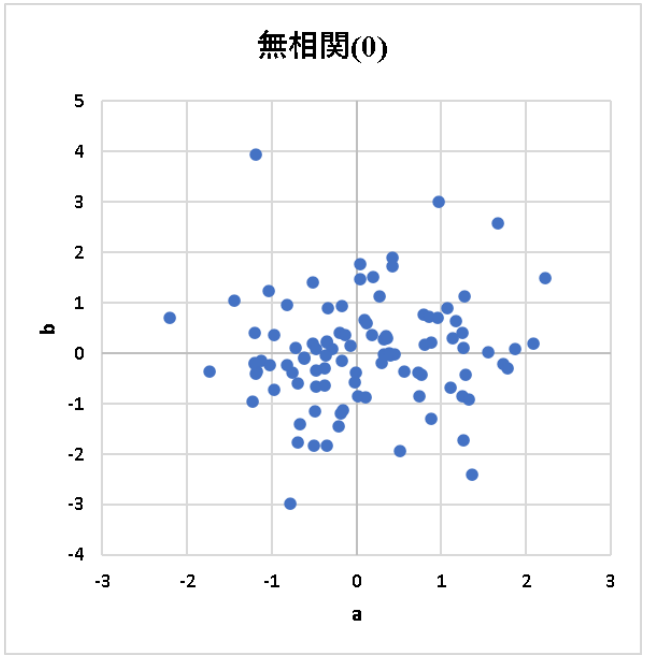
共分散同様に、要素が 2 つ以上ある場合には相関もマトリックスを用いて表現します。

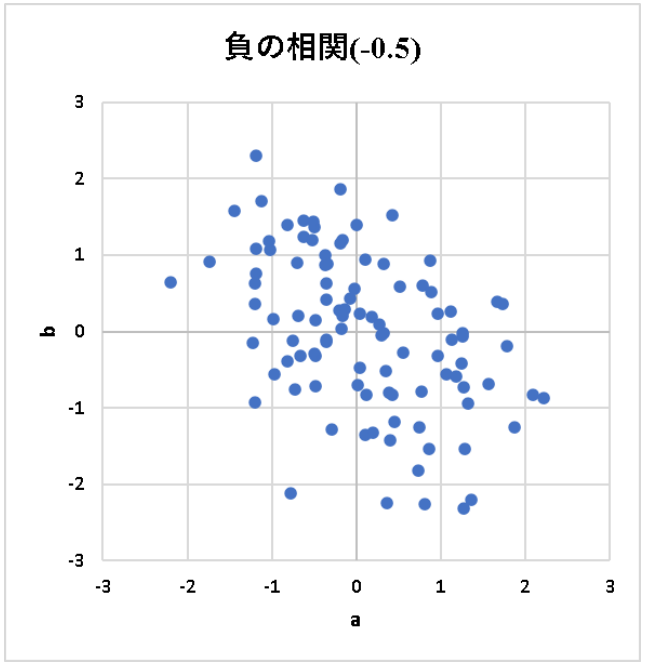
相関は便利で使いやすいのですが、使い方に注意が必要です。相関は単なる平均的な関係を示すだけで、たとえば A と B の相関が高いからといって、それが，A が B の原因であるとか、B が A の原因であるとか、事象の因果関係を示すことにはなりません。この点には注意が必要です。

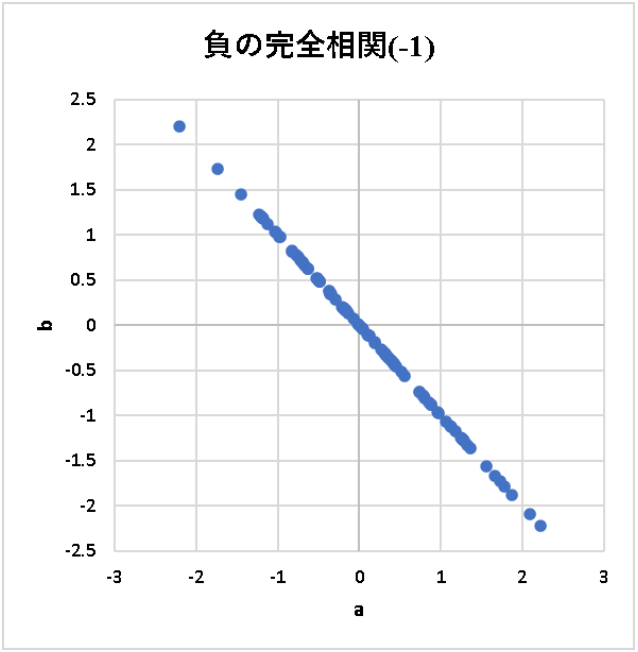
相関を、散布図を用いて可視化してみましょう。つぎの図は乱数を用いて確率変数 a, b を生成し、正の完全相関、正の相関、無相関、負の相関、負の完全相関を散布図として表現したものです。(練習問題 1.9)






無相関(0)
5
4
3
2
1
0
-1
-2
-3
-4
b
-3
-2
-1
0
1
2
3
a


負の相関(-0.5)
3
2
1
0
-1
-2
-3
b
-3
-2
-1
0
1
2
3
a


負の完全相関(-1)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
b
-3
-2
-1
0
1
2
3
a

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。

例題 1.5：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。

	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

1.3.4 分布の形状に関する要約統計量

観測値の分布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。

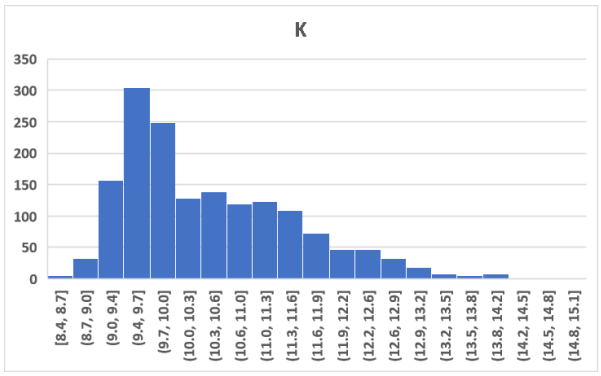
歪度

分布の歪の度合いを表す歪度(skew)は

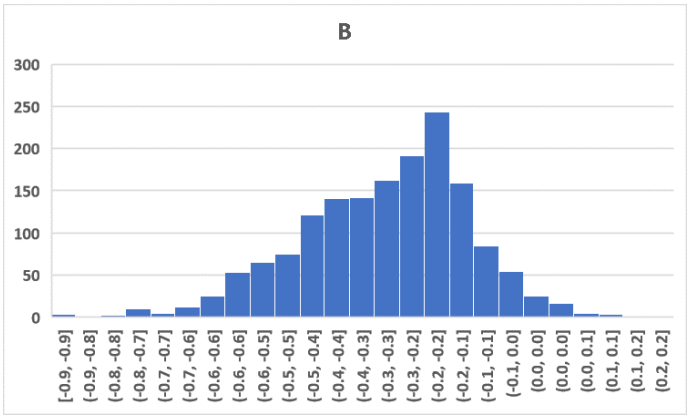
$$歪度 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n} \Big/ \sqrt{\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right]^3}$$

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。



－ (練習問題 1.1　歪度 0.86)



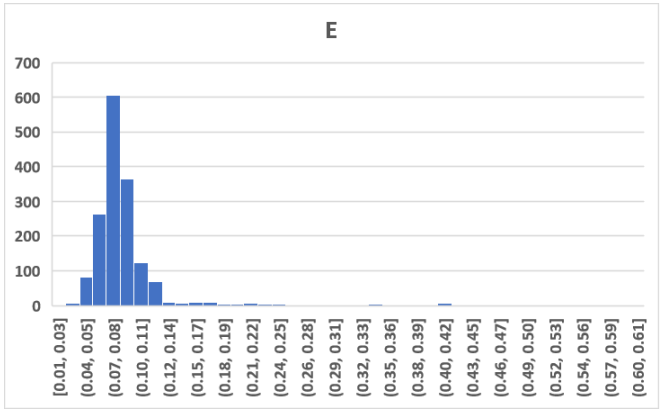
(練習問題 4.3 データの対数　歪度-0.43)

尖度

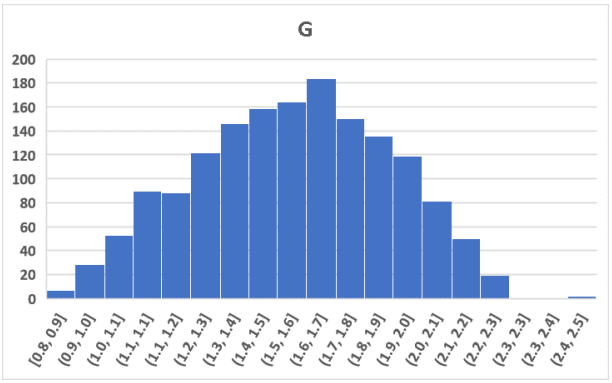
尖度(kurt)は分布の中心の尖り具合、すそ野の厚さを表します。

$$\text{尖度} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n} \Big/ \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right]^2$$

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。



(練習問題 1.1 尖度４１)



(練習問題 4.3 データの対数　尖度-0.67)

例題 1.6：ワインデータの歪度と尖度を計算してみましょう。

	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 確率と確率分布

２章では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ます。これは日常生活の中で育まれる確率です。分布も同様で左右対称のベル型の分布を思い浮かべます。偏差値の分布はその例です。本章では統計学でいう確率と分布についてサイコロとワインデータを用いて学んでいきます。

2.1 サイコロを投げる

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統計分析の対象にはなりません。

模型(モデル)

- **試行**
  - 試行とは、そのそれぞれの結果が偶然に左右される観測、または実験のことです。
- **根元事象**
  - 試行によって起こる個々の結果のことです。
- **事象**
  - 根元事象の集合のことです。
- **標本空間**
  - すべての根元事象の集合のことです。
- **確率**
  - 事象の起こりやすさのことです。

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いと置け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 $P(\Omega)=1$

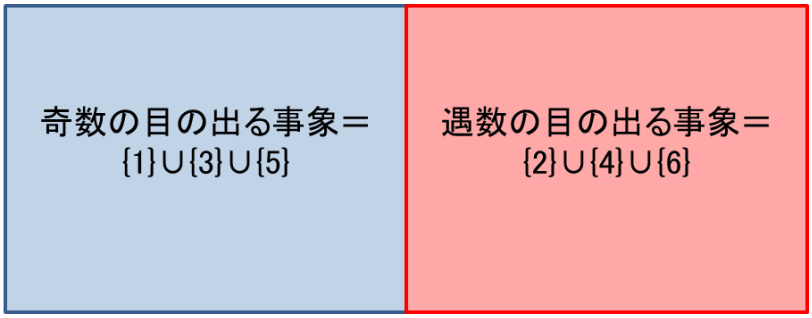
と定義されます。

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。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。

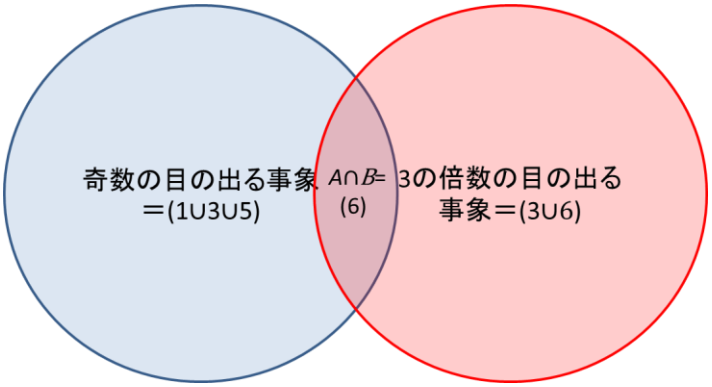


P(奇数)=P(1)+P(3)+P(5)=1/6+1/6+1/6=1/2　P(遇数)=P(2)+P(4)+P(6)=1/6+1/6+1/6=1/2

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。



２つの事象 A と B の関係

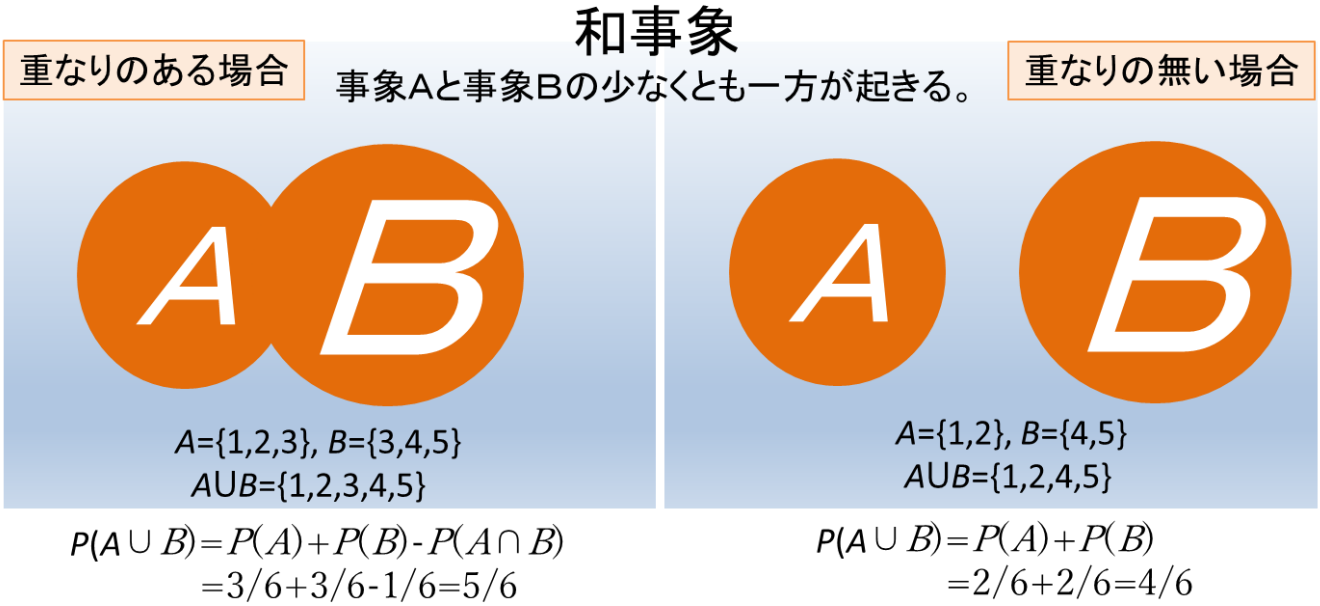
- 和事象($A \cup B$)：A と B の少なくとも一方が起こる
- 積事象($A \cap B$)：A と B が同時に起こる
- 余事象(^{c})：A^c、A が起こらない事象；B^c、B が起こらない事象
- 全事象(Ω)：標本空間全体の事象
- 空事象($\varnothing$)：何も起こらない事象
- 排反な事象($A \cap B=\varnothing$)：A と B が同時に起こらない事象

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、B は 3 の倍数の目の出る事象ですから、重なり合う事象があります。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

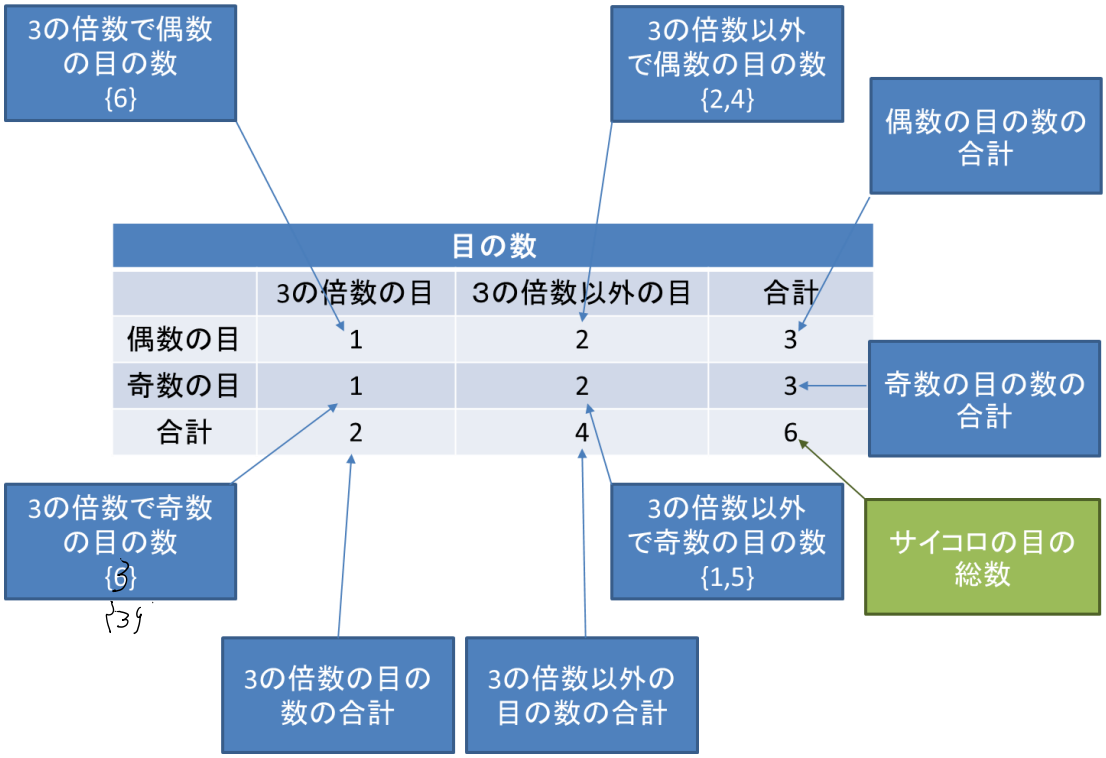
$A=(1 \cup 3 \cup 5)$、$B=(3 \cup 6)$ 、$P(A)=P(1)+P(3)+P(5)=3/6=1/2$、$P(B)=P(3)+P(6)=2/6=1/3$、$P(A \cap B)=1/6$、

$P(A \cup B)=P(A)+P(B)-P(A \cap B)=3/6+2/6-1/6=4/6=2/3$ となります。

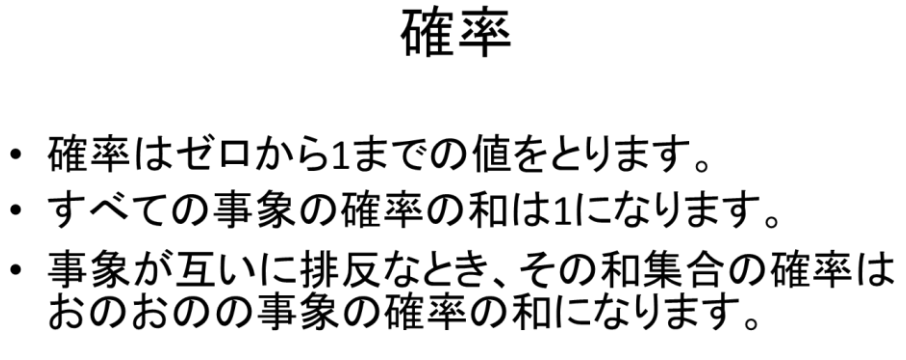


例題 2.4： A を奇数の目の出る事象、B を偶数の目の出る事象、C を 3 の倍数の目の出る事象、D を C の余事象とします。B の目が出たときにそれが D に含まれる目でもある確率はいくらでしょうか？これを $P(D|B)$ と書き、条件付確率といいます。サイコロの目の出方は等確率とします。

$A=(1\cup3\cup5)$、$B=(2\cup4\cup6)$、$C=(3\cup6)$、$D=(1\cup2\cup4\cup5)$となります。したがって、$P(A)=1/2$、$P(B)=1/2$、$P(C)=1/3$、$P(D)=2/3$ です。$P(B\cap D)$はサイコロの目が B と Dに含まれる確率ですから、Bに含まれる確率 $P(B)$ と B の目がでたときにその目が D にも含まれる確率 $P(D|B)$の積と等しくなります。B と D に共通の要素は 2 と 4 なので、$P(B\cap D)$は 2/6 です。$P(B)=1/2$ ですから $P(D|B)=2/6/(1/2)=2/3$ になります。クロス表を用いると見通しが良くなります。



目の数			
	3の倍数の目	3の倍数以外の目	合計
偶数の目	1	2	3
奇数の目	1	2	3
合計	2	4	6



2.1.3 試行と事象

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 回振ったとき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 2.5:　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。次の図の横軸は 1 番目のサイコロ、縦軸は 2

番目のサイコロの目とします。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						

２つのサイコロの目の組み合わせは全部で３６個あります。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							

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 2.5、例題 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。

事象の 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						

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。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。$P(A \cap B)=9/36=1/4$、一方で $P(A)P(B)=1/2 \cdot 9/36=1/8$ となります。

2.2 確率変数

2.2.1 確率変数

変数 X がどのような値を取るかは事前にはわからないのですが、その値の確率が与えられているとき、その変数 X は確率変数といいます。サイコロを振って出た目を観察する試行においてその結果を変数 X とします。X の根元事象は{1}, {2}, {3}, {4}, {5}, {6}で、その全体{1, 2, 3, 4, 5, 6}は標本空間です。それぞれの根元事象には確率が割り当てられます。したがって、この変数 X は確率変数です。この際にサイコロの出る目はとびとびの値でした。このような確率変数を離散型確率変数といいます。サイコロの生成する乱数は１から６までの整数です。

模型(モデル)

- **試行**
  - サイコロを振る
- **根元事象**
  - {1},{2},{3},{4},{5},{6}
- **事象**
  - A={1,3,5}など
- **標本空間(全事象)**
  - Ω={1,2,3,4,5,6}
- **確率**
  - {P(A)=3/6=1/2}

通常のサイコロは６つの４角形からなる６面体です。それぞれの面は正方形で、サイコロは立方体でもありま

す。６面体ダイズと呼ばれたりします。この面の数を増やしていくと、それを多面体ダイズといいます。たとえば、１２面ダイズ  は１から１２までの乱数を等確率で生成します。それぞれの面の確率は 1/12 となります。さらに面の数を増やしていき 120 面体  とするとそれぞれの面の出る確率は 1/120 となります。面の数を無限大に増やすとそれぞれの出る面の確率はゼロになってしまいます。

確率変数は
- 離散型確率変数
  - とびとびの値をとる確率変数
- 連続型確率変数
  - 連続的な値(実数値)をとる確率変数

に分類されます。

2.2.2 独立な確率変数

一方の事象の起こる確率が、もう一方の事象の起こる確率に影響されないとき、それぞれの事象は独立であるといいます。これは事象 A, B について $P(A\cap B)=P(A)P(B)$ が成り立つということです。$\cap$ は A と B が同時起こることを表しています。たとえば、確率変数 X と Y が独立であると、その分散では $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)$ が成り立ちます。X と Y が独立でなければ $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\mathrm{cov}(X, Y)$ となります。

独立と無相関は混同されやすいのですが、独立は関係のことであり、無相関は平均的な性質のことです。したがって、独立ならば無相関になりますが、無相関であれば独立というわけではありません。

例題 2.8：サイコロをふる試行が独立だとします。サイコロを２回投げたときに事象 A、B を $A\in\{1,2,3\}$、$B\in\{3,4,5\}$ とすると２回とも $A\cap B$ となる確率はいくらでしょうか？

１回目の試行の結果は 1, 2, 3, 4, 5, 6 のどれかです。したがってそれぞれの試行が独立であれば、その確率はそれぞれ 1/6 です。1 が出れば事象 A です。3 が出れば $A\cap B$ です。6 がでれば$\varnothing$となります。1, 2, 3, 4, 5, 6 のどれかが出た場合、それぞれの試行は左から A、A、$A\cap B$、B、B、$\varnothing$となります。したがって $A\cap B$ の確率は 1/6 です。つぎに 1 回目の試行が３として、2 回目の試行で出る目を考えてみます。これは 1, 2, 3, 4, 5, 6 のどれかです。したがって、２回目に $A\cap B$ が出る確率も 1/6 です。したがって、$A\cap B$ が２回続けて出る確率は 1/6・1/6=1/36 となります。これをさらに確かめてみましょう。すべての組み合わせを書いてみます。(1 回目の結果, 2 回目の結果)とします。

$(A, A), (A, A), (A, A\cap B), (A, B), (A, B), (A,\varnothing)$

$(A, A), (A, A), (A, A\cap B), (A, B), (A, B), (A,\varnothing)$

$(A\cap B, A), (A\cap B, A), (A\cap B, A\cap B), (A\cap B, B), (A\cap B, B), (A\cap B,\varnothing)$

$(B, A), (B, A), (B, A\cap B), (B, B), (B, B), (B,\varnothing)$

$(B, A), (B, A), (B, A\cap B), (B, B), (B, B), (B,\varnothing)$

$(\varnothing, A), (\varnothing, A), (\varnothing, A\cap B), (\varnothing, B), (\varnothing, B), (\varnothing,\varnothing)$

すべてで 36 組あります。この中で$(A \cap B,\ \ A \cap B)$となっているのは 1 つなのでその確率は 1/36 です。

2.3 確率分布

確率変数がとりえる値とそれに対応する確率を確率分布といいます。

2.3.1 主な離散型確率分布

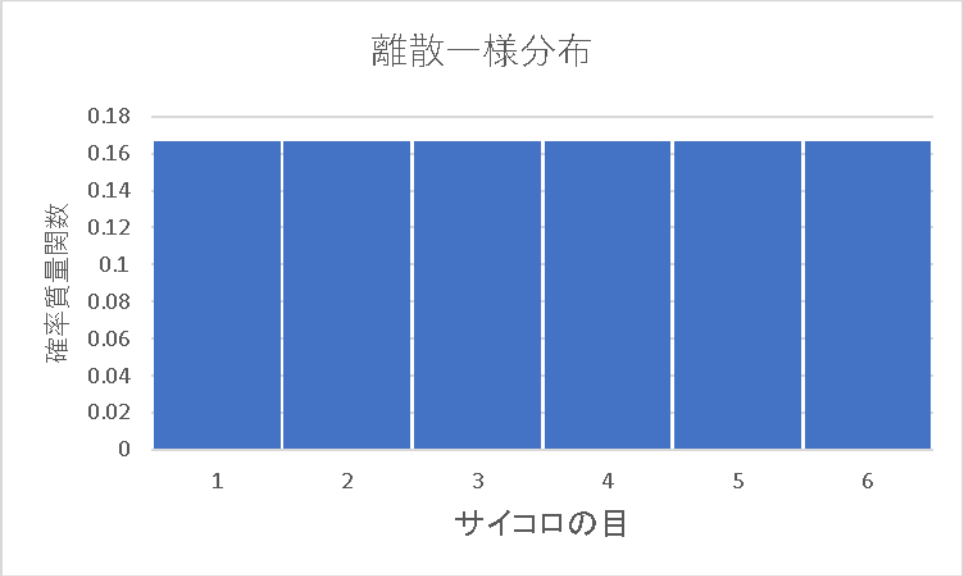
離散的確率変数の作る分布を離散型確率分布といいます。

離散一様分布

確率変数が離散値$x = 1,2,3, \ldots, N$で、それぞれが一様に同じ確率をもつとき、それらは離散一様分布にしたがうといいます。その確率は

$$f(x) = \frac{1}{N}, x = 1,2,3, \ldots, N$$

となります。サイコロの目では$x = ,2,3,4,5,6$ですから確率は 1/6=0. 167 となります。



すべての事象の確率を足すと 1/6+1/6+1/6+1/6+1/6+1/6= 1 になります。

ベルヌーイ分布

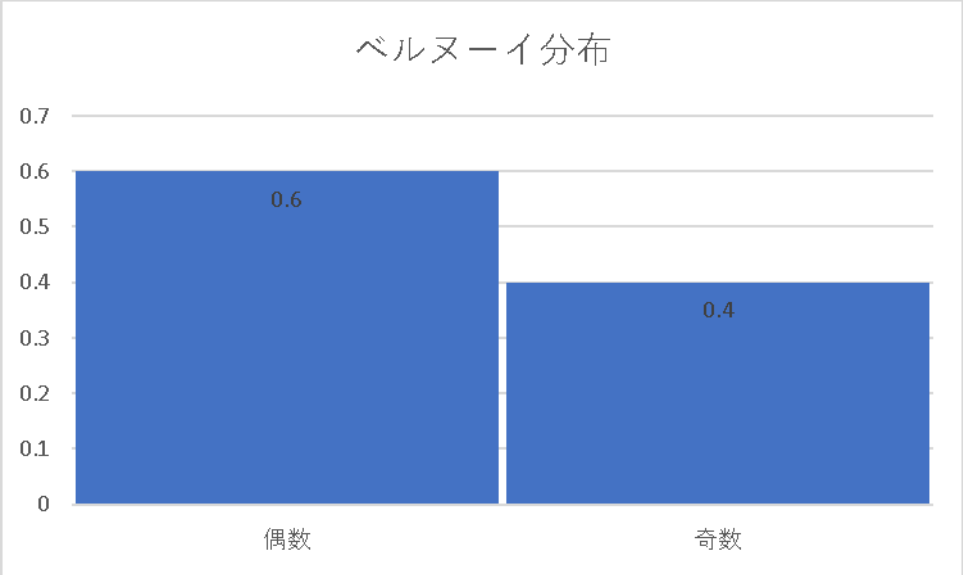
サイコロを投げたとき、その出る目を偶数と奇数に分けることができます。このような 2 値で表される事象が起こる行為をベルヌーイ試行といいます。この場合に、確率 p で奇数が出て、確率 $1-p$ で偶数が出ます。その分布はベルヌーイ分布となります。結果が起こる確率は、一定かつ独立である必要があります。

[表, 裏]、[1, 0]、[上がる、下がる]など試行の結果が 2 値になるものはベルヌーイ試行です。

ベルヌーイ分布の確率分布は

$$f(X=1)=p, f(X=0)=1-p$$

で与えられます。平均は p、分散は $p(1-p)$ となります。サイコロの目の偶数、奇数がそれぞれ 0.6 と 0.4 とするとベルヌーイ分布は



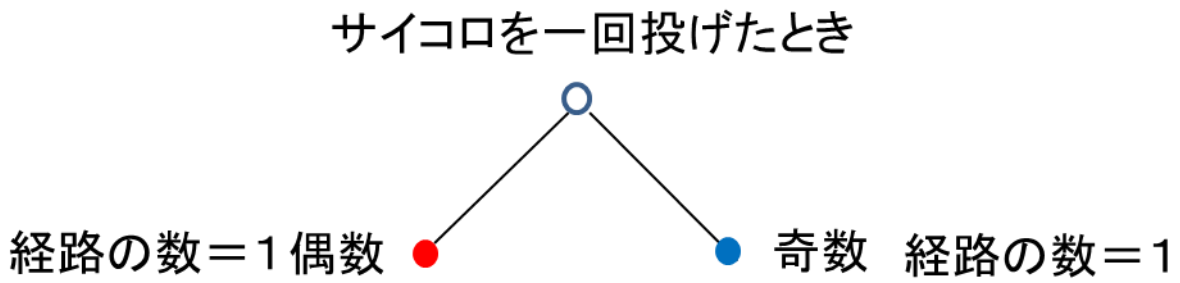
となります。全ての確率を足すと 0.6+0.4=1 になります。

ベルヌーイ分布にしたがう事象をくり返すと 2 項分布になります。

二項分布

サイコロの出る目を偶数と奇数に分ける場合を考えてみましょう。

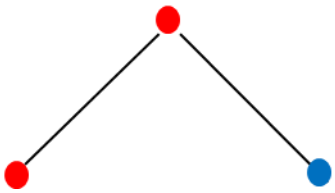
サイコロを一回投げたときの結果はつぎのようになります。奇数と偶数の出る確率をそれぞれ p と 1−p とします。赤🔴が偶数、青🔵が奇数とします。まず一番下の赤🔴と、青🔵に到達する経路の数を数えます。



それぞれ 1 です。○と赤🔴と、青🔵を結ぶ線が経路です。

つぎに再度サイコロを投げてみましょう。まず、一回目の結果が偶数の場合を考えます。その結果は

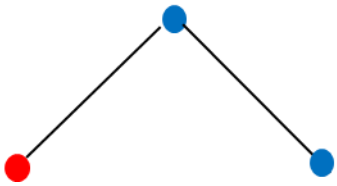
偶数が出た後に
サイコロを一回投げたとき



となります。赤🔴から赤🔴と赤🔴から青🔵の２つの可能性があります。赤🔴と、青●に到達する経路の数は１ずつです。

つぎに一回目の結果が奇数の場合を考えます。

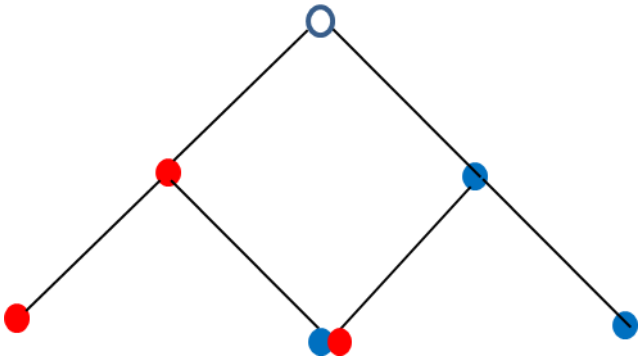
奇数が出た後に
サイコロを一回投げたとき



となります。青🔵から赤🔴と青🔵から青🔵の２つの可能性があります。赤🔴と、青●に到達する経路の数は１ずつです。

この２つの分岐を最初のグラフに書き加えます。

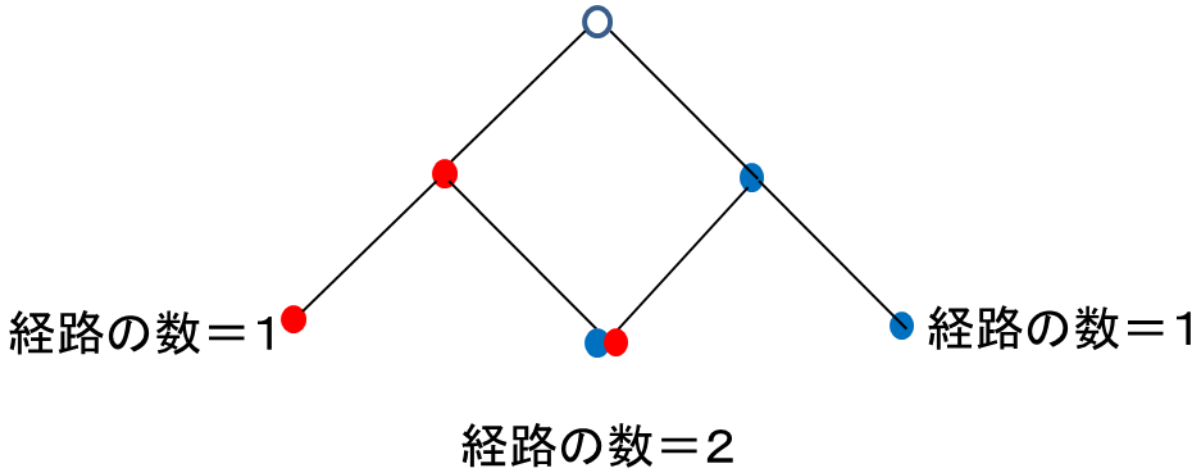
サイコロを二回投げたとき



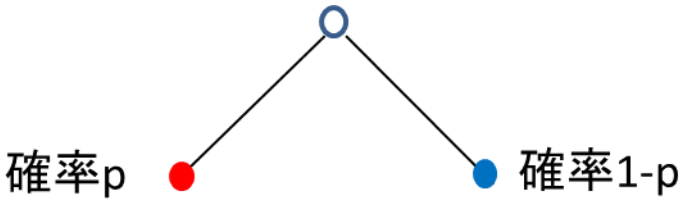
一番下の左赤🔴に到達する経路：○ → 🔴 →🔴
一番下の中央青🔵に到達する経路：○ → 🔴 →🔵
一番下の中央赤🔴に到達する経路：○ → 🔵 →🔴
一番下の左青🔵に到達する経路：○ → 🔵 →🔵
それぞれに至る経路は１つずつです。

しかし、○ → ● →●と○ → ● →●は赤丸と青丸の数は同じですので、同じと考えると中央に来る経路の数は２つです。

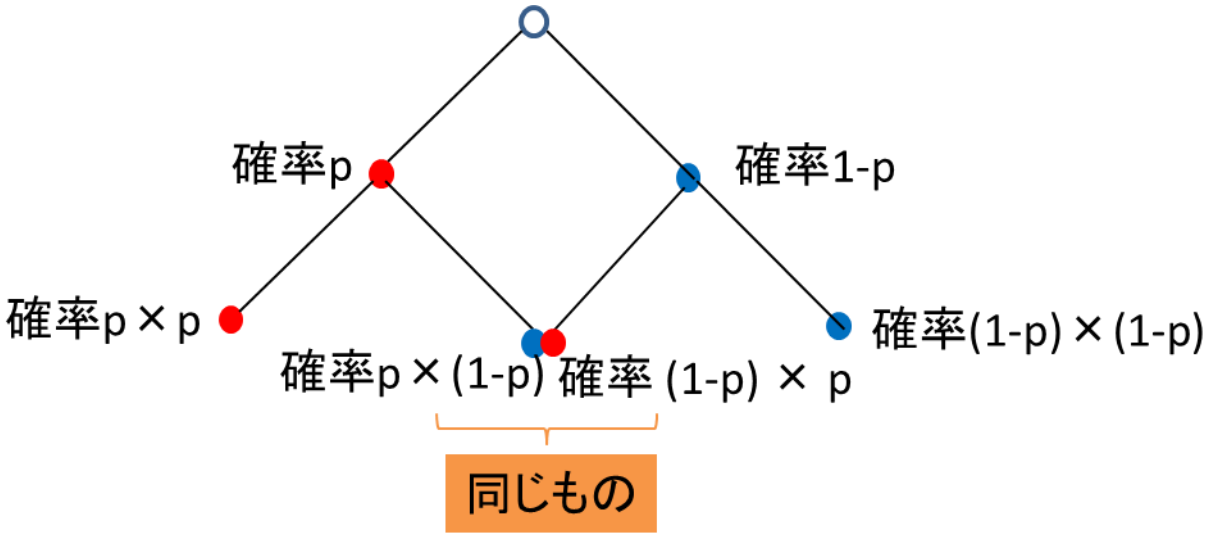
それぞれに到達する経路の数を数えます。



つぎに確率についても同様に考えてみましょう。同じグラフが使えます。サイコロを一回投げたときの結果はつぎのようになります。

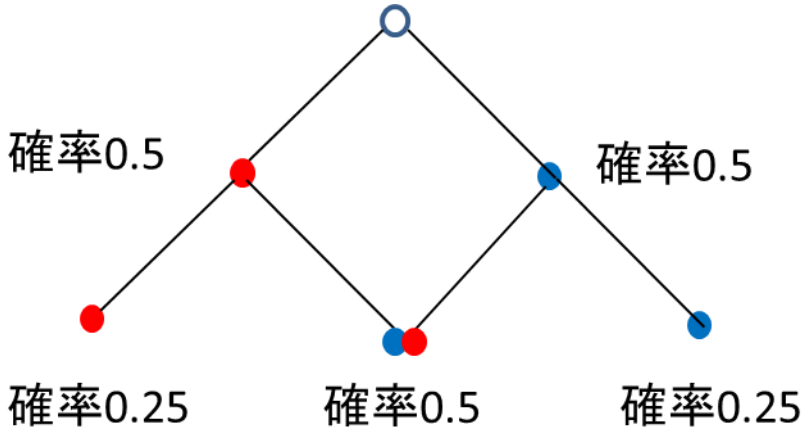


サイコロをニ回投げたときの結果はつぎのようになります。



サイコロを一回投げると
偶数の出る確率は
　　経路の数×確率 p
奇数の出る確率は
　　経路の数×確率 1−p
となります。

サイコロを二回投げると
偶数→偶数と出る確率は
　　経路の数×確率 p×p
偶数→奇数または奇数→偶数と出る確率は
　　経路の数×確率 p×(1−p)
となります。p=0.5 とすると



となります。これがサイコロを 2 回投げたときの二項分布です。

二項分布とは、結果が成功か失敗、裏か表、上昇か下落というような 2 値で表される試行を n 回行ったときに得られる離散型確率分布です。それぞれの試行は独立でなければなりません。p と n について確率質量関数は

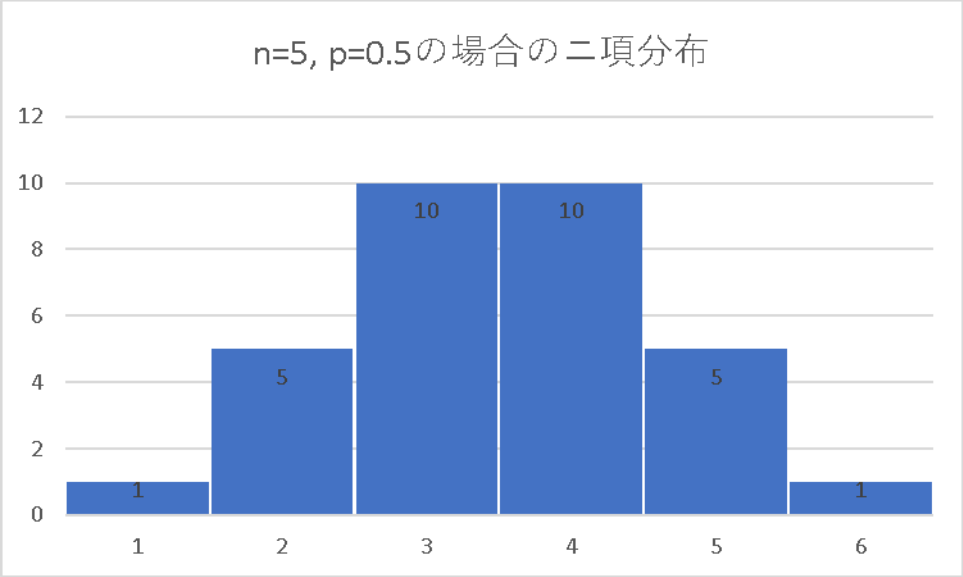
$$f(x) = \quad {}_nC_x p^x (1-P)^{(n-x)} = \frac{n!}{k!\,(n-x)!} p^x (1-p)^{(n-x)}$$

となります。ここで、${}_n C_k$ は n 個から k 個を選ぶ組み合わせの数です。p は成功確率です。2 項係数を表しています。また、

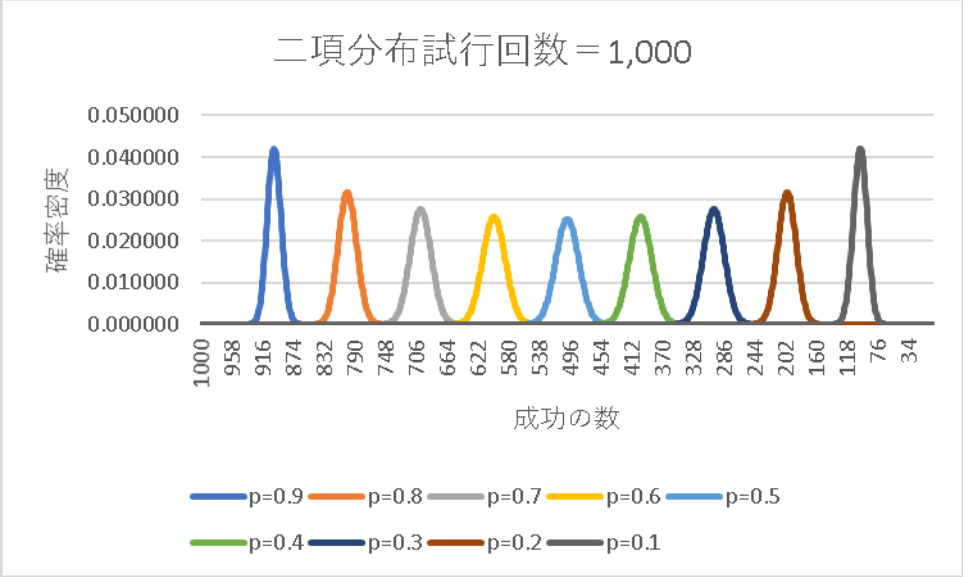
$${}_nC_x = \frac{n!}{k!\,(n-x)!}$$

です。ニ項分布では平均は $E(X) = np$、分散は $\mathrm{var}(X) = np(1-p)$ となります。n=1 のとき、2 項分布はベルヌーイ分布になります。

例題 2.9　n=5 で p=0.5 の場合の分布を計算してみましょう。
x=0 ${}_5C_0$=5!/0!/(5-0)!=1×2×3×4×5/(0!)/(1×2×3×4×5)=1
x=1 ${}_5C_1$=5!/1!/(5-1)!=1×2×3×4×5/(1)/(1×2×3×4)=5
x=2 ${}_5C_2$=5!/2!/(5-2)!=1×2×3×4×5/(1×2)/(1×2×3)=10
x=3 ${}_5C_3$=5!/3!(5-3)!=1×2×3×4×5/(1×2×3)/(1×2)=10
x=4 ${}_5C_4$=5!/4!(5-4)!=1×2×3×4×5/(1×2×3×4)/1=5
x=5 ${}_5C_5$=5!/5!(5-5)!=1×2×3×4×5/(1×2×3×4×5)/0!=1



例題 2.10: 二項分布を試行回数を 1000 回に固定して、成功確率を 0.1 から 0.9 まで変化させてグラフにしてその変化の度合いを確認してみましょう。



2.3.2 主な連続型確率分布

確率変数 Xが連続な値をとるとき、その分布は連続型確率分布となります。

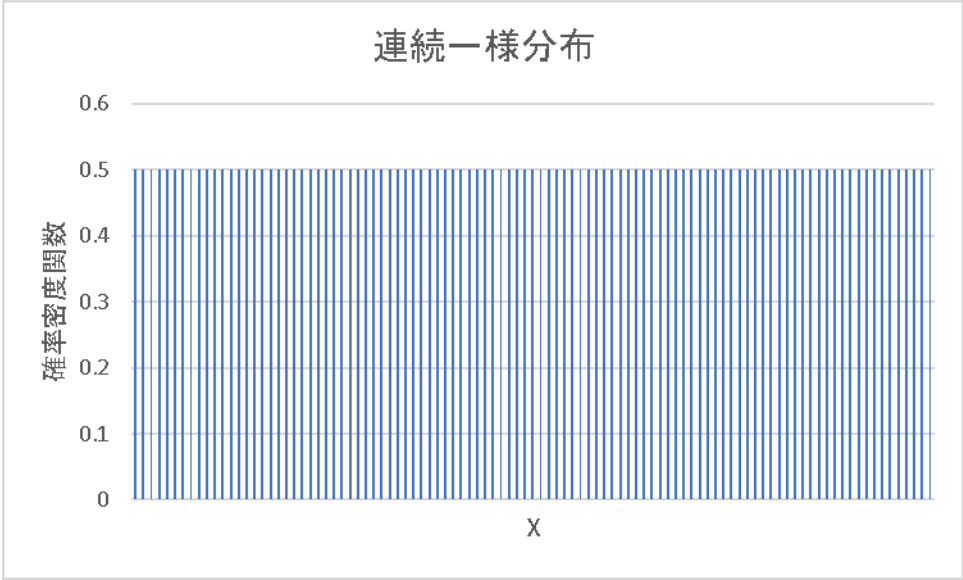
一様分布

確率変数の最小値と最大値を a, b としたときに、この区間で確率変数の生起する確率は等しいので、$\mathrm{U}(a,b)$ と書くことがあります。

連続一様分布の確率密度関数は

$$f(x)=\begin{cases}\dfrac{1}{b-a}、 & a\le x\le b\\ 0、それ以外のとき\end{cases}$$

となります。つぎに図は連続一様分布の様子です。



正規分布

平均に対して分布の形が対象で釣鐘の型をしていて、確率変数 X がとびとびの値ではなく連続な確率分布が正規分布です。正規分布では、分散は山の裾の広がり具合を表し、平均は分布の中心を示しています。正規分布の確率密度関数は平均と分散の関数として表されます。

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

ここでμは平均を、σ^2は分散を表します。平均ゼロ、分散１のとき標準正規分布といいます。

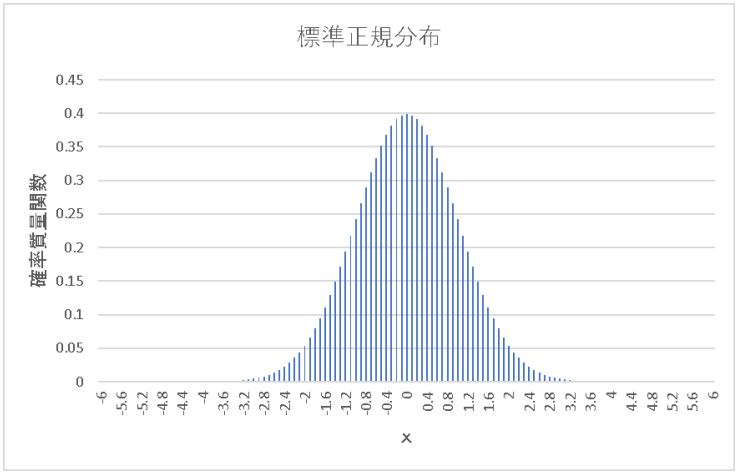
$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

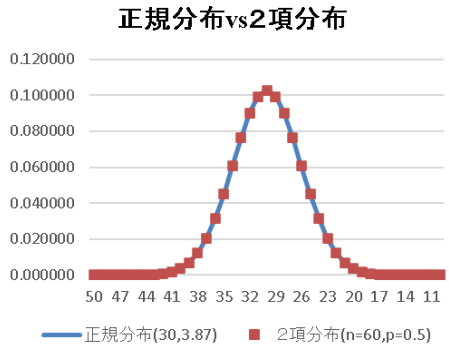
$$z=\frac{x-\mu}{\sigma}\sim N(0,1)$$

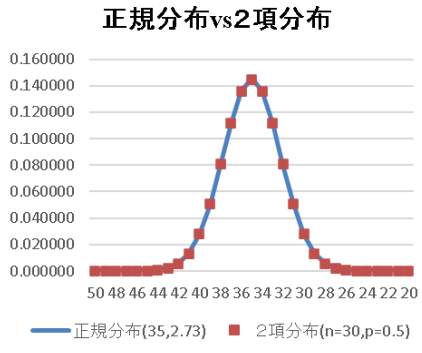
とすると、zは標準正規分布($N(0,1)$)にしたがいます。

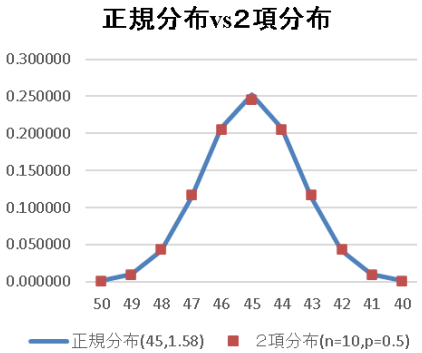
例題 2.11：標準正規分布を描いてみましょう。また、2 項分布が正規分布で近似できる自由度はどの程度か図

で確かめてみましょう。

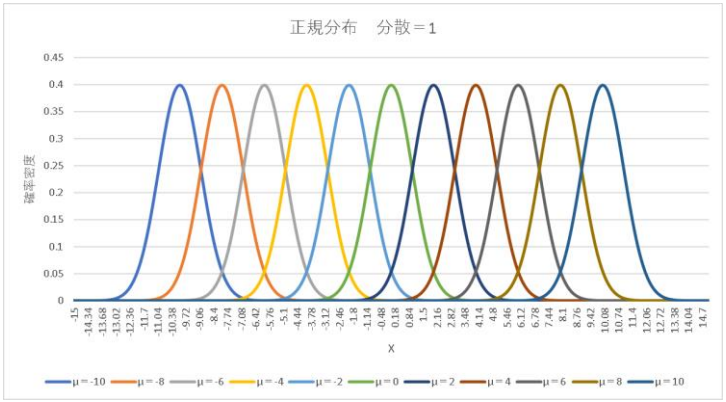


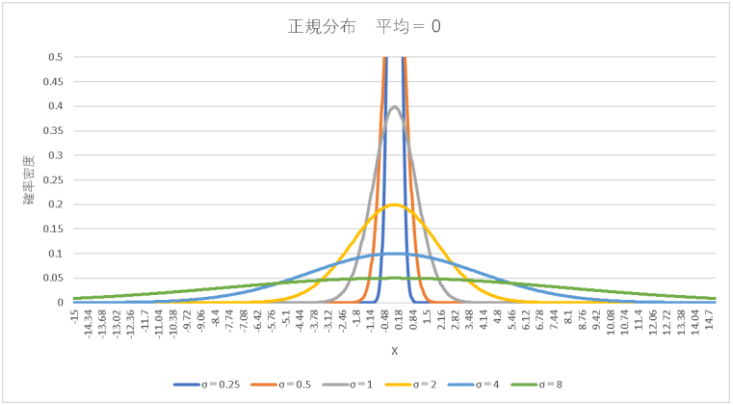






例題 2.12：平均を−10 から 10 まで変化させ、分散を 1 に固定して、正規分布を描いてみましょう。また、平均を 0 に固定して、分散を 0.1 から 100 まで変化させ正規分布を描いてみましょう。





2.4 期待値

離散値 $x_1, x_2, \cdots$の集合から得られる確率変数 Xの期待値は、

$$E(x) = \sum_{j=0}^{J} x_j f(x_j)$$

で表されます。jは根元事象の番号です。**x_jは確率変数の値で、$f(x_j)$はx_jの確率を表します。**

離散型確率分布にしたがう確率変数について考えてみましょう。x_jの相対度数をN_j/N、その確率をp_jとすると、その平均は

$$\overline{x_\iota} = \sum_{j=1}^{J} x_j \frac{N_j}{N}$$

で、期待値は

$$E(X) = \sum_{j=1}^{J} x_j p_j$$

です。

連続型の場合は、

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

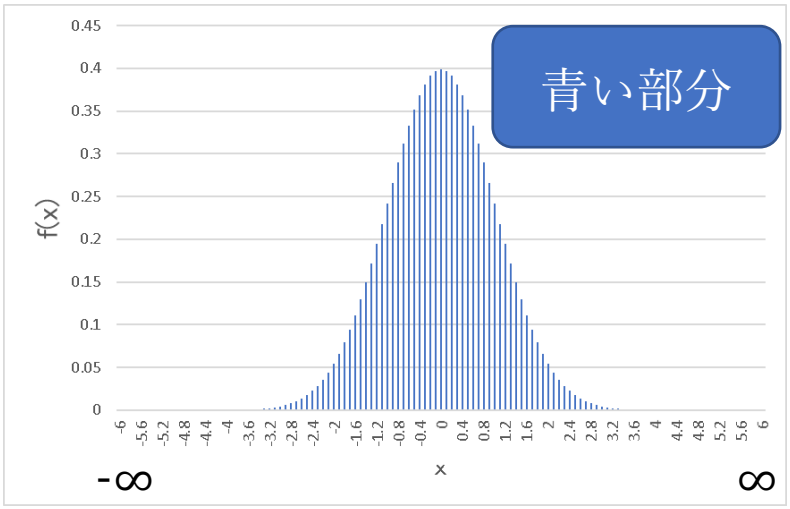
として定義されます。fは確率密度関数です。

は積分の範囲を指定してい

$$\int_{-\infty}^{\infty} xf(x)\mathrm{d}x$$

$xf(x)$を積分します

$\int$ は積分を表し



期待値には「期待される」値という意味もあり、予測に近い意味の場合もあります。

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、2 項分布、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。

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。

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？

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。

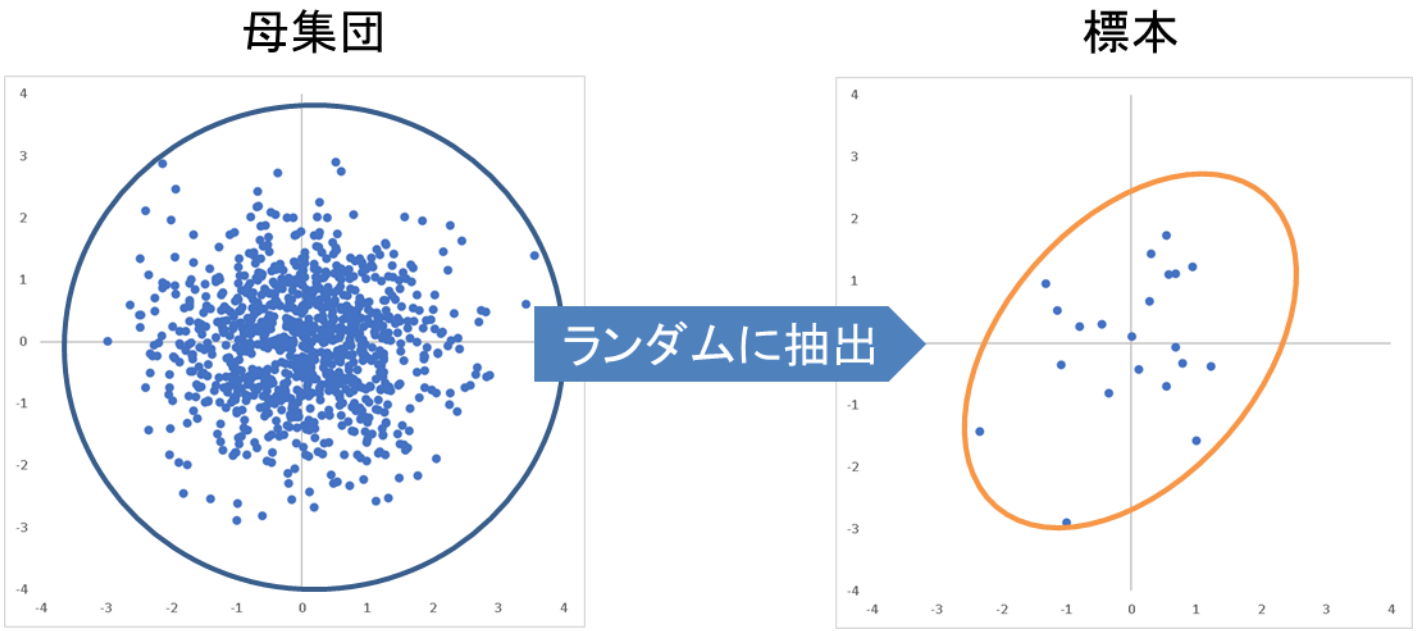
練習問題 2.5: A と B という事象があって、それが独立である場合と相関のない場合の違いについて説明してみましょう。

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。

練習問題 2.8: ベルヌーイ分布の例をあげてみましょう。

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？

データ全体を母集団と呼び、その母集団から抽出されたデータを標本、またはサンプルと呼びます。この２つは明確に区別される必要があります。



3.1 母集団

母集団とは、調査対象となっているデータのもととなる集合のことです。母集団というときには２つのタイプに分類されます。定義により母集団が確定している場合と、ある特定のモデルを前提としている場合があります。

例題 3.1：いくつかの身近な事例(調査・研究)を思い浮かべ、それらに関する母集団となる統計データと標本となる統計データについて記述してみましょう。

調査・研究	母集団	標本
選挙の当選予測	全有効票数	出口調査で得られた票数
製品満足度調査	製品を購入したすべてのお客様	アンケートに答えた一部のお客様
品質管理	製造したすべての商品	検査対処となった一部の商品
株価の予測	株価の予測モデル	入手可能な過去の株価

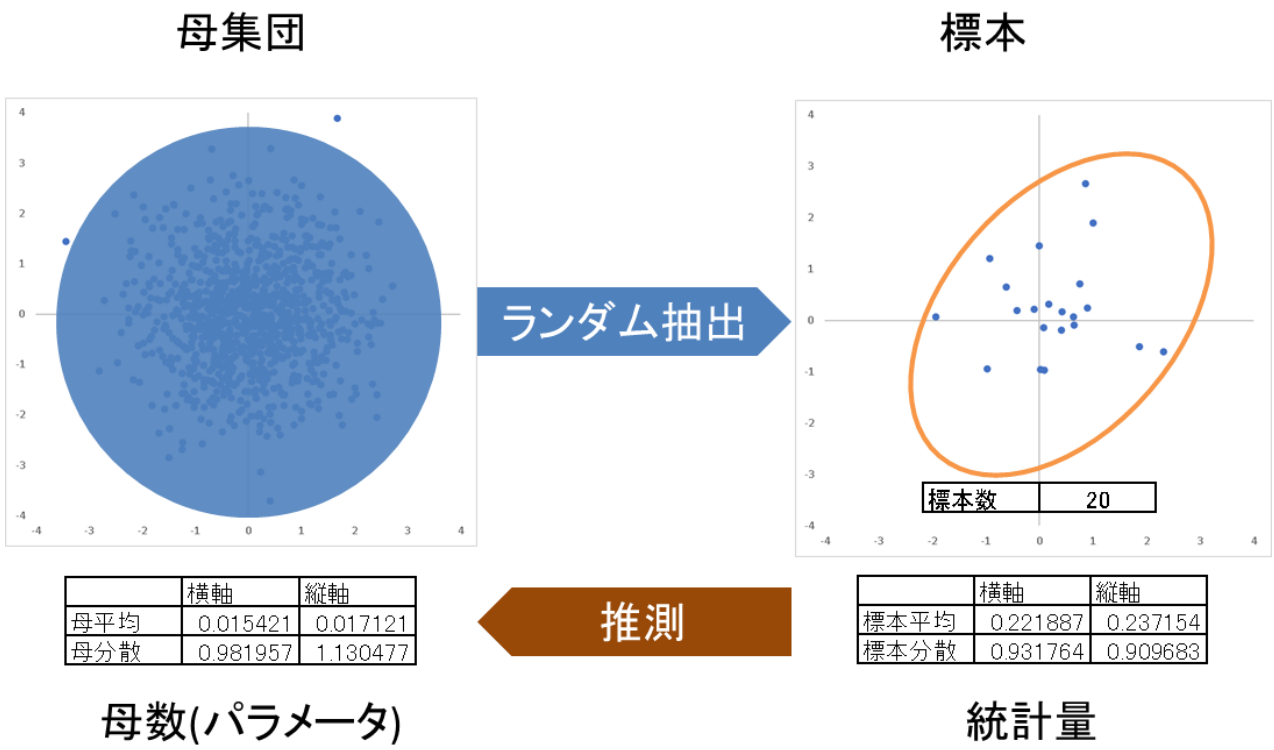
標本は、その母集団から抽出された個体の集合で、母集団の部分集合です。私たちは母集団について知りたいと思っているのですが、実際に知ることができるのは標本についてであって母集団についてではありません。したがって、推測統計では、部分集合である標本から集合全体の母集団を推測します。この過程では誤差が生じます。そこで、その大きさを確率の理論を用いて評価し、分析結果の信頼度を明らかにします。これが統計的手法を用いる１つのメリットです。定義により母集団が確定している場合は選挙の当選予測などに相当します。ある特定のモデルを前提としている場合は株価の予測などです。

繰り返しになりますが、母集団は様々な理由から母集団すべてを把握できないために、その母集団から標本を

得て、その標本を分析します。つまり、標本を分析しながら、母集団の特性を知ろうとしているのです。

母集団(確率分布)を特徴づける定数を母数(パラメータ)といいます。母平均、母分散は母数です。一方、標本に適用した統計的な関数を統計量といいます。標本平均、標本分散は統計量です。

例題 3.2：２組の正規乱数を 1000 個発生させそれを母集団とします。つぎにその母集団から２０個の標本を抽出し、母平均、母分散、標本平均、標本分散を計算してみましょう。



	横軸	縦軸
母平均	0.015421	0.017121
母分散	0.981957	1.130477

	横軸	縦軸
標本平均	0.221887	0.237154
標本分散	0.931764	0.909683

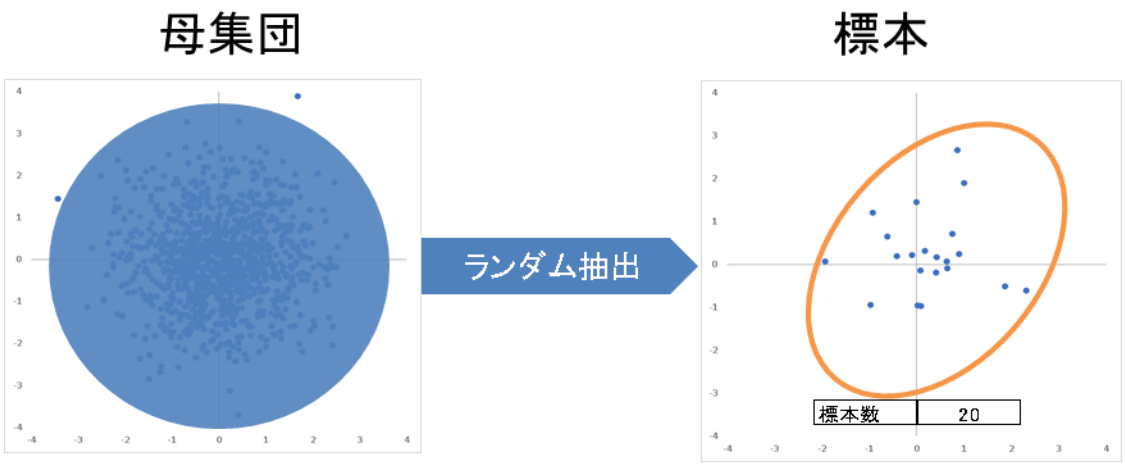
多くの調査・研究では母集団について知ることはできません。したがって、標本から母集団の統計的性質を推定するのです。

3.2 大数の法則と中心極限定理

データ全体を母集団と呼び、その母集団から抽出されたデータを標本といいます。標本の大きさが大きくなるとそれにともない、標本から得られる統計量は真の統計量(母数)に近づいていきます。

母集団が平均をもつときに、標本の大きさを大きくしていくと、母集団のもつ平均(母平均)、または真の平均に標本の平均は近づいていきます。これを大数の法則といいます。

例題 3.3：例題 3.2 で生成したデータを用いて、標本の大きさを２０，１００、５００、１０００と変えて標本分散、標本平均を計算してみましょう。

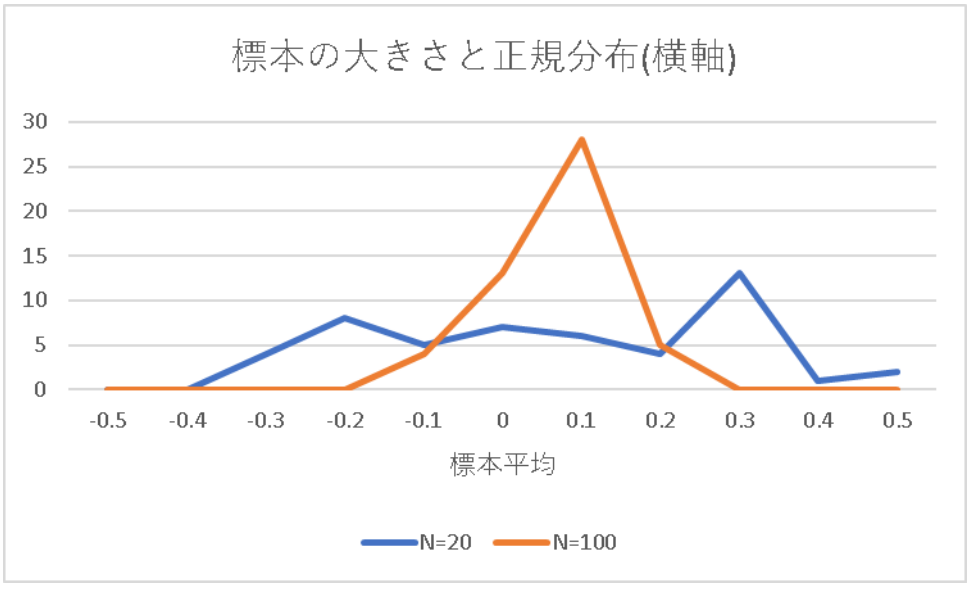


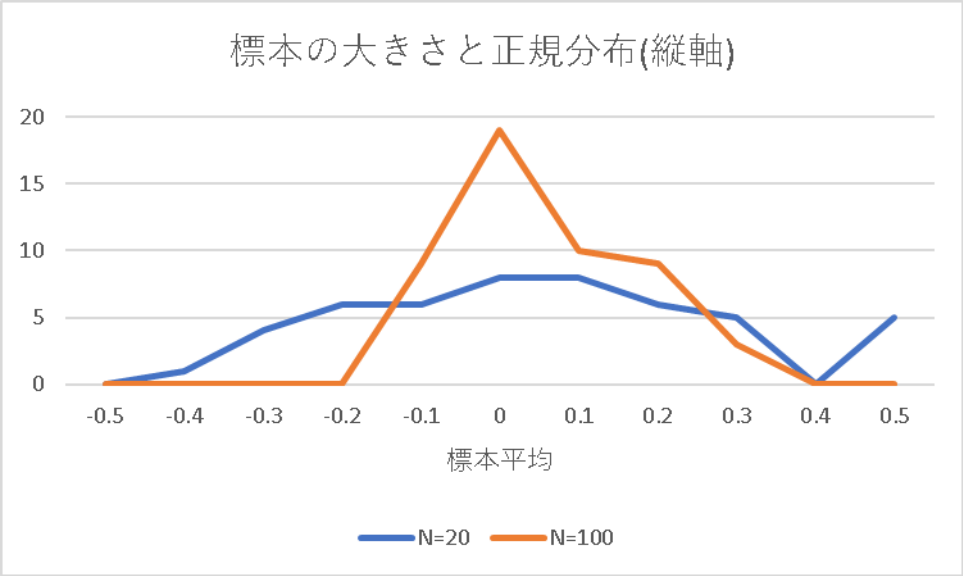
統計量

	標本平均		標本分散	
標本の大きさ	横軸	縦軸	横軸	縦軸
20	0.221887	0.237154	0.931764	0.909683
100	0.084487	0.068655	1.022771	0.929659
500	-0.00805	0.070463	1.000771	1.20215
1000	0.015421	0.017121	0.981957	1.130477

大数の法則により、Nが大きくなれば、観測データの平均$\bar{x}$は期待値 μ に近づきます。期待値は理論的な確率分布の平均と同じです。

例題 3. 4：例題２で生成したデータを用いて、標本の大きさが２０と１００の標本を母集団から複数抽出し正規性をグラフで表現してみましょう。





標本の平均は標本の大きさを大きくすれば正規分布に近づいていきます。これが中心極限定理です。

3.3 推定の性質

推測統計では、部分集合である標本から統計量を用いて母集団の母数を推定量として推測します。そこで推定量の性質について明らかにします。

3.3.1 一致性

ある母数の推定量がデータの数の増加にしたがい母数に収束するとき、それを一致性とよび、そのような推定量を一致推定量といいます。実際には標本の大きさは有限であり、推定量にはばらつきがあります。

3.3.2 不偏性

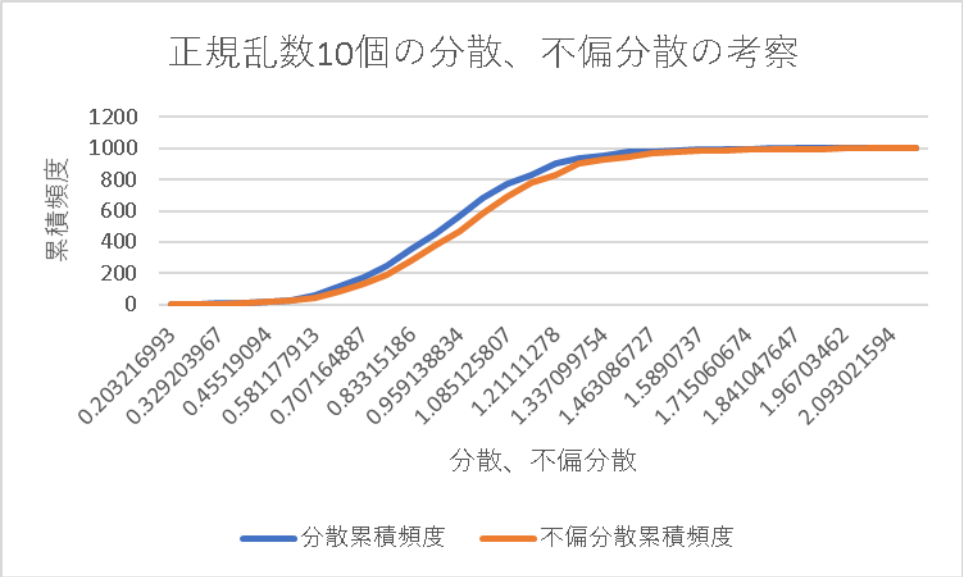
もう 1 つの推定量の基準に不偏性があります。推定量の期待値が母数に等しくなるとき不偏性があるといいます。その性質をもつ推定量を不偏推定量といいます。σ^2の不偏推定量は、得られたデータが$x_1, x_2, \ldots, x_n$のとき

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

という形で得ることができます。$\bar{x}$は得られたデータの平均値です。これを不偏分散とよびます。$n-1$は自由度といいます。x_iは自由にn個の値を取れるのですが、不偏分散の計算には平均値が含まれています。1 章では偏差の和がゼロになるように平均値を計算しました。平均が計算に含まれてしまうと、x_iは自由にn個の値を取れなくなってしまいます。自由に取れる値の数は$n-1$です。 1 つは平均と整合性が取れるように決まります。したがって不偏分散を得るには偏差平方和を$n-1$で割るのです。

練習問題 3.5: 正規乱数を 1 試行で 10 個発生させ、その分散、不偏分散を計算し、それを 1000 回繰り返して、

その特徴を可視化してみましょう。



標本の大きさが 10 個程度では自由度で割る効果が現れます。不偏分散の方がより母分散の 1 に近くなっています。

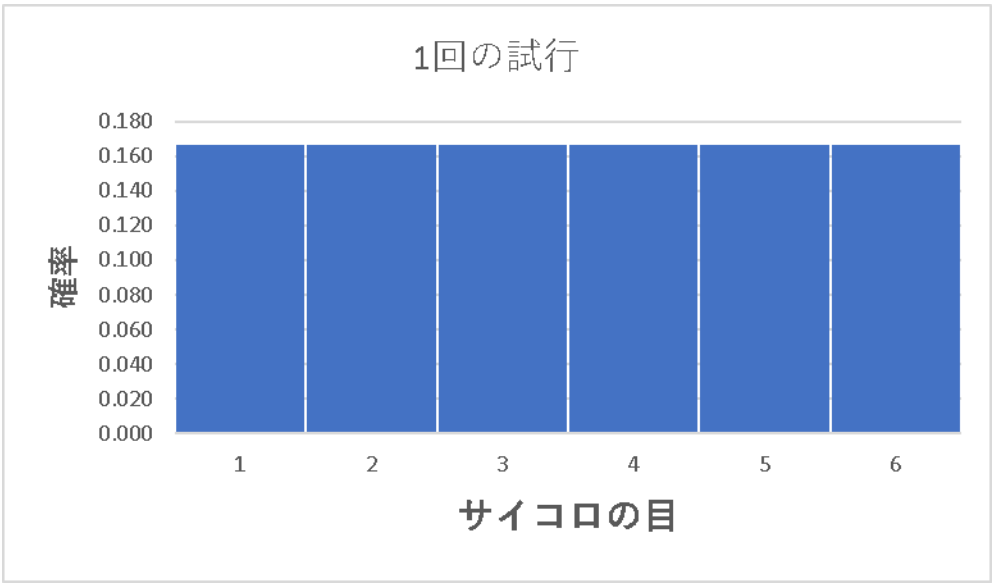
3.4 標本分布

例題 3.2〜4 で見たように、母集団から n 個の標本を繰り返し抽出すると、それぞれのデータ集合は，同じ値になるとは限りません。したがって、これらのデータ集合を確率変数と見なすことができます。

標本平均や標本分散などは統計量です。それぞれの標本抽出によって得られるデータ(情報)の値は同じになるとは限らないため、それぞれの統計量は，標本抽出の際にそれぞれが異なる数値となります。したがって、それぞれの標本抽出で得られた統計量から分布が得られます。このような、統計量の確率分布を標本分布といいます。

例題 3.6：サイコロを 1 回だけ振ることで得られた目の平均と 2 回だけ振ることで得られる目の平均を計算して、頻度図にしてみましょう。

1 回だけの試行：サイコロを一回だけ振ることを考えるとその出る目は 1, 2, 3, 4, 5, 6 のどれかです。したがってその目が出たときの平均はそれぞれ、1, 2, 3, 4, 5, 6 です。どの目も同じ確率で起こるとすると、平均が 1, 2, 3, 4, 5, 6 になる確率はそれぞれ 1/6 となります。



これは離散型一様分布になります。

2 回の試行：２度サイコロを投げるときには最初の結果と、２番目の結果が同じになるとは限りません。最初が 1 の場合を考えると、２番目の結果は 1, 2, 3, 4, 5, 6 の可能性があります。そこでこれらの結果をつぎに様に表現します。

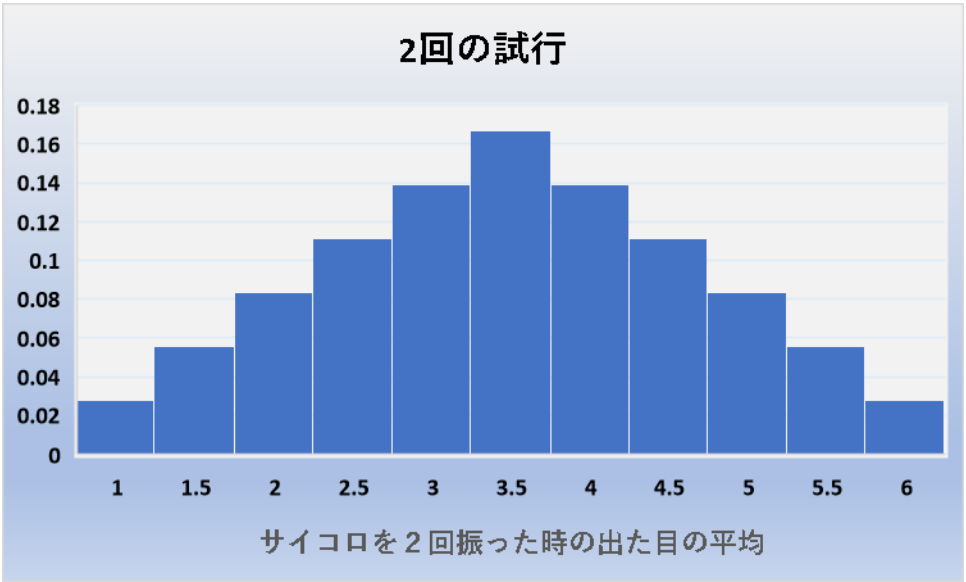
1 回目の試行の結果、2 回目の試行の結果の平均値をつぎの表に表しました。

		一回目の試行					
	平均値	1	2	3	4	5	6
2回目の試行	1	1.0	1.5	2.0	2.5	3.0	3.5
	2	1.5	2.0	2.5	3.0	3.5	4.0
	3	2.0	2.5	3.0	3.5	4.0	4.5
	4	2.5	3.0	3.5	4.0	4.5	5.0
	5	3.0	3.5	4.0	4.5	5.0	5.5
	6	3.5	4.0	4.5	5.0	5.5	6.0

そうすると平均の範囲は 1～6 となります。また、平均の標本空間Ωは

Ω＝{1.0，1.5，2.0，2.5，3.0，3.5，4.0，4.5，5.0，5.5，6.0}となります。その頻度を数えます。

頻度は{1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1}となります。これを頻度図としたものがつぎの図です。



平均はどのように標本が得られるかでばらつくことが分かります。そのばらつき具合はベル型の分布をしています。サイコロを振る回数を増やしていくと、この分布は正規分布に近づいていきます。それは中心極限定理を説明しています。

3.4.1 カイ二乗分布

確率変数 $X_1, X_2, \cdots, X_n$ が、互いに独立に、平均ゼロ、分散 1 の標準正規分布にしたがうとき、その統計量

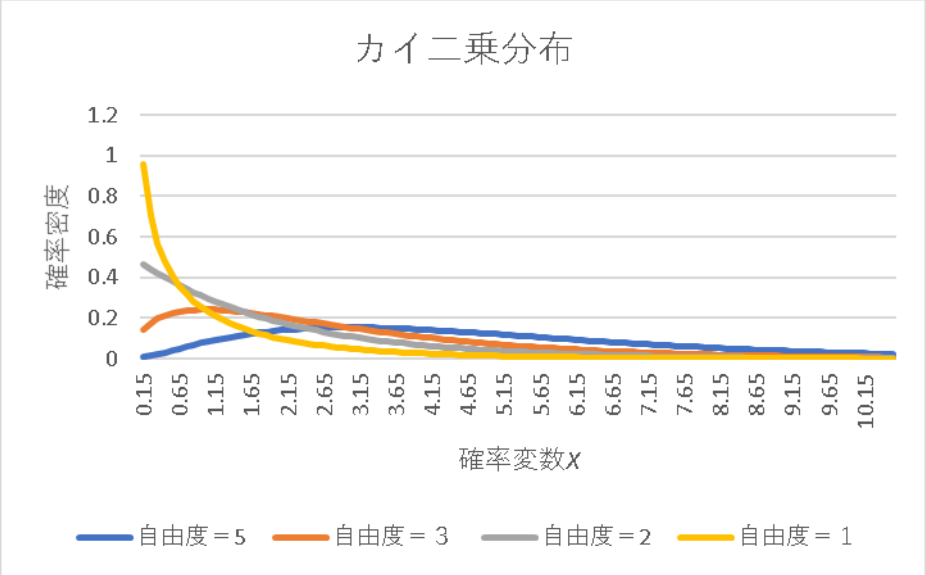
$$Z = \sum_{i=1}^{n} X_i^2$$

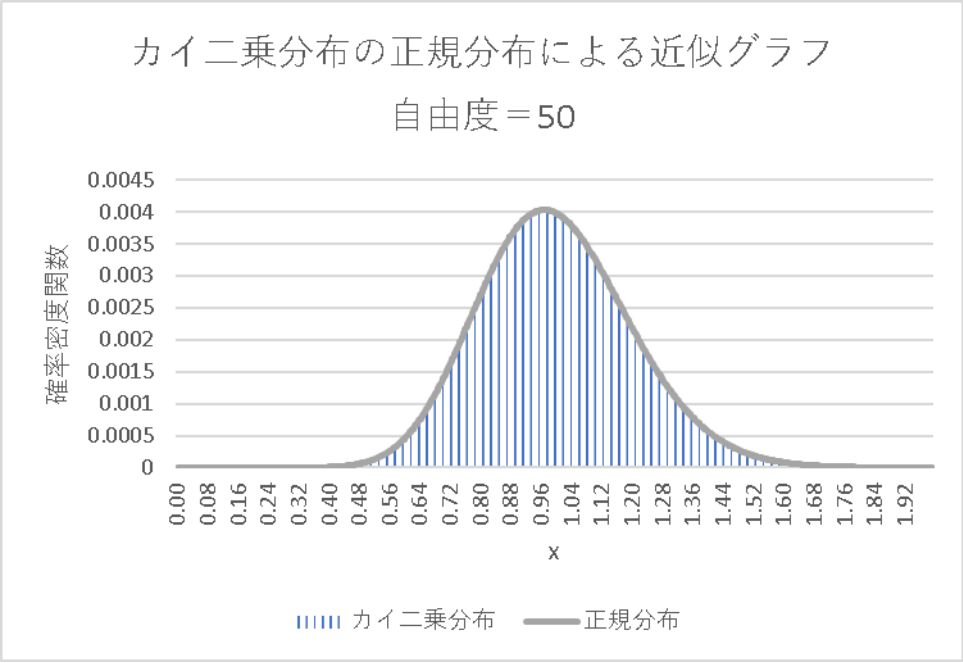
がしたがう分布を自由度 n のカイ二乗分布といいます。

$$Z \sim \chi^2$$

カイ二乗分布は n が大きくなると正規分布にしたがいます。

例題 3.7：カイ二乗分布の自由度を 1, 2, 3, 4, 5 と変えて図に描いてみましょう。また、見た目で正規分布といえるような標本の自由度を探しましょう。





標本の大きさが十分に大きければ Z は正規分布にしたがいますが、十分でなければカイ二乗分布にしたがいます。

例題 3.8：確率変数X_iが平均μ、分散σ^2に、したがうときその二乗和はどのような分布にしたがうでしょうか？

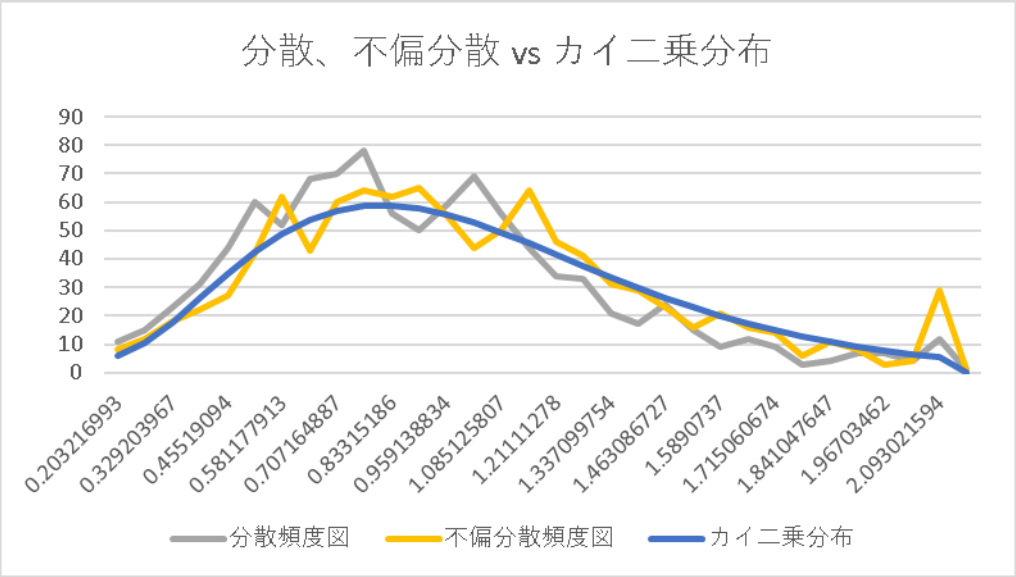
確率変数 $X_1, X_2, \cdots, X_n$ が、互いに独立に、平均ゼロ、分散 1 の標準正規分布にしたがうとき、その二乗の和$Z = \sum X^2$がカイ二乗にしたがうのでした。これを一般化してX_iが平均μ、分散σ^2にしたがうのですから、X_iを変換する必要があります。X_iから平均μを引き標準偏差σで割ってあげればX_iは標準正規分布にしたがいます。この統計量の二乗の和はカイ二乗分布にしたがいます。

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2=\sum_{i=1}^{n}Z^2\sim\chi^2_{(n-1)}$$

左辺を、不偏分散を含む形に変形します。

$$\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2(n-1)}{(n-1)}\frac{1}{\sigma^2}=\frac{s^2(n-1)}{\sigma^2}\sim\chi^2_{(n-1)}$$

つぎの図は例題 3.5 で作成したデータをもとに作成されています。例題３.５では１０個の乱数を生成し、その不偏分散と分散を計算しました。そしてその試行を 1000 回繰り返しました。x 軸は不偏分散です。ここでは不偏分散と分散の累積分布を計算し、それを利用して密度関数を計算し頻度をもとめています。そしてカイ二乗分布と不偏分散と分散の頻度を折れ線グラフで示しています。不偏分散の頻度図がカイ二乗分布の頻度図に近いことがみて取れます。



3.4.2 t分布

確率変数が正規分布にしたがうとき、その母集団の平均と分散が既知であるというような場合は、まれです。ステューデントの t 分布は、標本の大きさが小さいときに、そのような母集団の平均を推定するのに用いられます。

確率変数 $X_1, X_2, \cdots, X_n$ は平均 μ、分散 σ^2 の正規分布に独立にしたがいます。その標本平均が

$$\bar{X} = \frac{X_1 + X_2, \cdots + X_n}{n}$$

不偏分散が

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

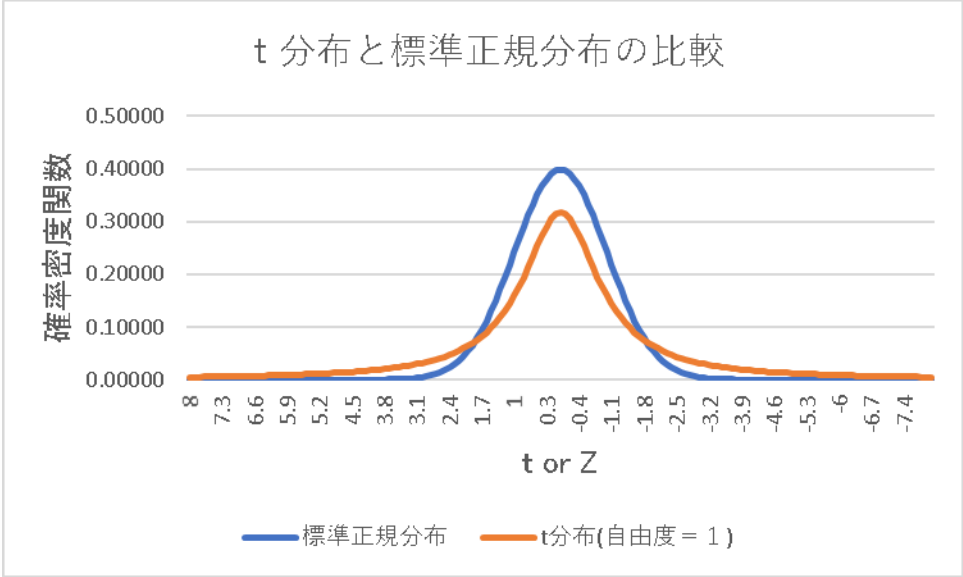
のとき、

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

は自由度 n の t 分布にしたがいます。標本の数が十分に大きければ、t 統計量は標準正規分布にしたがいます。

$\bar{X}$ の標準偏差 $S/\sqrt{n}$ を標本平均の標準誤差(standard error ， s.e.)といいます。

例題 3.9：自由度 1 の t 分布と正規分布を比べてみましょう。



確率分布の分類

連続　vs　離散

(正規分布)　(2項分布)

母集団 vs 標本

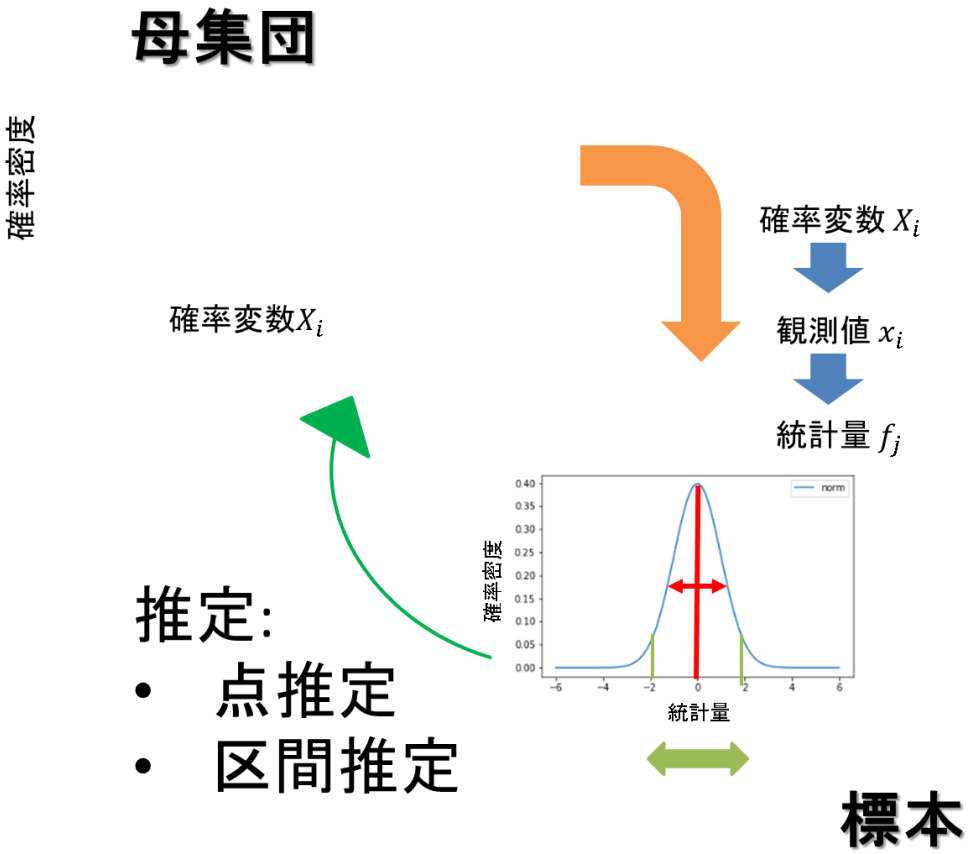
(正規分布)　(t-分布)

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。
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。
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。
練習問題 3.6: カイ二乗分布、ｔ分布が正規分と一致すると見えるデータ数を目視で確認してみましょう。

第４章　統計的推定

推測統計のはじめは統計的推定です。確率変数と確率分布、そして観測値を基礎とする推測統計を学びます。ここでは、未知の母数を、観測値(得られたデータ)をもとに推測していきます。これを統計的推定の問題といいます。得られたデータ x から未知の母数を考えるとき、**その統計量の推定値とともに、その信頼度も考える必要があります。**つまり、推定値がどの程度の範囲にあるかを考える必要があるのです。母集団からデータ $x_1, x_2, \ldots, x_n$ が得られるとすると母数の推定値は何度も計算することができ、かつその値はいつでも同じではありません。それらを確率変数ととらえるとき、推定量となります。

母数の推定値を表現する方法には２つあり、１つの値としてとらえるのが点推定、上限、下限の間の区間としてとらえるのが区間推定です。



4.1 点推定

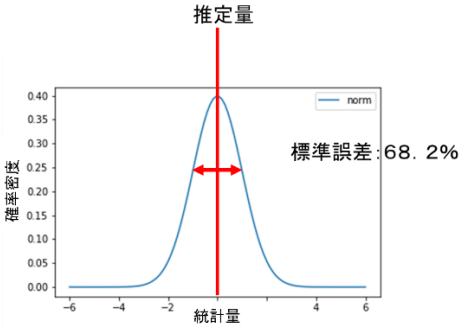
標本 $x_1, x_2, \ldots, x_n$ から算出される１つの値で、未知の母数を推定する方法を点推定といいます。
- 平均、分散など

母数 θ に対してその推定量は θ に $\hat{}$ をつけて表します。

4.1.2 標準誤差

母集団から得られた標本から統計量を推定するとき、そのばらつきの度合いを標準誤差といいます。これは標本のすべての組み合わせの標準偏差で表します。単に標準誤差といったときには平均のばらつきを表し、それは分散の推定量を標本の大きさで割り、その平方根をとったものです。推定量と標準誤差は組として示されます。

- **点推定**
  - **推定量　$\hat{\theta}$**
  - **標準誤差(推定量の標準誤差)　$se(\hat{\theta})$**



4.2 区間推定

標本から得られる統計量の上限と下限の 2 つの値を求めて、その間に母数がふくまれるという表現の方法が区間推定です。

- 信頼区間

確率変数を X、区間の上限を$U(X)$、下限を$L(X)$、そして、母数を M とすると、

$$L(X) \leq M \leq U(X)$$

と表現します。M は$U(X)$と$L(X)$の間に入ることを意味します。

- 信頼係数

この信頼区間の中に母数が入る確率が信頼係数で $1-\alpha$ で表します。したがって、

$$P[L(X) \leq M \leq U(X)] = 1 - \alpha$$

となります。$L(X)$、 $U(X)$の決め方が統計的推定に大きな影響を与えます。

4.2.1 母平均の区間推定

標本$x_1, x_2, \cdots, x_n$, は独立に平均 μ，分散 σ^2の正規分布にしたがうとします。

- 母分散が既知の場合
- 標準正規分布を用います。

母平均の区間推定を行ってみましょう。

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

$N(0,1)$は標準正規分布

- z_α：確率 α における標準正規分布の臨界値

1 - α は信頼係数で標本平均が信頼区間に入る確率になります。

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

標本$x_1, x_2, \cdots, x_n$, は独立に平均 μ，不偏分散s^2の正規分布にしたがうとします。

- 母分散は未知です。
- t 分布を用います。

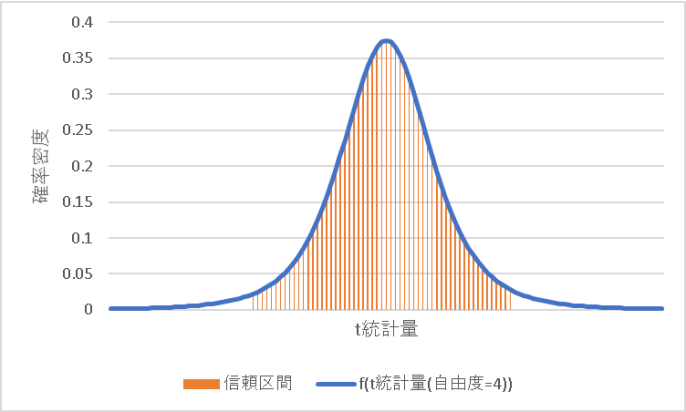
$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim T(n-1)$$

- $t_{(\alpha,n-1)}$：確率 α、自由度 n - 1 の t 分布の臨界値

$$\bar{X} - t_{(0.5\alpha,n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{(0.5\alpha,n-1)}\frac{s}{\sqrt{n}}$$

例題 4.1：ニキビの治療を受けに病院を訪れた５名の患者さんにそれぞれ A,B,C,D,E とローマ字を割り当てます。訪問時の患者 A のニキビの数は１１、B は９、C は１２、D は８、E は１０とします。その際の母平均の推定値を求めてみましょう。信頼係数は９５％とします。

患者のニキビの平均個数は１０個です。不偏分散は２、t 統計量の臨海値は±2.13 です。よって下限は 8.65、上限は 11.34 です。



例題 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。

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim T(n-1)$$

から

$$\bar{X} - t_{(0.5\alpha,n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{(0.5\alpha,n-1)} \frac{s}{\sqrt{n}}$$

が得られます。 α =0. 01 とすると$t_{(0.5\alpha,n-1)}$=2. 33 です。

	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	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
13	上限	8.43	0.54	0.28	2.63	0.09	16.55	48.59	1.00	3.32	0.67	10.49	5.69
14	下限	8.21	0.52	0.26	2.45	0.08	15.20	44.35	1.00	3.30	0.65	10.35	5.58
15		A	B	C	D	E	F	G	H	I	J	K	評価
16		7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	5

4.2.2 母分散の区間推定

分散の区間推定をする場合には、カイ二乗分布を用います。標本分散では

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \sigma^2 \chi^2_{(n-1)}$$

の関係があります。これを変形して、

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2_{(n-1)}$$

とします。そうすると z はカイ二乗分布にしたがいます。信頼係数 $1-\alpha$ の信頼区間は

$$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\chi^2_{(0.5\alpha,n-1)}}<\sigma^2<\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\chi^2_{(1-0.5\alpha,n-1)}}$$

となります。また

$$\frac{s^2(n-1)}{\chi^2_{(0.5\alpha,n-1)}}<\sigma^2<\frac{s^2(n-1)}{\chi^2_{(1-0.5\alpha,n-1)}}$$

標本の大きさが大きくなると信頼区間は狭くなります。

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。

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。

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。