金融財務研究会

㈱金融財務研究会
㈱経営調査研究会

確率・統計学入門

午後の部（エクセル編）

ZOOM ミーティング

2022 年 6 月 20 日(木)
13:30－16:30
講師：森谷博之 Quasars22

テレビや新聞、雑誌、インターネットには表やグラフが溢れています。確率統計はひとびとに商品・サービスの役割、良さ、危険性を理解してもらうために役立てられています。また、データをモデルで解析することで、いままで見えなかった特性や傾向を明確にできます。

午後の部ではエクセルを用いて、推測統計の基礎知識を身に着けていきます。乱数を用いて母集団と標本の関係を目でみて理解します。また、平均、分散についての点推定、区間推定を学びます。このような知識は、データのもつ特性の把握の基礎になります。顧客情報の管理・分析、財務情報の予測、確認、時系列データの各種分析の基礎になります。将来的に線形回帰、分散分析、統計的仮説検定、ポートフォリオ分析、投資戦略を構築する際の基礎となり、それらの専門知識を学ぶ際の時間と労力を低減します。

第１部：　統計的推定の基礎 13：30～15：00

- **母集団と標本、全数調査と標本調査、母集団とモデル**
- **ｔ分布とカイ二乗分布などの標本分布**

第２部：　統計的推定　15：10～16：30

- **記述統計と推測統計の役割**
- **真の値、点推定、区間推定**

■本セミナーに参加して修得できること

発生した問題についてデータをとおして理解する基本的な態度、得られた表やグラフの正しい解釈の仕方、手元のデータを表やグラフ、基本統計量を用いて表現する方法がわかります。疑似相関、疑似トレンドなどの誤解を招きやすい統計用語についても理解します。母集団と標本、統計的推定の役割について理解します。また、エクセルの活用方法が分かります。例題を通して理解を深めます。

■受講対象者

統計に関して「超初心者」の方を対象としています。

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

■使用ソフト：Excel。

■PC には事前に Excel がインストールされている必要があります。

参考文献：

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

第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。

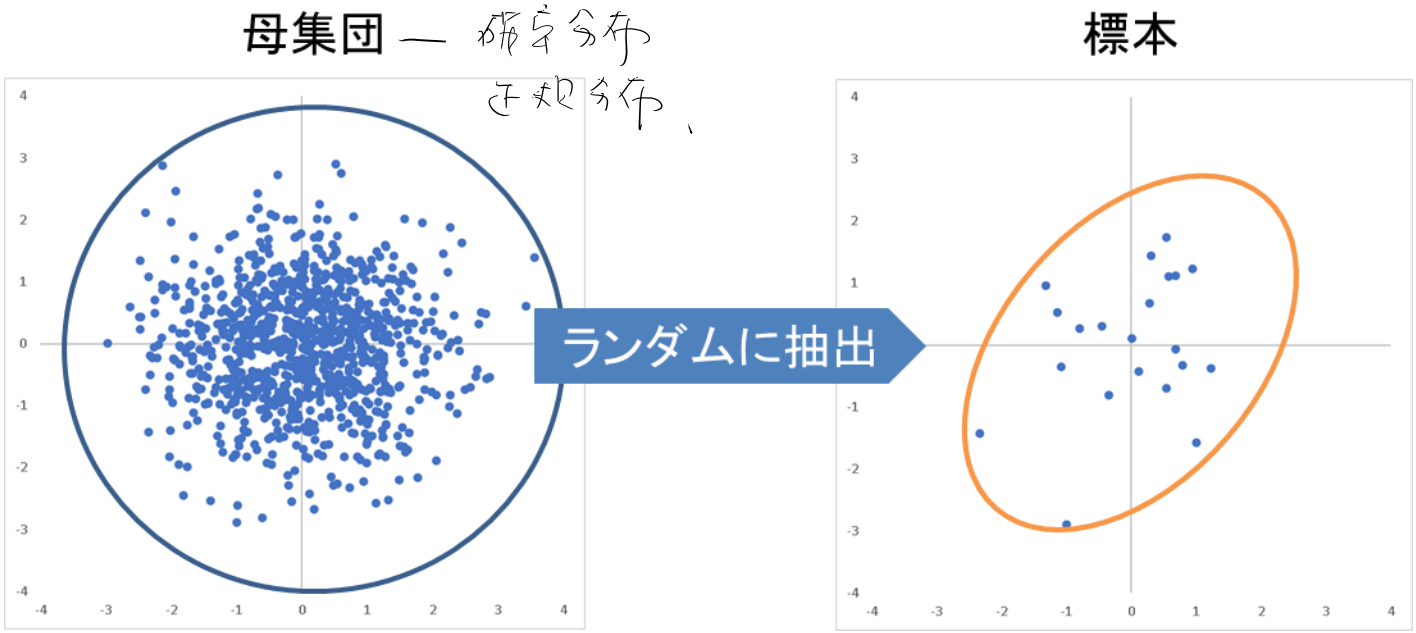


図　3.1　母集団と標本

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

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。)

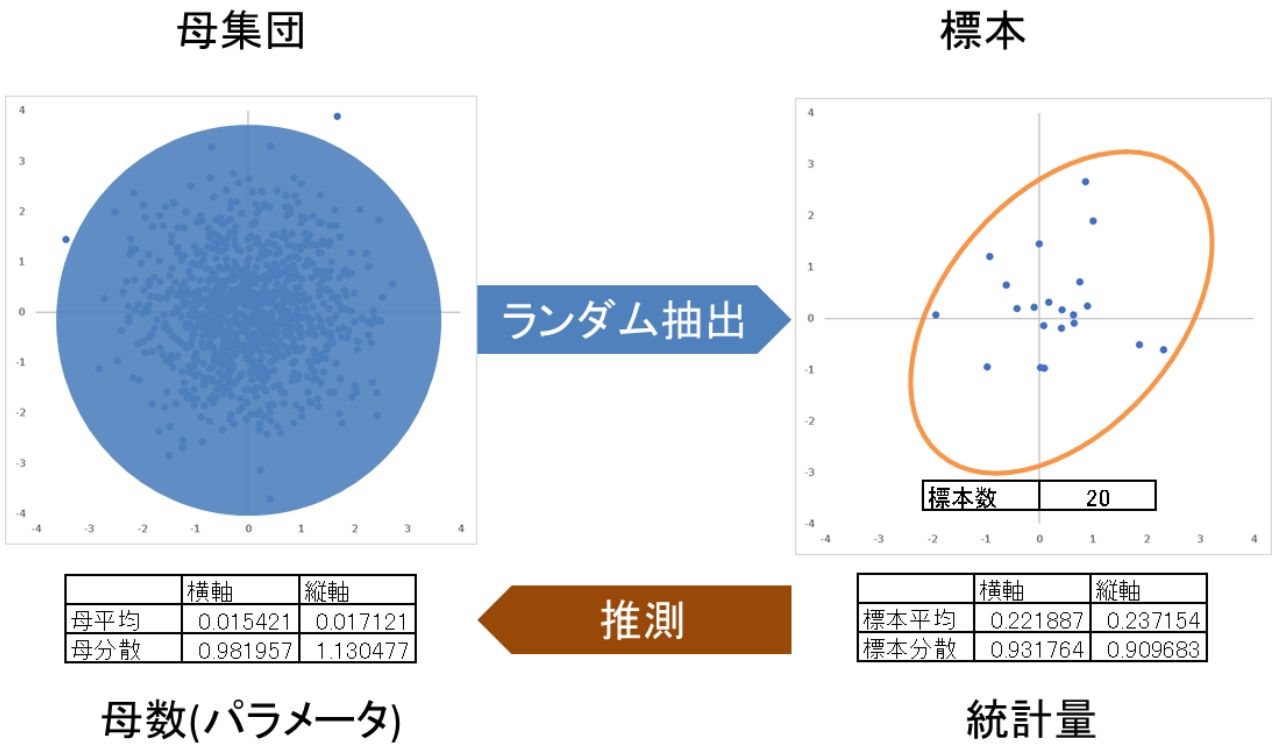


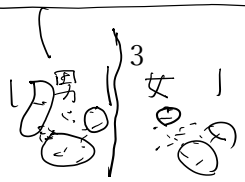
図 3.2 母集団と標本：母数と統計量の比較

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。

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。

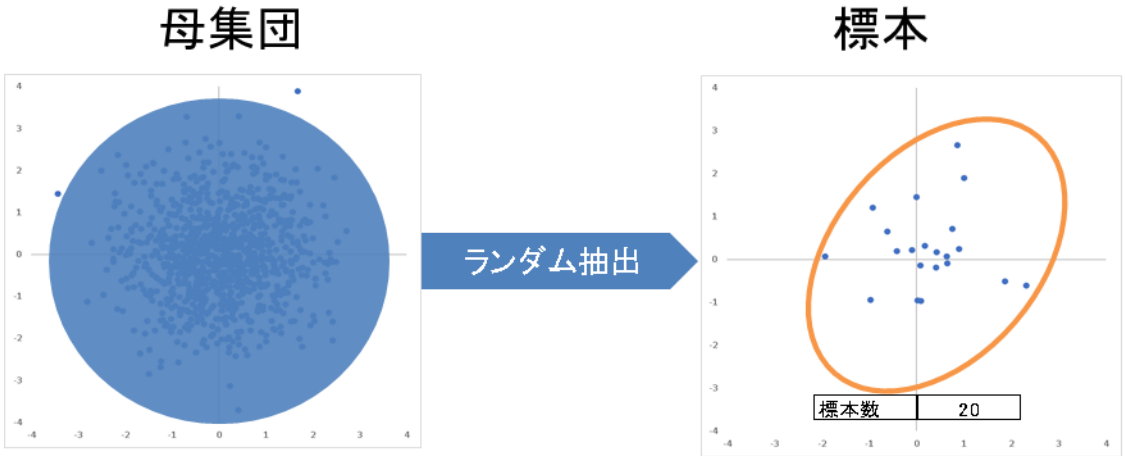
観察研究では特に無作為化が難しく、処理も被験者自らが選択しているために、処理の選択に偏りを生じる可能性があります。

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。



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

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 3.4：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。

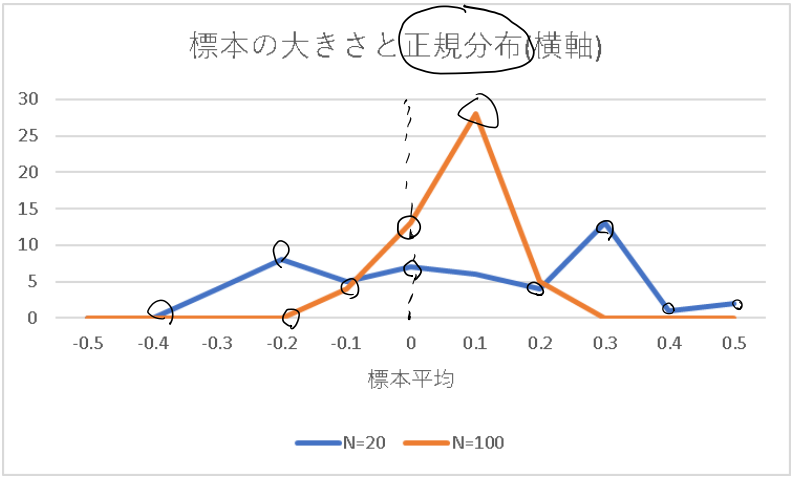


図　3.4　母集団と標本：標本の大きさと頻度図(横)

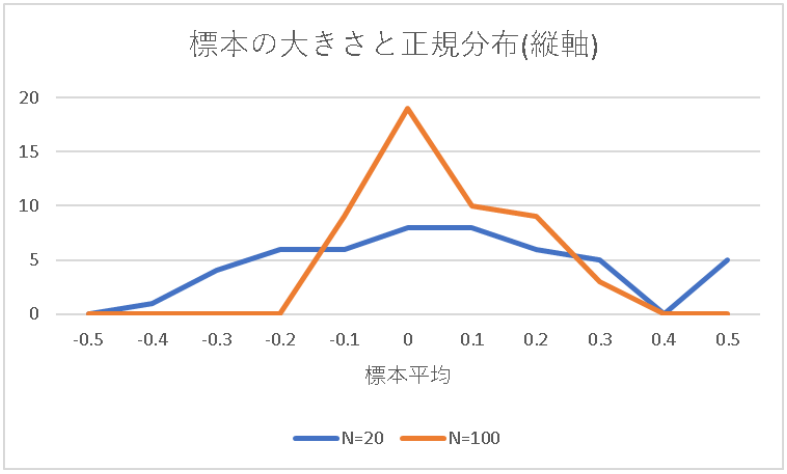


図　3.5 母集団と標本：標本の大きさと頻度図(縦)

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。

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。

3.4.2. 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$x_1, x_2, \ldots, x_n$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。

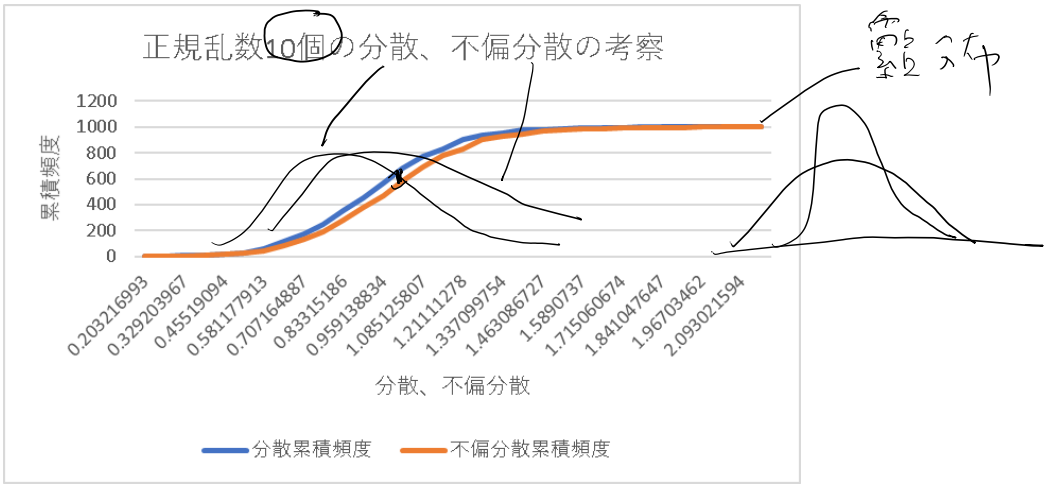


図 3.6 不偏分散

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。

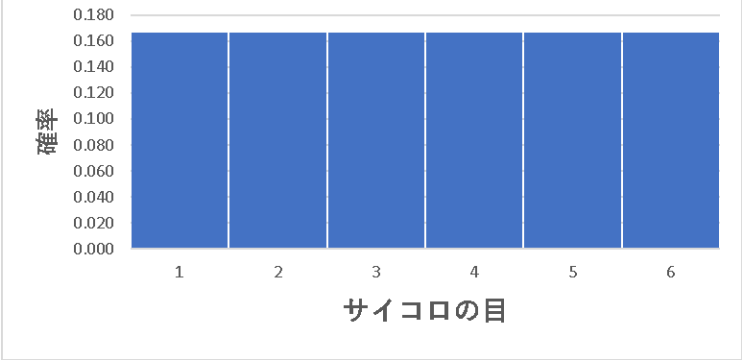


図　3.7　一回の試行と頻度図

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

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。

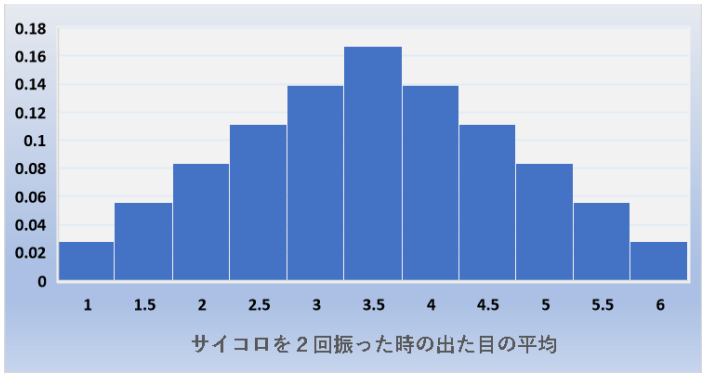


図　3.8　２回の試行と頻度図

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$X_1, X_2, X_3, \cdots, X_n$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

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

がしたがう分布は自由度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。

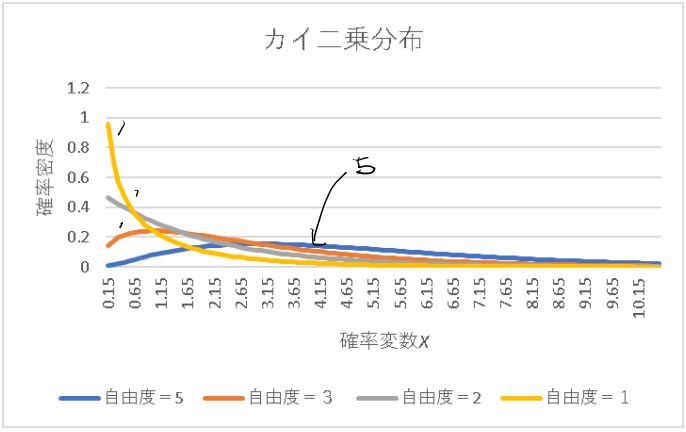


図　3.9　χ二乗分布と自由度

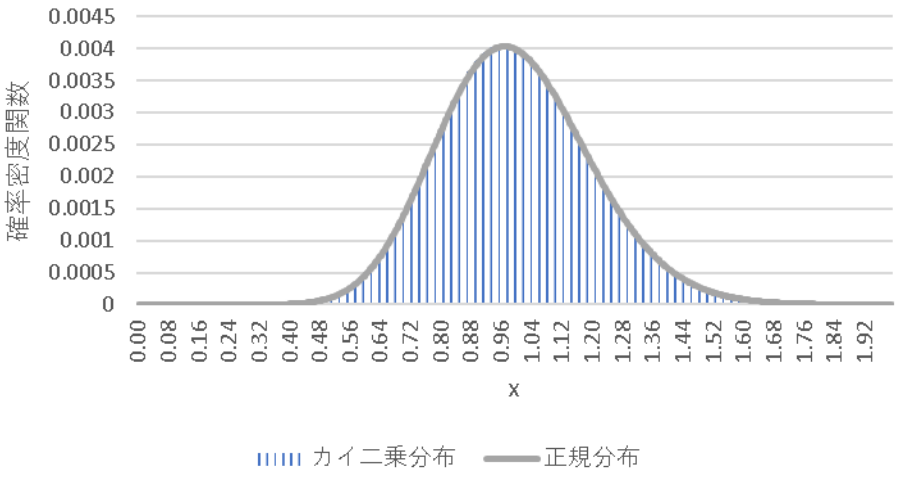


図　3.10　χ二乗分布の正規分布による近似：自由度=50

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 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, …, 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。

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

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。

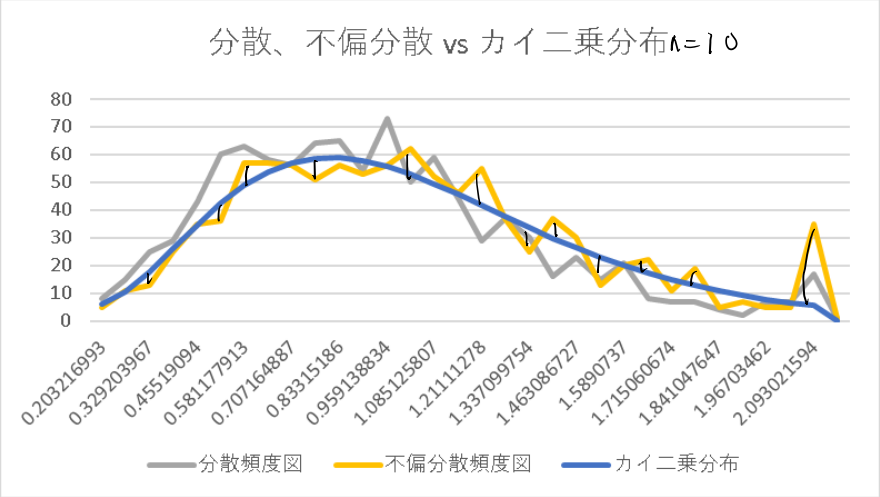


図 3.11　χ二乗分布と不偏分散 例題 3.8

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$X_1, X_3, X_3, \cdots, X_n$は平均$\mu$、分散$\sigma^2$の正規分布に独立にしたがいます。その標本平均が

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

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。

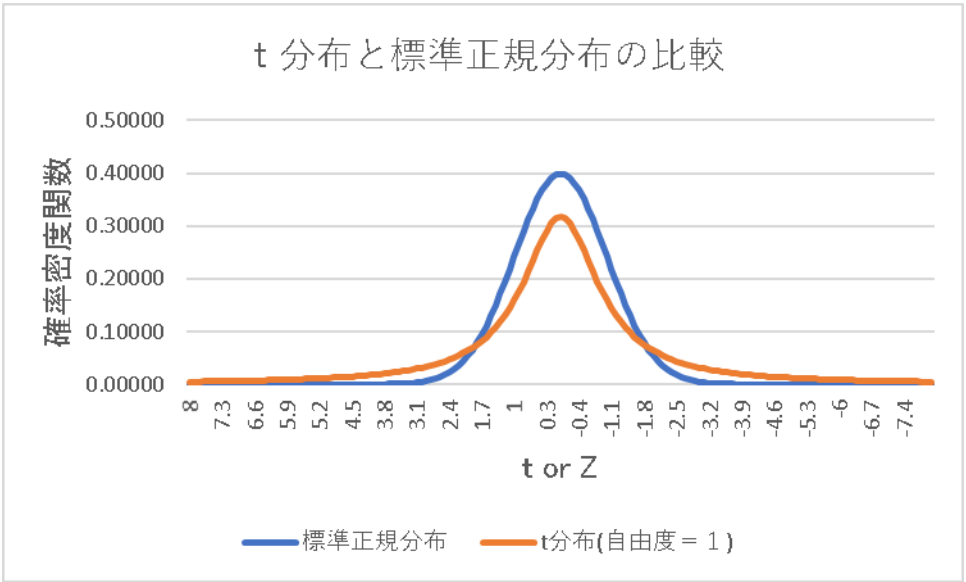


図 3.12　t 分布と標準正規分布

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。

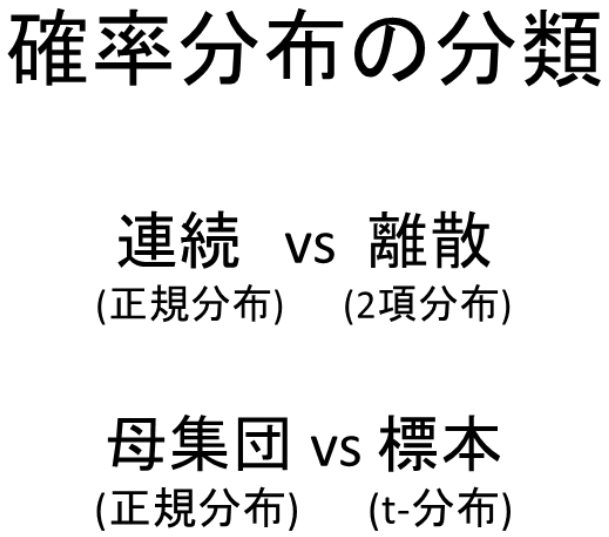


図 3.13 確率分布の分類

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るのです。母集団からデータ$x_1, x_2, \ldots, x_n$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るのです。**

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。

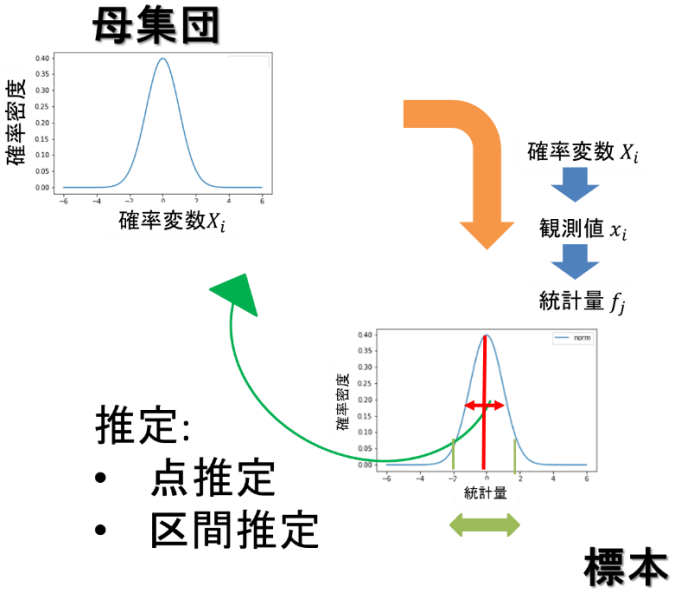


図 4.1 推定

4.1. 点推定

標本$x_1, x_2, \ldots, x_n$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 $\hat{\theta}$ — 標本平均
  - 標準誤差(推定量の標準誤差) se($\hat{\theta}$)

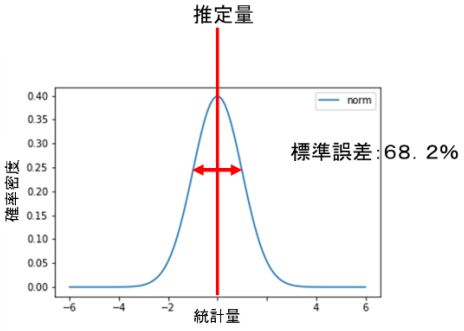


図 4.2 点推定

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もちます。

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

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

$N(0,1)$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

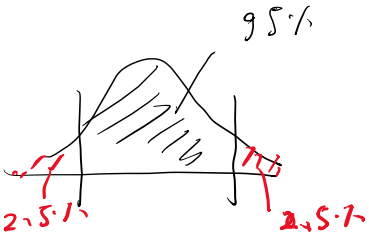
$1-\alpha$は信頼係数で母平均μが信頼区間に含まれる割合になります。

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

b) **母分散は未知**

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。母分散は未知なので、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。

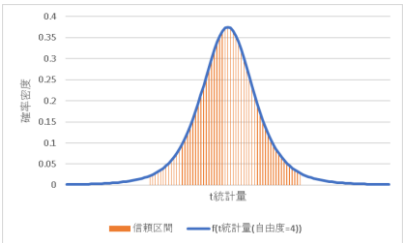


図 4.3 区間推定と分布

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。

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

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。$\alpha = 0.01$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

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

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

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

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、母数がそ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	0.585346	1.959964	1.62161	0.280391	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	0.11703	1.046123	1.07475	0.03818	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.inv(確率,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。

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、

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

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

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

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

となります。

（平均の差
（比率の差

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。