第２章　確率と確率分布

５つのキーワードを中心に確率と確率分布について理解していきます。

2.1 事象についての5つのキーワード

たとえば、サイコロを投げるとき、硬貨を投げるときなどのように、それぞれの結果が偶然に左右される観測、または実験のことを試行といいます。根元事象とは、試行によって起こる結果のことで、それ以上に分けられません。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。そして、このような事象の起こりやすさが確率です。

事象

- **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。

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。

1.2.1 確率の定義

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。

確率

- 確率はある事象の期待される割合を指します。
- その値はゼロから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。

ある確率法則にしたがう変数を確率変数といいます。統計学的には確率分布から得られる変数です。

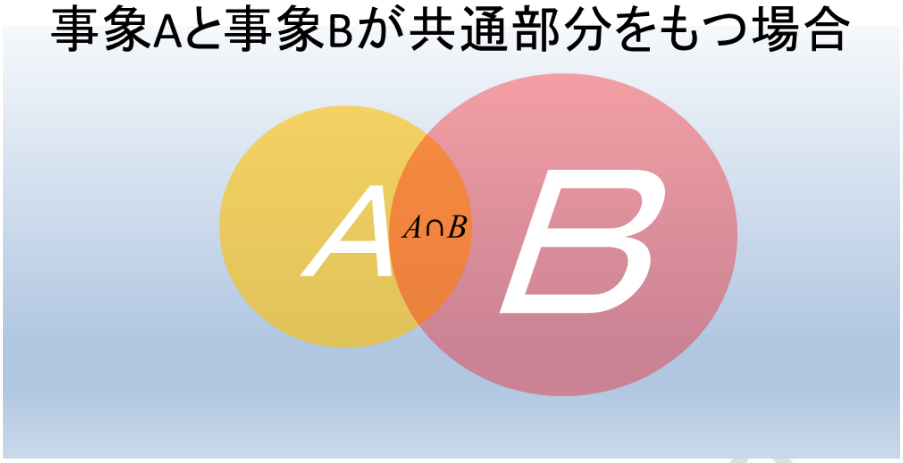
事象

- **試行**
  - サイコロを振る
- **根元事象**
  - {1},{2},{3},{4},{5},{6}
- **事象**
  - A={1，6}など
- **標本空間(全事象)**
  - Ω={1，2，3，4，5，6}
- **確率**
  - {P(A)=2/6}

２つの事象 A と B の関係について考えてみましょう。

積事象($A \cap B$)：A と B は同時に起こります。

共通部分をもつ事象
事象Aと事象Bが共通部分をもつ場合



例：さいころの目の出る確率を均等とします。そうするとそれぞれの根元事象の出る確率は 1/6 になります。A＝{1, 2, 3}、B={3, 4, 5}とすると $A \cap B$ の確率はいくらでしょうか？また、A={1, 2, 3, 4}、B={3, 4, 5, 6}の場合はどうでしょうか？

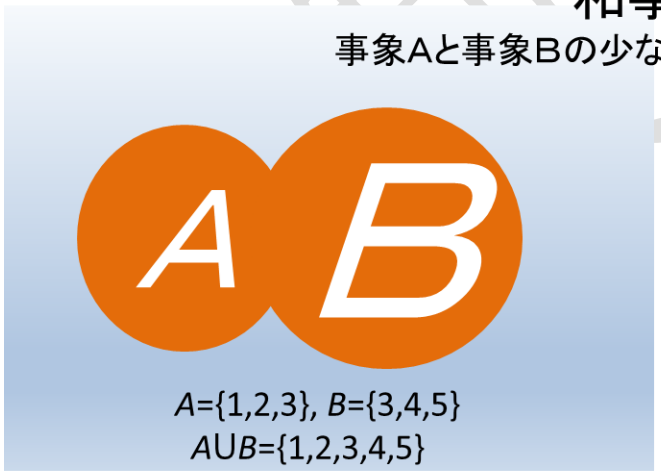
A＝{1, 2, 3}、B={3, 4, 5}のときは $A \cap B$ は{3}なので 1/6 になります。

A＝{1, 2, 3, 4}、B={3, 4, 5, 6}のときは $A \cap B$ は{3, 4}なので 2/6 になります。

和事象($A \cup B$)：A と B の少なくとも一方が起こる

例：A={1, 2, 3}, B={3, 4, 5}の和事象について考えてみましょう。また、A={1, 2}, B={4, 5}の和事象について考えてみましょう。

和事象
事象Aと事象Bの少なくとも一方が起きる。



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 3/6 + 3/6 - 1/6 = 5/6$$

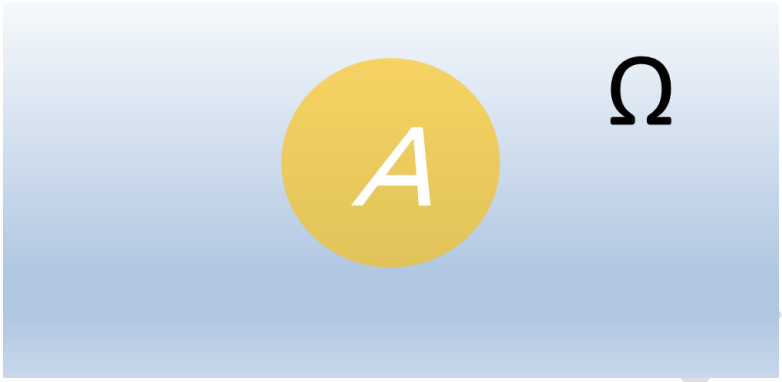
$$P(A \cup B) = P(A) + P(B)$$
$$= 2/6 + 2/6 = 4/6$$

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(∅)：何も起こらない事象

事象Aは標本空間の部分集合
事象$A \in \Omega$



排反事象
事象Aと事象Bが共通部分をもたない場合

2.1.2 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るとき、変数 X を確率変数といいます。これらは

- 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

からなります。

2.2 期待値

$x_1, x_2, \cdots$の集合から得られる確率変数 X は離散型で、その期待値は、

$$E(X) \equiv \sum_i x_i f(x_i)$$

で表されます。fは離散型の分布(確率質量関数)です。

連続型の場合は、

$$E(X) \equiv \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

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。

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。確率変数 X の実現値として得られるデータ、標本は頻度図として要約することができます。X=x_i の相対度数を N_i/N、その確率を p_i とすると、標本平均は $\bar{x} = \sum_i x_i N_i / N$で表せます。期待値は$\mu = \sum_i x_i p_i$ です。

同じ議論が分散についても成り立ちます。また、確率変数 Xの関数$f(X)$も確率変数なので、その期待値を考えることができ、同様の議論が成り立ちます。

2.2.1 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 $P(A∩B)=P(A)P(B)$が成り立つということです。∩は A と B が同時起こることを表しています。$P()$は確率を表します。たとえば、確率変数 X と Y が独立であると、その分散では $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)$が成り立ちます。$X$ と Y が独立でなければ $\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\mathrm{cov}(X,Y)$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:さいころ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 $A∈\{1, 2, 3\}$、$B∈\{3, 4, 5\}$ とすると $A∩B$ の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 $A∩B$ です。6 がでれば⌀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、$A∩B$、B、B、⌀となります。したがって $A∩B$ 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 $A∩B$ が出る確率も 1/6 です。したがって、$A∩B$ が２回続けて出る確率は 1/6・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。

$(A, A), (A, A), (A, A∩B), (A, B), (A, B), (A,⌀)$

(A, A), (A, A), $(A, A\cap B)$, (A, B), (A, B), $(A,\varnothing)$

$(A\cap B, A)$, $(A\cap B, A)$, $(A\cap B, A\cap B)$, $(A\cap B, B)$, $(A\cap B, B)$, $(A\cap B,\varnothing)$

(B, A), (B, A), $(B, A\cap B)$, (B, B), (B, B), $(B,\varnothing)$

(B, A), (B, A), $(B, A\cap B)$, (B, B), (B, B), $(B,\varnothing)$

$(\varnothing, A)$, $(\varnothing, A)$, $(\varnothing, A\cap B)$, $(\varnothing, B)$, $(\varnothing, B)$, $(\varnothing,\varnothing)$

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 確率分布

確率変数がとりえる値とそれに対応する確率を確率分布といいます。

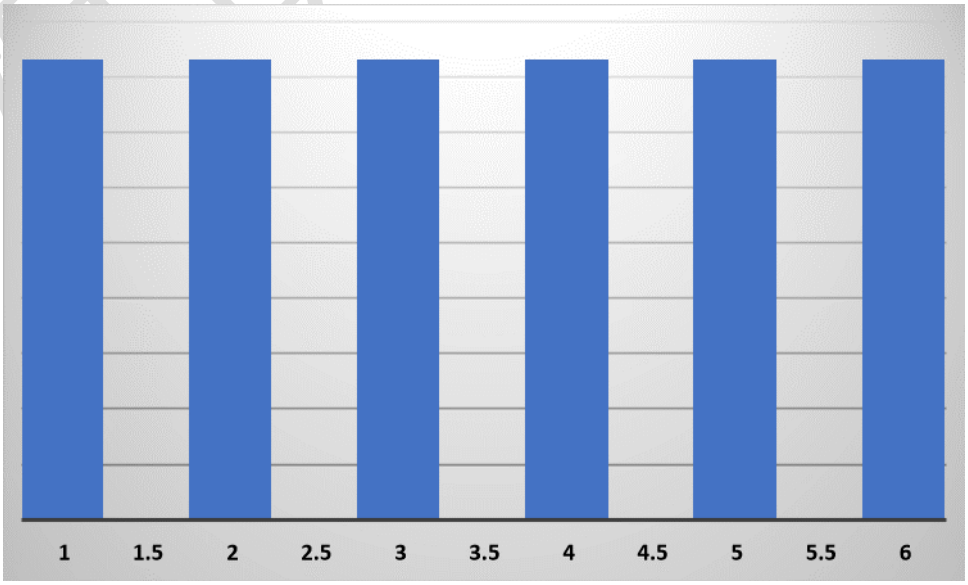
2.3.1 主な離散型確率分布

離散的確率変数の作る分布を離散型確率分布といいます。

離散一様分布

確率変数が離散値で、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。確率は $1/n$ となります。

$$f(x) = 1/N,\ x = 1, 2, \ldots, N$$



ベルヌーイ分布

コインを投げたときには、一般的には表と裏しか出ません。このような事象が起きる行為をベルヌーイ試行といいます。この場合に、確率 p で表が出て、確率 $1-p$ で裏が出るとき、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

ベルヌーイ分布の確率分布は

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

で与えられます。平均は p、分散は $p(1-p)$ となります。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。

二項分布

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

$$f(x)={}_nC_x p^x (1-p)^{(n-x)} = \frac{n!}{k!(n-x)!} p^x (1-p)^{(n-x)}$$

となります。ここで、${}_n C_k$ は n 個から k 個を選ぶ組み合わせの数です。p は成功確率です。2 項係数を表しています。また、

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

です。2 項分布では平均は $E(X)=np$、分散は $\mathrm{var}(X)=np(1-p)$ 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　n=5 で p=0.5 の場合の分布を計算してみましょう。

$x=0$ ${}_5C_0=5!/0!/(5-0)!=1\times2\times3\times4\times5/(0!)/(1\times2\times3\times4\times5)=1$

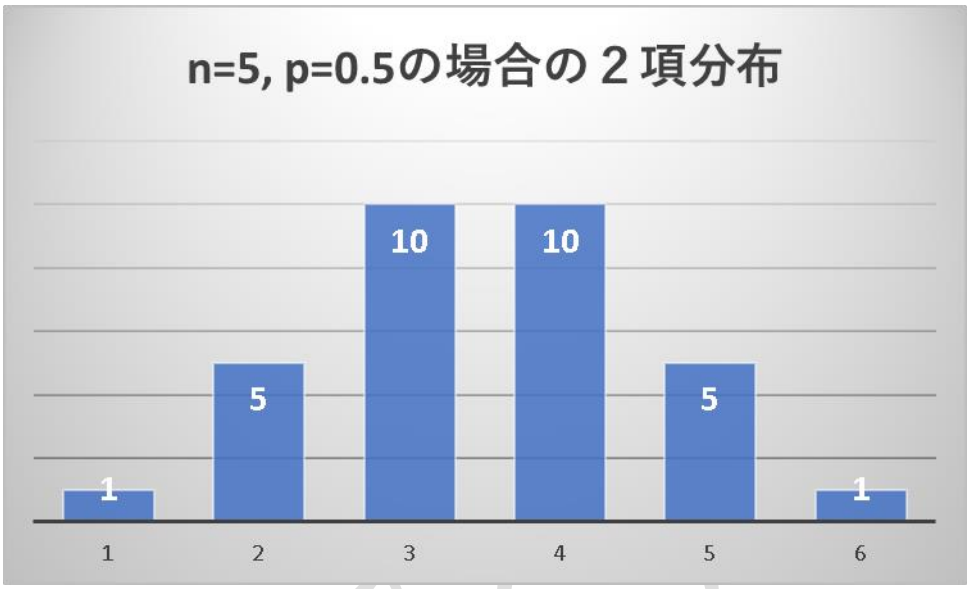
$x=1$ ${}_5C_1=5!/1!/(5-1)!=1\times2\times3\times4\times5/(1)/(1\times2\times3\times4)=5$

$x=2$ ${}_5C_2=5!/2!/(5-2)!=1\times2\times3\times4\times5/(1\times2)/(1\times2\times3)=10$

$x=3$ ${}_5C_3=5!/3!(5-3)!=1\times2\times3\times4\times5/(1\times2\times3)/(1\times2)=10$

$x=4$ ${}_5C_4=5!/4!(5-4)!=1\times2\times3\times4\times5/(1\times2\times3\times4)/1=5$

$x=5$ ${}_5C_5=5!/5!(5-5)!=1\times2\times3\times4\times5/(1\times2\times3\times4\times5)/0!=1$



2.3.2 主な連続型確率分布

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。これは全ての実数 x_i について、$X=x_i$ である確率がゼロである場合と同じです。

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、この区間の確率変数が生起する確率は等しくなります。U(a,b)と書くことがあります。

連続一様分布の確率密度関数は

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

となります。

正規分布

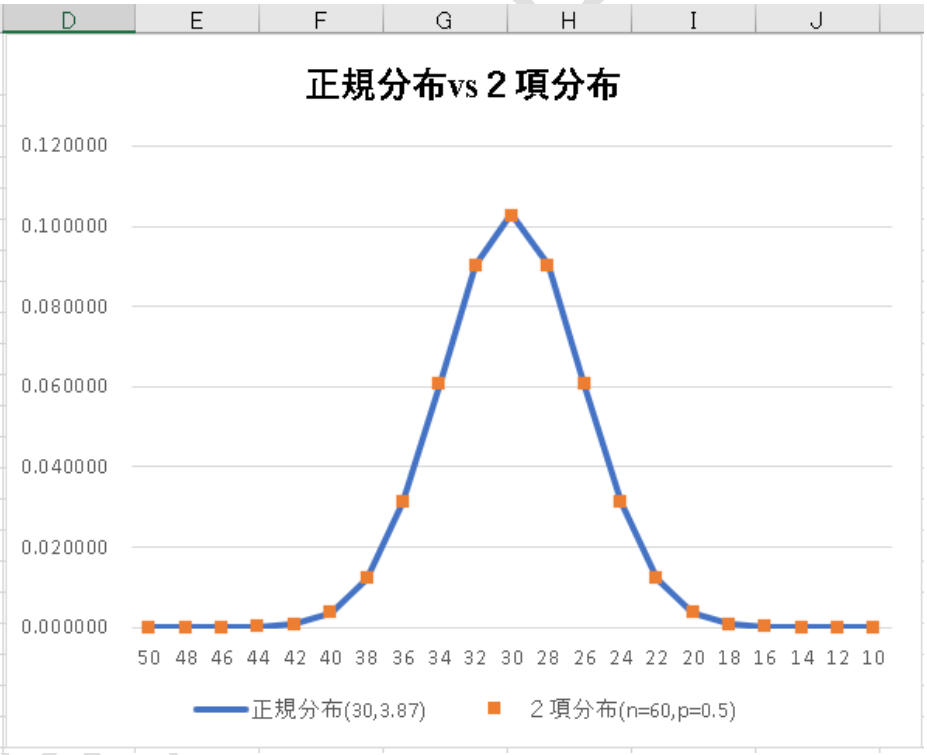
確率変数 X がとびとびの値ではなく連続していて、平均を頂点とする山が 1 つあり、山の裾のが左右対称で、ベル型をしている分布が正規分布です。正規分布では、分散は山の裾の広がり具合を表し、平均は分布の中心を示しています。正規分布の確率密度関数は平均と分散の関数として表されます。

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

ここで$\bar{X}$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。

例：エクセルで、n=60，p=0.5 のときの２項分布と正規分布を描いてみましょう。

	A	B	C
1		正規分布(30,3.87)	２項分布(n=60,p=0.5)
2	50	0.000000	0.000000
3	48	0.000002	0.000001
4	46	0.000020	0.000015
5	44	0.000150	0.000130
6	42	0.000848	0.000802
7	40	0.003675	0.003636
8	38	0.012200	0.012277
9	36	0.031025	0.031270
10	34	0.060428	0.060617
11	32	0.090149	0.089963
12	30	0.103006	0.102578
13	28	0.090149	0.089963
14	26	0.060428	0.060617
15	24	0.031025	0.031270
16	22	0.012200	0.012277
17	20	0.003675	0.003636
18	18	0.000848	0.000802
19	16	0.000150	0.000130
20	14	0.000020	0.000015
21	12	0.000002	0.000001
22	10	0.000000	0.000000



正規分布と２項分布についてのエクセル関数

	A	B	C
1		正規分布(30,3.87)	２項分布(n=60,p=0.5)
2	50	=NORM.DIST(A2,30,T3,0)	=BINOM.DIST(A2,60,0.5,0)
3	=+A2-2	=NORM.DIST(A3,30,T3,0)	=BINOM.DIST(A3,60,0.5,0)

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=5, 10, 20, 100 と変化させてみましょう。

練習問題 2.2:　赤ワインデータの要素の１つである硫化カリウムについて 10 段階の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を示せる前提条件は何ですか？

練習問題 2.4:　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 2.5: 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 2.6: 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 2.7: ベルヌーイ分布の例をあげてみましょう。

練習問題 2.8:　私たちは分布という言葉を日常よく使います。統計学ではより厳密に用語が用意されています。それは確率質量関数と確率密度関数です。その違いを説明してみましょう。

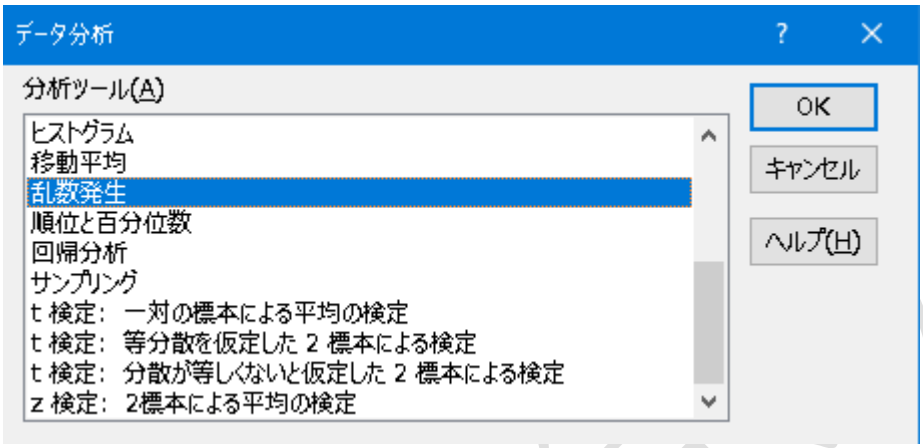
練習問題 2.9: 離散確率データの確率については理解しやすいです。頻度/頻度の総数で得られます。連続確率変数の場合には分母の頻度の総数は無限になってしまいます。そう考えると確率はゼロになってしまいます。正しいでしょうか？

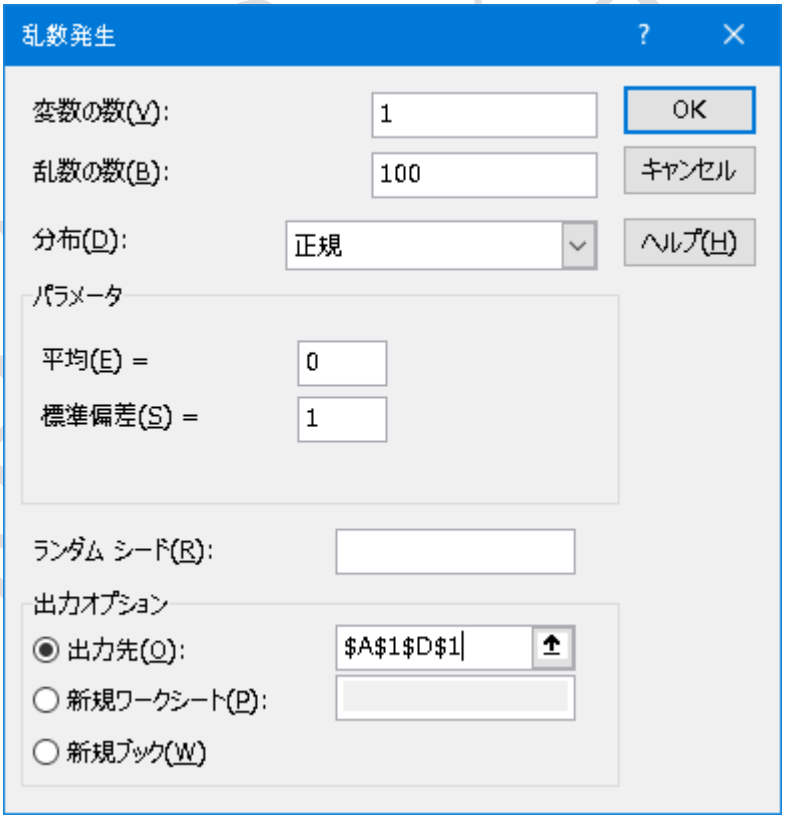
複製を禁ず

第2章

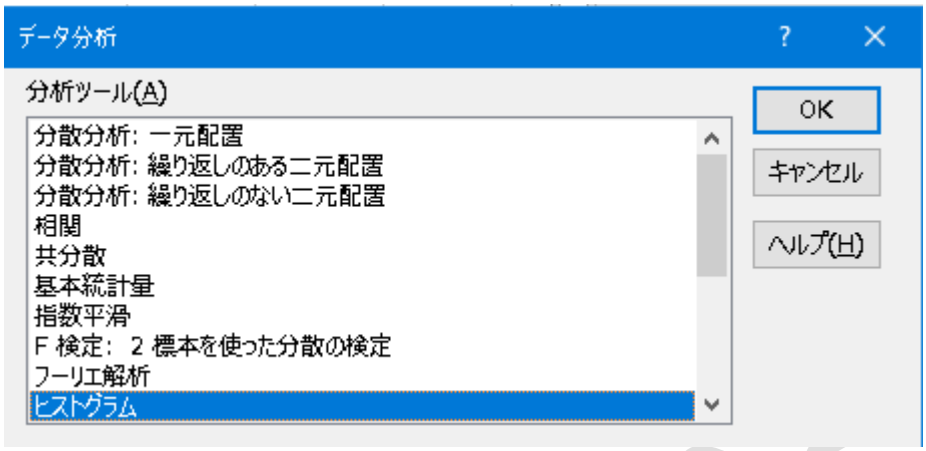
練習問題 2.1　エクセルを用いて乱数を発生させ、ヒストグラムを描きましょう。乱数は一様分布、ベルヌーイ分布、2 項分布、正規分布から発生させてみましょう。その際にデータ数を n=5, 10, 20, 100 と変化させてみましょう。

データ分析では乱数を発生されることができるために、母集団と標本の関係を理解するのに利用することができます。実際に行ってみるとその違いを実感できます。

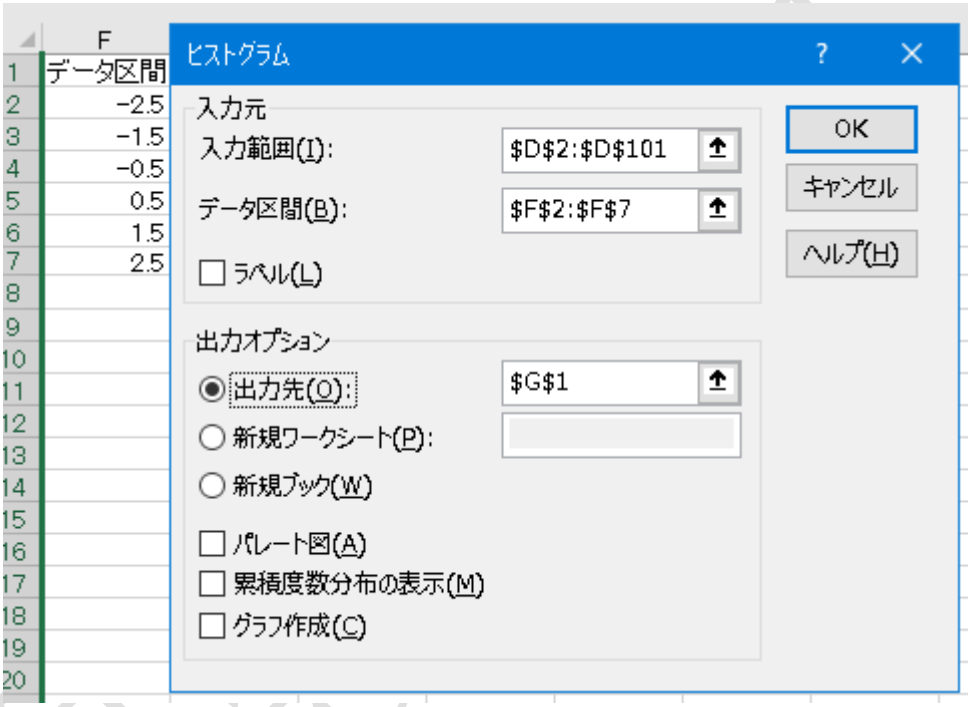




乱数の数は５，１０，２０，１００のものを生成します。つぎにヒストグラムを作ります。



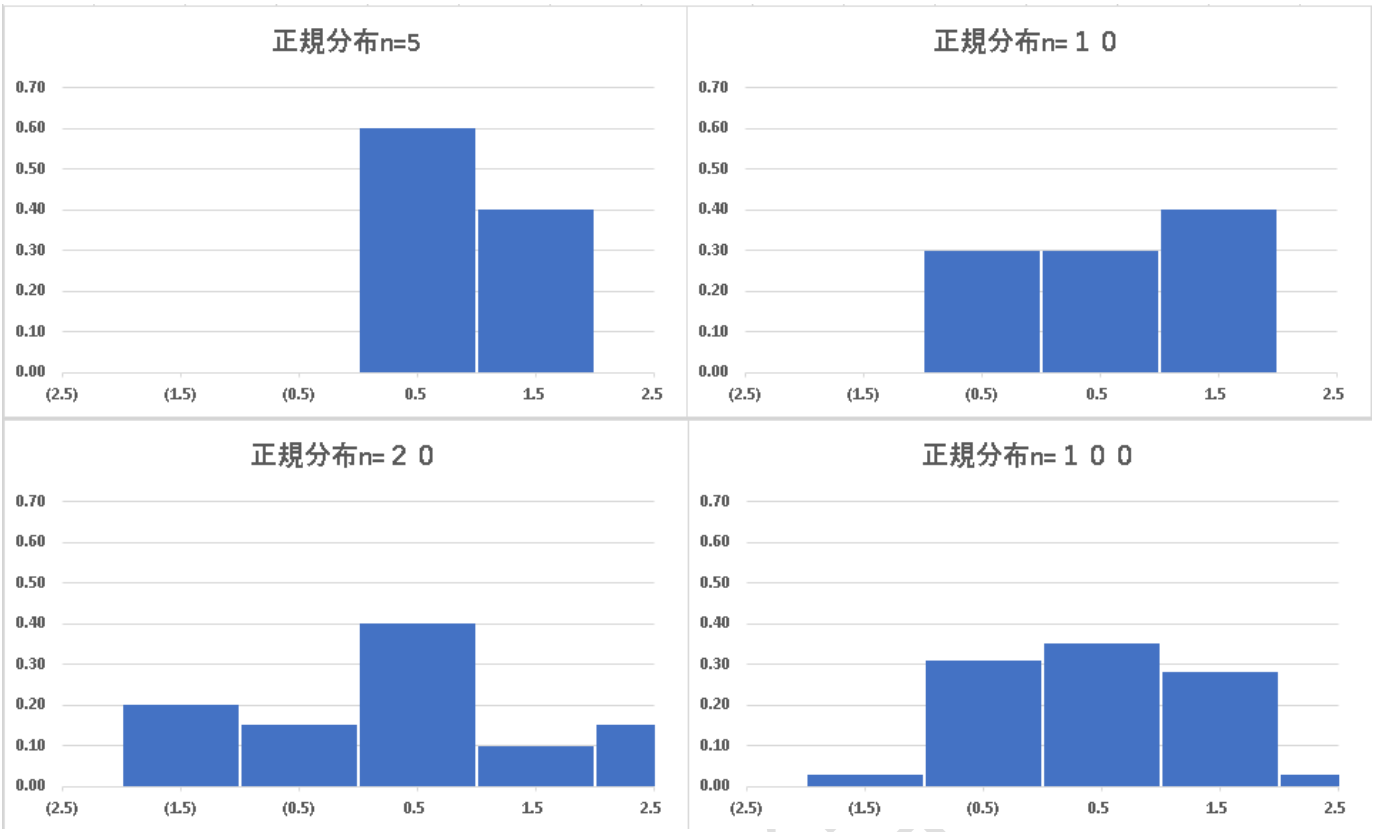
ヒストグラム生成時にデータの区間(F1:F7)を指定していることに注意してください。



つぎは正規分布の結果です。

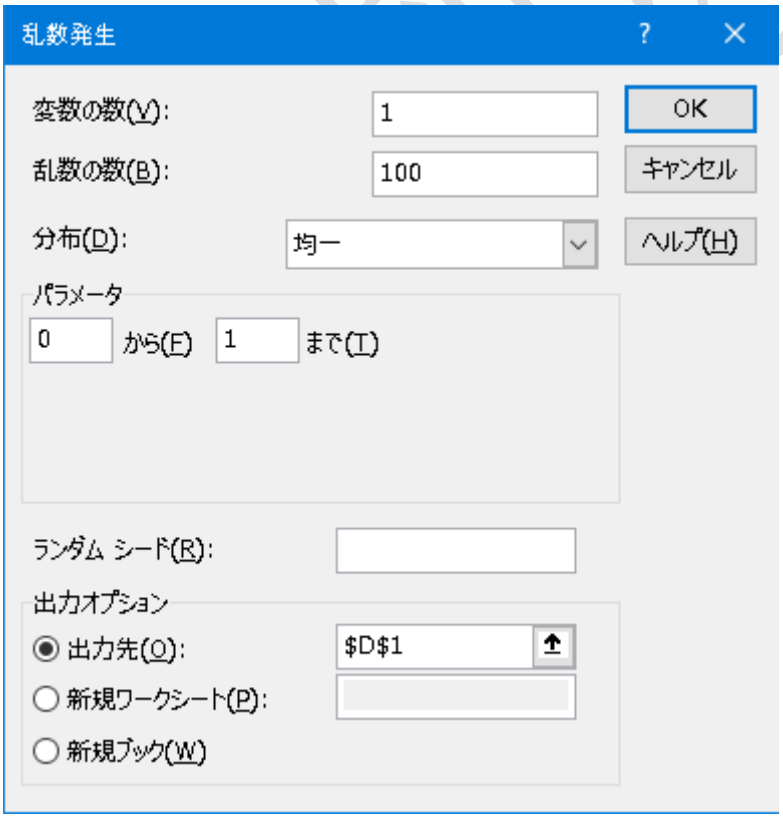
	F	G	H	I	J	K	L	M	N	O	P
1	データ区間	正規分布 n=5	正規分布 n=10	正規分布 n=20	正規分布 n=100		データ区間	正規分布 n=5	正規分布 n=10	正規分布 n=20	正規分布 n=100
2	-2.5	0	0	0	0		-2.5	0.00	0.00	0.00	0.00
3	-1.5	0	0	4	3		-1.5	0.00	0.00	0.20	0.03
4	-0.5	0	3	3	31		-0.5	0.00	0.30	0.15	0.31
5	0.5	3	3	8	35		0.5	0.60	0.30	0.40	0.35
6	1.5	2	4	2	28		1.5	0.40	0.40	0.10	0.28
7	2.5	0	0	3	3		2.5	0.00	0.00	0.15	0.03

標本の大きさが小さいと中央の標本が得られるとは限らずに左右にばらつくことが分かります。

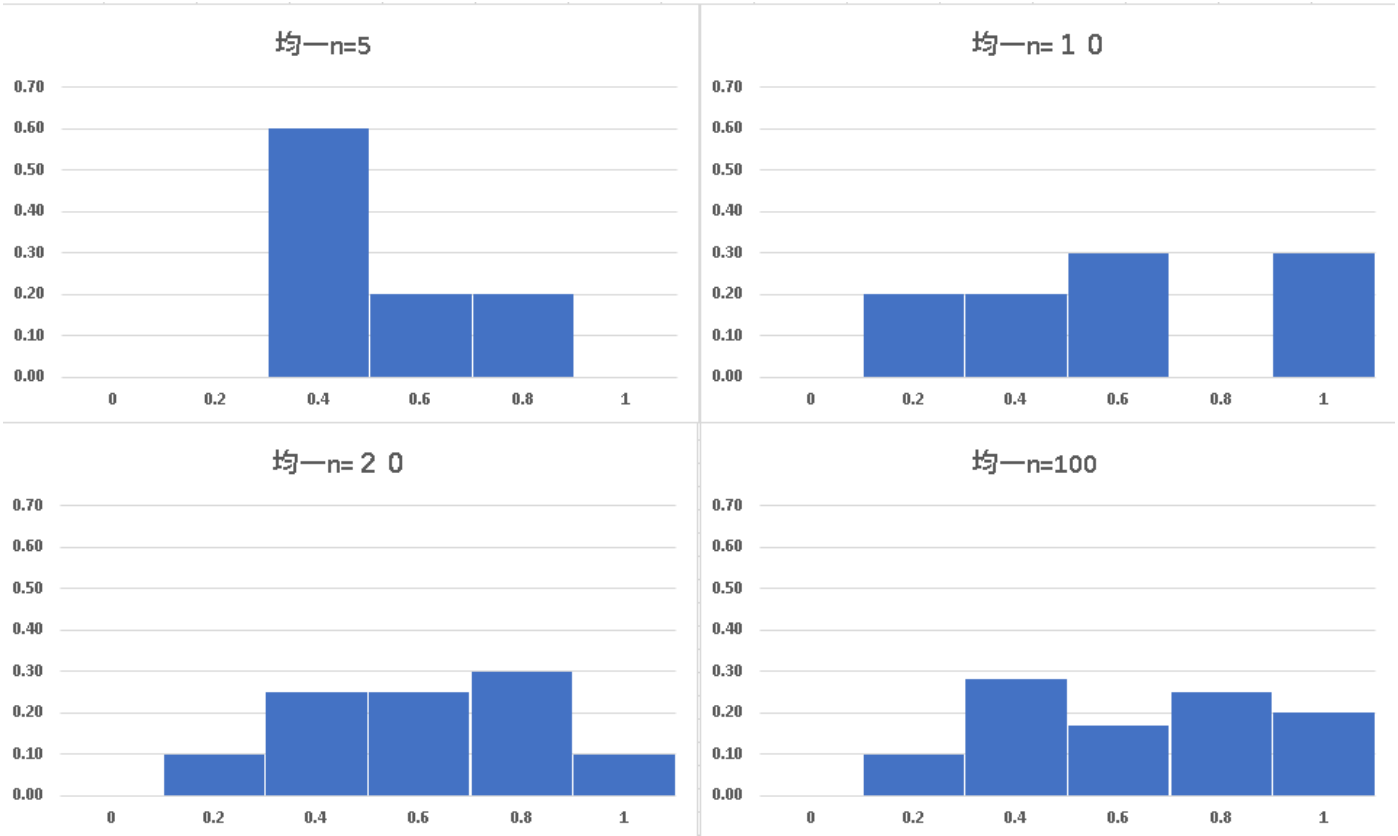


乱数＞正規分布

つぎに一様分布を作ってみます。分布は”均一”を選びます。パラメートはディフォルトの０から１を使います。

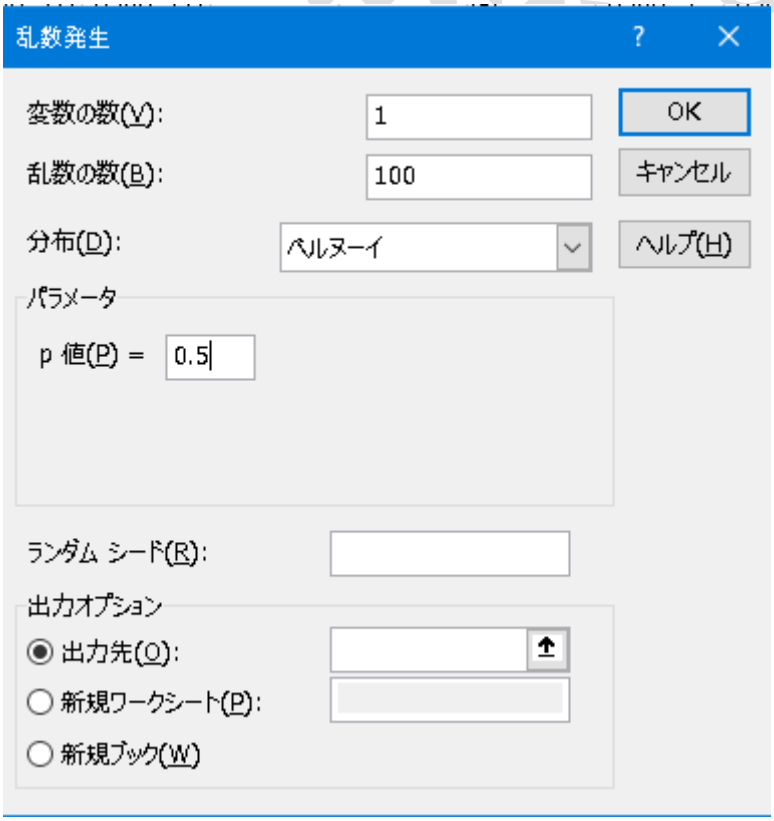


小さな標本数では左右にばらつく傾向は他と同じです。

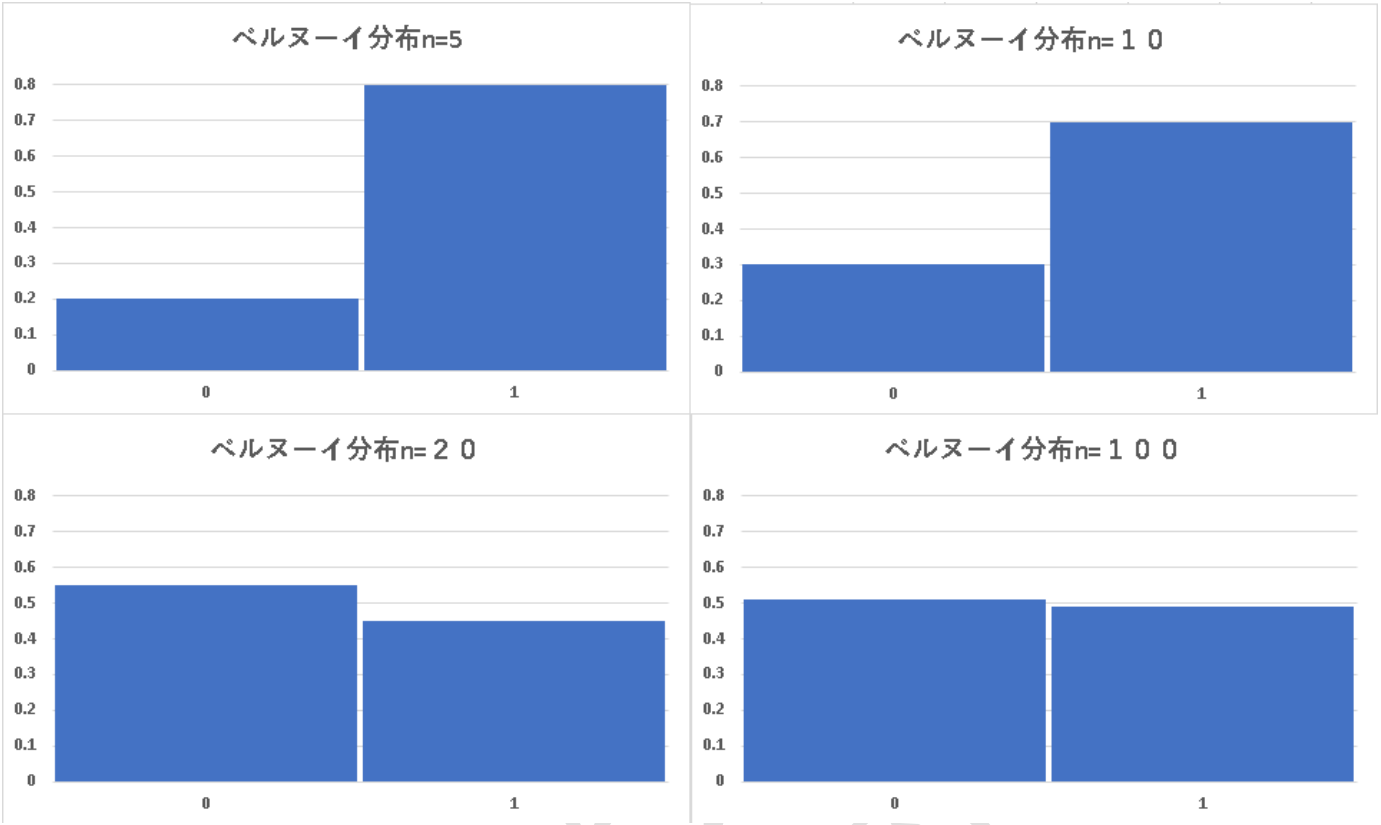


乱数＞一様乱数

つぎにベルヌーイ分布を作成します。ｐ値は 0.5 を指定します。



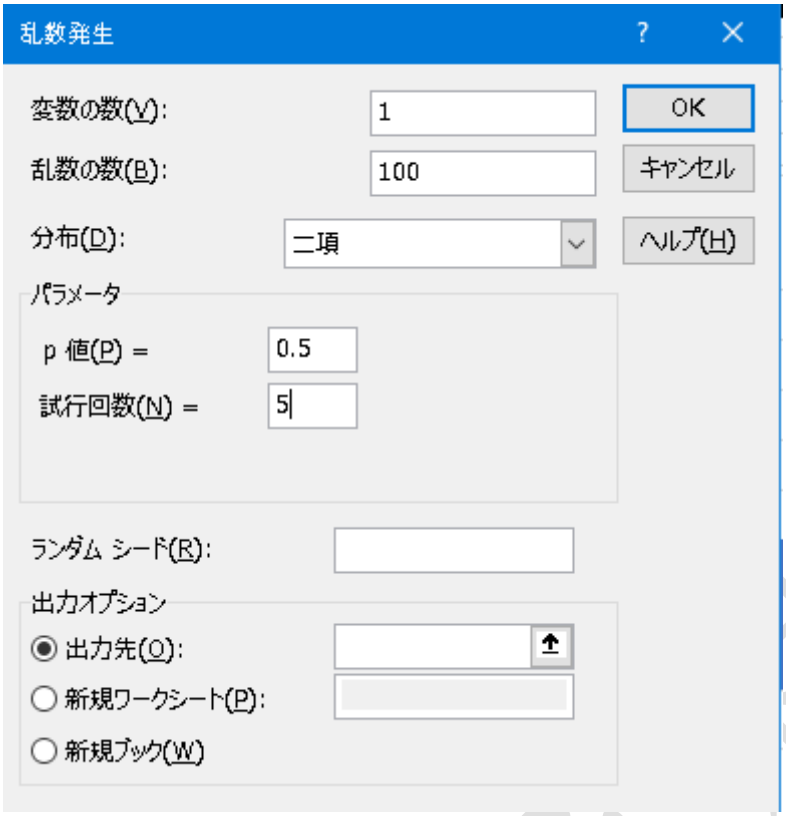
ベルヌーイ分布では標本数が小さいと頻度図の高さが大きく異なることが分かります。



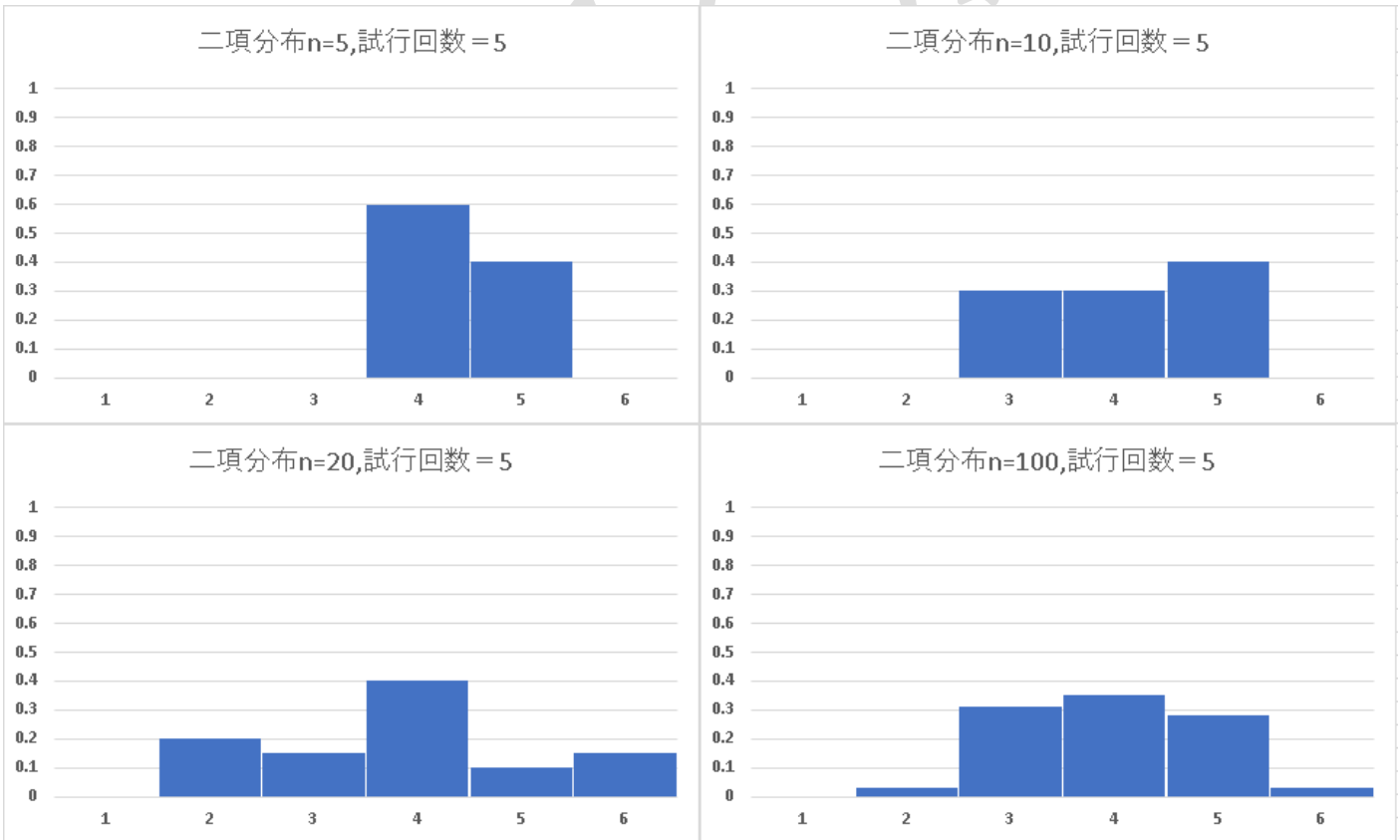
乱数＞ベルヌーイ乱数

複製など

２項分布では試行の回数を５回としました。この回数を増やすとすそ野が広がります。



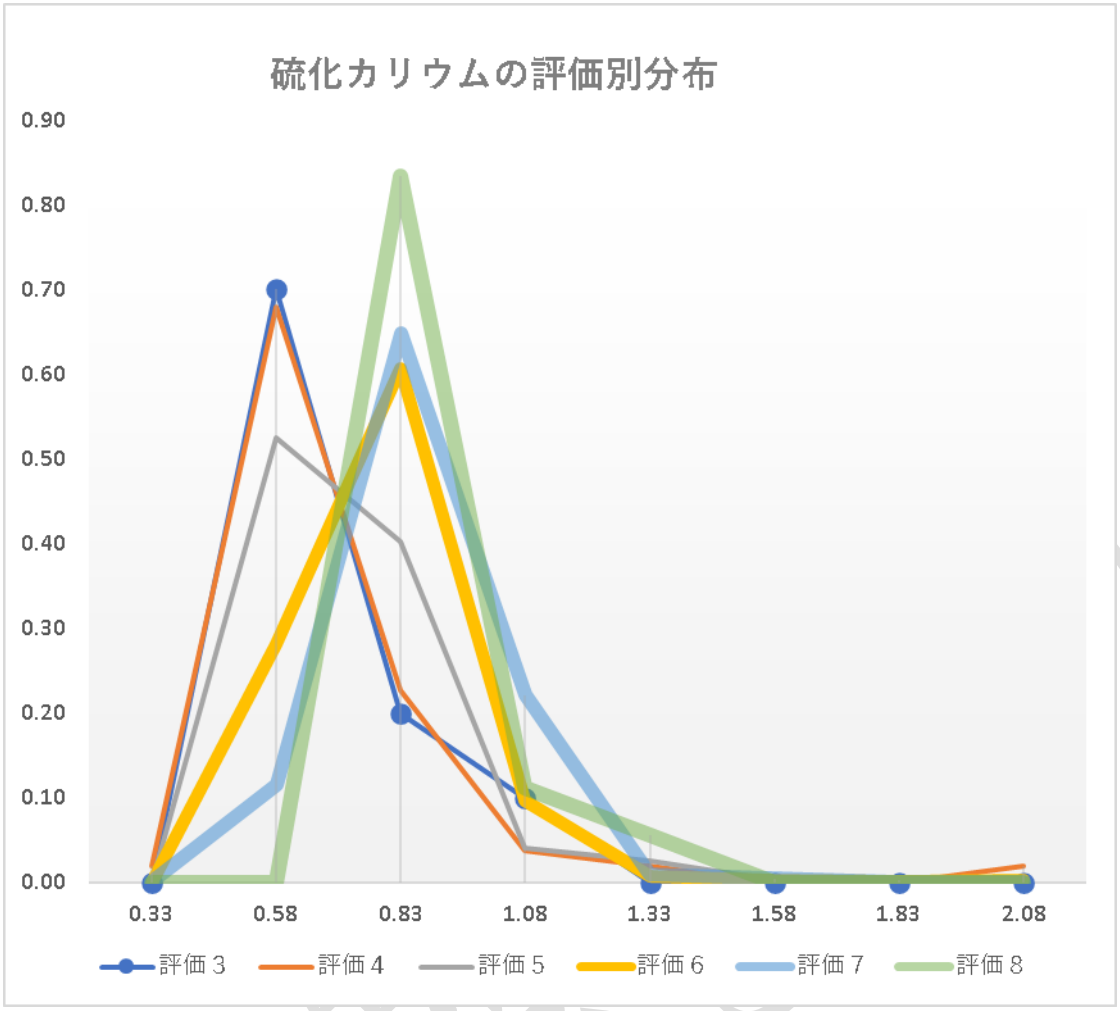
標本のサイズが小さいと左右にばらつく傾向は他と同じです。



乱数＞2 項乱数

練習問題 2.2：硫化カリウムについて評価別分布を作成してみましょう。

散布図から評価と硫化塩の間には特別な関係がありそうなので、評価別の頻度図を作成してみました。分布の山が２つありそうなことが見えてきました。



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を示せる前提条件は何ですか？

根元事象は１，２，３，４，５，６，７，８，９，１０それぞれの評価です。標本空間はこれらすべてです。違いは根元事象は最小単位の事象を表すのに対して、標本空間は全体の空間を表しています。

練習問題 2.4:　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。

確率変数であるのはそれぞれの化学成分の値ですが、観測された値は観測値となり確定値です。

練習問題 2.5: 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。

A と B が独立であるとは A と B が同時に生起する確率は P(A)×P(B)となる事象です。どちらがどちらかをおこすという因果関係はありません。一方で相関は単なる傾向を示しているだけです。

練習問題 2.6: 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。

一様分布も正規分布も確率変数のばらつきを表すので分散を用います。一様分布の分散は

$(b-a)^2/16$で与えられます。b は上限, a は下限です。

練習問題 2.7: ベルヌーイ分布の例をあげてみましょう。
コイン投げの表裏の結果、サイコロを振った時の６が出るときとそうでないときの結果、野球の勝敗など

練習問題 2.8:　私たちは分布という言葉を日常よく使います。統計学ではより厳密に用語が用意されています。それは確率質量関数と確率密度関数です。その違いを説明してみましょう。
確率質量関数は離散確率変数に用いられます。確率密度関数は連続型の確率変数に使われます。確率変数がある範囲でとる確率を表します。

練習問題 2.9: 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？
連続確率変数ではたとえば特定の一点の値が出る確率はゼロであるとしか言いようがありません。しかし、それがほんのわずかであるとしても範囲で指定されるとその範囲内で事象が生じる確率はありそうです。

複製を禁ず