

確率・統計学基礎

午前の部（エクセル編）

zoom ミーティング

2022年8月4日(木)

9:30-12:30

講師：森谷博之 Quasars22

■このセミナーは Zoom でご受講いただけます。お申し込み後に詳細をメールでお送りします。

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

本セミナーでは、Excel を用いて、確率変数と確率分布の関係、母集団と標本の違いについて学びます。Excel を用いて確率変数を発生させ、目で確率変数の性質を理解します。また、推定の性質について学びます。本セミナーは記述統計の知識を前提として、推測統計に必要な基礎知識を身につけていきます。より論理的にデータを処理する際に必要となる確率変数について学びます。Excel を用いて問題を表現したり、解いたりすることでデータの処理の仕方の基本を身につけます。

第1部：確率変数と確率分布 9:30~11:00

- ・ 確率変数とその性質：予測ができない？
- ・ 確率分布のタイプ：離散型確率分布、連続型確率分布
- ・ 期待値：平均と期待値と分散

第2部：母集団と標本 11:10~12:30

- ・ 母集団：2つのタイプの母集団
- ・ 大数の法則と中心極限定理
- ・ 推定の性質：一致性と不偏性
- ・ 標本分布：カイニ乗分布、t分布

■本セミナーに参加して修得できること

発生した問題についてデータをととして理解する基本的な態度、論理的な思考方法の基礎を身につけます。データの本質を理解する、将来を予測する、過去の現象を説明するなどの推測統計に必要な最低限の知識を身につけます。また、Excel を用いた演習を多く取り入れ、理解を深めます。

■受講対象者

「記述統計」を一通り学んだことのある方を対象としています。

学生時代に確率・統計を学んだが覚えていない、確率・統計学が日々の仕事、研究にどのように役立つかわからない、統計分析をしたいがどうすれば良いかわからない、部下の統計分析を理解したい、データ分析の本質を理解したい方など。記述統計については「確率・統計学入門」のセミナーを開催しています。

■使用ソフト：Excel

■PC には事前に Excel がインストールされている必要があります。

参考文献：

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

本セミナーは、早急にデータ分析の基礎を身につけたい人のために構成されています。そのために、確率変数の特性を徹底的に分析します。特に、独立な確率変数を重視します。それは多くの実務的に役に立つと考えられるモデルが、独立な確率変数を前提としているからです。この独立な確率変数の性質と観測値・実現値の性質を比べることで、データ分析を行う基礎を養います。

1. 確率：確率変数と確率分布

1.1. 確率入門

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。サイコロを振るとき、その結果は偶然に左右されます。つまり、試行の結果は、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つの面には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。根元事象の全体 $\Omega = \{1, 2, 3, 4, 5, 6\}$ を標本空間と呼びます。標本空間はすべての根元事象の集合です。つまりこれらはサイコロを振る前から確定しています。

サイコロが賭けに使われるときは、出た目の数よりは、目が奇数であるか偶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。これら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。

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。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 1.1 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。

1.1.1. 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になる確率は同じであるとは限りません。したがって、発生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。

1.1.2. 事象と確率

2つの事象 A と B の関係

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


1.2. 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 1.2 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。

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

1.2.1. 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 $P(A \cap B) = P(A)P(B)$ が成り立ちます。 \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 1.1 : 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。 $A \cap B$ 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。

1.3. 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。

1.3.1. 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。

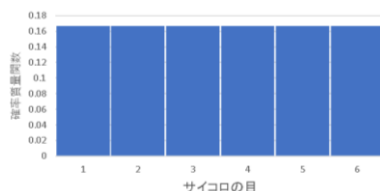


図 1.3 離散一様分布 (例題 3.5)

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。離散型確率分布の表記の仕方を見てみましょう。 $P(\cdot)$ は確率を表します。 $X = x$ は確率変数が離散値 x であることを示しています。このような関数を、確率質量関数といいます。

1.3.2. 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

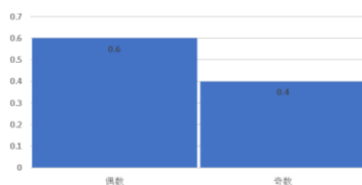


図 1.4 ベルヌーイ分布(練習問題 2.8)

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。

1.3.3. 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。奇数と偶数が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。

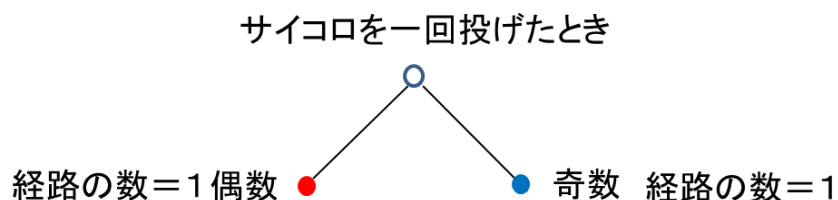


図 1.5 2 項分布

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、1回目の結果が偶数の場合を考えます。その結果は

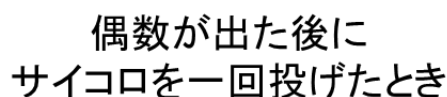


図 1.6 2 項分布

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1回目の結果が奇数の場合を考えます。

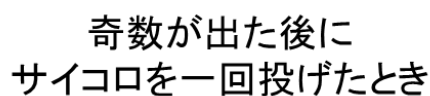


図 1.7 2 項分布

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

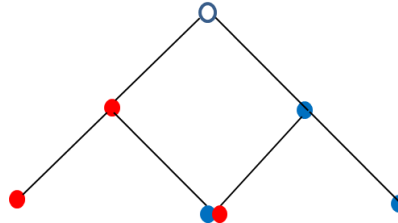


図 1.8 2 項分布

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、○ → ● → ●と○ → ● → ●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。

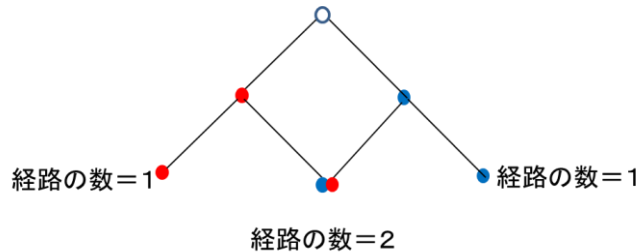


図 1.9 2 項分布

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1回投げたときの結果はつぎのようになります。

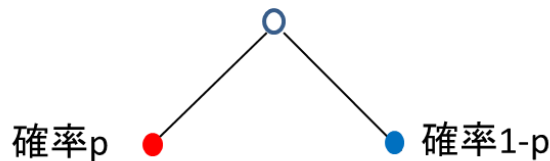


図 1.10 2 項分布

サイコロを2回投げたときの結果はつぎのようになります。

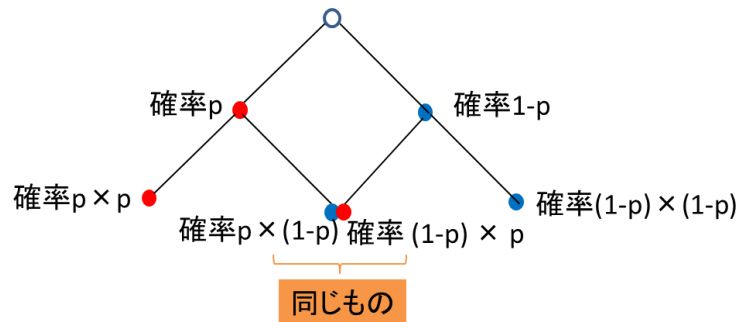


図 1.11 2 項分布：確率

サイコロを1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2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

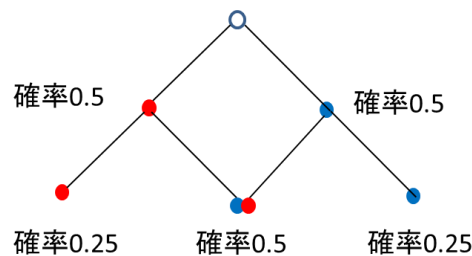


図 1.12 二項分布：確率

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

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

となります。ここで、 ${}_n C_x$ は n 個から k 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。 n 枚の札は、すべて色分けされていると考えます。つまり、1枚1枚の札は識別可能です。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いるのだから $n-1$ 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るその選び方は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なると、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 1.2: $n=5$ 、 $p=0.5$ の場合の分布を計算してみましょう。

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

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

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

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

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

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

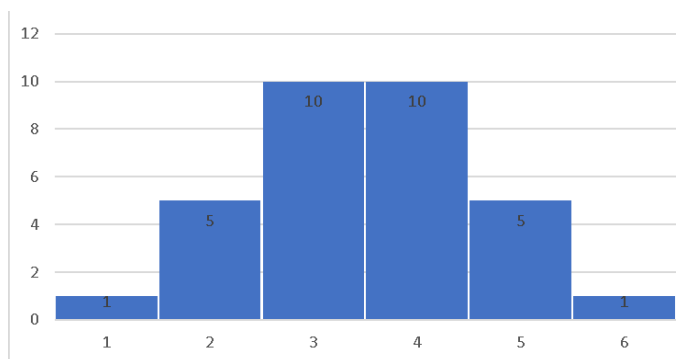


図 1.13 2 項分布: $p=0.5, n=5$

例題 1.3: 二項分布の試行回数を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。

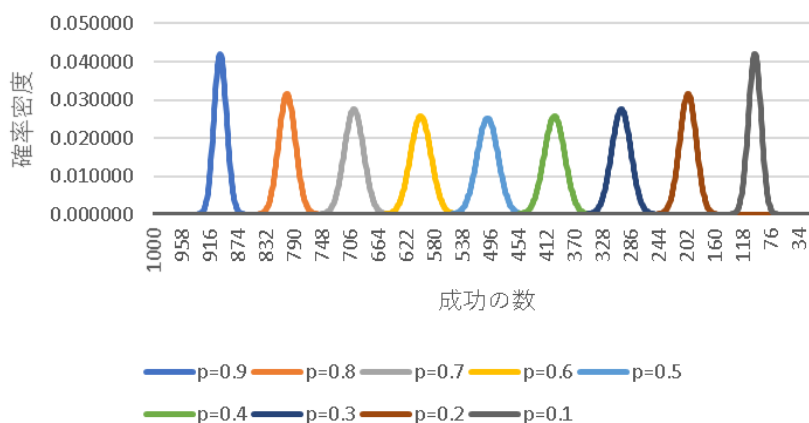


図 1.14 2 項分布: $p=0.1 \sim 0.9$, 試行回数=1000

$p=0.9$ 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。

1.4. 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。

1.4.1. 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。

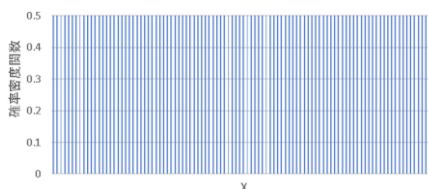


図 1.15 一様分布(練習問題 2.11)

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。

1.4.2. 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。

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

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。

$$g(z) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{z^2}{2} \right)$$
$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

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 1.4：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。

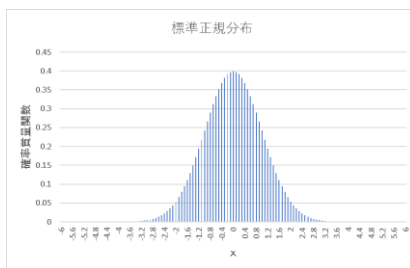


図 1.16 標準正規分布

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 1.17 を参照してください。

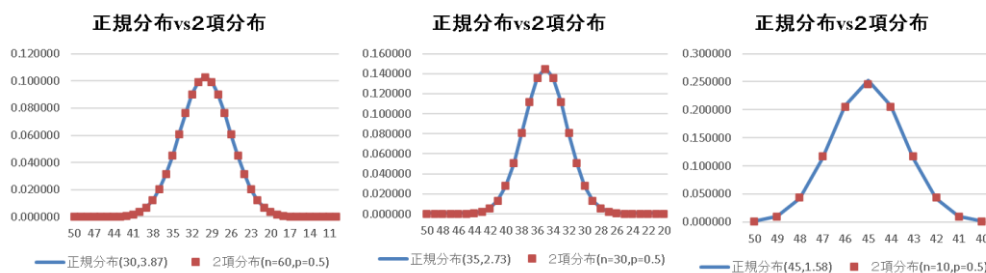


図 1.17 標準正規分布と二項分布

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 1.5：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 1.18 を、平均を固定した正規分布は図 1.19 を参照してください。

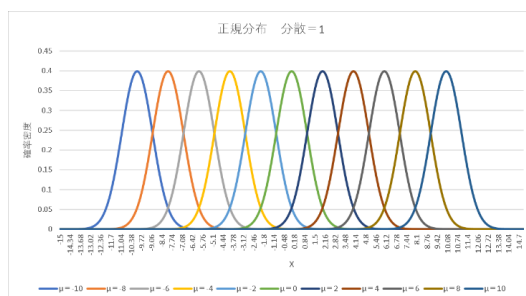


図 1.18 正規分布：平均を変化

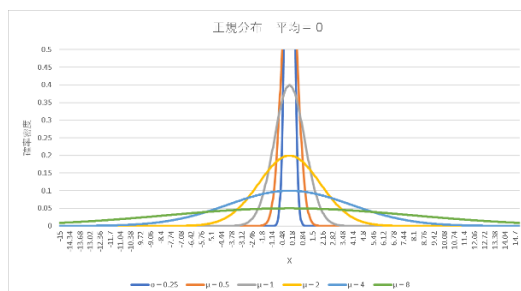


図 1.19 正規分布：分散を変化

1.5. 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、

$$E(X) = \sum_j x_j P(X = x_j)$$

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

$$\bar{x} = \sum_j x_j \frac{N_j}{N}$$

で、期待値は

$$E(X) = \sum_j x_j p_j$$

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。

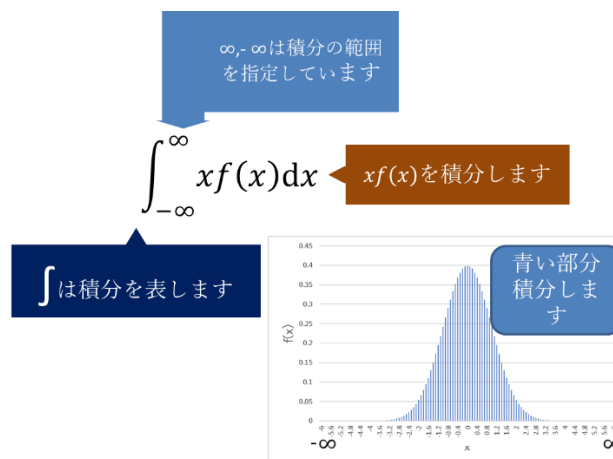


図 1.20 積分と期待値

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 1.1: 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 1.2: 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 1.3: 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 1.4: トランプの標本空間はなんでしょうか？

練習問題 1.5: 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 1.6: 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 1.7: 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 1.8: 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 1.9: ベルヌーイ分布の例をあげてみましょう。

練習問題 1.10: 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？

2. 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。

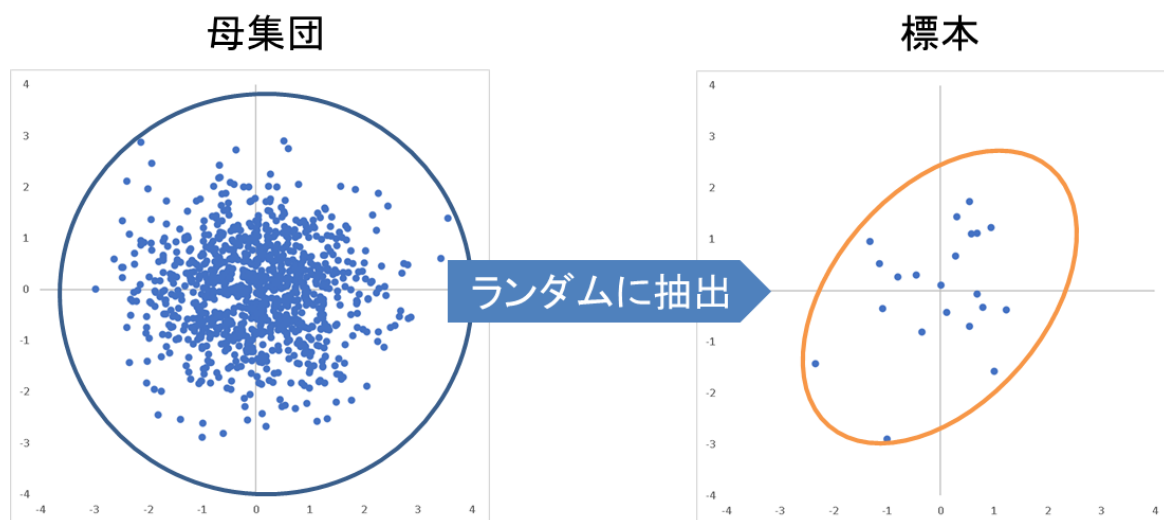


図 2.1 母集団と標本

2.1. 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 2.1：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 2-1 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 2.2 : 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。)

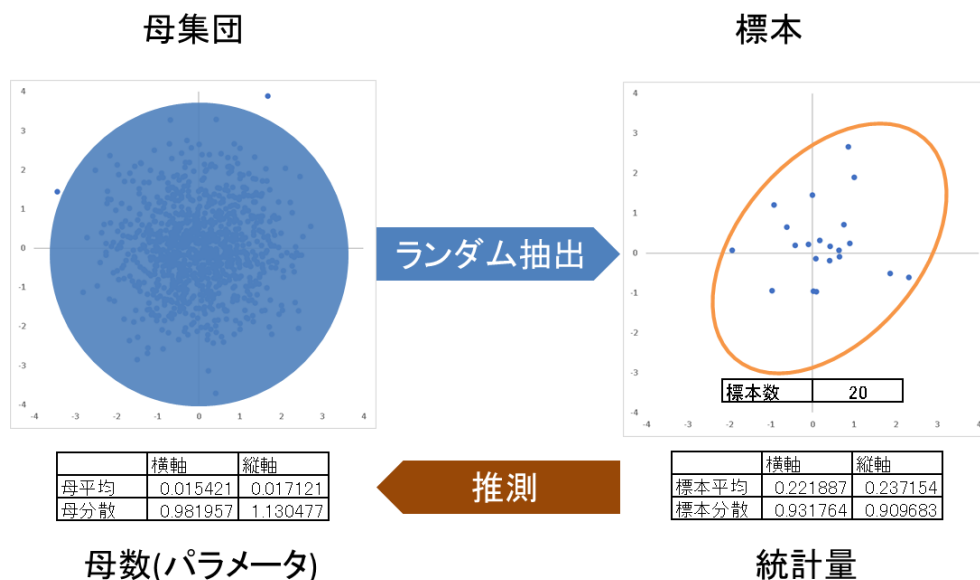


図 2.2 母集団と標本：母数と統計量の比較

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。

2.2. 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。

- 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。
- 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。
- 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。

2.3. 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 2.3 : 例題 2.2 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。

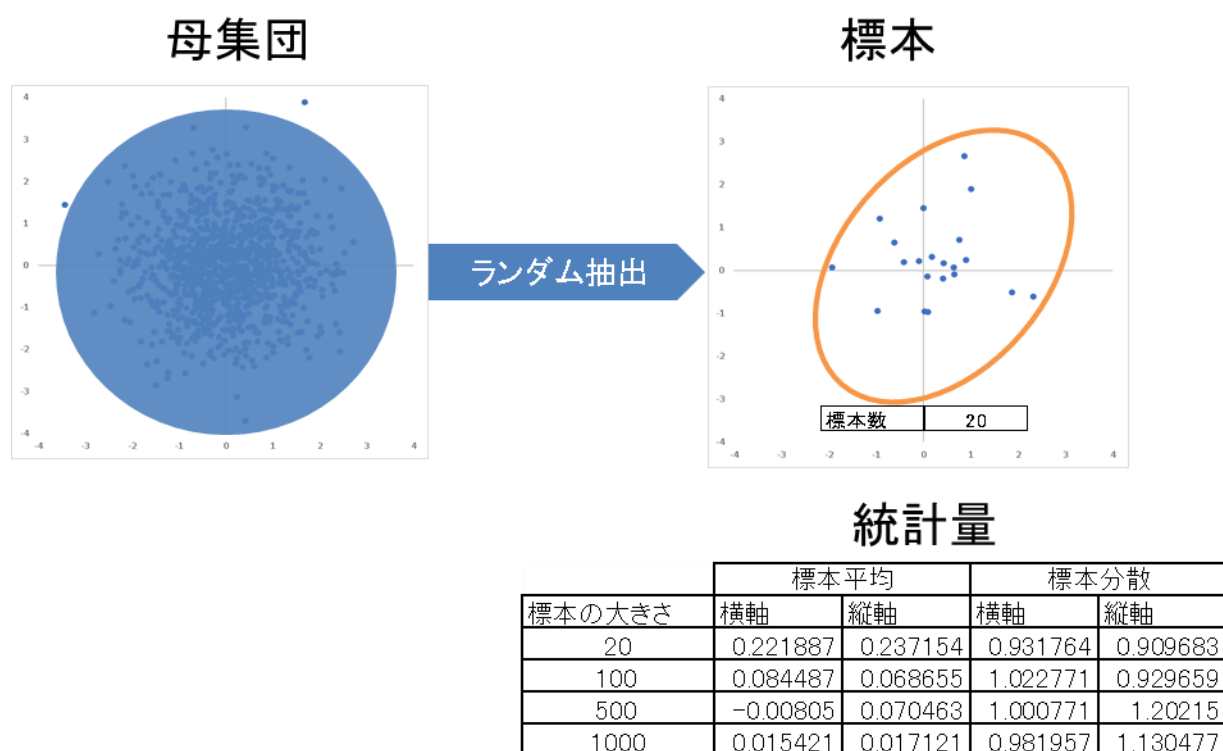


図 2.3 母集団と標本：サイズの違い

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 2.3 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 2.4：例題 2.2 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。

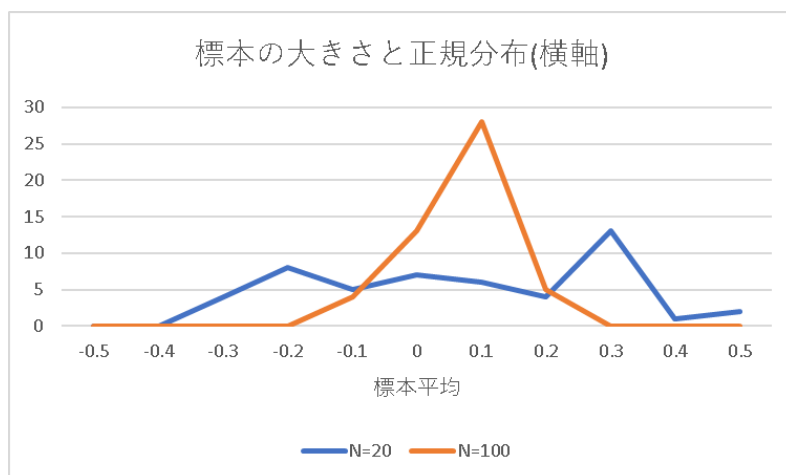


図 2.4 母集団と標本：標本の大きさと頻度図(横)

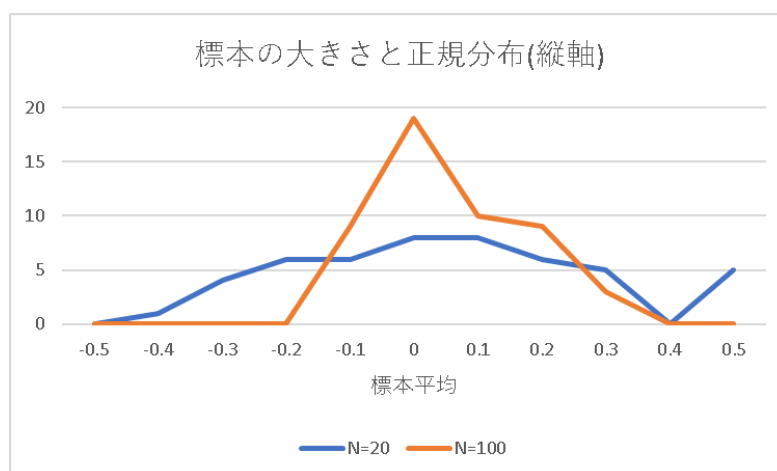


図 2.5 母集団と標本：標本の大きさと頻度図(縦)

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、重複がないようにとる必要があります。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。また、例題 2.4 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。

2.4. 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。

2.4.1. 一緻性

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、それを一緻性とよび、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。

2.4.2. 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。偏差の和がゼロになるように平均値を計算します。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 2.5: 30 個の正規乱数を発生させ、それを 1 試行として 1000 試行行い、それぞれの試行から乱数を 10、15、20、25、30 個選んで、それぞれを 1 グループとして分散と不偏分散を計算して、グラフとして表示してみましょう。また、グループごとの分散、不偏分散の平均、最大値、最小値を計算して、グラフにしてみましょう。

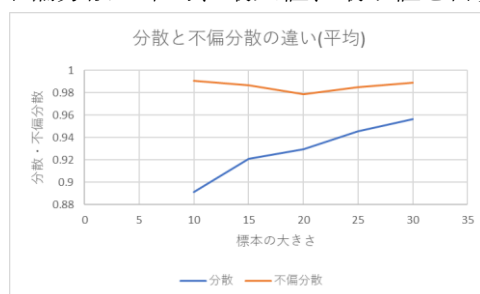


図 2.6 分散と不偏分散の違い(平均)

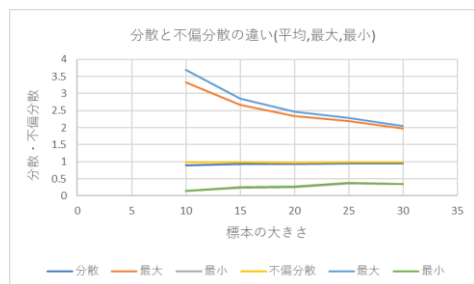


図 2.7 分散と不偏分散の違い(平均、最大値、最小値)

2.5. 標本分布

例題 2.2~2.4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。

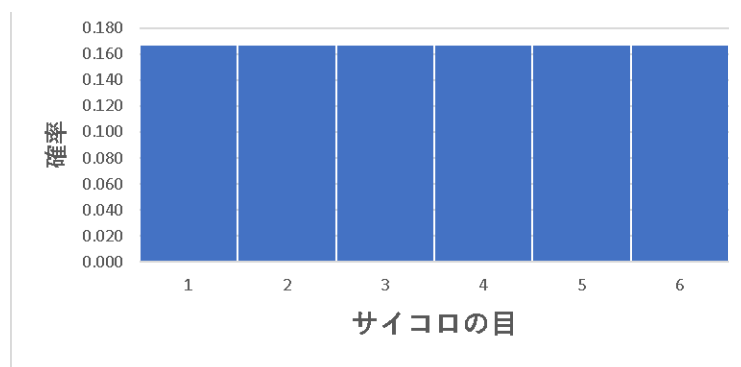


図 2.8 一回の試行と頻度図

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2-2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2-2に表しました。

		一回目の試行					
		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 2-2 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。

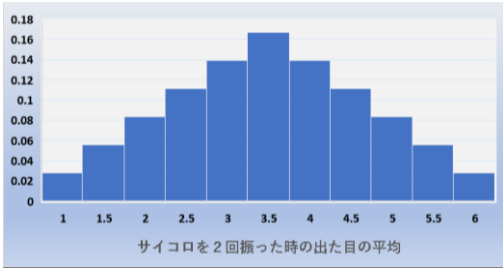


図 2.9 2回の試行と頻度図

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。

2.5.1. カイ二乗分布

確率変数 $Z_1, Z_2, Z_3, \dots, Z_n$ 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

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

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 2.7：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。

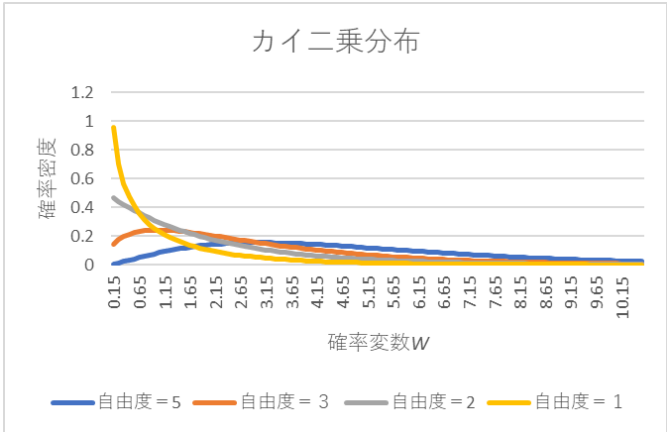


図 2.10 χ 二乗分布と自由度

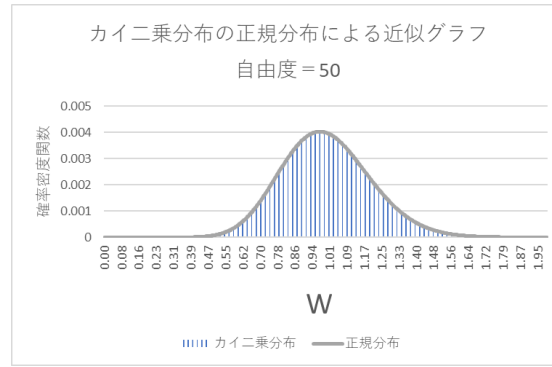


図 2.11 χ 二乗分布の正規分布による近似：自由度=50

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 2.8：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 Z_1, Z_2, \dots, Z_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 $W = \sum Z^2$ がカイ二乗にしたがうのでした。 X_i が平均 μ 、分散 σ^2 にしたがうのであれば、 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。

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

左辺を、不偏分散を含む形に変形します。

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

図 2.12 では、まず 10 個の乱数を生成し、その不偏分散と分散を計算します。そしてその試行を 1000 回繰り返します。 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。

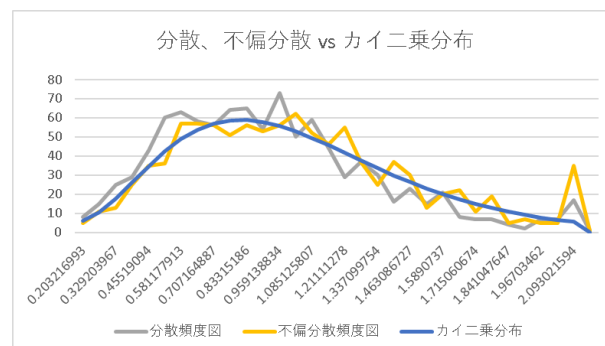


図 2.12 χ 二乗分布と不偏分散 例題 3.8

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。

2.5.2. 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

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

です。このとき、

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

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 2.9：自由度 1 の t 分布と正規分布を比べてみましょう。

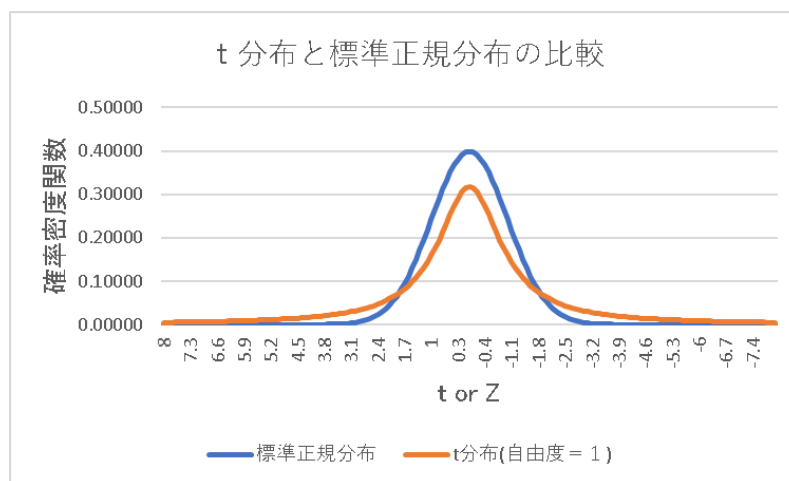


図 2.13 t 分布と標準正規分布

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。 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。

2.5.3. 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 W_1 と W_2 の比は F 分布にしたがいます。

$$F = \frac{W_1/d_1}{W_2/d_2}$$

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 2.14 確率分布の分類

練習問題 2.1: 平均と期待値の違いを説明してみましょう。

練習問題 2.2: 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 2.3: 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 2.4: 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 2.5: 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 2.6: 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。