

本書は、エクセルを用いてデータ分析に必要な統計学的手法を学ぶ入門書です。手元にあるデータの性質を、グラフとして表示して目視でとらえ、また数値として要約する記述統計を出発点として、データの真の姿を推測する点推定と区間推定までを学びます。点推定と区間推定を学ぶことは推測統計の基盤を学ぶことになります。これらの過程で、さまざまで、より複雑な推測統計の手法を独学で学ぶ、学び方を身に着けます。第1章では記述統計を簡単に復習します。第2章では、事象と確率の関係からはじめて、確率分布を理解します。確率変数は、確率分布で表現される確率にしたがい生起される乱数です。乱数ですので将来の値がいくつになるかの予測はできませんが、期待値は計算できます。つぎの第3章で、母集団と標本の間関係を理解します。2つの関係を支配しているものは標本の大きさです。この関係を、エクセルで乱数を発生させて理解します。例題のなかでF9(再計算)キーを押してくださいというものが、その関係を明確にしてくれるはずです。標本はいつも異なるもので、多くの場合、母集団ではないのだということ、たとえば母数のもつ平均と標本の平均が同じになることはまれなのだということを肌で感じられれば合格です。それが感じられて、はじめて、点推定と区間推定が理解できたことになり、そしてつぎの学習の段階に進むことができます。

目次	
第1章 記述統計	4
1.1. 変数の分類	4
1.2. 記述統計	5
1.2.1. データの可視化	5
1.2.2. 要約統計量	10
第2章 確率変数：確率と確率分布	21
2.1. 確率	21
2.1.1 確率の定義	22
2.1.2. 事象と確率	22
2.1.3. 条件付確率	24
2.1.4. 事象と試行	26
2.1.5. 事象と独立性	27
2.2. 確率変数	28
2.2.2. 独立な確率変数	29
2.3. 離散型確率分布	30
2.3.1. 離散一様分布	30
2.3.2. ベルヌーイ分布	31
2.3.3. 二項分布	31
2.4. 連続型確率分布	35
2.4.1. 連続一様分布	36
2.4.2. 正規分布	36
2.5. 期待値	38
第3章 母集団と標本	40
3.1. 母集団	40
3.2. 適切なデータ収集	41
3.3. 大数の法則と中心極限定理	42
3.4. 推定の性質	43
3.4.1. 一貫性	43
3.4.2. 不偏性	44
3.5. 標本分布	44
3.5.1. カイ二乗分布	46
3.5.2. t 分布	48
3.5.3. F 分布	49
第4章 統計的推定	51
4.1. 点推定	51
4.1.1. 標準誤差	51
4.1.2. 一致推定量と不偏推定量	52
4.2. 区間推定	52
4.2.1. 信頼区間	52

4.2.2.	信頼係数.....	52
4.2.3.	母平均の区間推定	52
4.2.4.	母分散の区間推定	54
4.2.5.	信頼区間の意味.....	54
4.2.6.	母比率の信頼区間	55

第1章 記述統計

統計データとはどのようなものなのでしょうか？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必要があります。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（CVRV）で検査されています。CVRV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>

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

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なることがあります。

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。

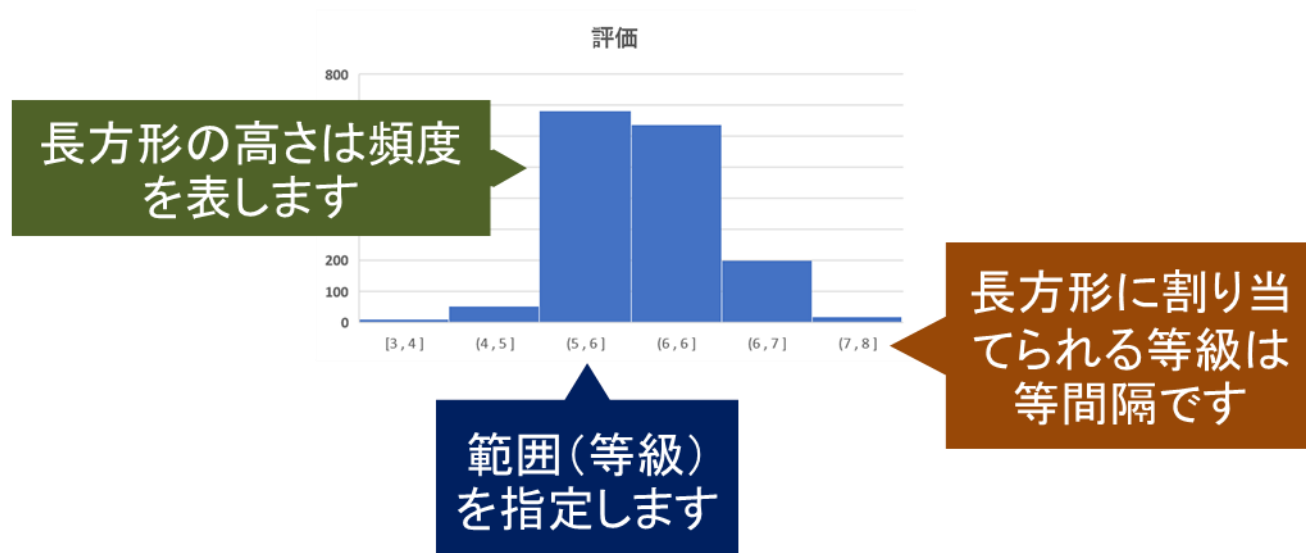


図 1.2 頻度図の製作ヒント

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なのが分かります。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。

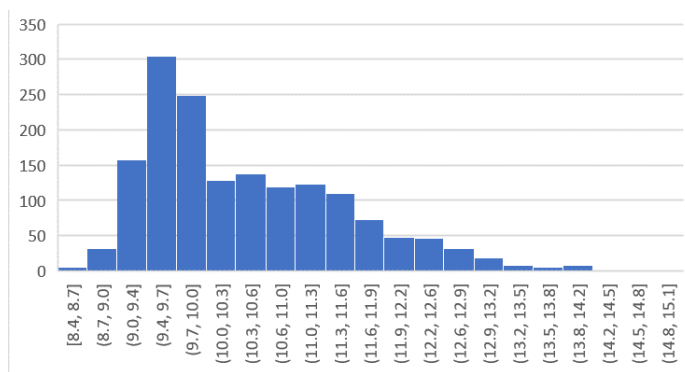


図 1.3 頻度図 化学成分 K

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。

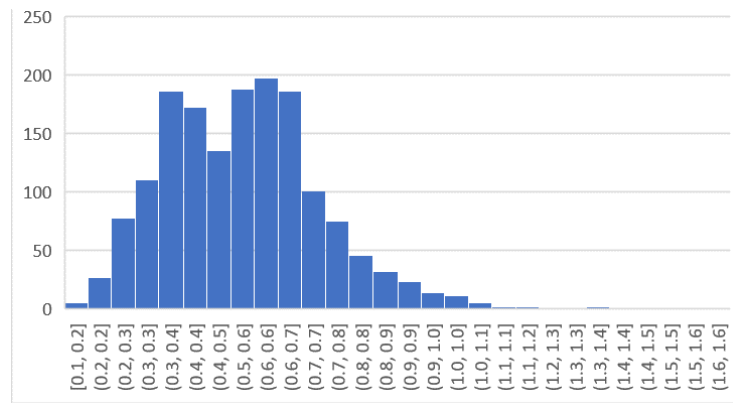


図 1.4 頻度図 化学成分 B

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。

同じデータでも等級のとりかたでイメージが変わる

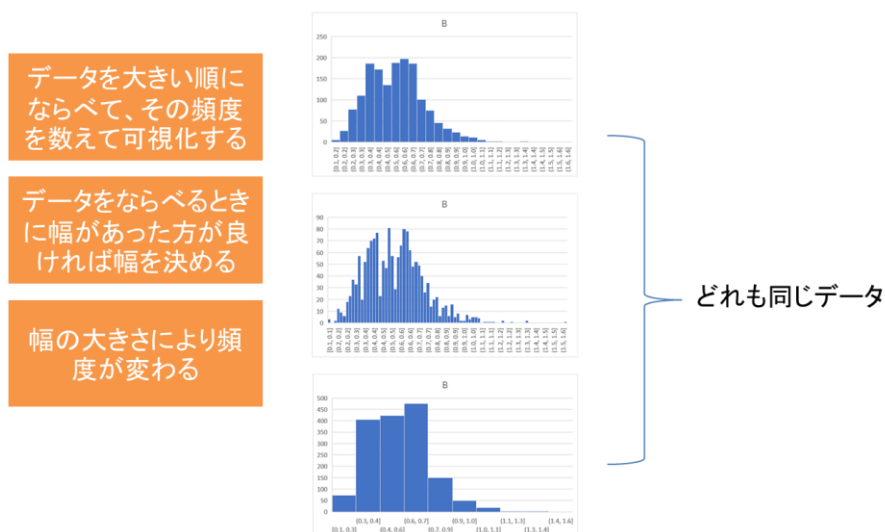


図 1.5 頻度図の等級と与える印象

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。

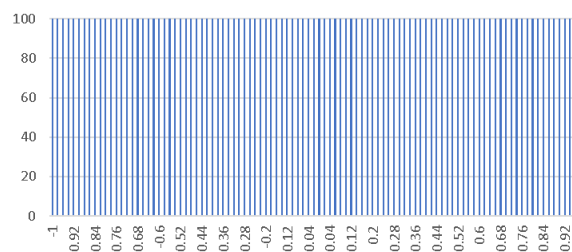


図 1.6 一様な頻度図(例題 2.6)

- － ベル型：頻度の高さは横軸に対してベル型。

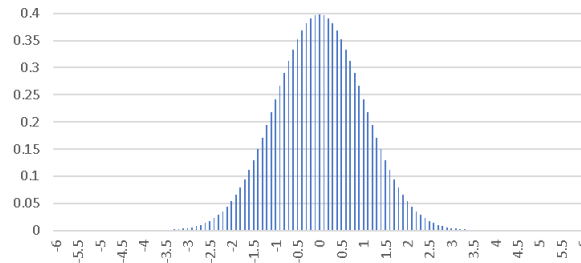


図 1.7 ベル型の頻度図(例題 2.7)

- 右に裾長：右にすそ野が長く、頻度が左寄り。

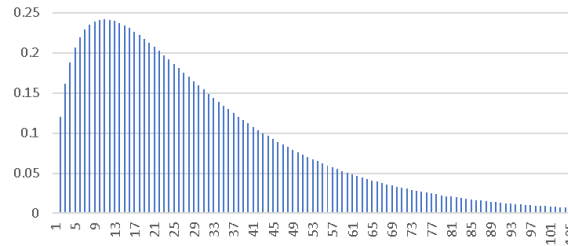


図 1.8 右にすそ長

- 左に裾長：左にすそ野が長く、頻度が右に寄り。

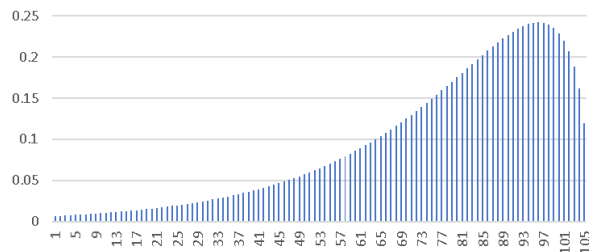


図 1.9 左にすそ長

- 複数のこぶ：複数頻度の山やコブ。

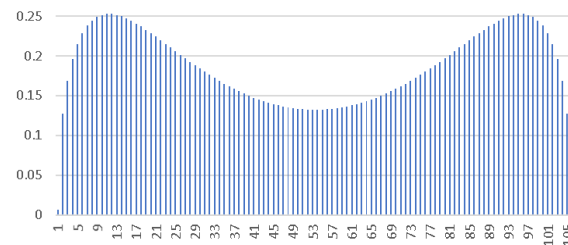


図 1.10 複数頻度の山に分けられます。

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。

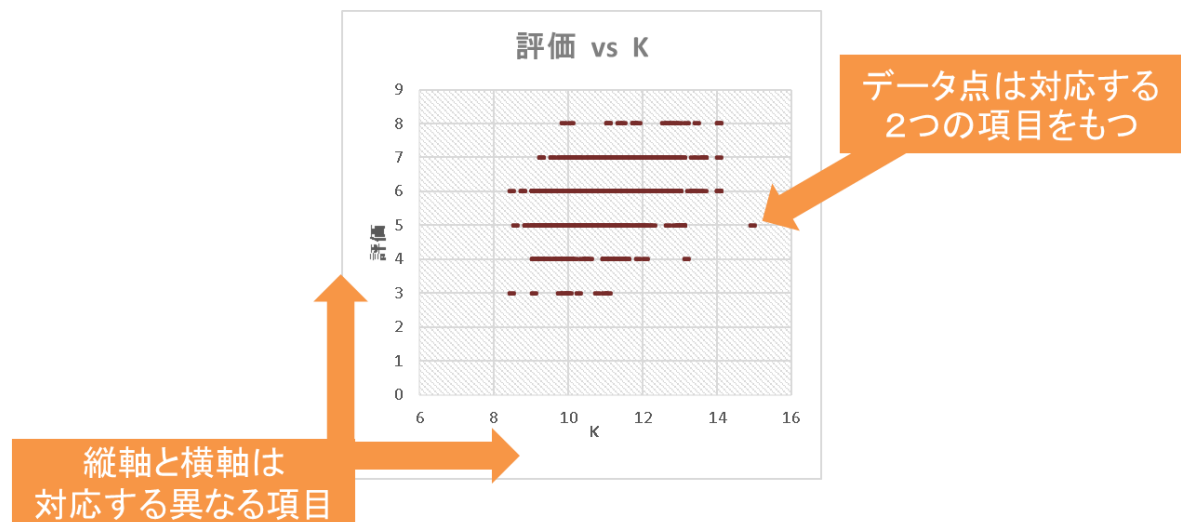


図 1.11 散布図

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。

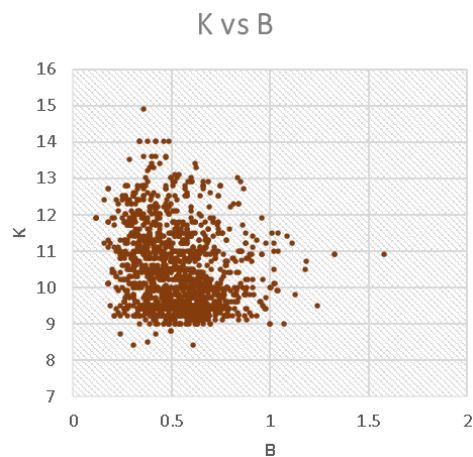


図 1.12 散布図：化学成分 K vs B

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。

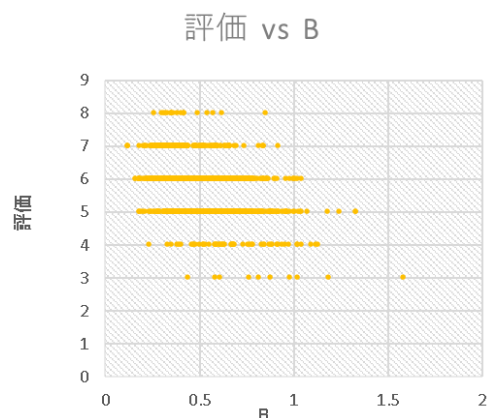


図 1.13 散布図: 評価 vs 化学成分 B

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ため、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きときもあります。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りますが、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。

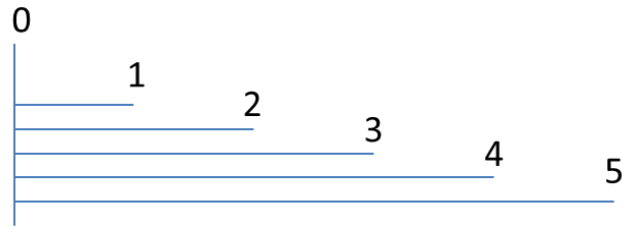


図 1.15 5つの数値

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。

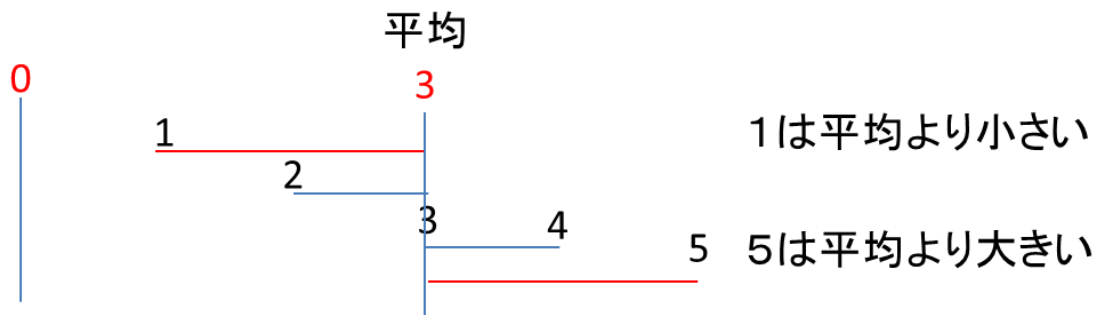


図 1.16 平均の分析

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

$$\frac{a_1 + a_2 + \dots + a_n}{n}$$

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。

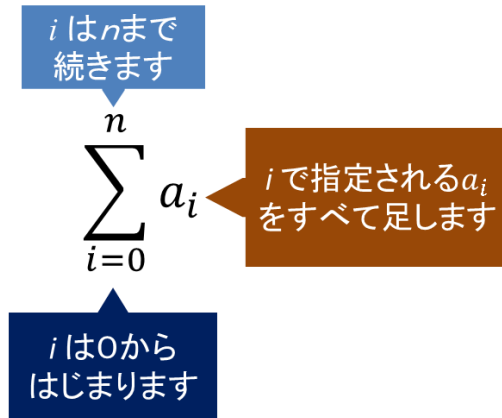


図 1.17 シグマの意味

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。

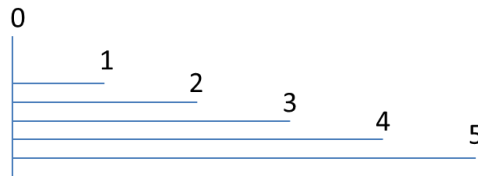


図 1.18 5つの数値

1 から 5 までの数値のばらつきを考えると、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。

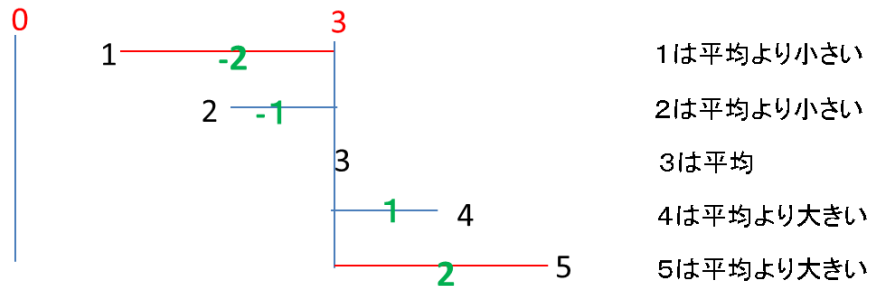


図 1.19 偏差

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。

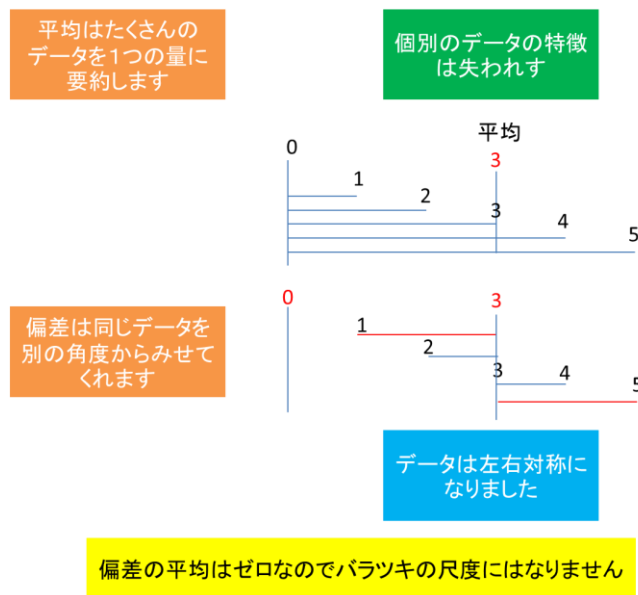


図 1.20 平均と偏差

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。

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

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。

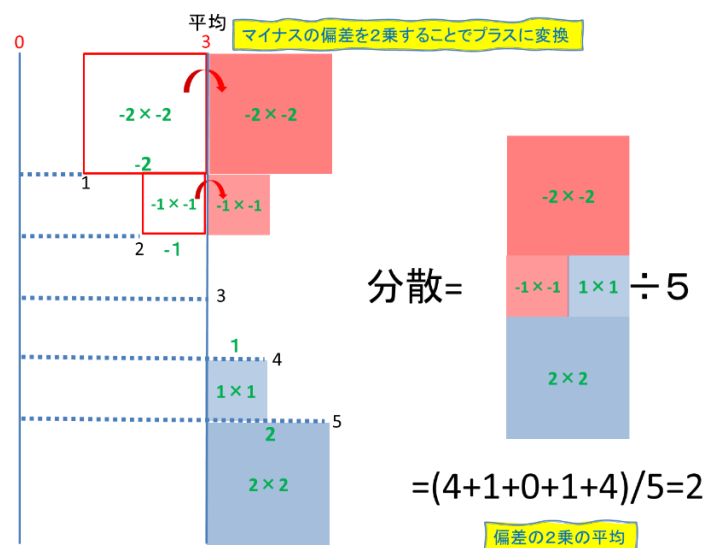


図 1.21 分散の理解

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。

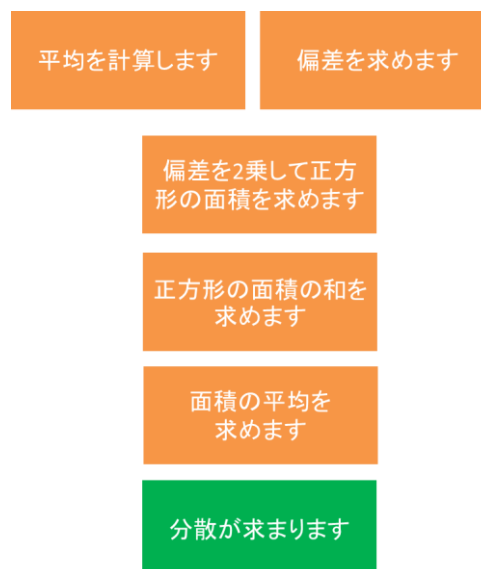


図 1.22 平均と偏差と分散の理解

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。

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

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。

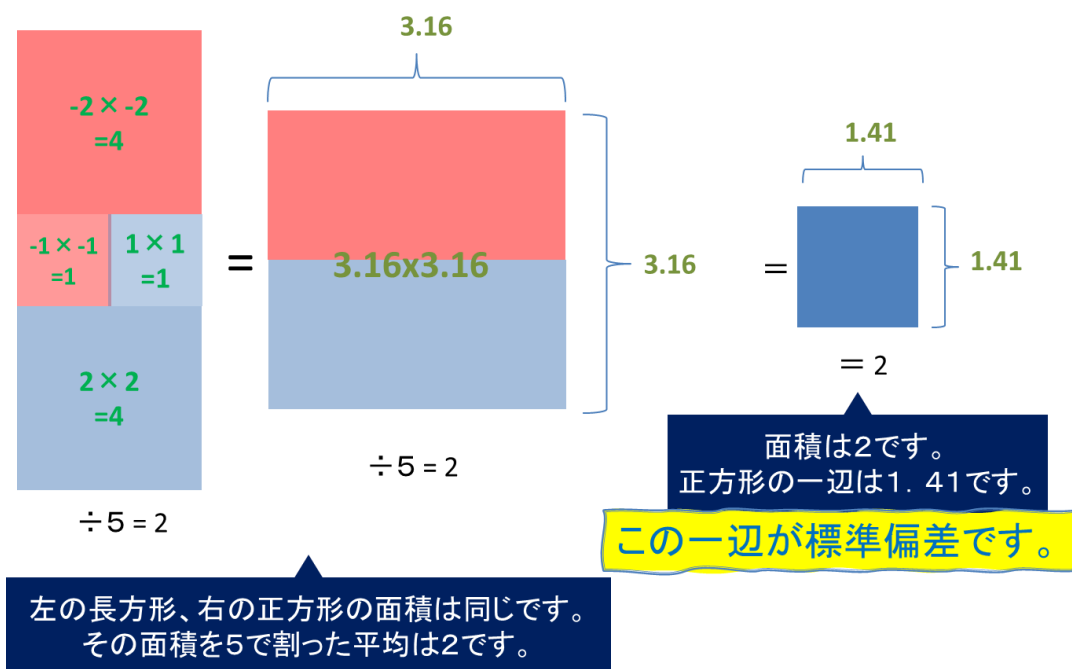


図 1.23 標準偏差のイメージ図

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。

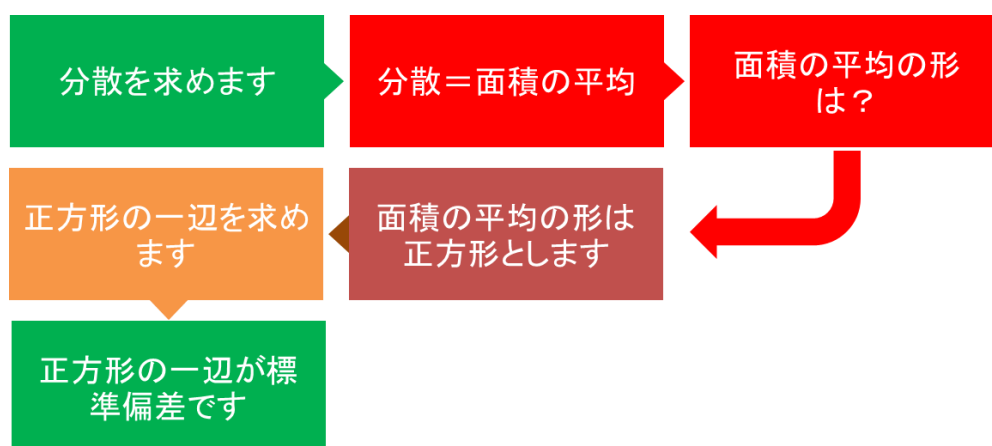


図 1.24 分散と標準偏差

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 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ の共分散は、つぎのように定義されます。

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

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 $\text{Cov}(a, a), \text{Cov}(b, b), \text{Cov}(c, c), \text{Cov}(d, d)$ 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	$\text{Cov}(a, a)$	$\text{Cov}(b, a)$	$\text{Cov}(c, a)$	$\text{Cov}(d, a)$
b	$\text{Cov}(a, b)$	$\text{Cov}(b, b)$	$\text{Cov}(c, b)$	$\text{Cov}(d, b)$
c	$\text{Cov}(a, c)$	$\text{Cov}(b, c)$	$\text{Cov}(c, c)$	$\text{Cov}(d, c)$
d	$\text{Cov}(a, d)$	$\text{Cov}(b, d)$	$\text{Cov}(c, d)$	$\text{Cov}(d, d)$

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。

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

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。

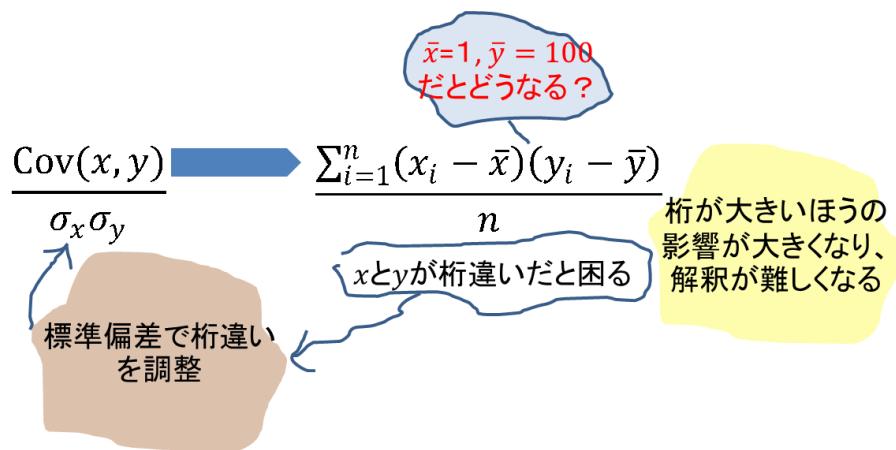


図 1.25 標準偏差と共分散

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)

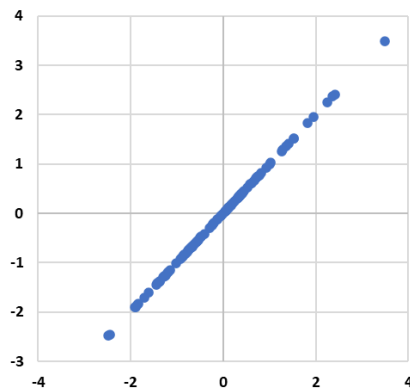


図 1.26 正の完全相関：相関=1

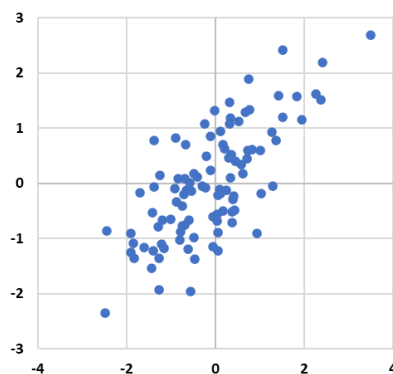


図 1.27 正の相関：相関=0.7

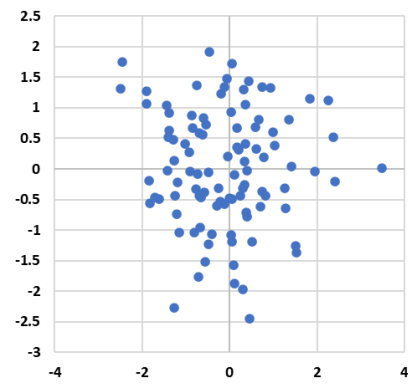


図 1.28 無相関：相関 = 0

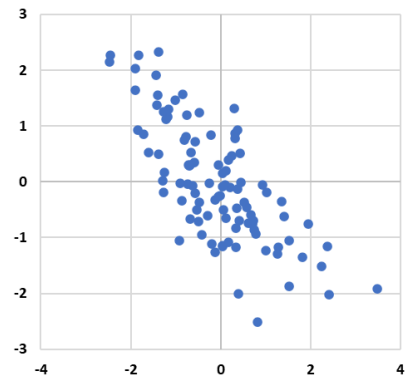


図 1.29 負の相関：相関 = -0.7

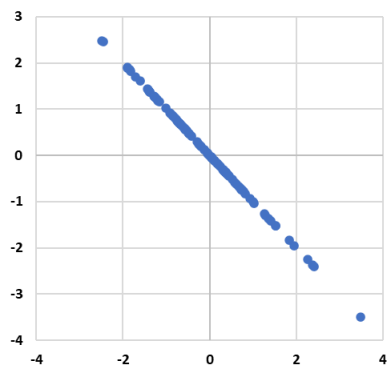


図 1.30 負の完全相関：相関 = -1

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

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

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。

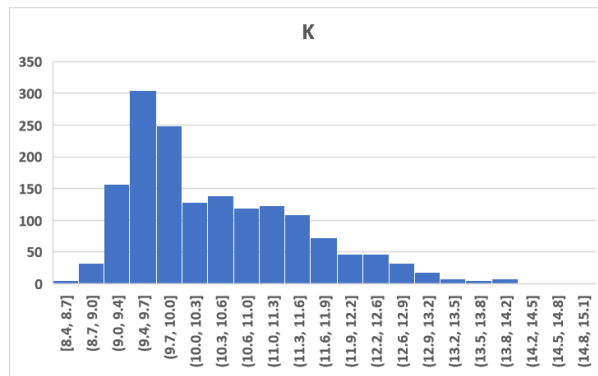


図 1.28 歪度=0.86 (練習問題 1.1)

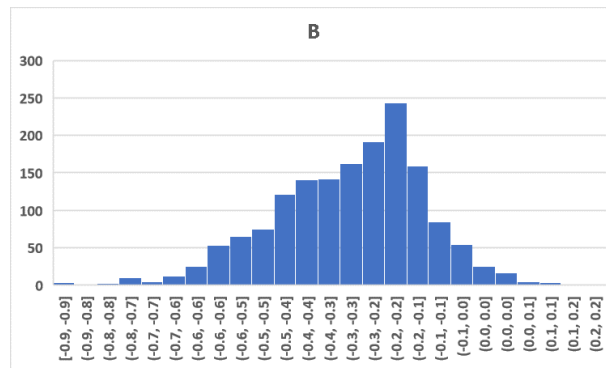


図 1.29 歪度=-0.43 (練習問題 4.3)

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。

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

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。

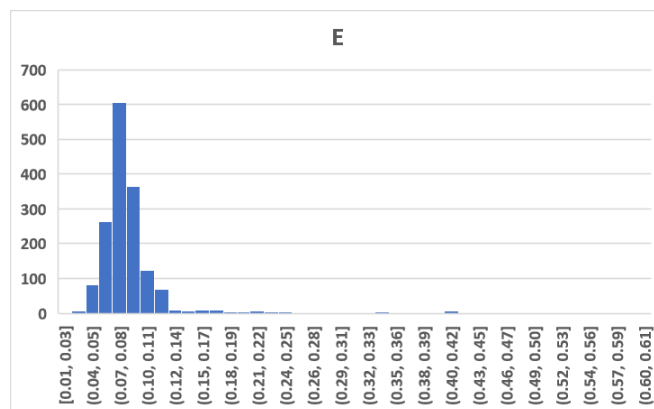


図 1.30 尖度=1.1 (練習問題 1.1)

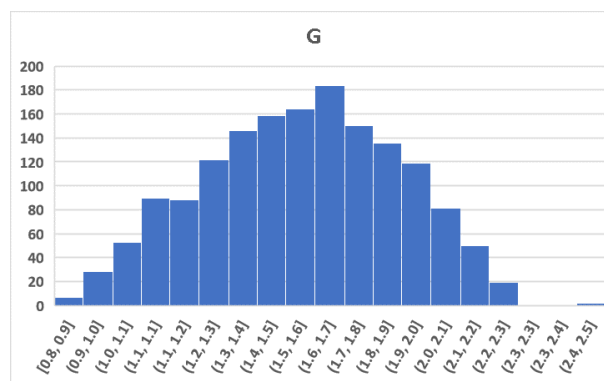


図 1.31 尖度-0.67 (練習問題 4.3 データの対数)

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？割合なののでしょうか？それとも何か別のものなののでしょうか？

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: 確率と確率分布

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。分布も同様に、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のでしょうか。本章では、統計学という確率と分布、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。

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。

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

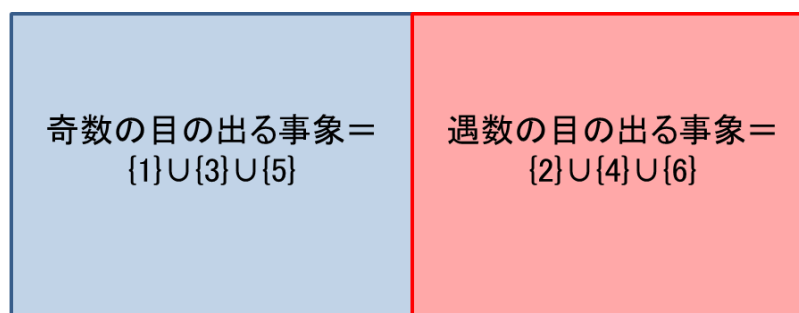
2つの事象 A と B の関係について考えてみましょう。典型的な例を2つ紹介します。

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。



$$P(\text{奇数})=P(1)+P(3)+P(5)=1/6+1/6+1/6=1/2 \quad P(\text{偶数})=P(2)+P(4)+P(6)=1/6+1/6+1/6=1/2$$

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とき $A \cap B$ の確率を求めましょう。サ

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。

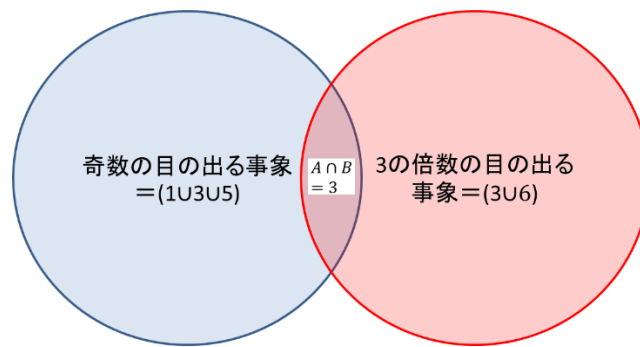


図 2.3 積事象

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(\complement) : 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 を奇数の目が出る事象、 B を3の倍数の目が出る事象とするととき $A \cup B$ 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 を奇数の目が出る事象 : $A = (1 \cup 3 \cup 5)$

B は3の倍数の目が出る事象 : $B = (3 \cup 6)$

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。

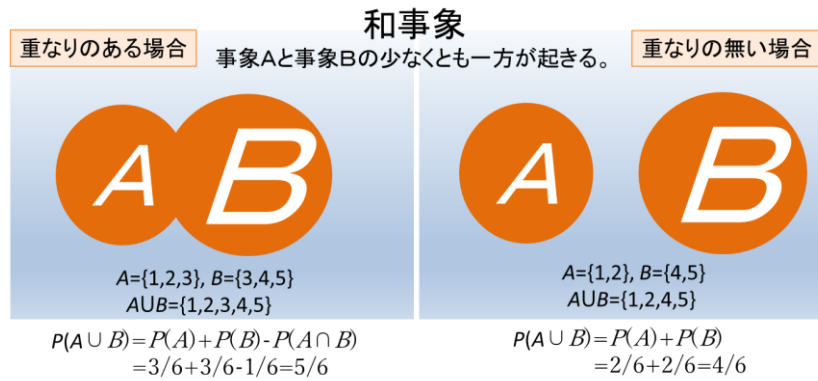


図 2.4 和事象：重なりのある場合とない場合

2.1.3. 条件付確率

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 は奇数の目の出る事象： $A = \{1 \cup 3 \cup 5\}$

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

B は偶数の目の出る事象： $B = \{2 \cup 4 \cup 6\}$

$$P(B) = P(2) + P(4) + P(6) = 1/2$$

C は3の倍数の目の出る事象： $C = \{3 \cup 6\}$

$$P(C) = P(3) + P(6) = 1/3$$

D は C の余事象： $D = \{1 \cup 2 \cup 4 \cup 5\}$

$$P(D) = P(1) + P(2) + P(4) + P(5) = 2/3$$

となります。

$P(B \cap D)$ は、事象 B と事象 D に含まれるサイコロの目の確率ですから、 $B \cap D = \{2, 4\}$ となります。よって、

$$P(B \cap D) = 2/6 = 1/3$$

です。これは、つぎのようにも考えられます。事象 B と D に含まれるサイコロの目は事象 D の一部が事象 B に含まれると考えることもできます。そこで、つぎのように確率を考えます。

- 事象 $B = \{2, 4, 6\}$ の確率： $P(B) = (\#\{2, 4, 6\})/(\#\{\Omega\}) = 3/6 = 1/2$
($\#$ は $\{\}$ 内の数値の数を表します.)
- 事象 $B = \{2, 4, 6\}$ の目がでたときに、その目が $D = \{1, 2, 4, 5\}$ にも含まれるのは、 $\{2, 4\}$ です。その確率は $P(D|B) = 2/3$ となります。これは、 $P(D|B) = (2/6)/(1/2) = 2/3$ とも計算できます。
- つまり、事象 B と事象 D に含まれるサイコロの目の確率は、確率 $P(B)$ の事象と確率 $P(D|B)$ の事象が同時に起こる確率なので、これらの確率の積になります。よって、

$$P(B \cap D) = P(B) P(D|B)$$

となります。 $P(B \cap D) = (1/2) \cdot (2/3) = 1/3$ です。ここまでくると、事象の関係が明確になったのではないのでしょうか？実は、すでにお気づきになった方もいると思うのですが、この議論は逆でも成り立ちます。すなわち、

- 事象 $D = \{1, 2, 4, 5\}$ の確率 $P(D) = (4/6) = 2/3$ です。

- 事象 $D = \{1,2,4,5\}$ の目がでたときに、その目が $B = \{2,4,6\}$ にも含まれるのは $\{2,4\}$ でその確率は $P(B|D) = 1/2$ となります。これは、 $P(B|D) = 2/6 / (4/6) = 1/2$ と計算できます。
- つまり、事象 D と B に含まれるサイコロの目の確率は、確率 $P(D)$ の事象と確率 $P(B|D)$ の事象が同時に起こる確率なので、これらの確率の積になります。よって、

$$P(B \cap D) = P(D) P(B|D)$$

と等しくなります。 $P(B \cap D) = (2/3) \cdot (1/2) = (2/6) = 1/3$ です。どちらの方法でも結果は同じになりました。したがって、

$$P(B)P(D|B) = P(D) P(B|D)$$

つまり、 $P(B)P(D|B) = P(D)P(B|D) = P(B)P(D)$ となります。

標本空間： $\Omega = \{1,2,3,4,5,6\}$

事象の確率

事象 A : 奇数の目 = $\{1,3,5\}$
事象 B : 偶数の目 = $\{2,4,6\}$

$$\text{確率 } P(A) = \frac{\#\{1,3,5\}}{\#\{\Omega\}} = \frac{3}{6} = 1/2$$

$$\text{確率 } P(B) = \frac{\#\{2,4,6\}}{\#\{\Omega\}} = \frac{3}{6} = 1/2$$

事象 C : 3の倍数の目 = $\{3,6\}$
事象 D : C の余事象 = $\{1,2,4,5\}$

$$\text{確率 } P(C) = \frac{\#\{3,6\}}{\#\{\Omega\}} = \frac{2}{6} = 1/3$$

$$\text{確率 } P(D) = \frac{\#\{1,2,4,5\}}{\#\{\Omega\}} = \frac{4}{6} = 2/3$$

条件付確率 $P(D | B)$: 事象 B が起きたときにそれに D が含まれる確率

事象 B : 偶数の目 = $\{2,4,6\}$

事象 D : C の余事象 = $\{1, \mathbf{2}, 4, 5\}$

$$\{\mathbf{2}, 4\} \text{ は } B \text{ に含まれる、確率 } P(D | B) = \frac{\#\{2,4\}}{\#\{2,4,6\}} = \frac{2}{3}$$

$$\text{確率 } P(D | B) \text{ は } \frac{\#\{2,4\}}{\#\{2,4,6\}} = \frac{\#\{2,4\} / \#\{\Omega\}}{\#\{2,4,6\} / \#\{\Omega\}} = \frac{P(A \cap B)}{P(B)} \text{ と書ける、}$$

図 2.5 にあるように、クロス表を用いると見通しが良くなります。

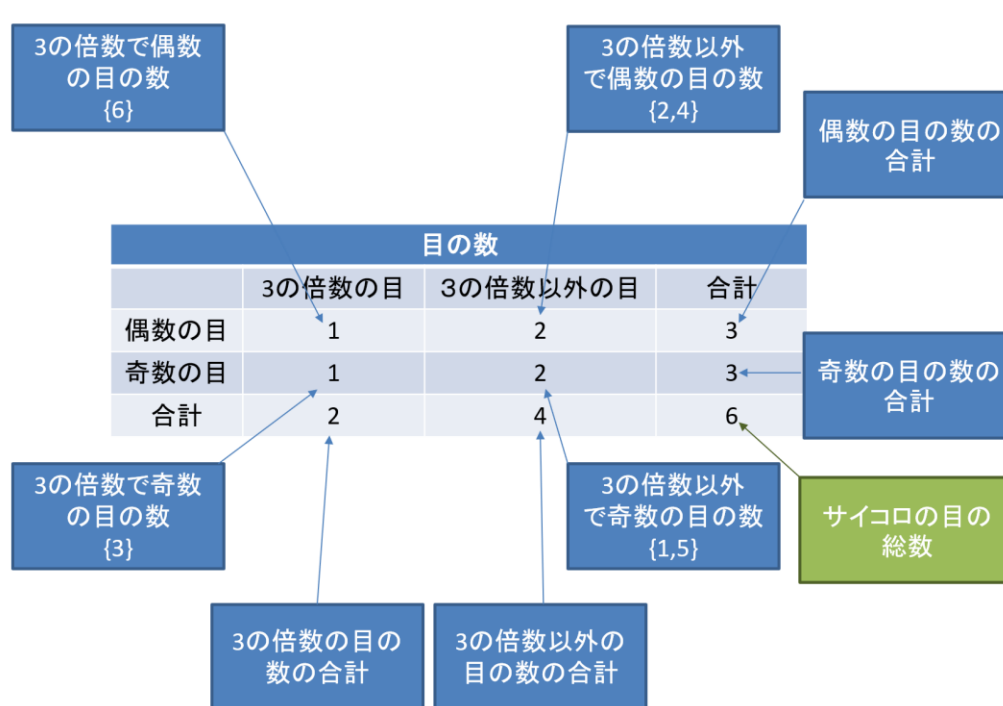


図 2.5 クロス表

確率

- 確率はゼロから1までの値をとります。
- すべての事象の確率の和は1になります。
- 事象が互いに排反なとき、その和集合の確率はおのこの事象の確率の和になります。

図 2.6 確率とは？

2.1.4. 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 を奇数の目が出る事象、 B を3の倍数の目が出る事象とするととき $A \cap B$ 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 2.5: 2つのサイコロを同時に振った時に両方とも偶数の目が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つのサイコロの目が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 2.7 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が出る事象をA、奇数の目が出る事象を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 2.8 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。

2.1.5. 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を奇数の目が出る事象、Bを3の倍数の目が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。奇数の目が出る事象は、3の倍数の目が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					
青のサイコロの目	積	1	2	3	4	5	6
	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 2.9 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。

2.2. 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 は確率変数です。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。


モデル(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)
 - $\Omega = \{1, 2, 3, 4, 5, 6\}$
- 確率
 - $\{P(A) = 3/6 = 1/2\}$

図 2.10 統計モデル：サイコロの例

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が出る確率は1/120となります。面の数を無限大に増やすとそれぞれの面が出る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。

今後、**確率変数に大文字(X, Y など)を使い、実現値に小文字(x, y など)を用います。**

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 が同時に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 $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ が成り立ちます。 X と Y が独立でなければ $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$ 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が出れば \emptyset 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, \emptyset 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 $1/6 \cdot 1/6 = 1/36$ 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, \emptyset)
 (A, A), (A, A), (A, $A \cap B$), (A, B), (A, B), (A, \emptyset)
 ($A \cap B$, A), ($A \cap B$, A), ($A \cap B$, $A \cap B$), ($A \cap B$, B), ($A \cap B$, B), ($A \cap B$, \emptyset)
 (B, A), (B, A), (B, $A \cap B$), (B, B), (B, B), (B, \emptyset)
 (B, A), (B, A), (B, $A \cap B$), (B, B), (B, B), (B, \emptyset)
 (\emptyset , A), (\emptyset , A), (\emptyset , $A \cap B$), (\emptyset , B), (\emptyset , B), (\emptyset , \emptyset)

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. 離散型確率分布

確率変数のとりえる値とそれらの確率との対応を示したものが確率分布です。このような分布は、実際には無数にあります。しかし、それらをいくつかの形に分類すると考えやすくなります。

離散型確率変数の作る分布を離散型確率分布といいます。

2.3.1. 離散一様分布

確率変数が離散値 $X = 1, 2, 3, \dots, N$ 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

$$P(X = x) = \frac{1}{N}, x = 1, 2, 3, \dots, N$$

となります。サイコロの目では $x = 1, 2, 3, 4, 5, 6$ ですから確率は $1/6 = 0.167$ となります。

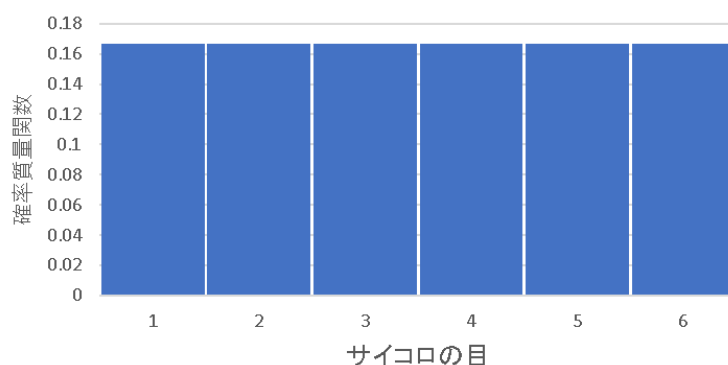


図 2.11 離散一様分布 (例題 3.5)

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。

2.3.2. 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。

[1,0]のベルヌーイ分布の確率分布は

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

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

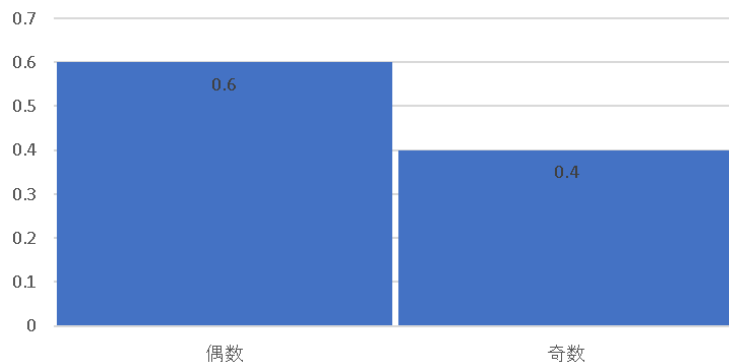


図 2.12 ベルヌーイ分布(練習問題 2.8)

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。

2.3.3. 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。

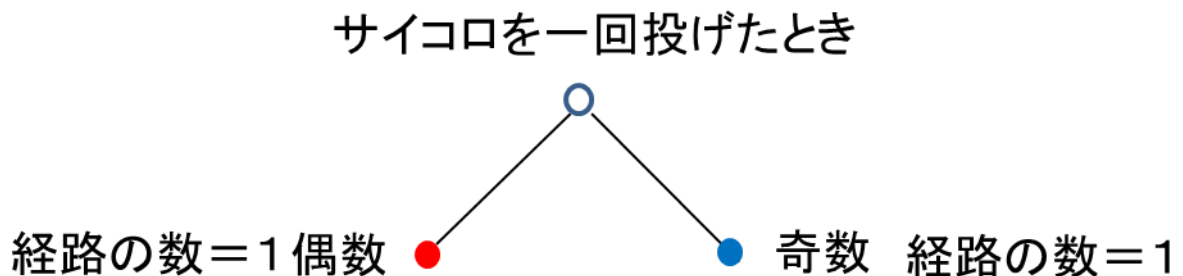


図 2.13 2項分布

それぞれ1です。○と赤●と、○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

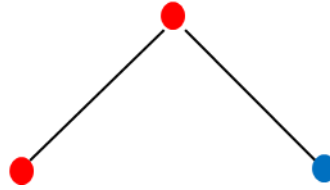


図 2.14 2 項分布

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

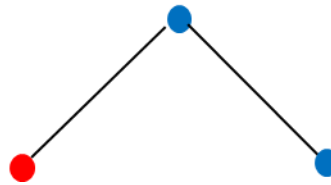


図 2.15 2 項分布

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

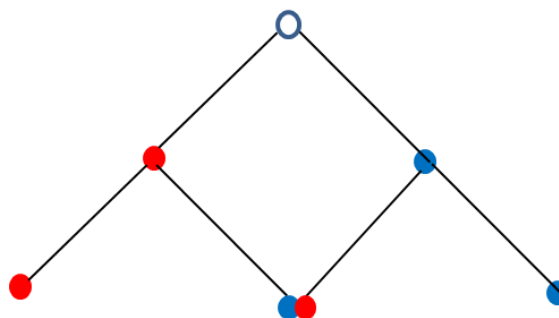


図 2.16 2 項分布

- 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、 $\bigcirc \rightarrow \bullet \rightarrow \bullet$ と $\bigcirc \rightarrow \bullet \rightarrow \bullet$ は出る順番は違いますが、赤丸と青丸の数は同じですので、同じと考えると経路の数は2つです。

それぞれに到達する経路の数を数えます。

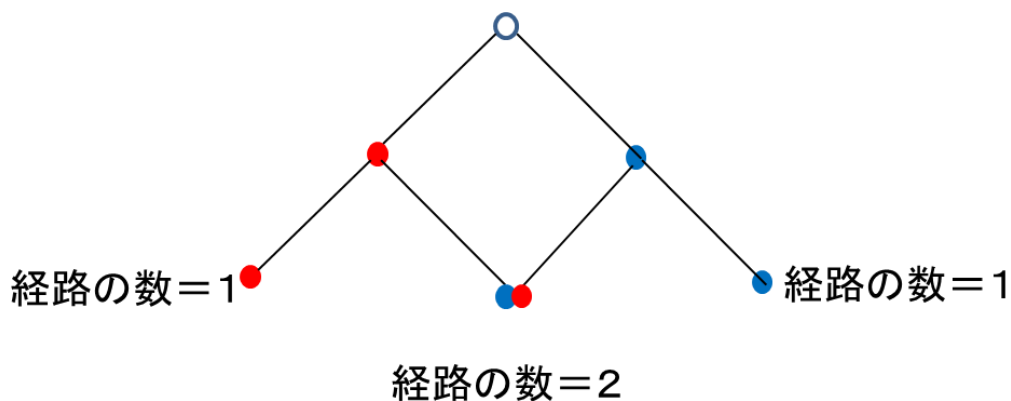


図 2.17 2 項分布

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。

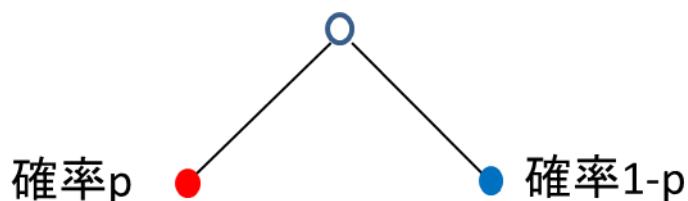


図 2.18 2 項分布：確率

サイコロを二回投げたときの結果はつぎのようになります。

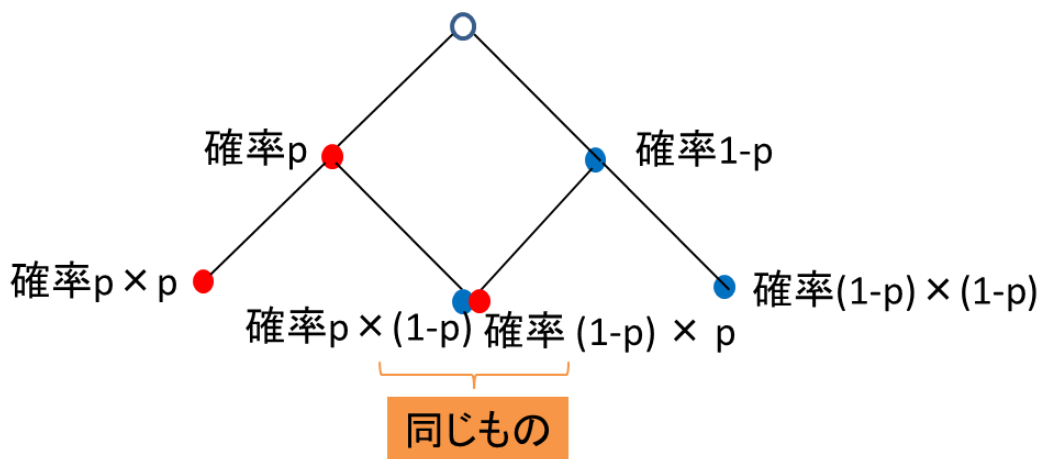


図 2.19 2 項分布：確率

サイコロを一回投げると

偶数の出る確率は

経路の数 \times 確率 p

奇数の出る確率は

経路の数 \times 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 $p \times p$

偶数→奇数または奇数→偶数と出る確率は

経路の数×確率 $p \times (1 - p)$

奇数→奇数と出る確率は

経路の数× $(1 - p) \times (1 - p)$

となります。 $p = 0.5$ とすると

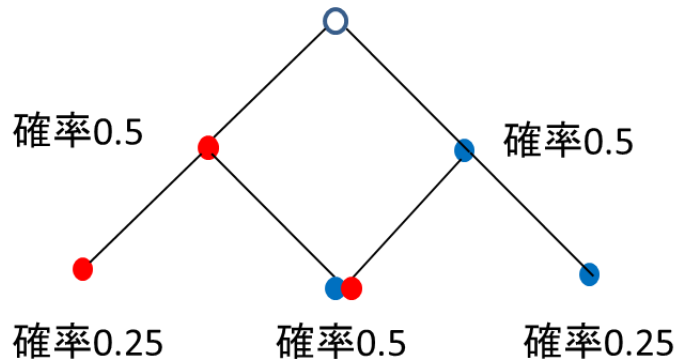


図 2.20 2 項分布：確率

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

$$P(X = x) = {}_n C_x p^x (1 - p)^{(n-x)} = \frac{n!}{x! (n - x)!} p^x (1 - p)^{(n-x)}$$

となります。ここで、 ${}_n C_x$ は n 個から x 個を選ぶ組み合わせの数です。2 項係数を表しています。 p は成功確率です。これを反復試行の確率ともいいます。

二項分布は統計学でも非常に重要な分布の 1 つです。この係数を n 枚の札の並べ方として考えてみましょう。最初の札は n 枚の札の中から一枚を選ぶので、その選び方は n 通りあります。つぎに 2 番目の札は一枚をすでに使ってしまったので $n - 1$ 枚の札の中から一枚を選ぶので、その選び方は $n - 1$ 通りあります。3 番目の札も同様に考えるとその選び方は $n - 2$ 通りあります。このように続けていくと最後には一枚の札が残りその選び方は 1 通りになります。つまり札の並べ方は $n \cdot (n - 1) \cdot \dots \cdot 2 \cdot 1$ 通りとなります。これを $n!$ で表します。札は赤と青に色付けされていて、その並べ方を数えるとすると、この数え方では赤と青の札の並べ方を重複して数えてしまっています。そこで赤の札を k 枚とすると青の札は $n - k$ 枚となります。赤の k 枚分の札と青の $n - k$ 枚分の札は重複して数えてしまっています。その場合の数はそれぞれ $k!$ 通りと $(n - k)!$ 通りです。これらを調整すると赤と青の札の並べ方は

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

通りとなります。

n 回の試行($n \geq 0$)の二項分布では平均 $E(X)$ は np 、分散 $\text{var}(X)$ は $np(1-p)$ 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。

$$x=0, {}_5C_0 = 5!/0!(5-0)! = 1 \times 2 \times 3 \times 4 \times 5 / (0!)(1 \times 2 \times 3 \times 4 \times 5) = 1$$

$$x=1, {}_5C_1 = 5!/1!(5-1)! = 1 \times 2 \times 3 \times 4 \times 5 / (1)(1 \times 2 \times 3 \times 4) = 5$$

$$x=2, {}_5C_2 = 5!/2!(5-2)! = 1 \times 2 \times 3 \times 4 \times 5 / (1 \times 2)(1 \times 2 \times 3) = 10$$

$$x=3, {}_5C_3 = 5!/3!(5-3)! = 1 \times 2 \times 3 \times 4 \times 5 / (1 \times 2 \times 3)(1 \times 2) = 10$$

$$x=4, {}_5C_4 = 5!/4!(5-4)! = 1 \times 2 \times 3 \times 4 \times 5 / (1 \times 2 \times 3 \times 4)1 = 5$$

$$x=5, {}_5C_5 = 5!/5!(5-5)! = 1 \times 2 \times 3 \times 4 \times 5 / (1 \times 2 \times 3 \times 4 \times 5)0! = 1$$

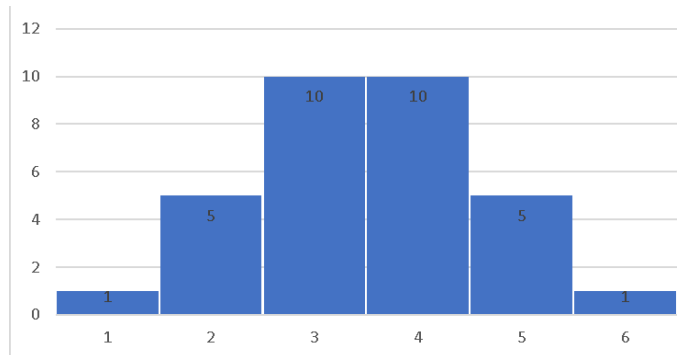


図 2.21 2項分布： $p=0.5, n=5$

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。

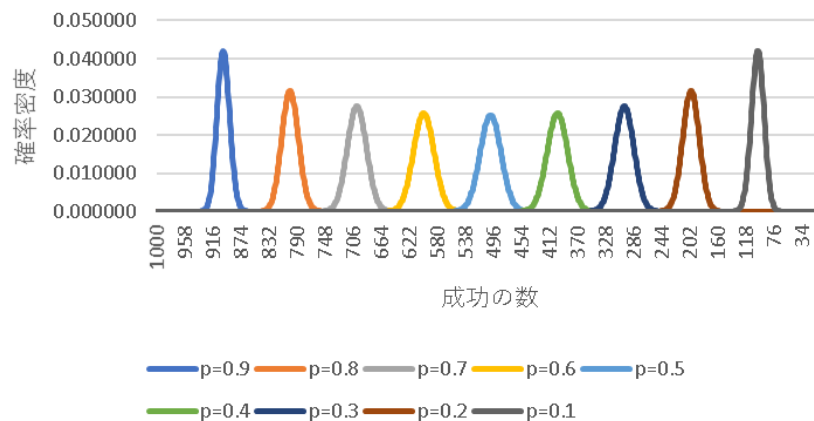


図 2.22 2項分布： $p=0.1 \sim 0.9$, 試行回数=1000

$p=0.9$ のときと $p=0.1$ のときの分布の幅が狭く、 $p=0.5$ のときの分布の幅が一番広がっています。これは分散が $np(1-p)$ であることから明確です。

2.4. 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。

2.4.1. 連続一様分布

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、連続一様分布の確率密度関数は

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{others} \end{cases}$$

となります。図 2.23 は連続一様分布の様子です。

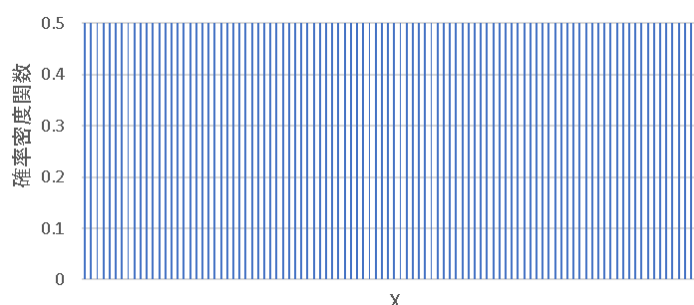


図 2.23 一様分布(練習問題 2.11)

多面体のサイコロで説明したように、面の数を無限にすると、出る面の確率はゼロになってしまいます。したがって、連続型の確率分布では離散型のように、 $f(X=x)$ のような書き方をしません。図 2.23 の確率変数の最大値は b 、最小値は a で、確率変数の生起する確率は一定ですから、確率の定義から青い部分の面積を 1 とすると、 $f(x) \cdot (b-a) = 1$ となり、 $f(x) = 1/(b-a)$ が得られます。これを x の確率密度関数とします。

2.4.2. 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となる確率分布が正規分布です。正規分布では、分散は山のすそ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。

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

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。

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

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

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。

標準正規分布は図 2.24 を参照してください。

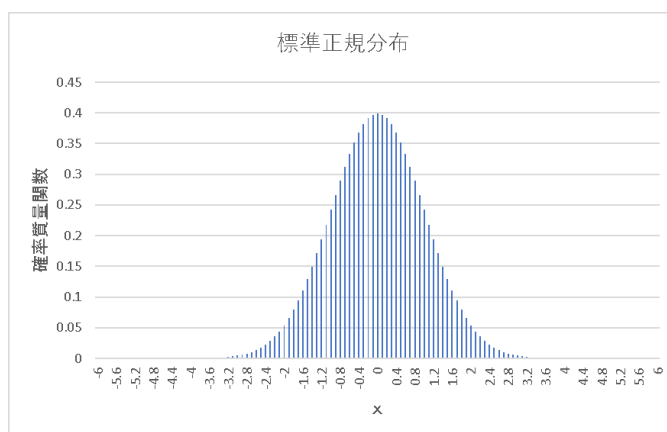


図 2.24 標準正規分布

2 項分布が正規分布で近似できる自由度はどの程度かは図 2.25 を参照してください。

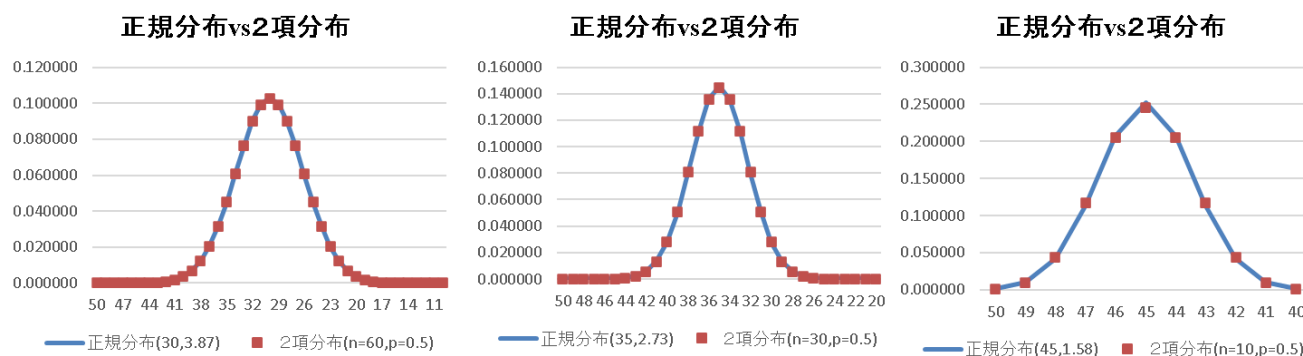


図 2.25 標準正規分布と二項分布

二項分布の平均は np 、分散は $np(1-p)$ なので n の大きさによって平均、分散が変化します。それに適合するように正規分布の平均と分散を調整しています。

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。

分散を固定した正規分布は図 2.26 を、平均を固定した正規分布は図 2.27 を参照してください。

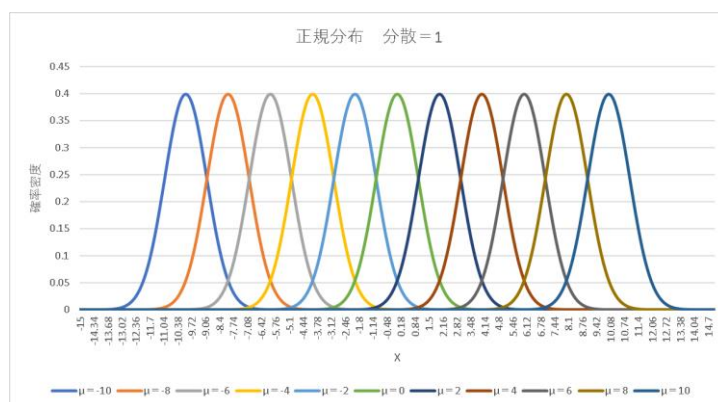


図 2.26 正規分布：平均を変化

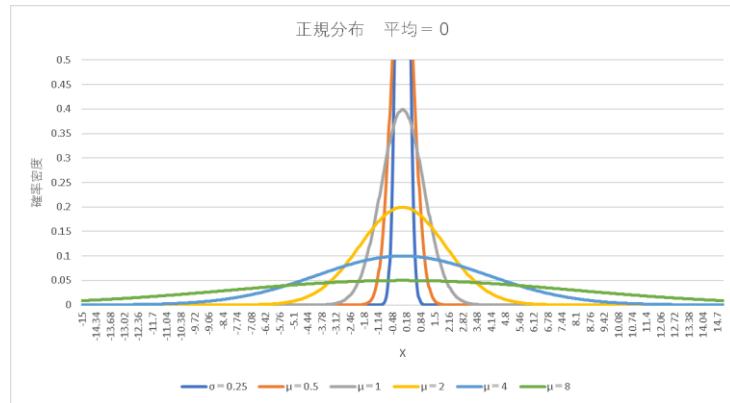


図 2.27 正規分布：分散を変化

2.5. 期待値

離散値 x_1, x_2, x_3, \dots の集合から得られる確率変数 X の期待値は、

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

で表されます。 j は根元事象の番号です。 x_j は根元事象の値で、 $f(x_j)$ 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

$$\bar{x}_l = \sum_{j=1}^J x_j \frac{N_j}{N}$$

で、期待値は

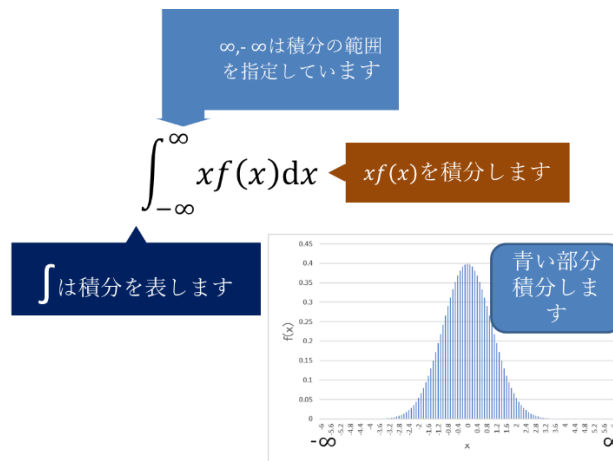
$$E(X) = \sum_{j=1}^J x_j p_j$$

です。

連続型の場合は、

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

として定義されます。 f は確率密度関数です。



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。また、期待値は推測統計で重要な役割を担います。宝くじを買うときも、株式に投資するときも、売り上げを予測するときも期待値を計算しているのです。

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 I 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: トランプの標本空間はなんでしょうか？

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。

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。

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。

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。

練習問題 2.9: ベルヌーイ分布の例をあげてみましょう。

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？

第3章 母集団と標本

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。これが推測統計の第一歩です。

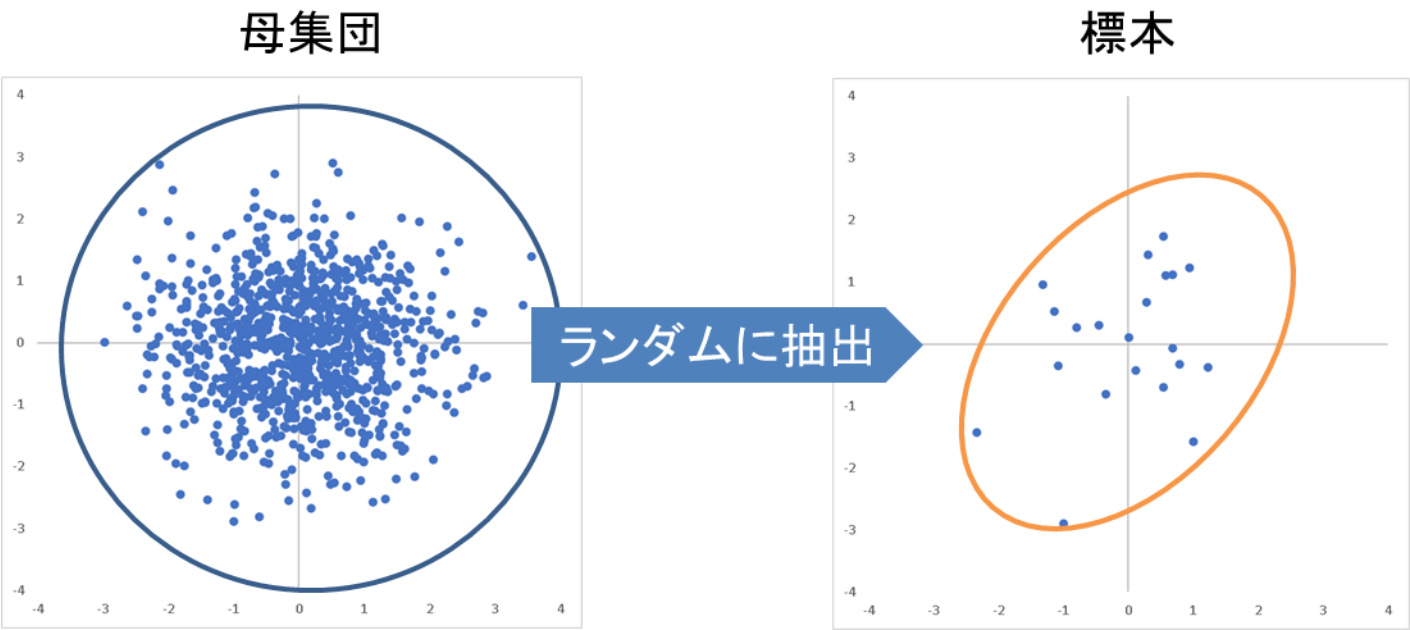


図 3.1 母集団と標本

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(模型)を前提としている場合があります。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。

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	検査対象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

表 3-1 母集団と標本

前者は選挙の当選予測などに相当します。後者は株価の予測などです。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。

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、標本を分析することで、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。(ここで母集団として乱数 1000 個は小さすぎます。)

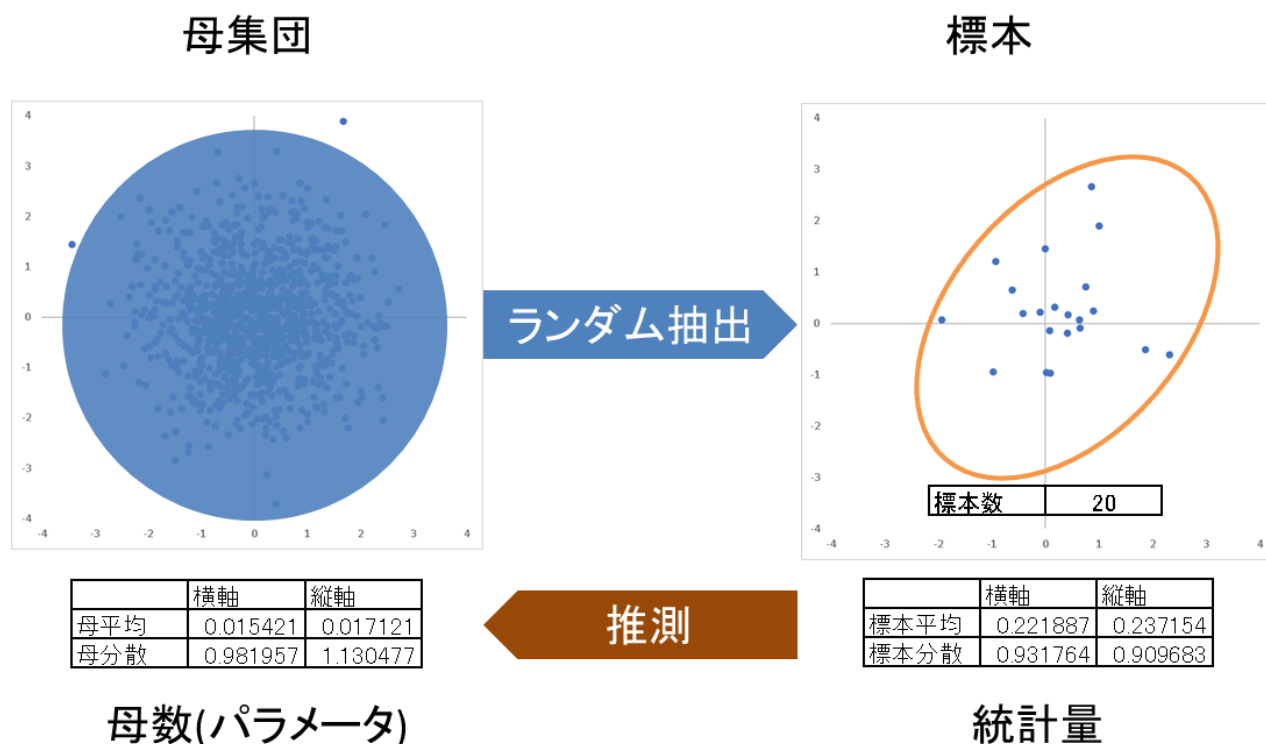


図 3.2 母集団と標本：母数と統計量の比較

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。例題 3.2 のエクセルシートを開いたら必ず F9 ボタンを押し、乱数を再度発生させ、違いを確認してください。

3.2. 適切なデータ収集

適切にデータを収集するためには、データの取得方法に注意を払う必要があります。物理実験や化学実験のように、実験室で環境を制御しながらデータ収集を行える場合と、観察研究のように、環境に介入することなく、自然の状態を観測して、必要なデータを集める場合があります。実験研究の場合には、実験単位で課される実験条件の処理に注目したフィッシャーの 3 原則に則ってデータを収集します。

- a) 局所管理：処理が均一な幾つかのブロックに分けて実験を行います。異なるブロックでは処理の違いを大きくします。

- b) 無作為化：処理以外の条件もできるだけ均一にする必要がありますが、均一にできない条件については偏りを排除するために、無作為に割り付けます。
- c) 繰り返し：処理を全く同じにしても、さまざまな理由によりデータには、ばらつきが生じます。このばらつきの大きさを見積もるために、実験を何度も繰り返す必要があります。

観察研究では特に無作為化が難しく、処理も被験者自らが選択しているために、処理の選択に偏りを生じる可能性があります。

3.3. 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。

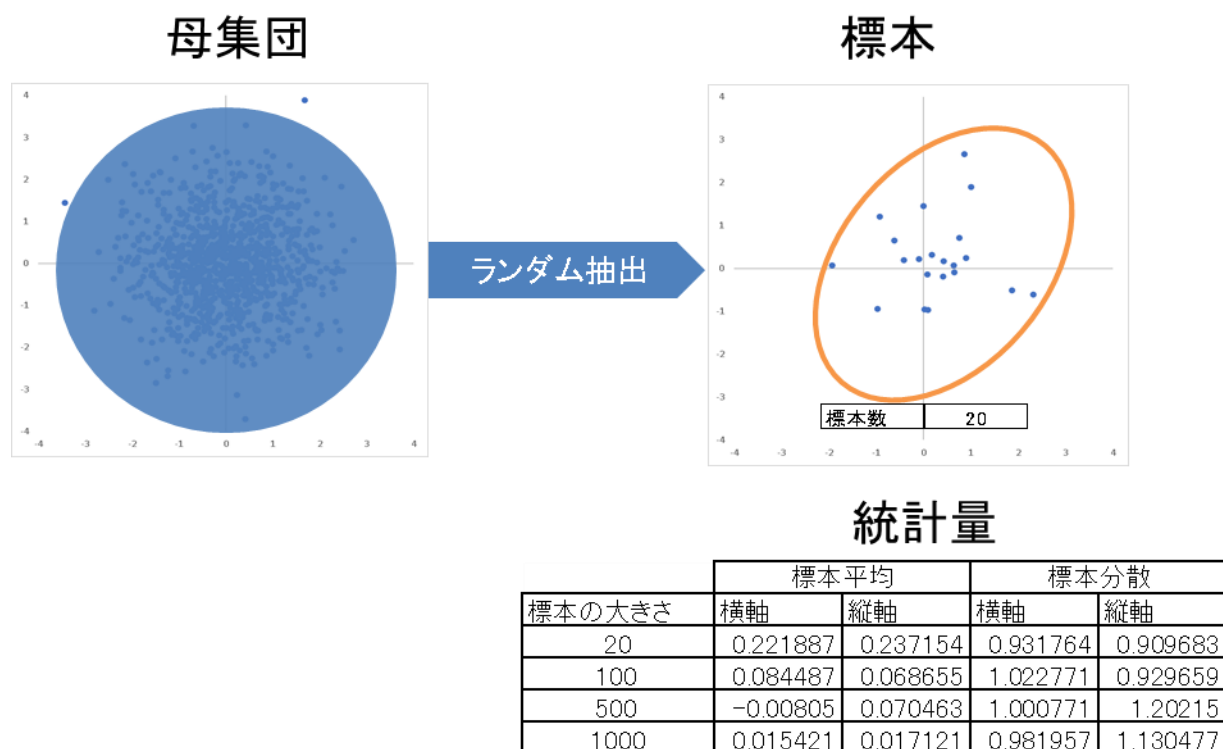


図 3.3 母集団と標本：サイズの違い

真の平均と標本の平均の誤差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。例題 3.3 のエクセルシートを開いたら必ず F9 ボタンを押し、乱数を再度発生させ、違いを確認してください。

例題 3.4：例題 3.2 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。

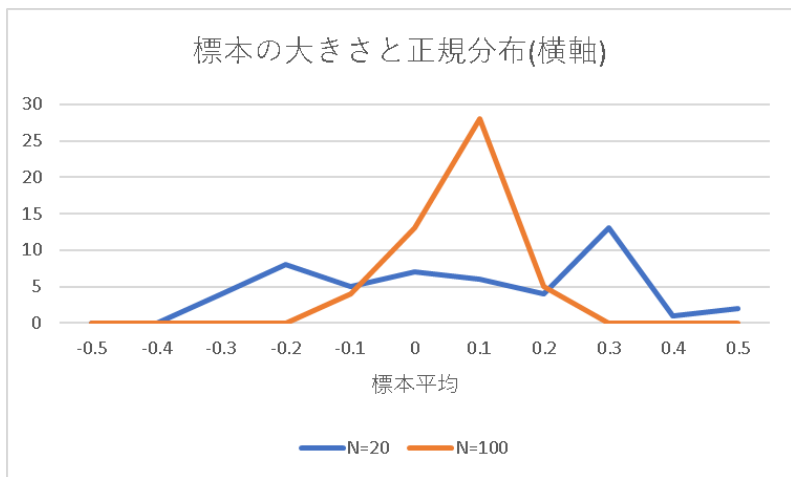


図 3.4 母集団と標本：標本の大きさと頻度図(横)

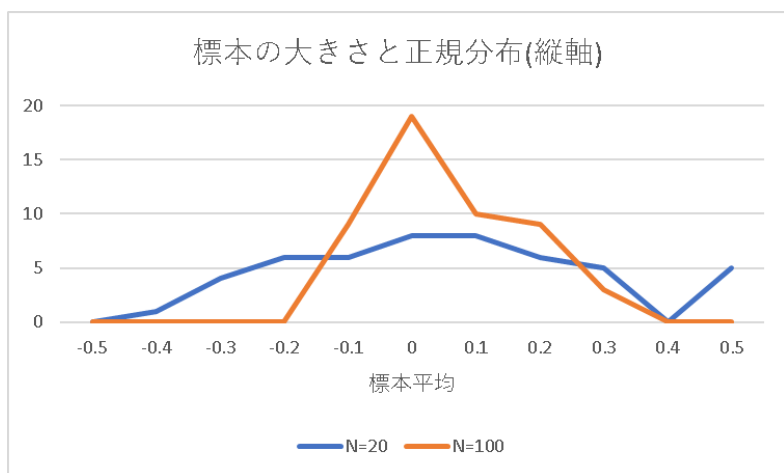


図 3.5 母集団と標本：標本の大きさと頻度図(縦)

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。期待値はしたがって、理論的な確率分布の平均のことです。この問題では、母集団のデータ数を $n=1000$ としているので、小さすぎます。そこから大きさ 20 の標本をとるのですが、重複がないようにする必要があります。それはデータを独立な確率変数にしたいからです。大きさ 100 の標本では、重なりを許していますが、理想的には重複がないようにしなければなりません。重複を許さないと標本の大きさが小さくなり、グラフをうまく描けないからです。この点に関しては、例題 2.7 の確率変数が独立にならない場合の問題を考えてみてください。また、例題 3.4 のエクセルシートを開いたら必ず F9 ボタンを押し、乱数を再度発生させ、違いを確認してください。

3.4. 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。推定量には母数の記号 θ に「ハット」を付けて $\hat{\theta}$ として示します。そこで推定量の性質について明らかにします。

3.4.1. 一貫性

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、それを一貫性とよび、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。

3.4.2. 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, \dots, x_n のとき

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

となります。 \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。

例題 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。

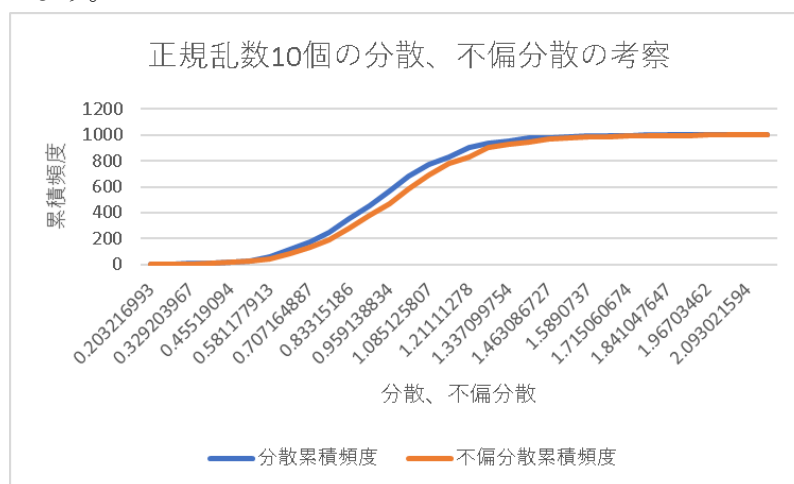


図 3.6 不偏分散

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。

3.5. 標本分布

例題 3.2~3.4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。

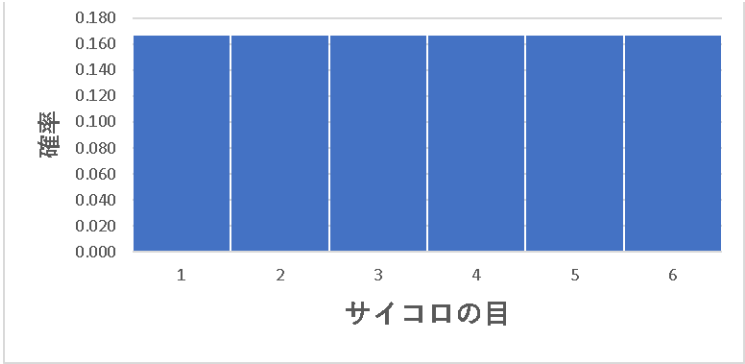


図 3.7 一回の試行と頻度図

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表 3-2 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表 3-2 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

表 3-2 2 回の試行と結果

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 $\Omega = \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0\}$ 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。

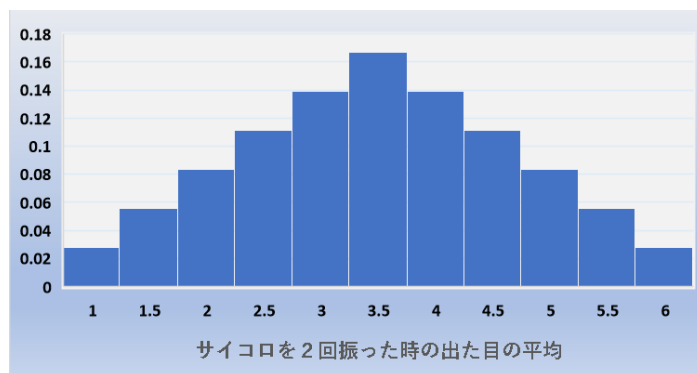


図 3.8 2回の試行と頻度図

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。

3.5.1. カイ二乗分布

確率変数 $X_1, X_2, X_3, \dots, X_n$ 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

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

がしたがう分布は自由度 n のカイ二乗分布といいます。

$$W \sim \chi^2$$

カイ二乗分布の平均は n 、分散は $2n$ になります。カイ二乗分布は n が大きくなると正規分布に近づきます。

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、見た目でも正規分布といえるような標本の自由度を探しましょう。

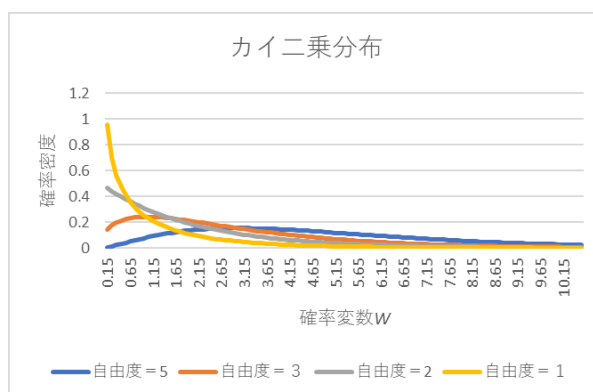


図 3.9 χ^2 二乗分布と自由度

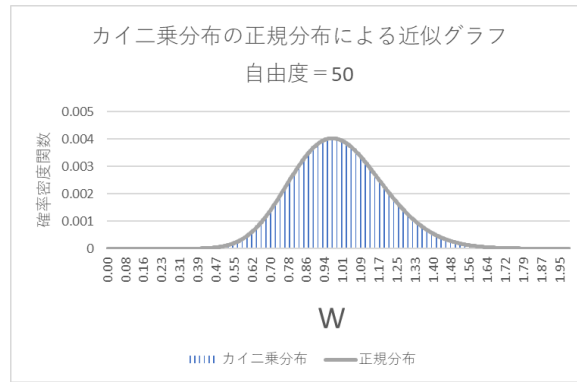


図 3.10 χ^2 二乗分布の正規分布による近似：自由度=50

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, \dots, 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。

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

左辺を、不偏分散を含む形に変形します。

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

図 3.11 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。不偏分散の分布がカイ二乗分布に近いことがみて取れます。

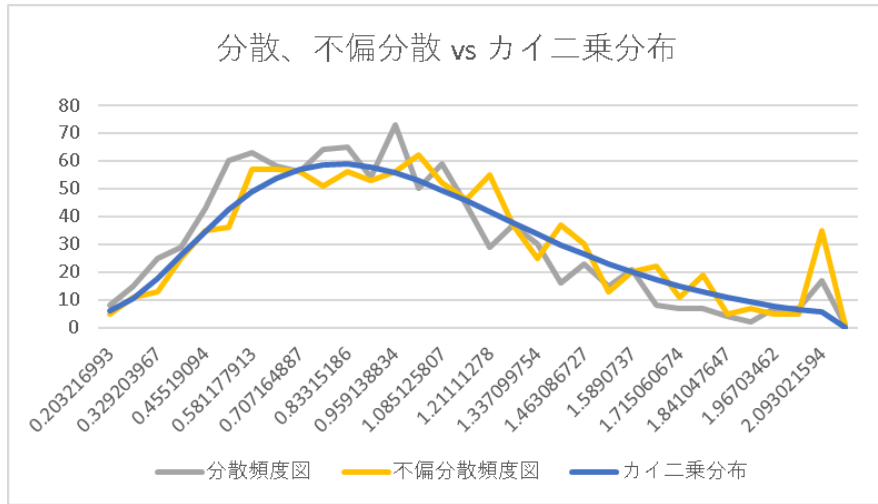


図 3.11 χ 二乗分布と不偏分散 例題 3.8

カイ二乗分布関数 $\text{CHISQ.DIST}(x, \text{自由度}, \text{関数形式})$ の x は $s^2 \cdot (10 - 1)$ となります。自由度は $10 - 1 = 9$ です。

3.5.2. 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 $X_1, X_2, X_3, \dots, X_n$ 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

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

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。

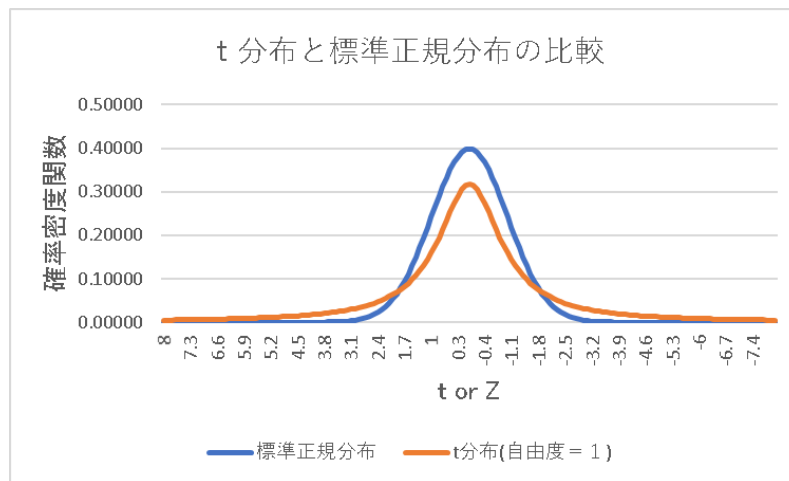


図 3.12 t 分布と標準正規分布

T 分布の期待値はゼロ、分散は $n - 1 > 2$ では $(n - 1)/(n - 2)$ となり、 $1 < n - 1 \leq 2$ では ∞ になります。したがって、 n が大きくなれば、 t 分布の分散は1に近づきます。それは橙色の曲線が正規分布に重なることを意味します。

3.5.3. F 分布

カイ二乗分布にしたがう自由度が d_1 と d_2 の2つの確率変数 Z_1 と Z_2 の比は F 分布にしたがいます。

$$F = \frac{Z_1/d_1}{Z_2/d_2}$$

ある模型(モデル)について複数の平均値が等しいかどうかを判定するときに F 分布は重要な役割をにないます。分散分析、線形回帰分析に使われます。

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)

図 3.13 確率分布の分類

- 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 分布が正規分布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

本章では、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、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。標本が母集団の一部である限り、推定値がどの程度の範囲にあるかを考える必要があります。母集団からデータ x_1, x_2, \dots, x_n が何度も得られるとすると、母数の推定値も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。得られたデータ x から母数を考えるとき、**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。**

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。

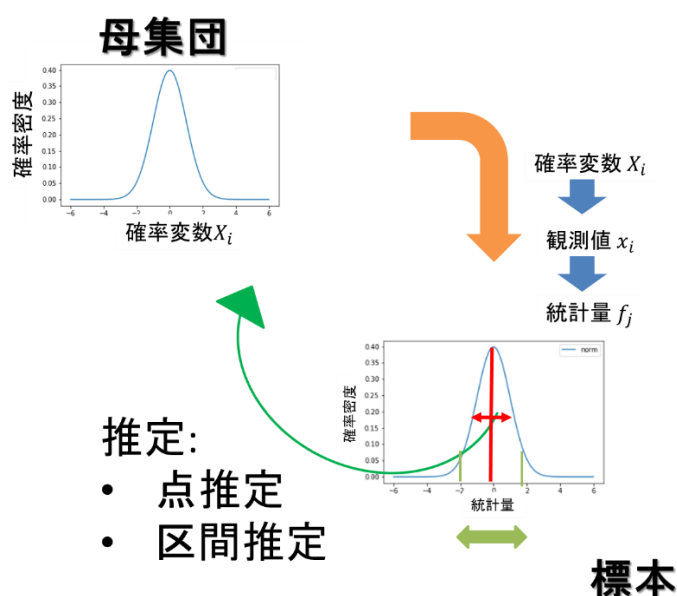


図 4.1 推定

4.1. 点推定

標本 x_1, x_2, \dots, 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、分散などの推定に用いられます。

母数 θ に対してその推定量は $\hat{\theta}$ として表します。

4.1.1. 標準誤差

母集団から得られた標本の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。単に標準誤差といったときには、平均についてのばらつきを表し、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。統計量により標準誤差の計算方法は異なります。

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

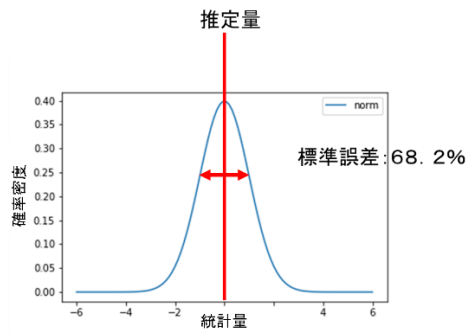


図 4.2 点推定

4.1.2. 一致推定量と不偏推定量

標本平均は一致推定量であり、不偏推定量でもあります。標本分散は一致推定量ではありますが、不偏推定量ではありません。不偏分散は一致性と不偏性を持ちます。

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つの値を決め、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。

4.2.1. 信頼区間

標本を X 、区間の上限を $U(X)$ 、下限を $L(X)$ 、そして、母数を μ とすると、

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

と表現します。 μ は $U(X)$ と $L(X)$ の間に含まれていることを意味します。(標本は観測値だけではなく、確率変数の場合もあります。) U, L は特定の関数です。

4.2.2. 信頼係数

この信頼区間の中に母数が含まれている割合が信頼係数で $1 - \alpha$ で表します。したがって、

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

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。 α を有意水準といいます。

4.2.3. 母平均の区間推定

母平均の区間推定を行ってみましょう。母分散が既知か未知かにより計算方法が異なります。

a) 母分散(σ^2)が既知

標本 x_1, x_2, \dots, 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(σ^2)が既知なので標準正規分布を用います。

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

$N(0,1)$ 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

$1 - \alpha$ は信頼係数で母平均 μ が信頼区間に含まれる割合になります。

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

b) 母分散は未知

標本 x_1, x_2, \dots, 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(σ^2)は未知なので不偏分散(s^2)を使います。そして t 分布を用います。

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

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

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

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です。不偏分散は2、 t 統計量の臨界値は ± 2.13 です。よって下限は8.65、上限は11.34です。

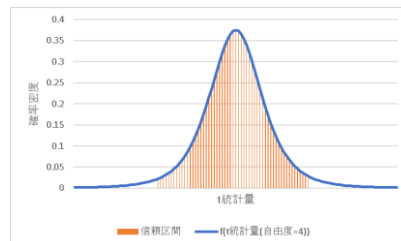


図 4.3 区間推定と分布

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。

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

から

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

が得られます。 $\alpha = 0.01$ とすると $t_{(0.5\alpha, n-1)} = 2.33$ です。結果は表 4-1 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

表 4-1 推定：上限と下限

4.2.4. 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

$$\sum_{i=1}^n (x_i - \bar{x})^2 \sim \sigma^2 \chi_{(n-1)}^2$$

の関係があります。 σ^2 は母分散、 $\chi_{(n-1)}^2$ は標本の大きさが $n-1$ のカイ二乗分布です。これを变形して、

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

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

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi_{(0.5\alpha, n-1)}^2} < \sigma^2 < \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi_{(1-0.5\alpha, n-1)}^2}$$

となります。また

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

と書くこともできます。標本の大きさが大きくなると信頼区間は狭くなります。

4.2.5. 信頼区間の意味

信頼区間95%の信頼区間とは、ある大きさの標本から信頼区間を推定したときに、その母数がこの信頼区間に含まれている割合が95%という意味であり、母数がこの区間に入る確率が95%という意味ではありません。

例題 4.3 標本の大きさが5の標準正規分布にしたがう確率変数を生成し、標本平均を計算し、95%の信頼区間をもとめ、それを1000回繰り返したときに、母数がこの区間に入る割合を計算してみましょう。

標準正規乱数を発生させるので、平均も分散も既知で、それぞれゼロと1です。したがって、スプレッドシートには `normsinv(確率)` を用いて、

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

を計算します。その結果は図 4.4 のとおりです。

	A	B	C	D	E	F	G	H	I	J	K
1			標本の大きさ=		5	平均	分散	z_0.95	最大値	最小値	0.106
2	0.097247	3.200368	1.187927	-0.73227	1.221934	0.995042	2.185444	1.959964	2.290827	-0.30074	0
3	-0.11558	1.974505	-1.3589	1.74651	0.113908	0.472088	1.926891	1.959964	1.688811	-0.74463	0
4	-2.20077	1.085602	0.868502	0.772527	-1.91333	-0.27749	2.662206	1.959964	1.707651	-1.15266	0
5	0.090692	-0.04376	0.036863	1.825016	0.232766	0.428315	0.619773	1.959964	1.118363	-0.26173	0
6	-1.14277	-0.00787	-0.69437	1.349172	2.334051	0.367642	2.095099	1.959964	1.636361	-0.90108	0
7	0.738407	0.808028	0.505071	2.288684	0.414812	0.951	0.585246	1.959964	1.62161	0.280301	1

図 4.4 標準正規乱数の信頼区間：母平均 = 0 が信頼区間に含まれる割合

平均も分散も既知で、それぞれゼロと 1 です。ファイル：例題 4.3 信頼区間の理解の sheet：N-Dist を開けてください。スプレッドシートには normsinv(確率)を用いて、 $z_{\alpha/2}$ をもとめています。 $z_{0.975} = 1.956$ を用いています。その結果の割合は k1 のセルにあり 0.106 です。

つぎに、sheet：T-Dist をあけてください。こちらでは、分散を未知として、 t 分布を用いています。

	A	B	C	D	E	F	G	H	I	J	K
1				n=	5	平均	分散	t_0.95	最大値	最小値	0.064
2	0.7533	-1.67232	-0.61499	0.994816	0.386271	-0.03058	1.219129	2.776445	1.401558	-1.34039	0
3	-1.13252	0.480196	0.756672	0.412497	-0.35687	0.031993	0.594536	2.776445	0.989393	-0.92541	0
4	-0.3182	0.169242	-0.5966	0.146198	-1.25449	-0.37077	0.3484	2.776445	1.103668	-0.36213	0
5	0.287845	0.518092	0.611778	1.68537	0.722445	0.765106	0.290253	2.776445	1.434054	0.096158	1
6	0.11702	1.046122	1.07475	0.02818	0.14707	0.24618	0.715486	2.776445	1.20646	0.8041	0

図 4.5 t 分布による信頼区間の作成：母平均=0 が信頼区間に含まれている割合

$t_{\text{inv}}(\text{確率}, \text{自由度})$ を用いています。 $t_{0.975} = 2.776$ となります。したがって、k1 のセルの母平均=0 が信頼区間に含まれる割合は 0.064 となっています。F9 を何度も押して確かめてください。信頼区間 95%の示す、母平均が信頼区間に含まれる割合が 95%(1-0.064=0.936)という数値に近くなっています。

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

4.2.6. 母比率の信頼区間

視聴率、支持率、財務比率など、推測統計では比率を扱うことがよくあります。ここでは母比率の信頼区間を求めます。

- 母比率を p とすると、標本の大きさが n の二項分布にしたがう確率変数 x の期待値と分散は

$$E[x] = np$$

$$V[x] = np(1 - p)$$

となります。

- p の推定量として、標本比率 $\hat{p} = x/n$ を用いると \hat{p} は p の不偏推定量です。
- \hat{p} の期待値と分散も不偏推定量となります。したがって、

$$E[\hat{p}] = p$$

$$V[\hat{p}] = \frac{p(1-p)}{n}$$

となります。

- n が大きいと中心極限定理により、 z は標準正規分布に近似的にしたがいます。

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

したがって、求める信頼区間は

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

となります。

練習問題 4.1: エクセル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。