11　正規分布

実験値の精度や誤差について考える場合に正規分布と呼ばれる確率分布が重要である。ここでは，乱数を用いたシミュレーションによって正規分布の性質について学ぶ。

11.1　一様乱数

ある測定を繰り返し行ったとき，偶然生じる様々な要因のために測定値は完璧に一致した値にはならない。ここでは，そのような偶然誤差のモデルとして乱数を用いることにする。無秩序に並んだバラバラの数列を「乱数」random number という。

Excel 2007

Excel には [0,1) (つまり $0 \le x < 1$) の範囲の一様乱数を発生させる関数「RAND」が用意されている。これを用いて 12000 の乱数を M3:X1002 の範囲に発生させる

1. Web ページから説明用ファイルをダウンロードし，Excel で開く。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1								この行まで	n=3	n=10			
2							行番号	平均	平均	平均	基準値		乱数
3		母集団	n=3の平均	n=10の平均	理想値		1						
4	n	12000	3	10			2						
5	平均						3						
6	分散						4						
7	母集団分散/n						5						

2. ワークシート名を「setsumei-11」から「説明」に変更する。
3. 課題提出用のファイル名で，Excel 形式で保存する。
4. 乱数の発生。
   (i) M3 に「=RAND()」と入力。
   (ii) M3 をクリップボードにコピーし，M3:X1002 の範囲に貼り付ける。
   - 関数「RAND」には引数はないがカッコは必要である。
   - ワークシートのどこかでセル入力があるたびに，乱数は再発生し値が変わる。
   - 最も簡単には，空のセルを選択した状態で Delete を押すと再計算される。
5. この乱数の分布を見る。
   (i) B5 に 12000 個の乱数の平均を計算する(「=AVERAGE(M3:X1002)」，計算に時間がかかるかも知れない)。
   (ii) B6 に 12000 個の乱数の分散を計算する(「=VARP(M3:X1002)」)。この分散は偏差二乗和をデータ数そのもので割った標本分散である。
   (iii) 度数分布計算ため，右の部分を利用する。
   (iv) C11:C32 の範囲を選択した状態で
   「=FREQUENCY(M3:X1002,A11:A31)
   /COUNT(M3:X1002)/0.05」
   確定するときは Ctrl+Shift+Enter。
   - 度数分布を全データ数で割ると，その範囲の値をとる確率が出る。
   - さらにひとつの範囲の幅 0.05 で割ることによって「確率密度」が出る。

	A	B	C	D	E	F
10	範囲	値	母集団	n=3	n=10	正規分布
11	0	0				
12	0.05	0.025				
13	0.1	0.075				
14	0.15	0.125				
15	0.2	0.175				
16	0.25	0.225				
17	0.3	0.275				
18	0.35	0.325				
19	0.4	0.375				
20	0.45	0.425				
21	0.5	0.475				
22	0.55	0.525				
23	0.6	0.575				
24	0.65	0.625				
25	0.7	0.675				
26	0.75	0.725				
27	0.8	0.775				
28	0.85	0.825				
29	0.9	0.875				
30	0.95	0.925				
31	1	0.975				
32		1				
33						
34				分布の総和		
35				x>0.6		
36				確率		

(v) B10:C32 の四角い範囲を選択。

(vi) メニュー [挿入] → [グラフ] →「散布図」で，形式は右下折れ線のみを選択 → [完了]。

(vii) X 軸にマウスを合わせダブルクリックし「目盛」で「最小値」を「0」に，「最大値」を「1」にして [OK]。

(viii) Y 軸にマウスを合わせダブルクリックし「目盛」で「最小値」を「0」に，「最大値」を「6」にして [OK]。

理想的な一様乱数は，平均 0.5，分散 1/12 = 0.08333··· となり，ここのやり方で分布図を書いた場合，0 と 1 以外の所はすべて 1 になる。

11.2 大数の法則

Excel で発生させた 12000 個の一様乱数について実際に求めた平均と分散は，理想的なものとは多少異なっている。何回か再計算しても様子は大して変わらないはずである。計算機で作り出される乱数は，統計的な意味で真にランダムではないので，「疑似乱数」と呼ばれる。ただし，平均と分散が理想値でないのは，このような「装置上」の問題の他に，12000 個という有限個のデータしか計算していないという，より本質的な問題に由来している。とはいえ，12000 は非常に大きい数である。現実の測定を 12000 回繰り返すことは，まずない。しかし，統計的にものを考えるためには，非常に大きい数を取り扱うことが大事である。その感覚をつかむために，今 12 個ずつ並んでいる乱数の 1 行目だけの平均，1〜2 行目の平均，1〜3 行目，···，という風に，1〜1000 行目までの平均をとり，データの数によって平均がどのように変化するかを確かめる。

Excel 2007

データ数による平均値の変化

1. G3 に「1」，H3 に「=AVERAGE($M$3:X3)」と入力。
2. G4 に「=G3+1」と入力。
3. H3 をコピーし H4 に貼り付け（H4 は「=AVERAGE($M$3:X4)」になったはず）。
4. G4:H4 の範囲をコピーし，G5:H1002 の範囲に貼り付け。
5. グラフを描くために G2:H1002 の範囲を選択。
6. メニュー [挿入] → [グラフ] →「散布図」で，形式は右下折れ線のみを選択 → [完了]。
7. X 軸上でマウスをダブルクリックし，「目盛」で「最小値」を「0」に，「最大値」を「1000」にして [OK]。
8. Y 軸上でマウスをダブルクリックし，「目盛」で「最小値」を「0.4」に，「最大値」を「0.6」にして [OK]。

データ数が少ないうちは平均値が 0.5 から大きくずれるが，データ数が大きくなると 0.5 に近づいていく。何度か再計算しても同様の結果になるはずである。一般に，繰り返し回数（データ数）が多いほど，分布は本来の分布に近づく。これを「大数の法則」という。

11.3 中心極限定理

原理的には無限に繰り返すことのできる測定も，手作業ならせいぜい数回繰り返すことができればよい方である。つまり，実際に得られるデータは，可能な無限のデータのうちホンの何個かを取り出したものに過ぎない。元のデータ全体を「母集団」population，実際に取り出した数個のデータを「標本」sample という。また，実際に取り出したデータの数を「標本数」または「サンプル数」という。これ以降，標本数を n と書く。ここでは，12000 個の乱数からなる母集団から n 個の標本をとりだすという操作（サンプリング）を繰り返し，n

個のサンプルの平均が示す性質を調べる。例として，$n = 3$ と $n = 10$ の場合を考える。

Excel 2007

標本平均の性質

1. I3 に「=AVERAGE(M3:O3)」と入力し，I3 をコピーし，I4:I1002 に貼り付け（各行のはじめ 3 個の平均）。
2. J3 に「=AVERAGE(M3:V3)」と入力し，J3 をコピーし，J4:J1002 に貼り付け（各行のはじめ 10 個の平均）。
3. C5 に「=AVERAGE(I3:I1002)」（“3 個の平均” の平均）。
4. C6 に「=VARP(I3:I1002)」（“3 個の平均” の分散）。
5. D5 に「=AVERAGE(J3:J1002)」（“10 個の平均” の平均）。
6. D6 に「=VARP(J3:J1002)」（“10 個の平均” の分散）。
7. 度数分布を求めるため，D11:D32 の範囲に
   「=FREQUENCY(I3:I1002,A11:A31)/COUNT(I3:I1002)/0.05」
   確定は Ctrl+Shift+Enter（次も同じ）。
8. E11:E32 の範囲に「=FREQUENCY(J3:J1002,A11:A31)/COUNT(J3:J1002)/0.05」。
9. 母集団の分布のグラフを消し B10:E32 の四角い範囲でグラフを書き直す。

母集団の中から数個のサンプルをとりだしたものでも，何回もサンプリングを繰り返し，平均の平均をとれば母集団の平均に近い数値が得られる。ただし，本来は無作為にサンプルの抽出を行わなければならないが，ここの例では厳密な無作為抽出にはなっていない。

平均の分散をとると，もうひとつの重要な性質が見えてくる。

Excel 2007

中心極限定理の確認

1. C7 に「=$B$6/C4」。
2. C7 をコピーし D7 に貼り付けると「=$B$6/D4」になる。

母集団の分散を標本数 n で割ると，標本平均の分散と一致する。ただし，この場合は，たかだか 1000 回のサンプリングなので，「ほぼ一致する」。

平均 μ，分散 σ^2 の母集団から n 個の標本を採り n 個の標本に関する平均 $\langle x \rangle$ を求める操作を繰り返す。このとき，n 個の標本に関する平均 $\langle x \rangle$ は，平均 μ，分散 σ^2/n の正規分布にしたがう。これを「中心極限定理」という。

11.4 正規分布 normal distribution

それでは，正規分布とはどのような分布か。変数 x が x から $x + dx$ の微小区間中の値をとる確率を $P(x)dx$ と書くとする。確率密度関数 $P(x)$ の区間長 dx 倍で確率がでる。$P(x)$ が次の関数で与えられるとき，x の確率分布は平均 μ，分散 σ^2 の正規分布であるという。正規分布はガウス分布 Gauss distribution とも呼ばれる。

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{11.1}$$

この関数は，次の式を満たす（規格化条件）。

$$\int_{-\infty}^{\infty} P(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = 1 \tag{11.2}$$

先ほどの $n = 10$ の標本平均の分布に合わせて正規分布のグラフを描いてみる。

Excel 2007

1. E5 に「0.5」，E6 に「=1/12」，E7 に「=E6/D4」と入力。
2. F11 に 「=1/SQRT(2∗PI()∗$E$7)∗EXP(−((B11−$E$5)^2)/2/$E$7)」と入力。
3. F11 をコピーし，F12:F32 に貼り付け。
4. 分布のグラフを消し B10:F32 の四角い範囲でグラフを書き直す。

正規分布で，x が a と b の間の値をとる確率 $P(a < x < b)$ は次のように計算できる。

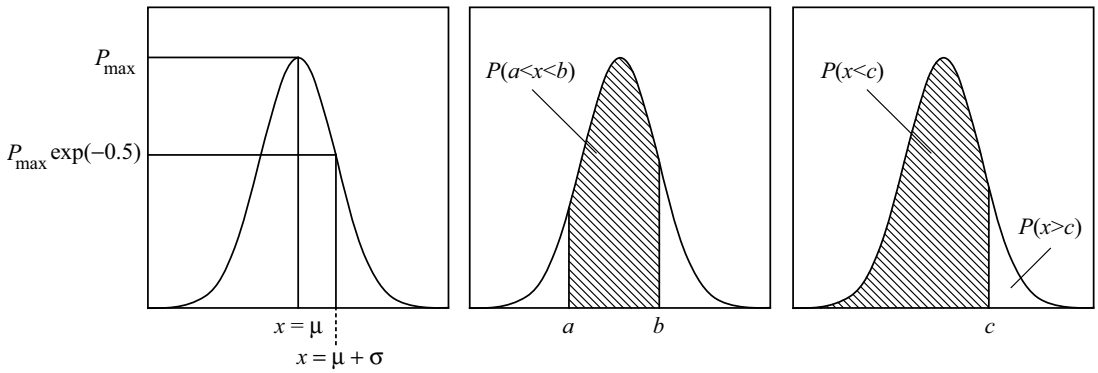
$$P(a < x < b) = \int_a^b P(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \tag{11.3}$$

x が $-\infty$ から c までの値をとる確率 $P(x < c)$ は累積確率と呼ばれ，次のように計算できる。

$$P(x < c) = \int_{-\infty}^c P(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^c \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \tag{11.4}$$

x が c から ∞ までの値をとる確率（上側確率）$P(c > x)$ は規格化条件と累積確率から計算できる。

$$P(x > c) = 1 - P(x < c) \tag{11.5}$$



先ほどの $n = 10$ の標本平均の分布で，平均値が 0.6 より大きい確率を求める。

Excel 2007

1. E34 に「=SUM(E11:E32)」。
2. E35 に「=SUM(E24:E32)」。
3. E36 に「=E35/E34」。
4. F36 に「=1−NORMDIST(0.6,E5,SQRT(E7),1)」。

   E36 が実測値から求めた確率，F36 が正規分布の場合の確率になる。

[Excel 関数の説明] 「NORMDIST」は正規分布の場合の累積確率を求める関数。引数は 4 つある。はじめは累積確率を求めたい x の値，次は正規分布の平均値 μ，3 番目は正規分布の標準偏差 σ（分散の平方根），4 番目は定数 1 を入れる。ちなみに 4 番目の引数を 1 ではなく 0 にした場合，累積確率ではなく正規分布の確率密度関数 $P(x)$ の値が計算される。

[Excel 関数の説明] Excel には，正規分布で累積確率がある値 p になるような x を求める関数「NORMINV」も用意されている。この関数は 3 つの引数をとる。はじめは，求めたい累積確率の値 p，2 番目は正規分布の平均 μ，最後は正規分布の標準偏差 σ である。

11.5 標準正規分布

x の確率分布が，平均 μ，分散 σ^2（標準偏差 σ）の正規分布であるとする。このとき，基準値 z を次のように定義する。

$$z = \frac{x-\mu}{\sigma} \tag{11.6}$$

基準値の確率分布は平均 0，分散 1（標準偏差 1 ）の正規分布になる。このような正規分布を「標準正規分布」という。式を書くと次のようになる。

$$P(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \tag{11.7}$$

先ほどの $n = 10$ の標本平均について，基準値を求めて分布を描き，標準正規分布のグラフと重ねる。

Excel 2007

1. K3 に「=(J3−$E$5)/SQRT($E$7)」と入力し，K3 をコピーして K4:K1002 の範囲にペースト（平均と標準偏差には理想値を用いた）。
2. 度数分布を求める準備のため，次のように入力する。ダウンロードしたファイルではすでに入力されている。

	A	B	C	D
39	基準値			
40	範囲	値	n=10	正規分布
41	−5	−5		
42	−4.5	−4.75		
43	−4	−4.25		
44	−3.5	−3.75		
45	−3	−3.25		
46	−2.5	−2.75		
47	−2	−2.25		
48	−1.5	−1.75		
49	−1	−1.25		
50	−0.5	−0.75		
51	0	−0.25		
52	0.5	0.25		
53	1	0.75		
54	1.5	1.25		
55	2	1.75		
56	2.5	2.25		
57	3	2.75		
58	3.5	3.25		
59	4	3.75		
60	4.5	4.25		
61	5	4.75		
62		5		

3. C41:C62 に「=FREQUENCY(K3:K1002,A41:A61)/COUNT(K3:K1002)/0.5」，Ctrl+Shift+Enter。
4. D41 に「=1/SQRT(2∗PI())∗EXP(−(B41^2)/2)」と入力。
5. D41 を D42:D62 にコピー & ペースト。
6. B40:D62 の四角い範囲を選択。
7. メニュー [挿入] → [グラフ] →「散布図」で，形式は右下折れ線のみを選択→ [完了]。
8. X 軸にマウスを合わせダブルクリックし「目盛」で「最小値」を「-5」に，「最大値」を「5」に，「Y/数値軸との交点」を「-5」にして [OK]。
9. Y 軸にマウスを合わせダブルクリックし「目盛」で「最小値」を「0」に，「最大値」を「0.5」にして [OK]。

[Excel 関数の説明] 「NORMSDIST」は，標準正規分布の場合のある基準値 z の値に対して累積確率を求める関数。引数は 1 つだけで，累積確率を求めたい z の値のみ。

[Excel 関数の説明] 「NORMSINV」は，反対に標準正規分布で累積確率がある値 p になるような基準値 z の値を求める。引数は 1 つだけで，基準値を求めたい累積確率 p の値のみ。この関数は次回用いる。

11.6　正規乱数

一様乱数から正規乱数（正規分布している乱数）を作り出す方法はいくつか考えられている。一例は今回用いたような一様乱数の和を利用する方法である。これには，多数の一様乱数を発生させる必要がある。もっと一様乱数の数が少なくてすむ方法に Box-Muller の方法がある。

Box-Muller 法では，0 以上 1 未満の 2 つの一様乱数 U_1, U_2 から，平均 0，標準偏差 1 の 2 つの正規乱数 X_1, X_2 を発生させる。

$$X_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2) \tag{11.8}$$

$$X_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2) \tag{11.9}$$