統計・多変量解析と
ソフトコンピューティング

第2章　確率試行のシミュレーション

本稿掲載のWebページ

http://mybook-pub-site.sakura.ne.jp/Statistics_Multivariate/index.html

古橋　武

目次

第2章

確率試行のシミュレーション

2.1　乱数について

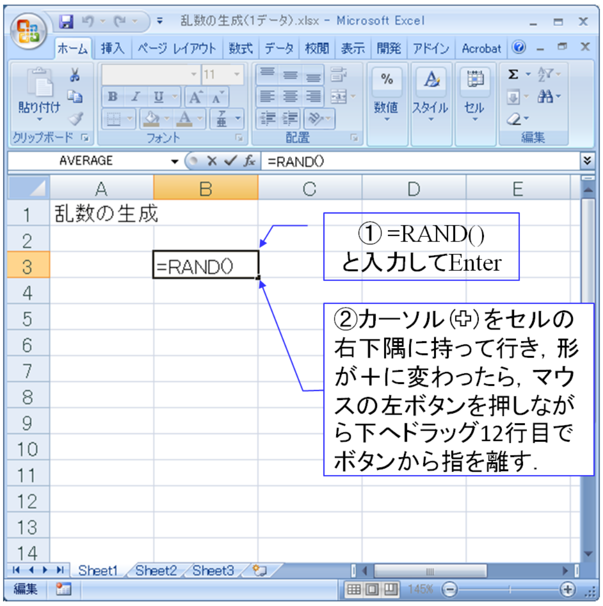


図 2.1: RAND() 関数の入力 (乱数の生成.xlsx)

まず，RAND() 関数を体験しよう．図 2.1 は Microsoft Excel 2007 を立ち上げて，最初に現れる画面 Sheet1（シートと呼ばれる）に RAND() 関数を入力したところを示す．図説の「乱数の生成.xlsx」は対応する Excel ファイルのファイル名である．これら Excel ファイルは以下の URL(http://www.kyoritsu-pub.co.jp/bookdetail/9784320122680) からダウンロードできる．シート上の一つ一つのますはセルと呼ばれる．セルは列をアルファベットで表し，行を数字で表している．図ではセル B3 に

$$= \mathrm{RAND}() \tag{2.1}$$

と入力してある．セルに文字・記号などを打ち込むには入力したいセルにカーソルを持っ

ていって，マウスの左ボタンを 2 度連続してクリックすることで入力可能となる．本書ではマウスの左ボタンをクリックすることを左クリック，右ボタンの場合を右クリック，2 度連続してクリックすることをダブルクリックとよぶ．RAND につづくカッコの中には何も打ち込む必要はない．上式を打ち込んだ後に Enter キーを押すと，RAND() 関数の入力が確定される．次にカーソルをセル B3 の右下隅に持って行くと，カーソルの形が細い十字に変わるので，マウスの左ボタンを押しながらカーソルを移動させると太い黒枠がカーソルについてくる（ドラッグすると呼ばれる）．図 2.2 は 12 行目までドラッグして，左ボタンから指を離した結果である．太枠内のどのセルでもよいから選んで左ダブルクリックすると RAND() 関数がコピーされていることを確認できる．これで 10 個の乱数を一度に生成できる．図 2.3 はこれらの乱数の折れ線グラフを作画する手順を示す．図中の数字の示す順番に操作を行う．まず，(1) セル B3 を左クリックし，次に Shift キーを押しながらセル B12 を左クリックする．これでグラフに表示したいデータの範囲を指定できたので，その後に (2) 挿入→ (3) グラフ→ (4) 折れ線→ (5) 折れ線と左クリックしていくことで作画できる．図 2.4 は折れ線グラフのできあがりの画面である．なお，シート上の 10 個の RAND() 関数を再計算させるには，F9 ボタンを押せばよい．F9 ボタンを押す度に，シート上の全関数が再計算される．

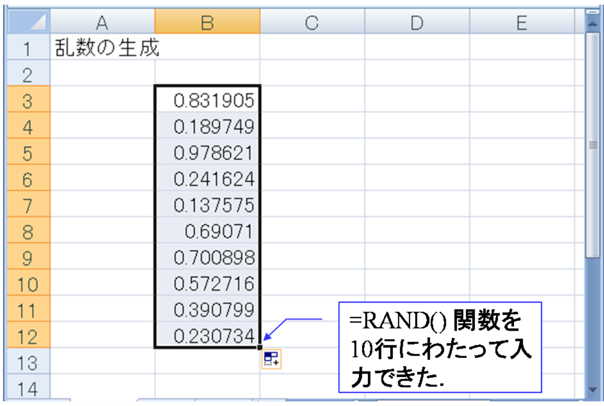


図 2.2: RAND() 関数のコピー結果 (乱数の生成.xlsx)

RAND() 関数は，その出力を x とすると $0 < x < 1$ の区間の擬似一様乱数を生成する．本書ではこの区間を $(0, 1)$ と表記する．なお，$0 \leq x \leq 1$ の場合は $[0, 1]$ として区別して表記する．$0 < x \leq 1$ の場合には $(0, 1]$ とする．ここで，乱数列とは，要素間に規則性のない数列のことである．乱数は乱数列の要素である．一様乱数とは，ある有限の区間を区切ったときに，その区間内の全ての要素が同じ確率で現れるような乱数のことである．また，擬似乱数とは，規則に基づいて生成される数列が，実用の範囲内では乱数列と見なせる数列の要素のことである．RAND() 関数は区間 $(0, 1)$ を約 27 兆分割した各分割点の値を出力する．本書で用いる限りでは，ほぼ一様と言ってよい乱数（擬似一様乱数）で

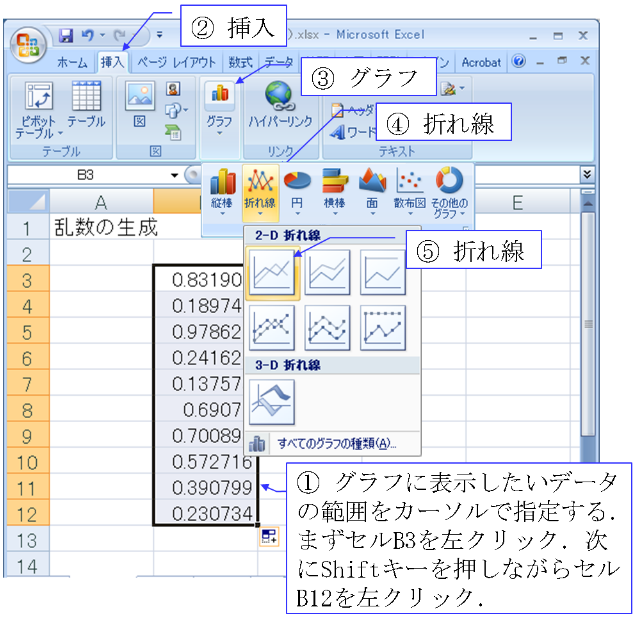


図 2.3: 折れ線グラフの作画

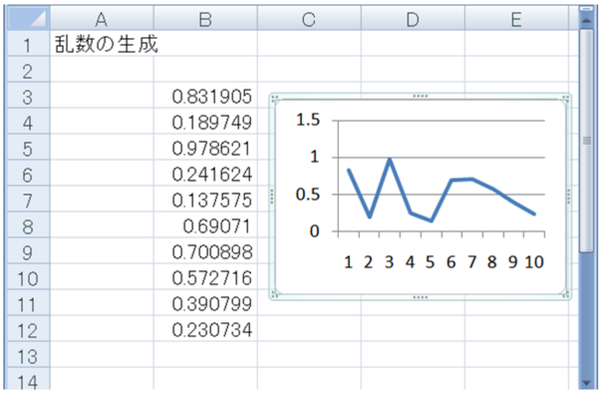


図 2.4: 折れ線グラフの出来上がり (乱数の生成.xlsx)

ある．

RAND() 関数の入力されているいずれのセルでもよいので，カーソルを持って行って，右クリック→セルの書式設定→表示形式→数値→小数点以下の桁数=17 と設定するとこの乱数は小数点以下 15 桁まで生成されていることが分かる．

RAND() 関数のアルゴリズムは次の通りである [1]．ここでは Excel のマクロのプログラミング言語である Excel VBA(Visual Basic for Application) により記述する．なお，マクロについては Excel VBA の解説書が数多く出版されているのでそちらを参照されたい．本書では??節にマクロの使用例を紹介する．さらに，マクロは第??章以降で多用する．IX, IY, IZ の初期値をそれぞれ $[1, 30268]$, $[1, 30306]$, $[1, 30322]$ の間の整数値に設定し

て，以下の計算を繰り返す．

$$IX = \mathrm{MOD}(171 \times IX, 30269) \tag{2.2}$$
$$IY = \mathrm{MOD}(172 \times IY, 30307) \tag{2.3}$$
$$IZ = \mathrm{MOD}(170 \times IZ, 30323) \tag{2.4}$$
$$RANDOM = \mathrm{MOD}(IX/30269 + IY/30307 + IZ/30323, 1) \tag{2.5}$$

式中の 30269, 30307, 30323 は素数である．MOD(I, J) 関数は I を J で割った余りを出力する関数である．上式の計算を繰り返すと，式 (2.2) は 1〜30268 の全ての整数値を順不同に出力した後，ふたたび初期値にもどり，以降同じ整数値列を出力する．図 2.5 はこのことを確かめている Excel の画面である．セル B5 の 10000 を初期値として，セル B6 以下では一つ上のセルの値を入力として式 (2.2) の計算結果を出力している．セル B5 から B30272 までで 1〜30268 の間の全ての整数値を一度ずつ出力した後，セル B30273 にて初期値の 10000 を出力している．このことを確認するために B 列の 5 行目から 30272 行目までの数値を D 列にコピーして昇順に並べ替えている．図 2.6 は IX のコピーと並べ替えの手順を示す．コピー元の範囲指定にはセル B5 を左クリックした後に Shift キーと Ctrl キーを同時に押しながら↓キーを押す操作を行う．これにより先頭のセルから最下行のセルまでを少ない操作回数で指定できる．以降，図中の (2) → (3) → ⋯ (7) と操作を進めることで，IX のデータのコピーと昇順の並べ替えを行っている．(4) 番目の操作では「値のみ」を選択する．こうしないとセル D5 以下には IX を生成する式がコピーされてしまう．図 2.5 では，さらに E 列で D 列の上下の隣り合うセル同士の差を求めている．この計算式はセル E5 からセル E30271 まで記述されている．もし，D 列にて数値の重複もしくはジャンプがあれば，この差分値は 1 以外となる．セル G5 にて E 列の 1 の数をかぞえたところ 30267 個であった．並べ直した 30268 個の IX の値の間には重複もジャンプもなかった．

同様にして，式 (2.3)，(2.4) の計算式がそれぞれ 1〜30306, 1〜30322 の全ての整数値を順不同で出力することを確認できる．ある初期値から始めて，IX, IY, IZ の関数はそれぞれ異なる周期で整数値を出力する．3 つの関数が同時に初期値と同じ値を出力する周期は約 7 兆である [2]. これはそれぞれの素数から 1 を引いた値 30268, 30306, 30322 の最小公倍数 (=6953607871644) である．

式 (2.5) は IX, IY, IZ の値をそれぞれの素数で割った後に和を求め，さらに 1 で割った余りを出力する関数である．$IX/30269, IY/30307, IZ/30323$ は 1 より小さい有理数であり，それぞれ分母に異なる素数を持つので，決して同じ値になることはない．また，

$$0 < IX/30269 + IY/30307 + IZ/30323 < 3 \tag{2.6}$$

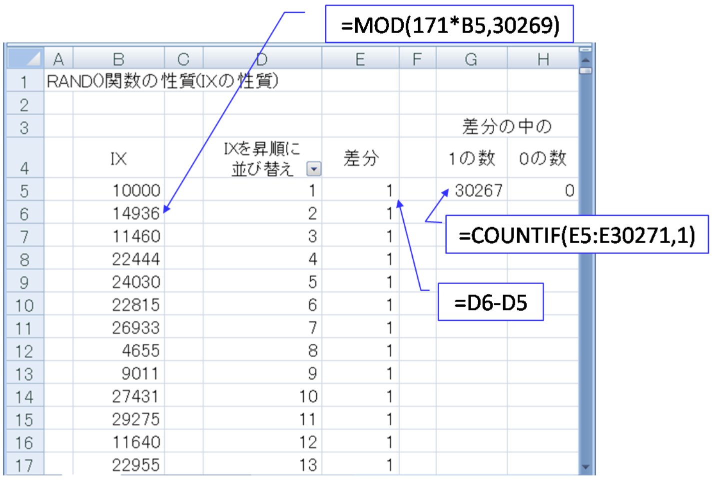


図 2.5: RAND() 関数の IX の性質 (IX の性質.xlsx)

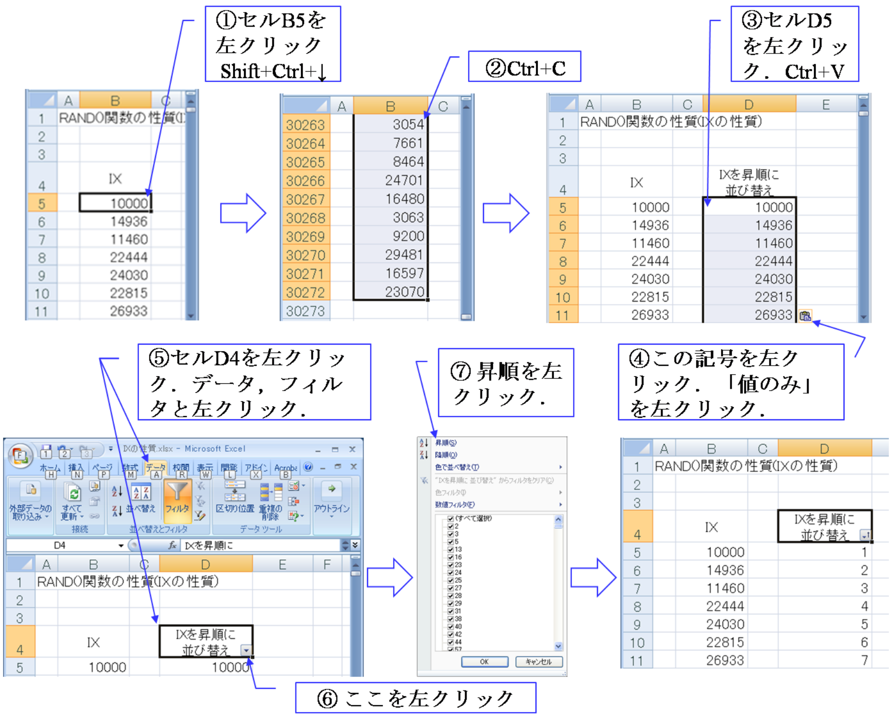


図 2.6: IX のコピーと並べ替え

である．(0, 3) の区間が約 27 兆分割されている．$IX/30269 + IY/30307 + IZ/30323$ を 1 で割って余りをとることは，$(0, 1), (1, 2), (2, 3)$ の各区間の分割点を (0, 1) 区間上に重ね合わせることに相当する．この場合に一致する分割点はない．なぜならば x, y を素数，a, b, c, d を $0 < a, c < x, 0 < b, d < y$ である整数として

$$\frac{a}{x} + \frac{b}{y} = 1 + \frac{c}{x} + \frac{d}{y} \tag{2.7}$$

が成立するとする．$c/x+d/y$ は $a/x+b/y$ を 1 で割った余りに相当する．上式は

$$(b-d)\frac{x}{y}=x-(a-c) \tag{2.8}$$

と変形できるが，x,y は異なる素数であり，$b-d<y$ であるので，左辺は 0 もしくは割り切れないで小数点以下の値を持つ．一方，右辺は 0 より大きい整数であるので矛盾する．すなわち，式 (2.5) の RANDOM の値は $(0,1)$ の区間を約 27 兆分割する有理数の点からなる．図 2.7 (a) は式 (2.2)〜(2.4) の計算を一億回繰り返した時の $IX/30269+IY/30307+IZ/30323$ の値の出現頻度のグラフである．$[0,3]$ の区間を $(0,0.0001],(0.0001,0.0002]$ と 0.0001 刻みの小区間に 30000 等分し，$IX/30269+IY/30307+IZ/30323$ の値が各区間内に入る頻度を求めた結果である．この結果を $(0,1],(1,2],(2,3]$ の 3 区間に分け，全てを $(0,1]$ 区間に重ね合わせた結果を図 2.7 (b) に示す．これは式 (2.5) の出力値 RANDOM の出現頻度に相当する．初期値を変えて何度同じグラフを描いても同様の結果を得ることができる．RAND() 関数が $(0,1)$ 区間の数値をまんべんなく出力することが分かる．約 7 兆個の周期の乱数列に対して 1 億個の乱数列は一部でしかないが，RAND() 関数の性能の一端を見ることができる．

図 2.8 は式 (2.5) の出力値 RANDOM の自己相関係数の計算例を示す．式 (2.2)〜(2.5) を用いて，ある初期値から始めて百万個の乱数を生成し，先頭から 1000 個ずつの組に分け，先頭の第 1 組と以降の第 i 組のデータとの自己相関係数 R_{1i} を求めた結果である．自己相関係数 R_{1i} は

$$R_{1i}=\frac{1}{nv_e^2}\sum_{j=1}^{n}(r_{1j}-\overline{r})(r_{ij}-\overline{r}) \tag{2.9}$$

により定義される．ここで，r_{ij} は第 i 組の j 番目のデータであり，$\overline{r},v_e^2$ はそれぞれ全乱数の平均値と不偏分散である．$n=1000$ である．不偏分散 v_e^2 は

$$v_e^2=\frac{1}{n^2-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(r_{ij}-\overline{r})^2 \tag{2.10}$$

により与えられる．第 1 組のデータ同士の自己相関係数はほぼ 1 であり，第 1 組と以降の組との係数の絶対値はほとんどが 0.1 以下である．分布形状においても述べたように，約 7 兆個の周期の乱数列に対して百万個程度の乱数列はほんの一部でしかないが，RAND() 関数の性能の一端を別の角度から見ることができる．

さて，Excel のシート上では RAND() 関数は図 2.1, 2.2 のように各セルに RAND() を記入して用いられる．Excel の RAND() 関数に式 (2.2)〜(2.4) の IX,IY,IZ の初期値を指定する機能はない．RAND() 関数の初期値設定は Excel 任せである．そこで，百万個のセルに=RAND() を記入し，(0, 1) の区間を百万分割して，そのうちのある区間を指定

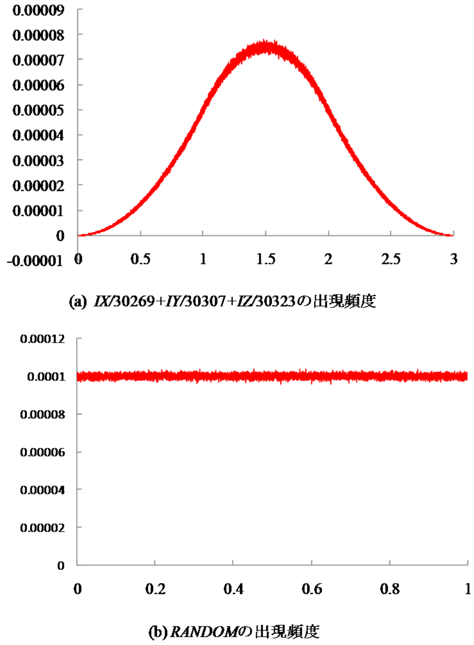


図 2.7: RAND() 関数の出力値の分布形状

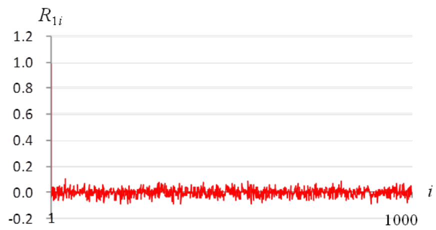
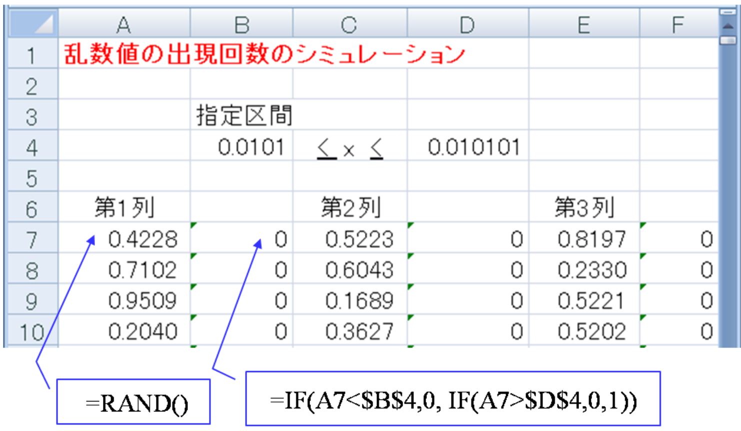


図 2.8: RAND() 関数の出力値の自己相関係数

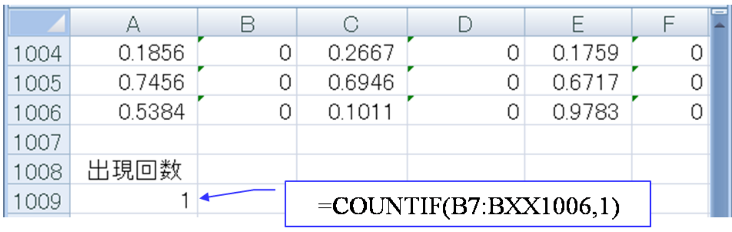
し，その区間内の値の出現回数を求めるシミュレーションを実施した．図 2.9，2.10 はそのシミュレーション画面である．第 1 列に RAND() 関数による乱数を 1000 個生成して，それぞれの右隣のセルで，セル B4 と D4 で指定された範囲内の数値が出現した場合に 1 を出力する設定である．IF(a, b, c) は，a の条件が成立したときに b の値を出力し，不成立のときに c を出力する関数である．すなわち，セル B7 ではセル A7 の値が 0.0101 より小さいときに 0 を出力し，0.0101 以上の場合には，次の IF 文により 0.010101 より大きい場合に 0 を出力し，0.010101 以下の場合に 1 を出力する．セル A7 の値が $[0.0101, 0.010101]$ の区間にあるとき 1 が出力される．これを右方向に第 1000 列までコピーして，百万個の乱数を生成している．セル A1009 では以上の全データの中で 1 の数を数えている．この例では百万回に 1 回だけ指定区間内の値が生成されたという結果が得られている．この計算を 1000 回繰り返したところ，出現回数の平均値は 0.989 であり，真の平均値は 95% の確率で $[0.925, 1.053]$ の区間内にあるという結果が得られた．Excel の RAND() 関数の

性能の一端を見ることができる．



	A	B	C	D	E	F
1	乱数値の出現回数のシミュレーション					
2						
3		指定区間				
4		0.0101	≦x≦	0.010101		
5						
6	第1列		第2列		第3列	
7	0.4228	0	0.5223	0	0.8197	0
8	0.7102	0	0.6043	0	0.2330	0
9	0.9509	0	0.1689	0	0.5221	0
10	0.2040	0	0.3627	0	0.5202	0

図 2.9: 乱数値の出現回数のシミュレーション（乱数値のシミュレーション.xlsx）



	A	B	C	D	E	F
1004	0.1856	0	0.2667	0	0.1759	0
1005	0.7456	0	0.6946	0	0.6717	0
1006	0.5384	0	0.1011	0	0.9783	0
1007						
1008	出現回数					
1009	1					

図 2.10: 乱数値の出現回数のシミュレーション（その 2）

2.2　多数回試行のシミュレーション (大数の法則)

大数の法則とは次のことをいう．

> ある独立試行において事象が起きる確率（数学的確率）が p であるとする．このような前提条件の下で、その事象が起きる比率が試行回数を増やすにつれて近づく値（経験的確率）は p である．

ここで，試行とはコインを投げたり，さいころ投げたりする行為をいう．独立試行とは，繰り返し行ったとしてもある回の試行が他の回の試行に影響を及ぼすことがない試行をいう．そして，事象とは試行の結果によって定まる事柄をいう．さいころ投げでは 1〜6 の目のどれかが出るという事象が起きる．さいころの目が偶数であるという事象は集合 $\{2, 4, 6\}$ と表される．

2.2.1 コイン投げのシミュレーション

コインを 10 回投げて，表裏の出る回数を数えるシミュレーションを体験しよう．そのために RAND() 関数を用いる．RAND() 関数は 0 より大きく 1 より小さい擬似一様乱数を生成する．この RAND() 関数を用いてコイン投げのシミュレーションを行う．図 2.11 ではセル B7〜B16 に関数

$$= \mathrm{INT}(2 * \mathrm{RAND}()) \tag{2.11}$$

を入力してある．この関数は 2×RAND() により $(0, 2)$ の範囲の擬似一様乱数を生成し，INT() 関数により小数点以下を切り捨てている．コインの表を 1，裏を 0 の値に対応させれば，セル B7〜B16 の値はコインを 10 回投げたときの結果を表している．セル B4 の関数

$$= \mathrm{COUNT}(\mathrm{B7} : \mathrm{B16}) \tag{2.12}$$

はセル B7〜B16 の範囲内で数値が含まれるセルの個数を数える．コインを投げた回数を表示している．次にセル C19 では

$$= \mathrm{COUNTIF}(\$\mathrm{B}\$7 : \$\mathrm{B}\$16, \mathrm{B19})/\$\mathrm{B}\$4 \tag{2.13}$$

により，セル B7〜B16 の範囲内でセル B19 と同じ数値を含んでいるセルの個数を数え，その結果をコインを投げた総数で割っている．結果はコインの裏の出た比率（経験的確率）となる．セル B7, B16, B4 をそれぞれ B7，B16，B4 と $ 記号を付けて表記しているのは，式 (2.13) をセル C20 にコピーした際に，いずれのセル番号も B7, B16, B4 のままで変化しないことを指定している．一方，B19 には $ 記号がついていないので，式 (2.13) が一つ下のセルへとコピーされた際に自動的に B19 から B20 への変換がなされる．C20 の値はコインの表の出た比率である．図 2.11 ではコインの表裏の比率を棒グラフに表してある．試行回数が少ない場合は，表裏の生起する比率は 0.5 から大きく異なることが多い．

図 2.12 は棒グラフを作成する手順を示す．図中の番号順に手続きを進める．まず，(1) 棒グラフに表示したい範囲を指定し，(2) 挿入，(3) 縦棒，(4)2-D 縦棒の一番左のボタンを左クリックをする．表示された棒グラフを図 2.13 に示す．ただし，このままでは横軸のラベルがデータ区間と合っていない．そこで，横軸ラベルの設定を行う．横軸の数値の辺りを右クリックすると，図 2.14 の左上のメニューが現れるので図中の番号順に手続きを進める．(1) データの選択，(2) 横軸ラベルの編集，(3) データ区間 (セル B19〜B20) を選択，(4)(5) ＯＫ．以上により，図 2.11 の棒グラフが得られる．

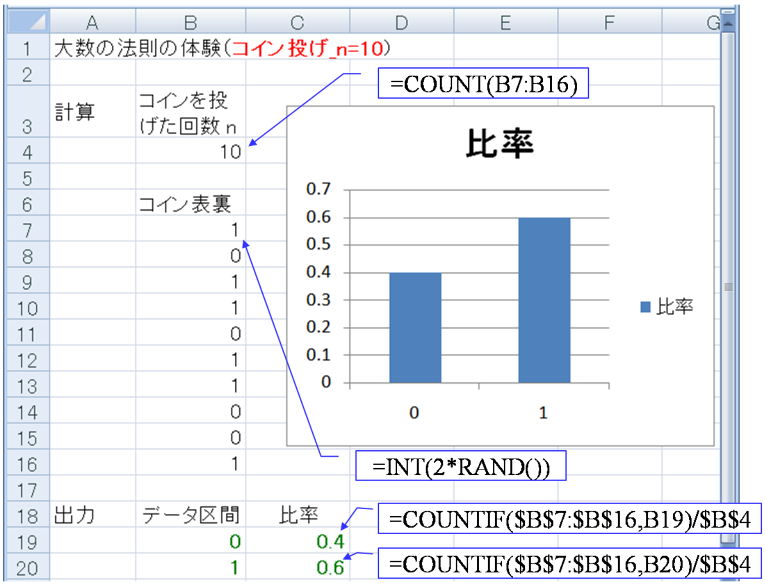


図　2.11:　コイン投げのシミュレーション (n=10)(コイン投げのシミュレーション（n=10）.xlsx)

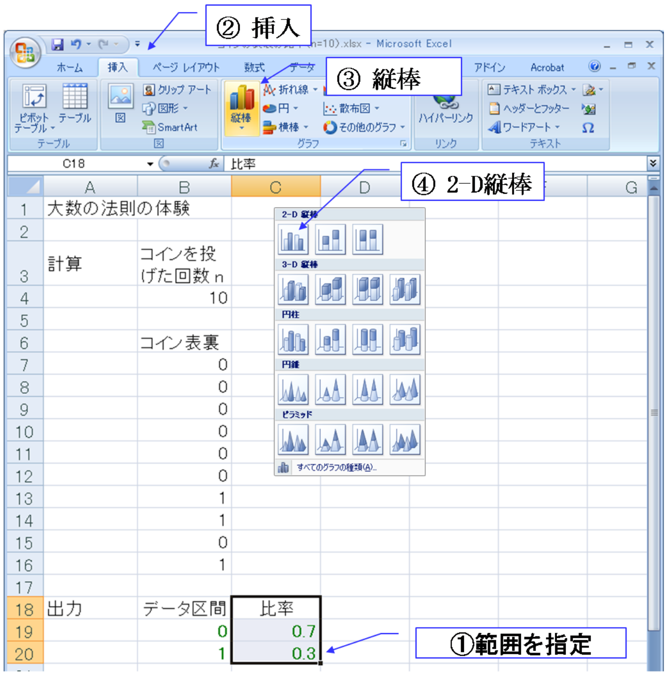


図 2.12: コイン投げ (n=10) グラフ作成

図 2.15 は，コインを 1000 回投げたときの表裏の比率を示す．比率は 0.5 に近い値を示している．コインの表／裏の出る確率はそれぞれ 0.5 の設定であった．コイン投げの試行回数が 10 回のときには，シミュレーションを繰り返す度に表裏の比率は大きく変化した．試行回数を 1000 回に増やしたことで，表裏の比率（経験的確率）が 0.5 に近づいた．

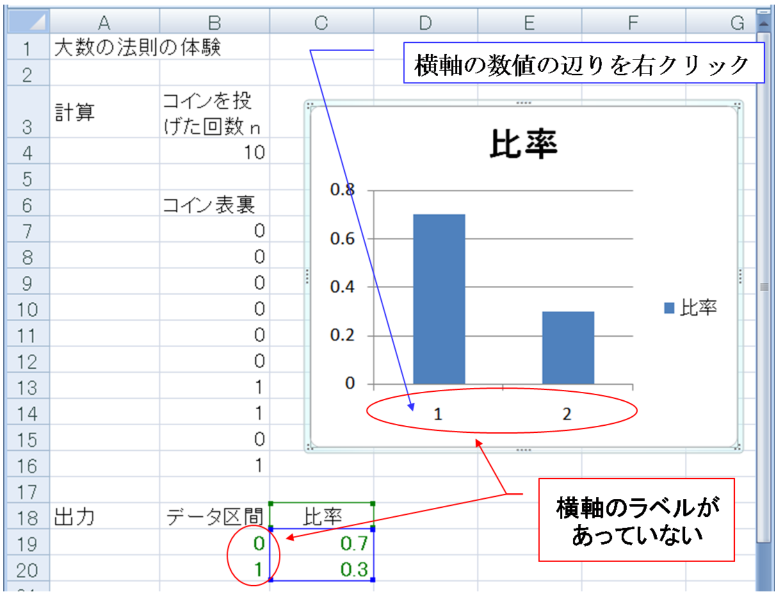


図 2.13: コイン投げ (n=10) グラフの横軸ラベルの設定

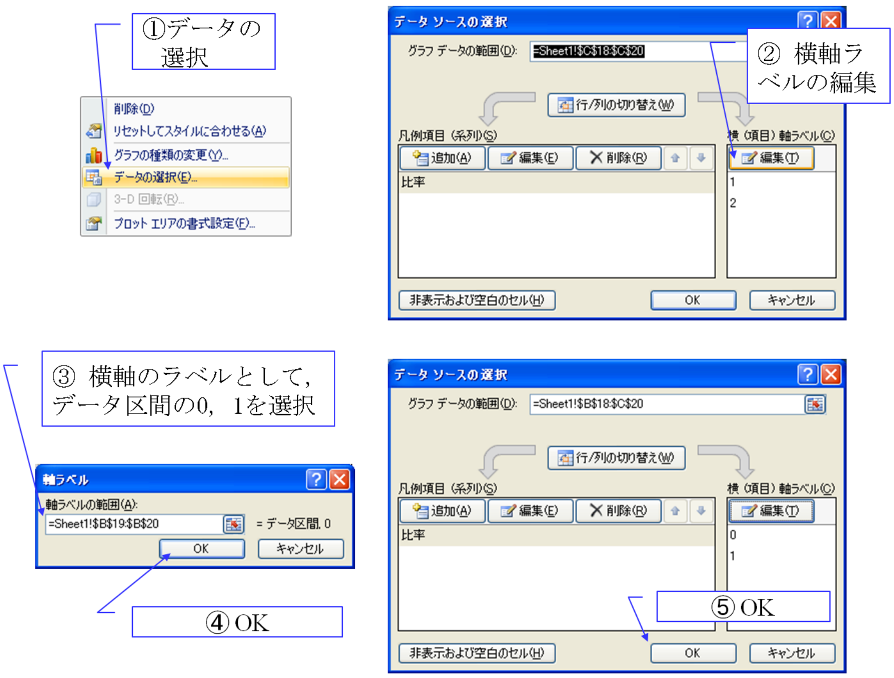


図 2.14: コイン投げ (n=10) グラフの横軸ラベルの設定手順

これは大数の法則のシミュレーションによる体験の一つである．

図 2.11 から図 2.15 への書き換えは次の手順による．まず，図 2.16 に示すように出力の比率を表示する場所を移動する．次に，図 2.17 に示すように，セル B7〜B16 のコインの表裏を出力する INT(2*RAND()) 関数を B7〜B1006 の範囲にコピーする．そして，セル B4 の COUNT(B7:B16) を COUNT(B7:B1006) としてコイン投げの回数 n をカウントす

る範囲を拡げる．同様に，セル F19, F20 のカウント範囲も拡げる．これにより，同図中の棒グラフが得られる．このままでは縦軸のラベルの範囲が自動調整されているために，必ずしも見易くはない．そこで縦軸のラベルを設定する．縦軸の数値の辺りを右クリックすると図 2.18 の左のメニューが現れる．ここで，軸の書式設定を選択し，次に現れる右の画面において軸のオプションの最小値を 0 に固定する．得られた棒グラフが図 2.15 中のものである．

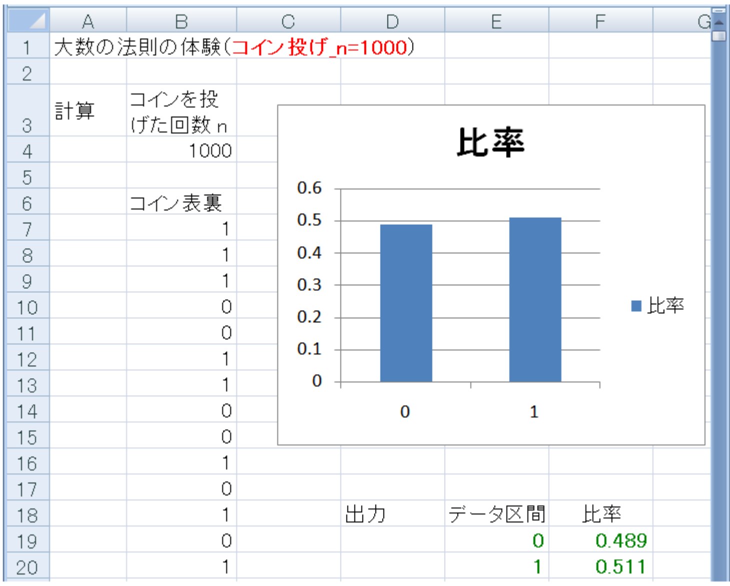


図 2.15:　コイン投げのシミュレーション (n=1000)(コイン投げのシミュレーション（n=1000）.xlsx)

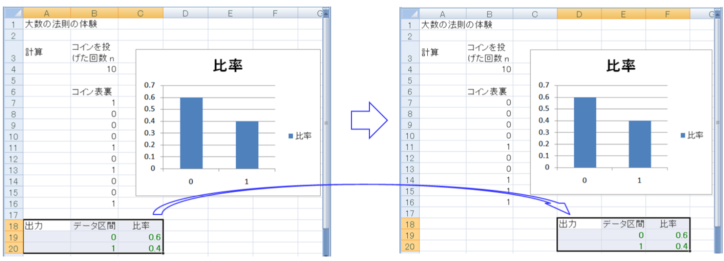


図 2.16: コイン投げ (n=1000) 出力範囲の移動

2.2.2　さいころ投げのシミュレーション

さいころ投げのシミュレーションを Excel を用いて実行する．1〜6 の各目の出る確率をいずれも 1/6 とし，各目の出る比率を棒グラフに表す．図 2.19 はその画面である．さ

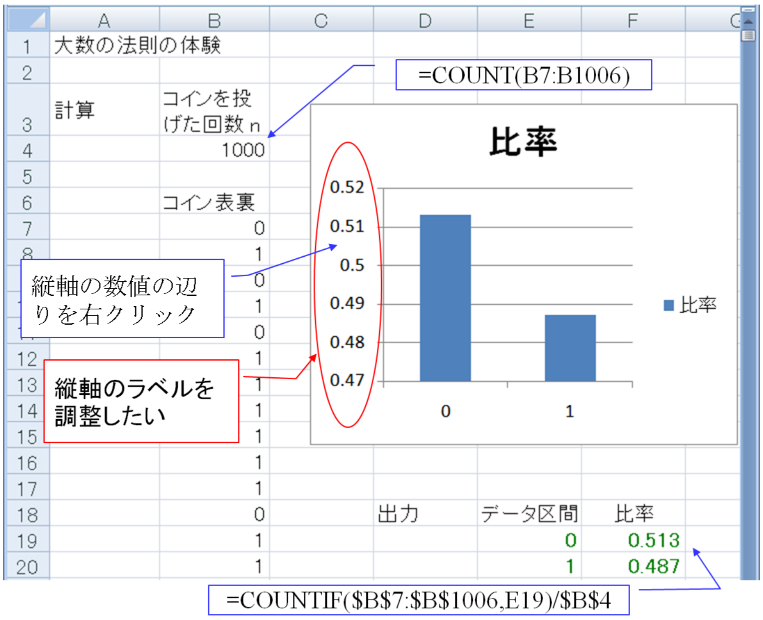


図 2.17: コイン投げ (n=1000) グラフ作成

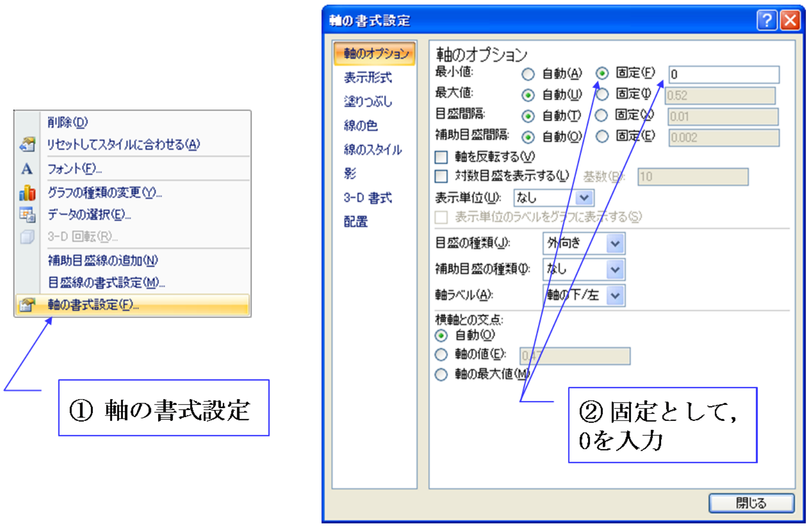


図 2.18: コイン投げ (n=1000) グラフの縦軸ラベルの設定手順

いころの目は

$$= \mathrm{INT}(6 * \mathrm{RAND}()) + 1 \tag{2.14}$$

により得る．この関数は，6*RAND() により $(0, 6)$ の範囲の擬似一様乱数を得て，INT() 関数により小数点以下を切り捨て，1 を足すことで，1〜6 の目をほぼ均等の確率で出力する．この関数を B7〜B5006 までコピーし，セル B4 にてその範囲のデータ数と，セル F19〜F24 にて各目の出現する比率を得ている． 試行回数が 5000 回と多いことで各目の比率が 1/6 に近い値となっている．

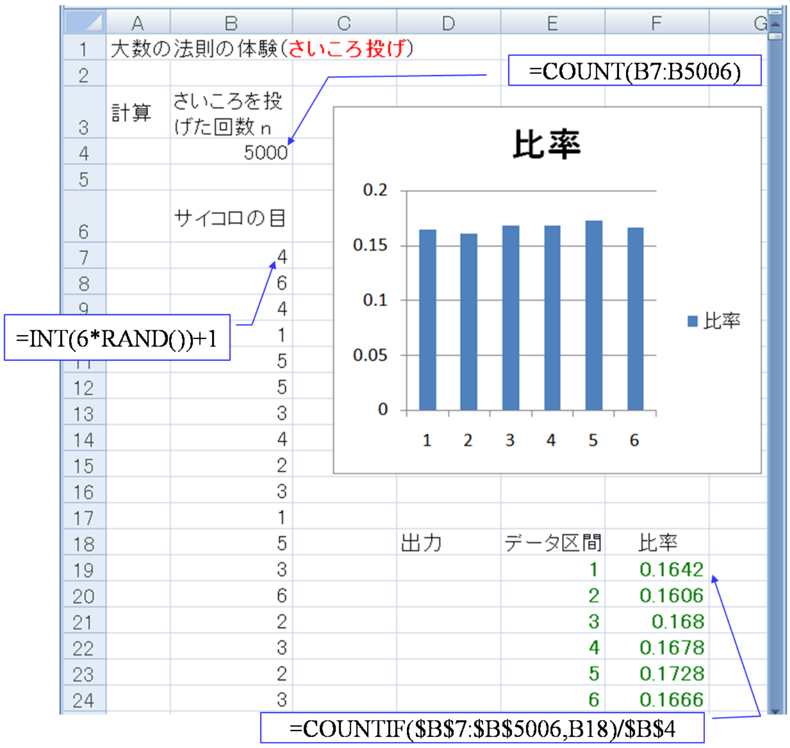


図 2.19: さいころ投げのシミュレーション (n=5000)(さいころ投げのシミュレーション（n=5000）.xlsx)

2.2.3　正規分布の確率密度関数と確率分布関数

正規分布の確率密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.15}$$

により与えられる．ここで，μ は平均，σ は標準偏差である．$\mu = 0, \sigma = 1$ とすると，標準正規分布の確率密度関数

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{2.16}$$

が得られる．標準正規分布の確率分布関数 $F(x)$ は，上式の $f()$ を用いると

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{2.17}$$

により与えられる．

標準正規分布の形状の作図の様子を図 2.20 に示す．まず，変数 x の値を入力する．セル B4 に x の最小値（ここでは-4 としている）を入力する．次に，セル B5 に関数

$$= \mathrm{ROUND(B4 + 0.1, 1)} \tag{2.18}$$

を入力する．ROUND(a, b) 関数は a の値の小数点以下 b+1 桁目を四捨五入する関数である．上式はセル B4 の値に 0.1 を足して，小数点以下 2 桁目を四捨五入する．この理由は以下の通りである．2 進数で小数を表す場合には丸め誤差が存在する．0.1 を足し続けていくと丸め誤差が累積されて，小数点以下 15 桁目にわずかなずれが現れてしまうことがあるからである．

10 進数の 0.1 を 2 進数で表現すると

$$0.1_{(10)} = 0.0001100110011\cdots_{(2)} \tag{2.19}$$

と循環小数で表される．添え字の (i) はその数字が i 進数であることを示す．2 進数の数字を 10 進数に換算するには

$$\begin{aligned} 0.000110011\cdots_{(2)} &= 2^{-4}\times 1 + 2^{-5}\times 1 + 2^{-8}\times 1 + 2^{-9}\times 1 + \cdots_{(10)} \\ &= 0.099609375 + \cdots_{(10)} \end{aligned} \tag{2.20}$$

とする．2 進数で $0.1_{(10)}$ を厳密に表すことはできない．

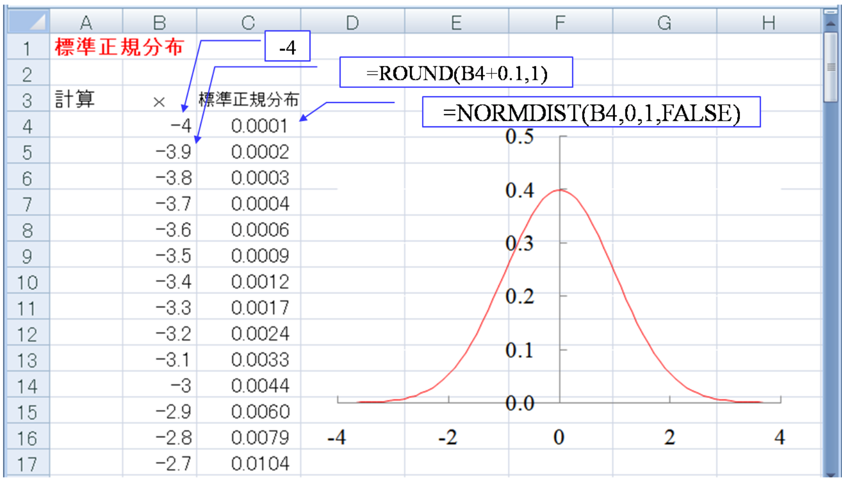


図 2.20: 標準正規分布の作図 (標準正規分布の作図.xlsx)

次に，標準正規分布の確率密度関数の値を求める関数を入力する．図 2.20 に示すように，セル C4 に

$$= \mathrm{NORMDIST}(\mathrm{B4}, 0, 1, \mathrm{FALSE}) \tag{2.21}$$

と入力している．NORMDIST(a, b, c, d) 関数は d が FALSE のとき，平均 b, 標準偏差 c の正規分布の $x = a$ における確率密度関数の値をを与える．$b = 0$, $c = 1$ であるので，式 (2.21) は B4 における標準正規分布の確率密度関数の値を与える．

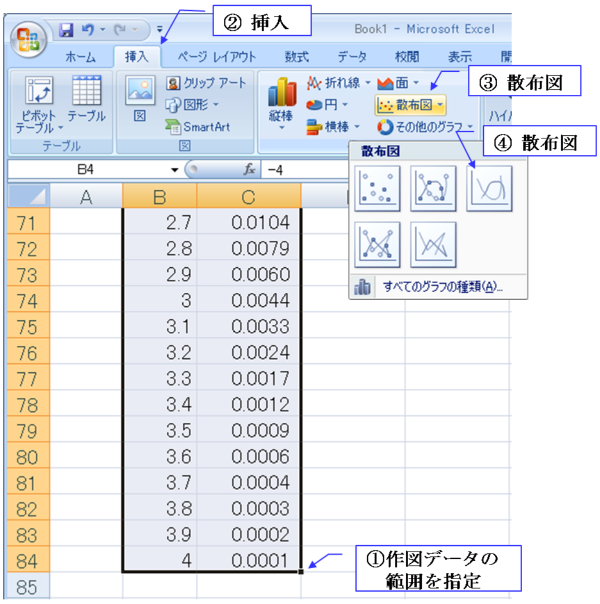


図 2.21: 標準正規分布の作図手順

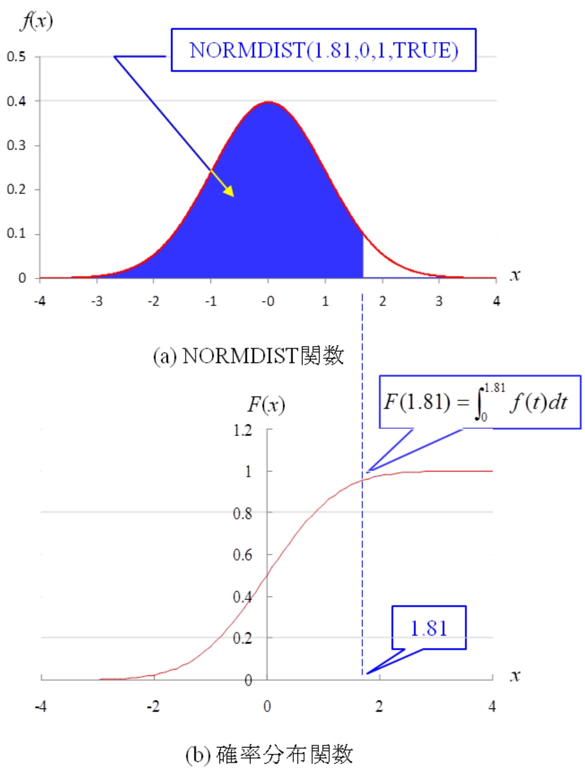


図 2.22: 標準正規分布の確率分布関数

グラフの作図手順を図 2.21 に示す．(1) 表示したいデータの範囲をカーソルで選択し，(2) 挿入，(3) 散布図、(4) 散布図（平滑線）と左クリックしていくことで実行できる．

標準正規分布の確率分布関数を与える NORMDIST$(x, 0, 1,$TRUE$)$ のイメージを図 2.22

に示す．式 (2.17) より，確率分布関数 $F(x)$ は確率密度関数 $f(x)$ の積分値であるので，同図 (a) の確率密度関数と $x = 1.81$ の直線と x 軸で囲まれた塗りつぶされた領域の面積が $x = 1.81$ のときの確率分布関数の値である．同図 (b) はこの面積値 $F(x)$ を縦軸にとったグラフ（確率分布関数）を示す．

2.2.4　正規分布のシミュレーション

図 2.23:　標準正規分布の再現のシミュレーション (n=5000)（標準正規分布の再現（n=5000）.xlsx）

標準正規分布に従う乱数（標準正規乱数）を生成して標準正規分布を再現するシミュレーションを行う．標準正規乱数は標準正規分布の確率密度関数に基づく乱数であり，大数の法則によると，この乱数を生成して出現頻度分布を求めると，生成する乱数の数を増やすにつれて出現頻度分布は標準正規分布に近づく．

図 2.23 はそのシミュレーション画面である．標準正規乱数は小数点以下を四捨五入して整数値で生成している．各整数値の出現回数を求めて，全データ数で割ることで出現頻度を求めている．標準正規乱数は

$$= \mathrm{NORMSINV(RAND())} \tag{2.22}$$

により生成している．NORMSINV() 関数は標準正規分布の確率分布関数の逆関数を与える．

RAND() は (0, 1) の区間の擬似一様乱数を与えるので，この乱数を入力とする逆関数の出力は擬似標準正規乱数となる．この様子を図 2.24 に示す．図中の上の曲線は標準正規分布の確率分布関数である．RAND() 関数の出力は縦軸の $F(x)$ の値として与えられる．NORMSINV() 関数の出力は横軸の x の値である．RAND() 関数の出力は $(0, 1)$ 区間にほぼ一様に分布している．0.5 付近の RAND() の値に対する NORMSINV() 関数の値は 0 付近に集中し，0 付近の x の出現頻度は高くなる．RAND() が 0 もしくは 1 に近い辺りでは，NORMSINV() は広い範囲に分散し，絶対値の大きな x の出現頻度は低くなる．理想的な出現頻度を縦軸にとって同図の下に示す．これは標準正規分布の確率密度関数である．擬似一様乱数の入力に対する出力は擬似標準正規乱数となる．

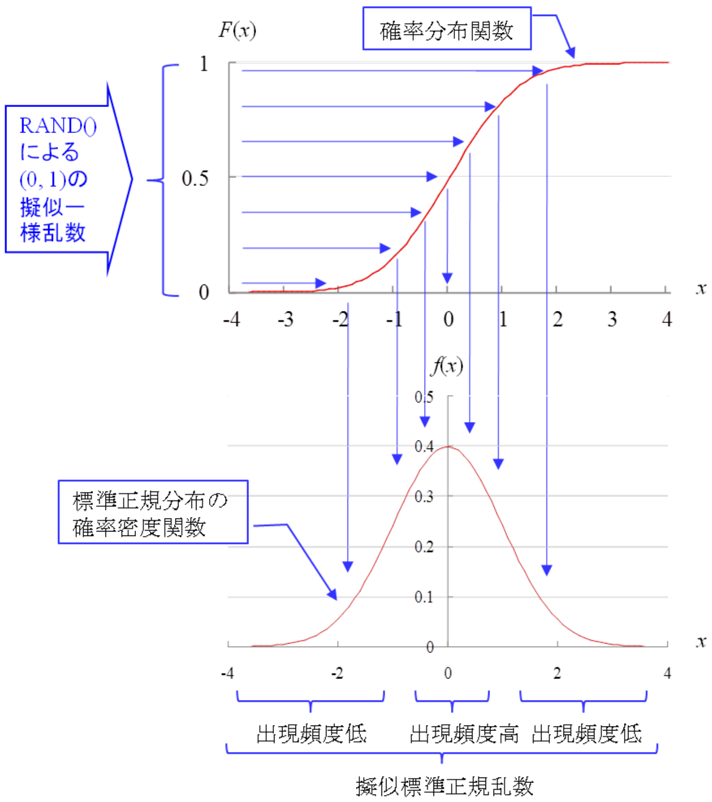


図 2.24: 確率分布関数の逆関数演算による擬似標準正規乱数の生成

図 2.23 では

$$= \mathrm{ROUND}(\mathrm{NORMSINV}(\mathrm{RAND}()), 0) \tag{2.23}$$

とすることで，標準正規乱数の小数点以下を四捨五入している．同図では各整数値の出現頻度を棒グラフに表示している．得られたグラフは正規分布に近い形状をしている．こ

の棒グラフに正規分布のグラフを並べて表示すれば，頻度分布と正規分布とを対比して見ることができる．図 2.25 は対比結果である．セル G19 では

$$= \mathrm{NORMSDIST}(\mathrm{E19} + 0.5) - \mathrm{NORMSDIST}(\mathrm{E19} - 0.5) \tag{2.24}$$

により，-3 ± 0.5 の範囲の確率を求めている．NORMSDIST(x) 関数は標準正規分布において x より小さな値の確率を出力する．以下のセルにはセル G19 の内容をコピーするだけでよい．なお，セルの表示桁数を図のように小数点以下 4 桁に設定するには，まず桁数を設定したいセルの範囲を選択し，その範囲内にカーソルを持ってきて右クリックすると，図 2.26 の左のメニューが現れる．(1) セルの書式設定，(2) 表示形式，(3) 数値，(4)-1234...，(5) 小数点以下の桁数を 4 に設定すればよい．次に，既作成のグラフに標準正規分布の棒グラフを追加する．グラフエリア内を右クリックすると，図 2.27 のメニューが現れる．(1) データの選択、(2) 追加，(3) 系列名を「標準正規分布」とし，(4) データの範囲を指定し，(5) ＯＫにより図 2.25 の図が得られる．ダウンロードサイトにある Excel ファイル（標準正規分布との対比.xlsx）では棒グラフがカラーで表示されている．このファイル内の赤の棒グラフが標準正規分布の理論式に基づく確率分布である．青の棒グラフが概ね理論値に沿っていることが分かる．頻度分布を繰り返し再計算させてみると，青の頻度分布が理論値に近い分布をしていることを見て取ることができる．

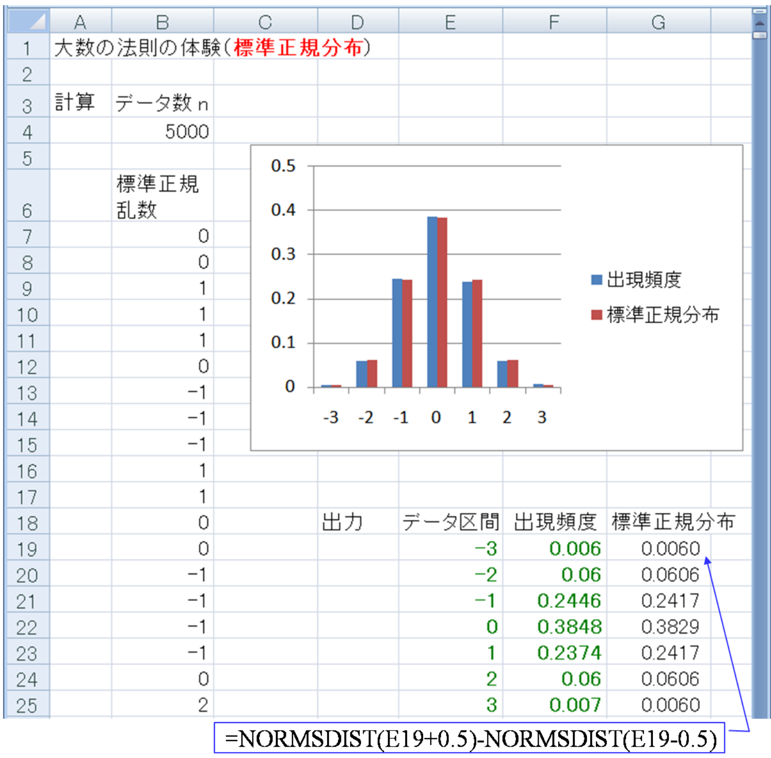


図 2.25: 標準正規分布との対比（標準正規分布との対比.xlsx）

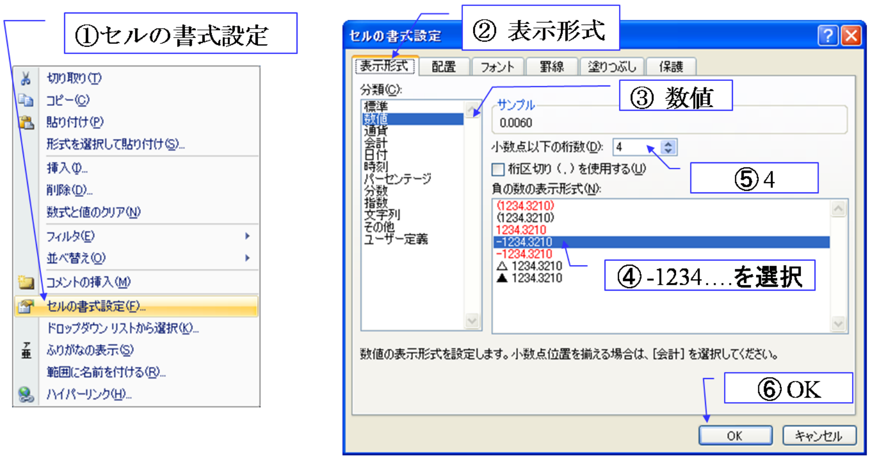


図 2.26: セルの表示桁数の設定

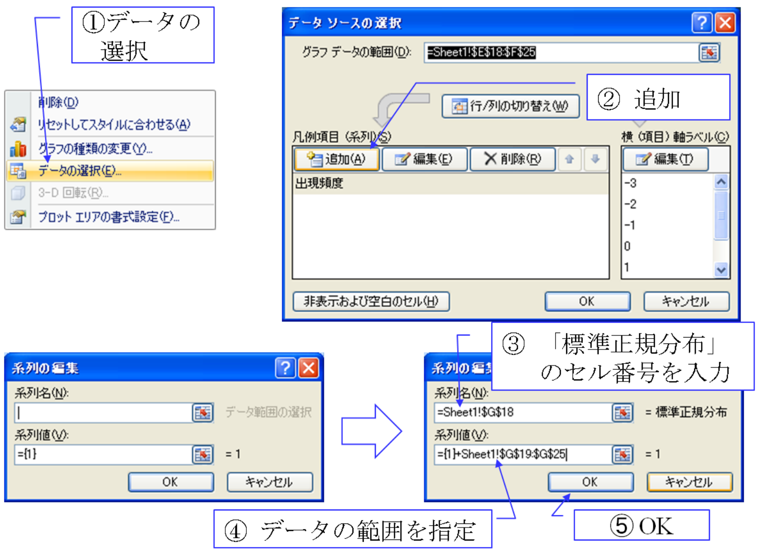


図 2.27: 標準正規分布との対比 棒グラフの追加手順

2.3　平均のシミュレーション (中心極限定理)

中心極限定理とは次のことをいう．

> 母集団の分布がどんな分布であっても，標本平均 $\bar{x}$ の分布は標本の大きさ n を増やしたとき近似的に正規分布に従う．

すなわち，コインを n 回投げたときに表の出る回数を q とすると，標本平均 $\bar{x} = q/n$ の分布は n を大きくすると正規分布に近づく．

2.3.1 コイン投げのシミュレーション

中心極限定理を Excel のシミュレーションにより体験できる．図 2.28 においてコインを $n=2$ 回投げ，表の出た回数を q として，比率 q/n を求めることを 1 組とする．第 i 組において表の出た回数を q_i とし，比率を $\overline{x}_i = q_i/n$ と表す．図 2.28 では $N=5000$ 組についてそれぞれの比率 $\overline{x}_i, i=1,2,\cdots,N$ を求めている．各組の比率 $\overline{x}_i$ は AVERAGE() 関数により求めている．セル B10 の AVERAGE(B7:B8) はセル B7, B8 の数値の平均値を出力する．第 i 組の j 番目のコインの表裏の値を x_{ij} とすると，第 i 組の表の比率 $\overline{x}_i$ は

$$\overline{x}_i \;=\; \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{2.25}$$

と表される．セル B12 では 5000 組の比率の平均値 $\overline{x}$ を求めている．この平均値 $\overline{x}$ は以下の式により与えられる．

$$\overline{x} \;=\; \frac{1}{N}\sum_{i=1}^{N} \overline{x}_i \tag{2.26}$$

セル C12 では不偏分散 v_e^2 を求めている．VAR(B10:GJI10) 関数は次式で与えられる不偏分散 v_e^2 を出力する．

$$v_e^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N} (\overline{x}_i - \overline{x})^2 \tag{2.27}$$

図 2.29 は 5000 組において表の出た比率の出現頻度および正規分布に基づく確率分布を求めている画面の抜粋である．セル C15 ではセル B10〜BJI10 の 5000 個のセルの中から，セル B15 で指定された値 (=0) の出現回数を数え，セル C4 の全組数 (=5000) で割ることで，0 の出現頻度を求めている．

確率分布を求める際に，NORMDIST(x,μ,σ,TRUE) と設定すれば式 (2.17) の $f()$ に式 (2.15) を用いた場合の値を出力できる．ここでは，μ,σ の代わりにセル B12, C12 で求めた 5000 組の比率の平均値 $\overline{x}$ と不偏分散 v_e^2 の平方根 $\sqrt{v_e^2}$ を用いて，NORMDIST($x+0.25,\overline{x},\sqrt{v_e^2}$,TRUE)-NORMDIST($x-0.25,\overline{x},\sqrt{v_e^2}$,TRUE) により $x\pm 0.25$ の区間の確率を得ている．なお，TRUE は NORMDIST(x,μ,σ,TRUE) が確率分布関数の x における値を出力することを指定するパラメータである．これを FALSE とすると，x における確率密度関数の値の出力を指定する．図 2.30 は得られた棒グラフである．シミュレーションを繰り返すと，ほとんどのときに，表の出た比率 $\overline{x}_i=0$ および 1 の出現頻度は対応する正規分布に基づく確率より大きいこと，また，$\overline{x}_i=0.5$ の出現頻度は対応する正規分

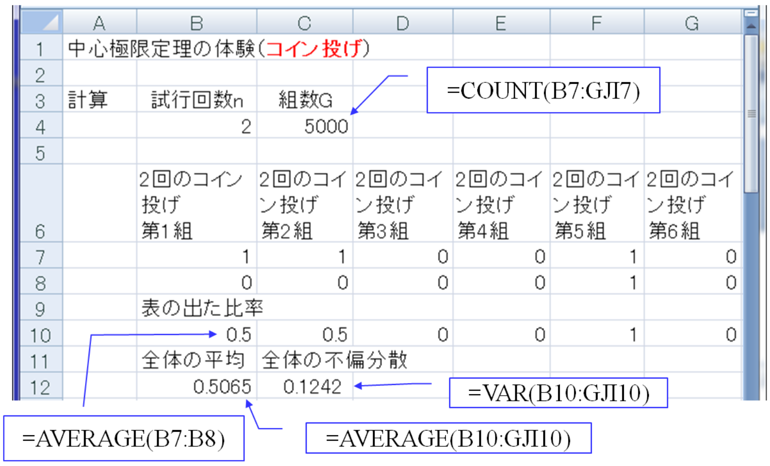


図 2.28: コイン投げのシミュレーション（中心極限定理の体験, n=2）（コイン投げ（中心極限定理）n=2.xlsx）

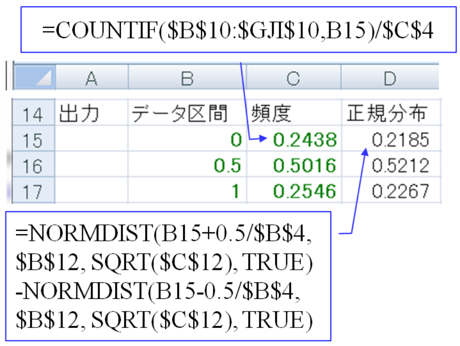


図 2.29: コイン投げ（頻度/確率分布の計算, n=2）（コイン投げ（中心極限定理）n=2.xlsx）

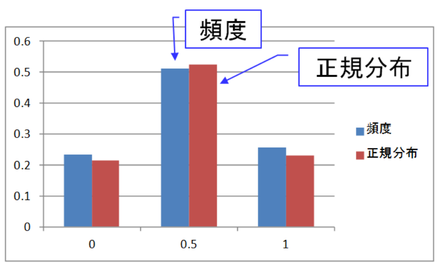


図 2.30: コイン投げ（頻度/確率分布，n=2）

布に基づく確率より小さいことを確認できる．標本の大きさ n が小さい場合は，標本平均の分布は正規分布とは異なっている．

図 2.31～図 2.33 は標本の大きさ n=10 として，5000 組において表の出た比率 $\bar{x}_i$ の出現頻度および正規分布に基づく確率分布を求めている画面の抜粋である．シミュレーションを繰り返すと，表の出た比率の出現頻度は正規分布に基づく確率の値の周辺にばらつくことが分かる．標本を n=10 と大きくしたことで，表の出た比率の出現頻度は正規分

布に近づいている．これは中心極限定理のシミュレーションによる体験の一つである．

	A	B	C	D	E	F	G
1	中心極限定理の体験(コイン投げ)						
2							
3	計算	試行回数n	組数G				
4		10	5000				
5							
6		10回のコイン投げ第1組	10回のコイン投げ第2組	10回のコイン投げ第3組	10回のコイン投げ第4組	10回のコイン投げ第5組	10回のコイン投げ第6組
7		0	0	1	1	1	1
8		1	1	1	0	1	1
9		1	1	0	1	1	1
10		0	0	1	0	1	1
11		1	0	1	0	0	0
12		1	1	1	1	1	1
13		1	1	0	1	0	0
14		1	1	1	1	1	0
15		1	1	1	1	1	0
16		0	0	0	0	0	0
17		表の出た比率					
18		0.7	0.6	0.7	0.6	0.7	0.5
19		全体の平均	全体の不偏分散				
20		0.49816	0.0249				

=VAR(B18:GJI18)

=AVERAGE(B7:B16)

=AVERAGE(B18:GJI18)

図 2.31: コイン投げのシミュレーション（中心極限定理の体験, n=10）（コイン投げ（中心極限定理）n=10.xlsx）

	A	B	C	D
22	出力	データ区間	頻度	正規分布
23		0	0.0008	0.0020
24		0.1	0.0084	0.0114
25		0.2	0.0466	0.0442
26		0.3	0.1188	0.1160
27		0.4	0.216	0.2062
28		0.5	0.2284	0.2486
29		0.6	0.212	0.2033
30		0.7	0.1176	0.1127
31		0.8	0.0412	0.0424
32		0.9	0.0092	0.0108
33		1	0.001	0.0019

図 2.32: コイン投げ（頻度/確率分布の計算, n=10）（コイン投げ（中心極限定理）n=10.xlsx）

2.3.2 指数分布のシミュレーション

確率分布が平均値の左右で非対称である例として，指数分布を取りあげる．まず，この分布のシミュレーションを図 2.34〜図 2.36 に示す．指数分布の確率密度関数は

$$f(x) = \lambda e^{-\lambda x} \quad (x \geq 0) \tag{2.28}$$

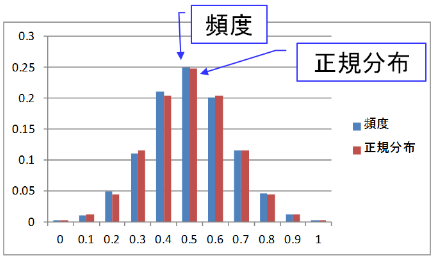


図 2.33: コイン投げ（頻度/確率分布，n=10）

と与えられる．この関数を $[0,\ x]$ の区間について積分することで，次式の確率分布関数 $F(x)$ が得られる．

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda x} \\ &= \left[-e^{-\lambda x}\right]_0^x \\ &= 1 - e^{-\lambda x} \end{aligned} \tag{2.29}$$

図 2.24 の疑似標準正規乱数の生成法と同様にして疑似指数乱数を生成する．すなわち，$F(x)$ の値として，RAND() 関数により (0, 1) 区間の疑似一様乱数を与える．

$$\mathrm{RAND}() = 1 - e^{-\lambda x} \tag{2.30}$$

次に，$F(x)$ の逆関数により疑似指数乱数を得る．

$$x = -\frac{1}{\lambda} ln(1 - \mathrm{RAND}()) \tag{2.31}$$

ここで，1-RAND() も (0, 1) 区間の疑似一様乱数なので，これを RAND() で置き換えても (0, 1) 区間の疑似一様乱数を得ることができる．よって，疑似指数乱数を生成する関数は次式となる．

$$x = -\frac{1}{\lambda} ln(\mathrm{RAND}()) \tag{2.32}$$

図 2.34 では，セル B7 において式 (2.32) の疑似指数乱数を生成している．ここではセル B4 にて $\lambda = 2$ としている．セル B9 では小数点以下 2 桁目を切り捨てている．このエクセルの画面では，行方向に 5000 組（ただし，1 組には 1 個の指数乱数しか入っていない）の指数乱数を生成している．図 2.35 のセル C14 ではセル B9〜GJI9 の 5000 個のセルの中から 0 の値の出現頻度を求めている．図 2.36 は得られた頻度分布を示す．横軸の値が 0 の棒グラフは [0, 0.1) の区間の値の出現頻度を表している．

図 2.37〜図 2.39 は標本の大きさ n=4 として，5000 組における出現頻度分布と正規分布の対比を示している．4 個の指数関数値の平均値は小数点以下 2 桁目を四捨五入してい

	A	B	C	D	E	F	G
1	指数分布のシミュレーション						
2							
3	計算	λ	試行回数n	組数G			
4		2	1	5000			
5							
6		1個の指数乱数第1組	1個の指数乱数第2組	1個の指数乱数第3組	1個の指数乱数第4組	1個の指数乱数第5組	1個の指数乱数第6組
7		0.42835776	0.223582	0.600337	0.455578	0.69152	0.166551
8		小数点以下第2位を切り捨て					
9		0.4	0.2	0.6	0.4	0.6	0.1
10		全体の平均	全体の不偏分散				
11		0.44182	0.2457				

= - (1/B4)*LN(RAND())

=ROUNDDOWN(B7,1)

図 2.34: 指数分布のシミュレーション（指数分布のシミュレーション.xlsx）

	A	B	C
13	出力	データ区間	頻度
14		0	0.1876
15		0.1	0.152
16		0.2	0.1238
17		0.3	0.0968
18		0.4	0.0812
19		0.5	0.0654
20		0.6	0.058
21		0.7	0.046
22		0.8	0.0342
23		0.9	0.0286
24		1	0.0238

図 2.35: 指数分布（頻度分布の計算, n=10）（指数分布のシミュレーション.xlsx）

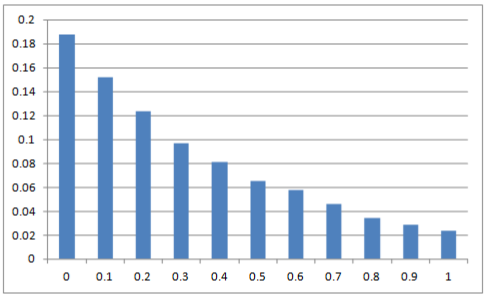


図 2.36: 指数分布（頻度分布）

る．データ区間は 0〜1 の間を 0.1 刻みとし，それぞれの値の出現頻度を求めている．図 2.39 より，n=4 の場合は，平均値の出現頻度分布は正規分布とは大きくずれていることが分かる．

図 2.40 は標本の大きさ n=100 として，5000 組における出現頻度分布と正規分布の対比を示している．標本を大きくすることで頻度分布が正規分布に近づいている様子が分かる．

	A	B	C	D	E	F	G
1	中心極限定理の体験（指数分布）						
2							
3	計算	λ	試行回数n	組数G			
4		2	4	5000			
5							
6		4個の指数 乱数 第1組	4個の指 数乱数 第2組	4個の指 数乱数 第3組	4個の指 数乱数 第4組	4個の指 数乱数 第5組	4個の指 数乱数 第6組
7		0.83221285	0.002913	0.375626	0.719842	0.02108	0.131087
8		0.35067523	0.596938	0.613515	0.343544	0.88907	0.16808
9		0.45997919	0.169374	0.3179	0.679007	0.57458	0.14775
10		0.24161403	1.062557	0.578632	0.518291	1.94294	0.013759
11		平均値					
12		0.5	0.5	0.5	0.6	0.9	0.1
13		全体の平均	全体の不偏分散				
14		0.49412	0.0607				

=ROUND(AVERAGE(B7:B10),1)

図 2.37: 指数分布のシミュレーション（中心極限定理の体験, n=4）（指数分布（中心極限定理）n=4.xlsx）

	A	B	C	D
16	出力	データ区間	頻度	正規分布
17		0	0.001	0.0219
18		0.1	0.0306	0.0452
19		0.2	0.106	0.0791
20		0.3	0.1698	0.1178
21		0.4	0.1816	0.1492
22		0.5	0.1638	0.1606
23		0.6	0.1208	0.1470
24		0.7	0.0852	0.1145
25		0.8	0.0574	0.0758
26		0.9	0.0318	0.0427
27		1	0.0204	0.0204

=NORMDIST(B17+0.05, B14, SQRT(C14), TRUE)
-NORMDIST(B17-0.05, B14, SQRT(C14), TRUE)

図 2.38: 指数分布（頻度/確率分布の計算, n=4）（指数分布（中心極限定理）n=4.xlsx）

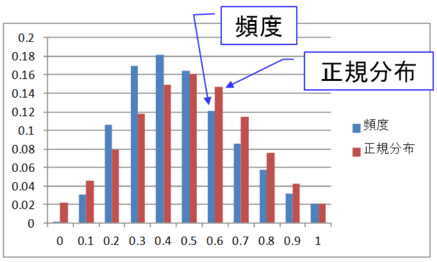


図 2.39: 指数分布（頻度/確率分布，n=4）

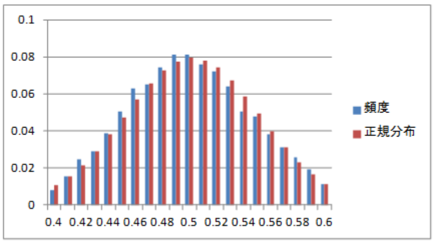


図 2.40: 指数分布（頻度/確率分布，n=100）（指数分布（中心極限定理）n=100.xlsx）

2.4　不偏分散のシミュレーション

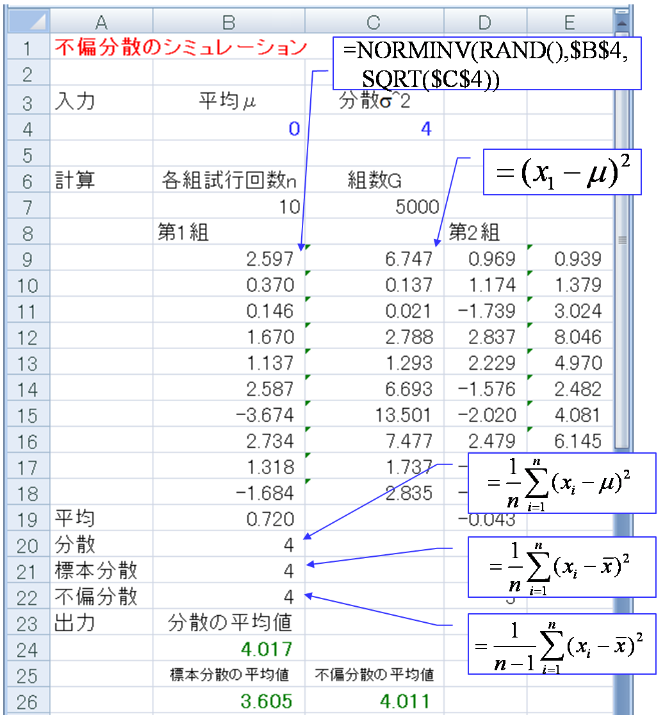
さて，コイン投げやさいころ投げにより得られる値を $x_i, i = 1, 2, \cdots, n$ とする．この値の不偏分散 v_e^2 と標本分散 s^2 は

$$
\begin{aligned}
v_e^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2
\end{aligned} \tag{2.33}
$$

と表される．v_e^2 と s^2 の違いは $n-1$ で割るか n で割るかである．不偏分散はその期待値が母分散に等しくなるような補正がなされている．標本分散の期待値は母分散よりは小さな値にずれる．このため標本から分散を求める演算には不偏分散が用いられる．Excel の VAR() 関数は式 (2.27) にて示したように不偏分散の値を出力する．ここで母分散ということばが初めて登場した．統計解析においては母集団を規定するパラメータである母平均，母分散の検定・推定を行うことが基礎にある．母集団とは調査対象となるデータの全体をいう．データは数値データであったり，属性などのカテゴリデータであったりするが，本書では数値データを扱う．母平均，母分散は母集団の平均値，分散値である．通常これらの値は未知であり，母集団から抽出された標本の平均値，分散値をもとに母平均，母分散の検定・推定がなされる．

本節ではシミュレーションにより，不偏分散 v_e^2 の期待値が標本分散 s^2 の期待値よりも母分散に近い値となることを確かめる．図 2.41 は標本分散，不偏分散の分布をシミュレーションにより求めている例である．セル B9〜B18 では平均 $\mu = 0$，分散 $\sigma^2 = 4$ の正規分布に基づく乱数（正規乱数）を 10 個生成している．ここで，$\mu = 0$，$\sigma^2 = 4$ は標本の母集団の分布を規定するパラメータであり，それぞれ母平均，母分散と呼ばれる．セル B9〜B18 の値は，母集団から得られた標本である．セル C9〜C18 では母平均 $\mu = 0$ が分かっているとして，各乱数値 x_i と母平均との差の 2 乗値 $(x_i - \mu)^2$ を求め，セル B20 でセル C9〜C18 の平均値を求め，小数点以下を四捨五入している．ここでは，これを分散とよんでいる．セル B21 で標本分散，セル B22 で不偏分散を求め，それぞれ小数点以下

を四捨五入している．以上 10 個の正規乱数を用いた計算結果を一組として，5000 組について同じ計算を繰り返し，各分散について平均値を求めたところ，セル B24 の分散の平均値とセル C26 の不偏分散の平均値が母分散 $\sigma^2 = 4$ に近い値となった．一方，セル B26 の標本分散の平均値は母分散よりも小さな値となった．組数を 5000 組としたことで不偏分散の平均値は期待値に近づき，これが母分散に近い値となった．図 2.42 は，20 行目の分散値，21 行目の標本分散値，22 行目の不偏分散値の出現頻度を棒グラフに表示した結果である．分散と不偏分散の棒グラフが類似の傾向を示し，標本分散のグラフだけが他のグラフと比べて分散値の小さな方へと分布の偏りが見られる．



	A	B	C	D	E
1	不偏分散のシミュレーション				
2					
3	入力	平均μ	分散σ^2		
4		0	4		
5					
6	計算	各組試行回数n	組数G		
7		10	5000		
8		第1組		第2組	
9		2.597	6.747	0.969	0.939
10		0.370	0.137	1.174	1.379
11		0.146	0.021	−1.739	3.024
12		1.670	2.788	2.837	8.046
13		1.137	1.293	2.229	4.970
14		2.587	6.693	−1.576	2.482
15		−3.674	13.501	−2.020	4.081
16		2.734	7.477	2.479	6.145
17		1.318	1.737		
18		−1.684	2.835		
19	平均	0.720		−0.043	
20	分散	4			
21	標本分散	4			
22	不偏分散	4			
23	出力	分散の平均値			
24		4.017			
25		標本分散の平均値	不偏分散の平均値		
26		3.605	4.011		

図 2.41: 不偏分散のシミュレーション（不偏分散のシミュレーション.xlsx）

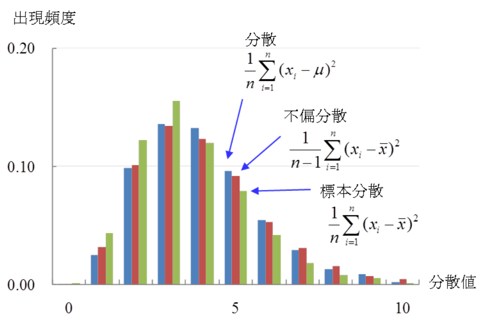


図 2.42: 分散の平均値の出現頻度の棒グラフ

参考文献

[1] B. A. Wichmann and I. D. Hill, "Algorithm AS 183: An Efficient and Portable Pseudo-Random Number Generator." Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 31, No. 2, pp. 188-190, 1982.

[2] B. A. Wichmann and I. D. Hill, "Correction: Algorithm AS 183: An Efficient and Portable Pseudo-Random Number Generator." Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 33, No. 1, p. 123, 1984.

索引

著者
古橋　武
名古屋大学工学研究科計算理工学専攻

本稿の内容は，

古橋武・宮本定明著
「統計・多変量解析とソフトコンピューティング —超多自由度系解析を目指して—」
金田・笹井監修，計算科学講座 第 3 巻，共立出版, 2012
http://www.kyoritsu-pub.co.jp/bookdetail/9784320122680