入門 統計学　第2章

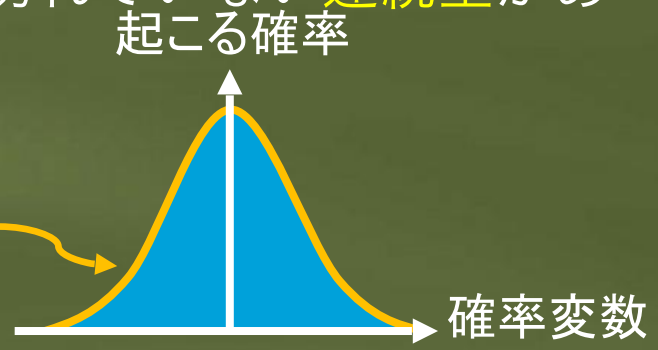
確率分布

『入門 統計学 第2版　一検定から多変量解析・実験計画法・ベイズ統計学まで一』(オーム社)

※注:本書を購入された方へのサービスですので，教科書指定(参考図書は不可)していない授業での使用はお控えください。

2.1　確率分布

- 確率分布とは，確率変数が取る値と，その起こる確率（生起確率）との対応関係を図や表，関数で示したもの
- 確率変数とは，ある値を取る確率が決まっている変数（値がとびとびの離散型と切れていない連続型がある）
- 正規分布などいろいろある

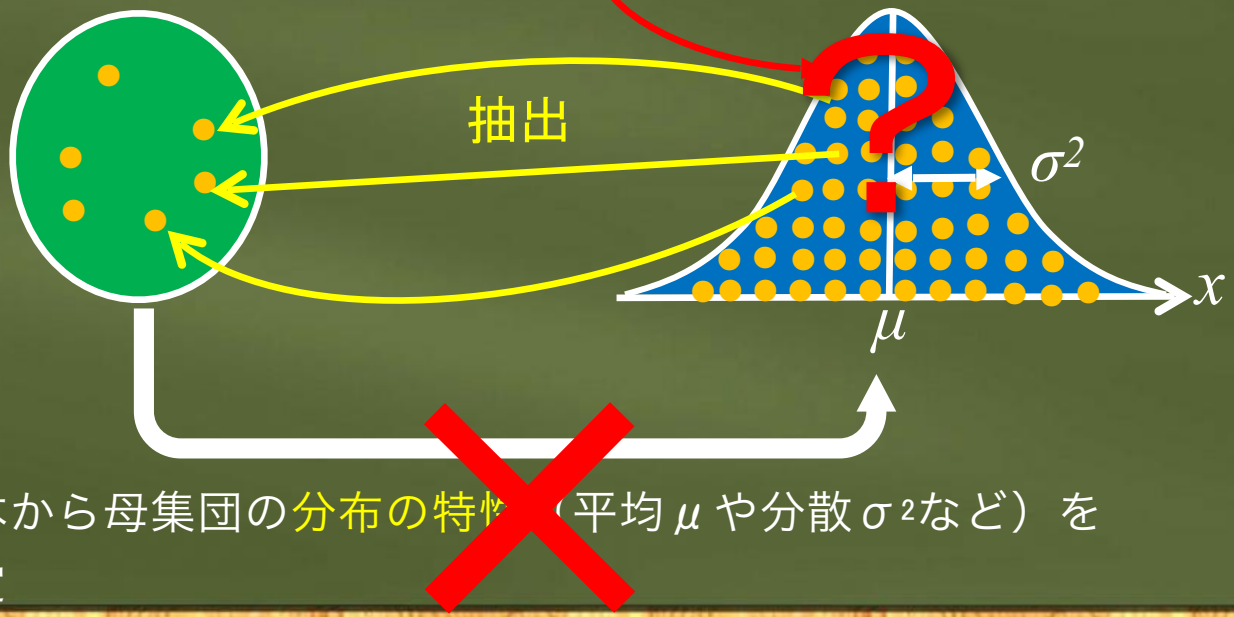


推測統計学（第3章以降）における確率分布の重要性

母集団が従う確率分布がわからなければ特性を推測できない

標本
（観測データ）

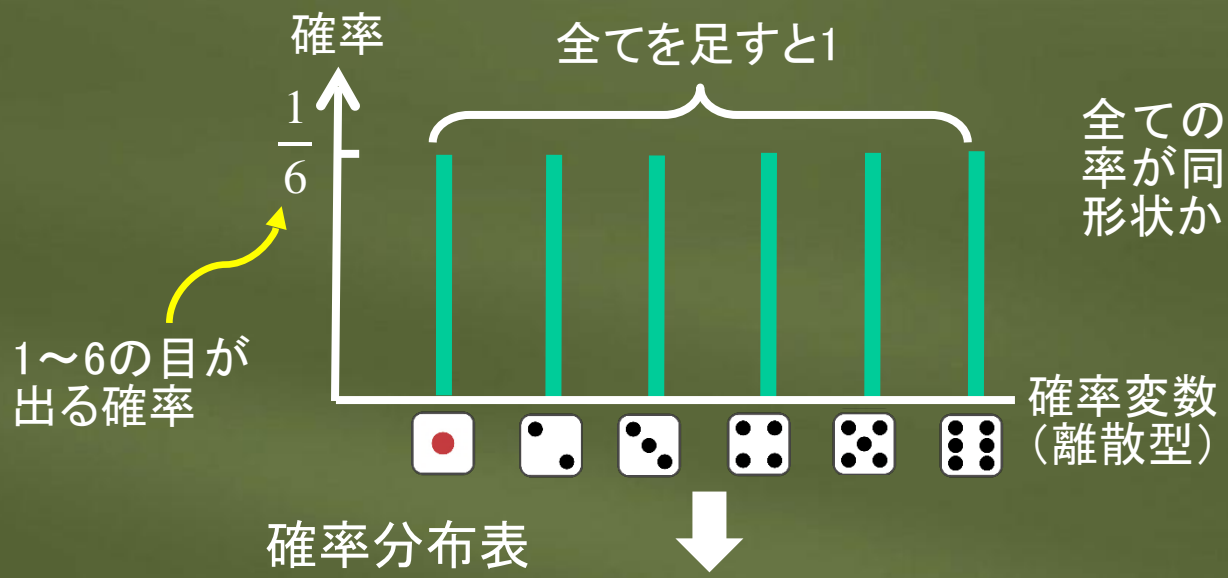
母集団は何らかの確率分布に従っていると仮定



標本から母集団の分布の特性（平均μや分散σ^2など）を推定

一番簡単な確率分布（離散型）

1個のサイコロ振りを考える



全ての確率変数の生起確率が同じ確率分布は，その形状から一様分布と呼ぶ

確率分布表

確率変数（サイコロの目）	1	2	3	4	5	6
（生起）確率						

ところで確率とは？

- ある事象の起こりやすさ（偶然性）を数値化したもの

ある試行の結果，事象Aが起こる確率

$$P(A) = \frac{\text{事}}{\text{起こり}} = \frac{n(A)}{n(U)}$$

試行：試しに何か行うこと（サイコロ振りが該当）

事象：試行で偶然に決まる結果（出たサイコロの目が該

例題：2つのサイコロ振り（試行）で，6の目（事象）が出る確率は？

P(6 / 出る)

確率変数と確率の正式な表記法

確率変数Xが実現値x_iを取る確率を表す関数

$$P(X = x_i) = p_i$$

実現値x_iを取る確率の値

確率変数（値の入れ物）

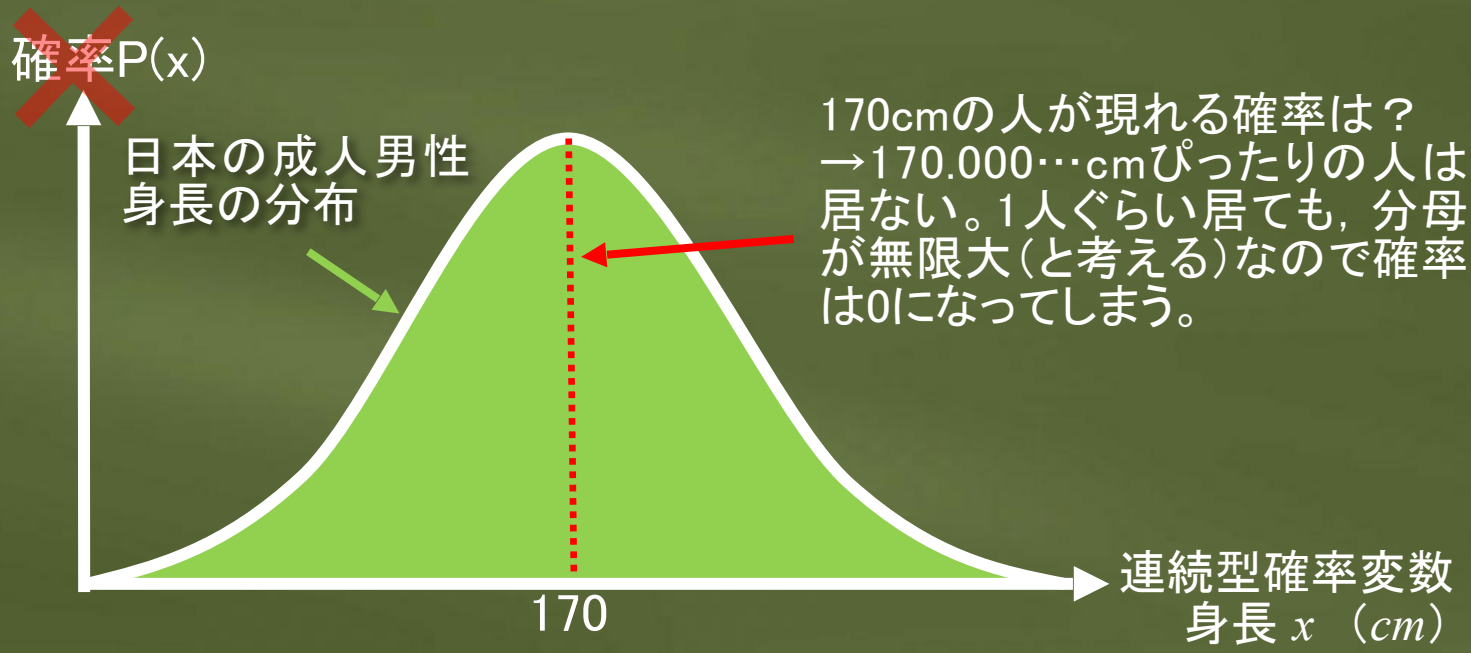
確率変数の実現値（実際に事象iが起きたとき観察される値）

事例：2つのサイコロ振りで6の目が出る確率は$P(X=6)$，その確率の値は$p_6=0.14$

注：大文字や小文字が混在すると初学者にはわかり難いので，本書では（第15章を除いて）大文字の確率変数Xは使用しません。確率の関数は単に$P(x)$と表します。

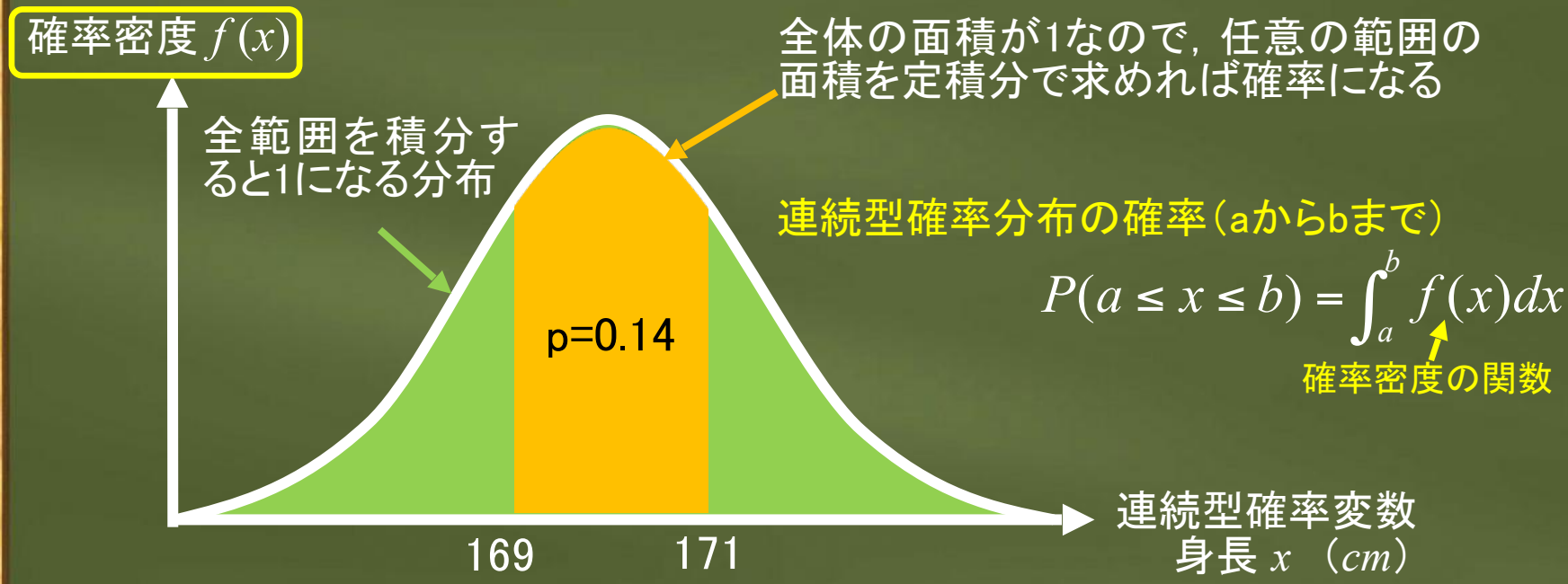
連続型確率分布の確率①

- 確率変数が連続型の分布では，ある値1点の確率は求められない



連続型確率分布の確率②

- 確率変数の任意の範囲の面積を求める（分布全体の面積を1に正規化しておけば確率になる)



確率分布の平均

（確率変数の平均）

- 試行の偶然の結果として期待される値なので期待値とも
- 内容：実現値xiと生起確率値piの積の総和（積分）

離散型確率分布の平均（期待値）$\mu = E(x) = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_{i=1}^{n} x_i p_i$

母集団の平均※　xの期待値　確率値

事例：サイコロ1個を振る試行で出る目の平均　$1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=3.5$

連続型は…

総和Σの代わりに積分

連続型確率分布の平均（期待値）$\mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx$

積分する範囲は分布によって異なる

確率piの代わりに確率密度f(x)

※確率分布は母集団に仮定するので，$\bar{x}$ではなく，母集団の平均を意味するμを使っている。

確率分布（確率変数）のバラツキ

- 分散：偏差平方と生起確率の積の総和（積分）

離散型確率分布の分散

分散記号　偏差平方　確率値

$$\sigma^2 = V(x) = (x_1 - \mu)^2 p_1 + \cdots + (x_n - \mu)^2 p_n = \sum_{i=1}^{n} (x_i - \mu)^2 p_i$$

母集団の分散　母集団の平均

事例：サイコロ1個を振る試行で出る目の分散

$$(1-3.5)^2 \times \frac{1}{6} + \cdots + (6-3.5)^2 \times \frac{1}{6} = 2.9$$

連続型は…

総和Σの代わりに積分∫

連続型確率分布の分散

$$\sigma^2 = V(x) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

- （確率分布の）標準偏差は，離散も連続も平方根を取るだけ

例題：サイコロを2回振る試行の平均と分散の計算

確率変数 x_i （サイコロの目）	2	3	4	5	6	7	8	9	10	11	12
生起確率 p_i											

平均$=\Sigma x_i p_i = 2\times1/36+3\times2/36+\ldots+11\times2/36+12\times1/36=7$

分散$=\Sigma(x_i-\mu)^2 p_i=(2-7)^2\times1/36+\ldots+(12-7)^2\times1/36=5.83$

標準偏差$=\sqrt{\text{分散}}=\sqrt{5.83}=2.45$

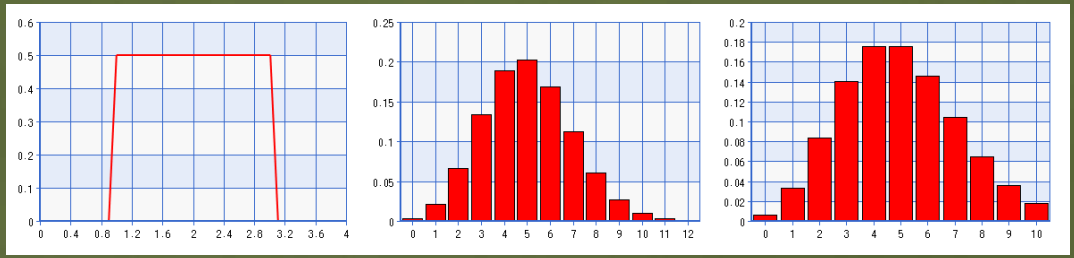
最初に5，次に6の目が出る確率$=\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

注：試行は独立しており，順列の確率である（表の11の目が出る確率2/36は最初に6，次に5の目が出る組合せを含めた確率）。

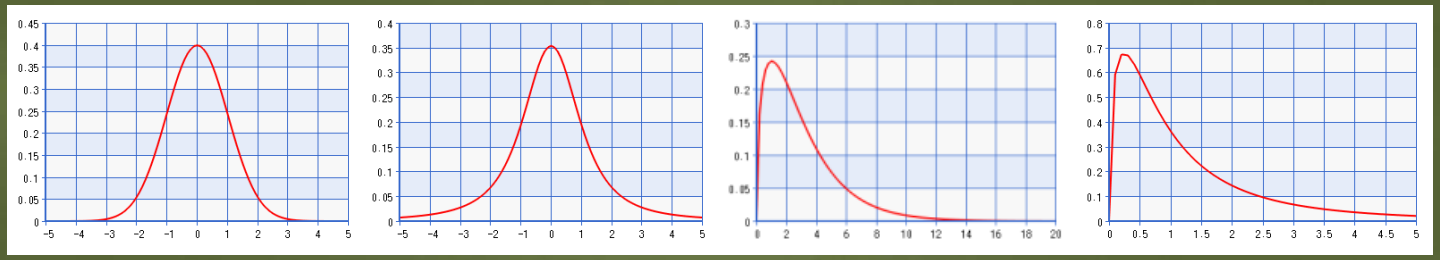
主な(本書で扱う)確率分布

- 主な確率分布
  - 離散型
    - 一様分布：全変数の値の確率が一定の分布(連続型もある)
    - 二項分布：ベルヌーイ試行(後述)の成功回数を確率変数とした分布
    - ポアソン分布：試行回数が多く，確率が小さい事象を対象とした分布
  - 連続型
    - 正規分布：試行回数が多い二項分布に近似させた分布
    - z分布(標準正規分布)：正規分布の平均を0，分散を1とした分布
    - t分布：母分散が未知の場合にz分布の代わりに用いる分布
    - χ^2分布：データの平方和が従う分布
    - F分布：分散の比が従う分布

主な確率分布の形状例



一様分布（連続型）　二項分布　ポアソン分布



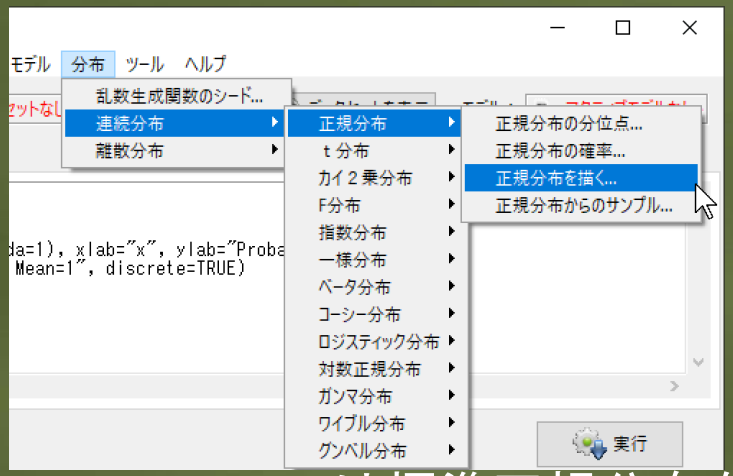
標準正規分布　t分布　χ^2分布　F分布

注1：ke!$^{+}$san（http://keisan.casio.jp/）上で描写

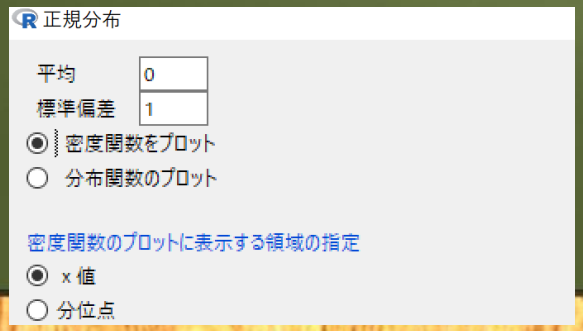
2：分布の形状は母数（試行回数や平均，分散など）によっても変化する

Rコマンダーで確率分布を描いてみよう！

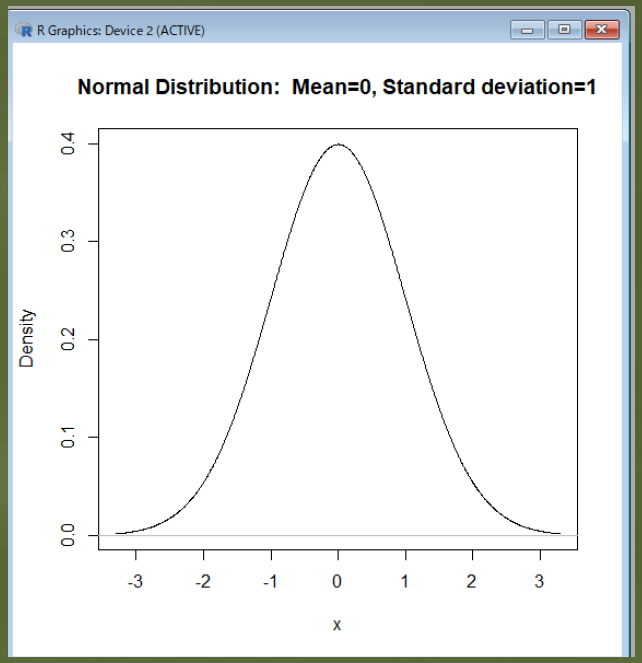
メニュー[分布]の中から描きたい分布を選ぶ



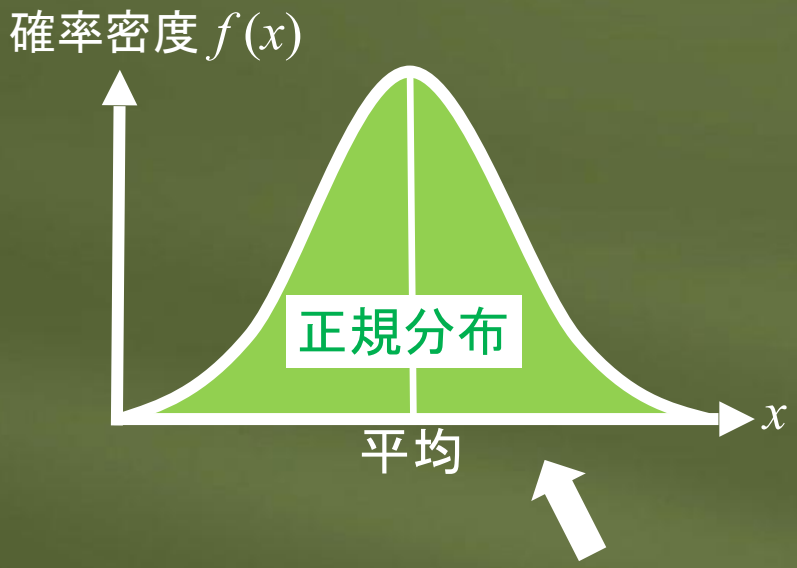
ここでは標準正規分布を描いてみる



R Guiの中のR Graphicsに描かれる



2.2　二項分布から正規分布へ



最も重要な確率分布が正規分布

推測統計学（第3章以降）では，標本の抽出元である母集団は正規分布に従うと仮定することが多い

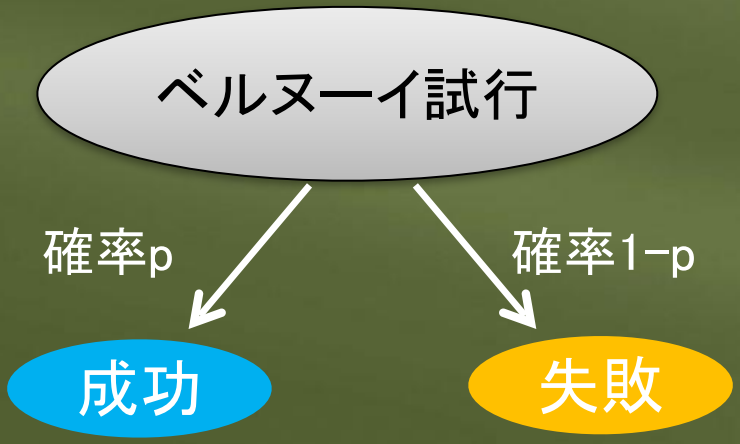
→理由：生物統計（体長，体重など）や社会現象（犯罪率，結婚率など）も正規分布に近似的に従うことが経験的に知られている

ド・モアブル（右）が，試行が多い二項分布を近似するために考え出した確率分布なので，まずは二項分布から解説


A. de Moivre
(1667～1754)

二項分布とベルヌーイ試行

- 二項分布：ベルヌーイ試行を繰り返したときの成功回数を確率変数とした確率分布
- ベルヌーイ試行：結果が成功か失敗しかなく，互いの試行が独立していて，個別の成功確率pが一定の試行



とか

サイコロ振り
（1の目が出るか出ないかで，常にp=1/6）

コイン投げ
（表が出るか出ないかで，常にp=1/2）

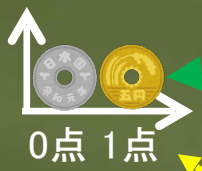
簡単なコイン投げで考えてみよう

コイン投げの二項分布

コイン投げの実験（ベルヌーイ試行）：表 が出れば1点，裏 は0点

a) コインが1枚の場合（1回のベルヌーイ試行，n=1）

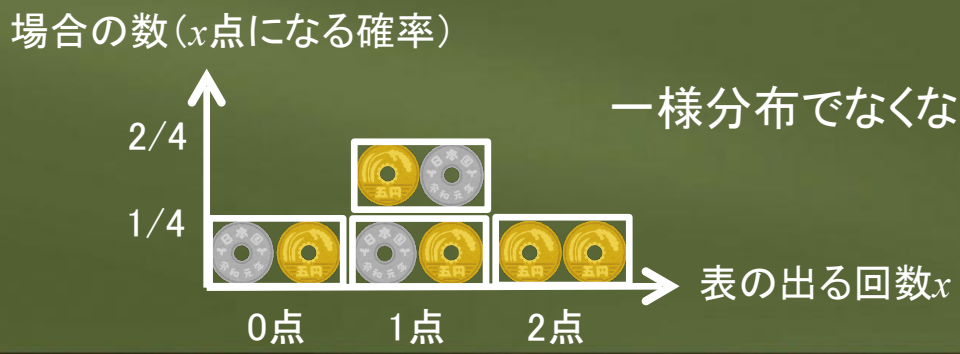
個別のコインの表が出る確率は常に0.5



n=1の二項分布をベルヌーイ分布と呼ぶ

確率変数の値（点数）は成功（表の出る）回数x

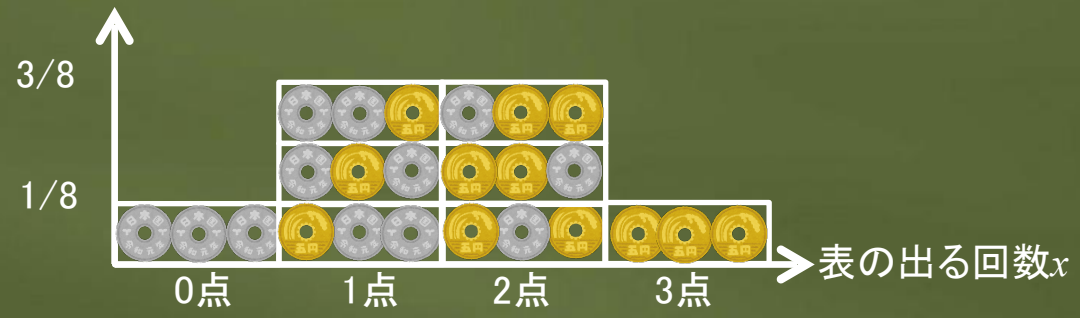
b) コインが2枚の場合（n=2，1枚を2回投げると考えても良い）



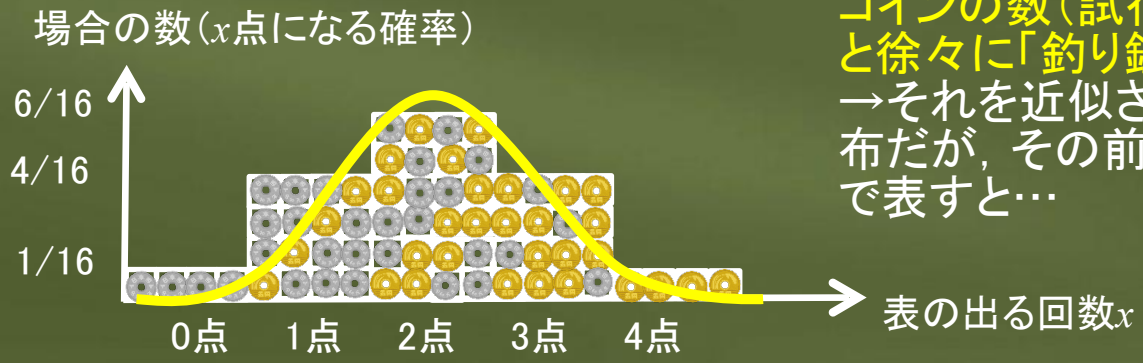
一様分布でなくなる

試行回数を増やしていくと…

c) コインが3枚の場合(n=3)



b) コインが4枚の場合(n=4)



コインの数(試行回数)を増やすと徐々に「釣り鐘型」に近づく。
→それを近似させたのが正規分布だが，その前に二項分布を式で表すと…

二項分布の確率関数式と母数

- n回の試行でx回成功する確率の分布

x回成功する確率　　個別の試行が成功する確率

二項分布の確率質量関数　$P(x) = {}_nC_x p^x (1-p)^{n-x} = \dfrac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$

連続型の確率密度に対する用語　　組合せの記号※　　分布の形を決める（母数）

※組合せ ${}_nC_x$：異なるn個のものからx個を取り出す組合せが何通りあるかを表す。例えば${}_4C_3$ならば4!÷3!で4になる。

- 分布の形状を決める母数（パラメータ）：試行回数nと　試行が成功する確率pの2つ→二項分布を$B(n, p)$と表す

Excel関数=BINOM.DIST（成功数，試行回数，成功確率，FALSE）

二項分布の平均と分散

注：分散の式の展開はやや難しいので飛ばしても結構です

- 二項分布はp（個別試行の成功確率）が一定なので，既述の離散型分布より単純になる

試行が成功する確率pは全て同じ

二項分布の平均（期待値）　$E(x)= x_1 p +\ldots+x_n p= np$

偏差平方の期待値（平均）

二項分布の分散

$$V(x) = E[(x-\mu)^2] = E[x^2 - 2\mu x + \mu^2] = E(x^2) - E(2\mu x) + E(\mu^2)$$

$$= E(x^2) - 2\mu E(x) + \mu^2 = E(x^2) - 2\mu\mu + \mu^2 = E(x^2) - \mu^2$$

$$= E(x^2) - [E(x)]^2 = \{E[x(x-1)] + E(x)\} - [E(x)]^2$$

np × (n−1)p

$$= [n(n-1)p^2 + np] - (np)^2 = n^2 p^2 + np(1-p) - (np)^2$$

$$= np(1-p)$$

x²の期待値とxの期待値の2乗の差になる

注：n=1のベルヌーイ分布：平均p，分散$p(1\text{-}p)$

例　題

- 1個のサイコロを3回振る試行(n=3)で，3の倍数が出る回数を確率変数xとすると，
  x は0（1回も出ない)，1，2，3の4種類になります。
  それぞれの生起確率P(x)はいくつになるでしょうか？

サイコロ6面のうち1/3の2面は3の倍数

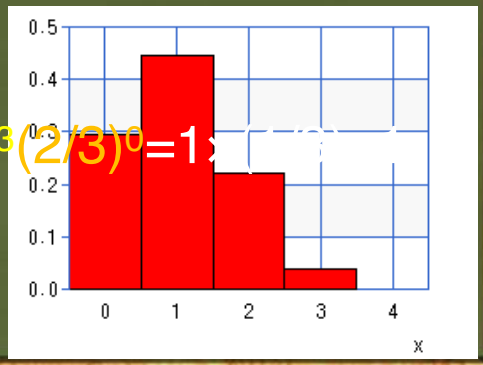
答え：3の倍数が出る確率pは1/3なので，出ない確率(1-p)は2/3となる。

$P(0)={}_3C_0(1/3)^0(2/3)^3=1\times1\times(2/3)^3$　　$P(1)={}_3C_1(1/3)^1(2/3)^2=3\times(1/3)\times(2/3)^2$

=BINOM.DIST（2,3,1/3,FALSE)

$P(2)={}_3C_2(1/3)^2(2/3)^1=3\times(1/3)^2\times(2/3)$　　$P(3)={}_3C_3(1/3)^3(2/3)^0=1\times$

確率変数 X (3の倍数が出る回数)	0回	1回	2回	3回
確率P(x)	0.296	0.444	0.222	0.037



正規分布への近似

二項分布の短所：①試行nが大きいと階乗計算が大変

②正の値の離散値しか扱えず，上限（n）もある

→（上下限のない）連続型分布関数で近似できれば解決！

→それが正規分布 $N(\mu,\ \sigma^2)$

正規分布の確率密度関数 $f(x)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

連続型確率変数　円周率　分散　ネイピア数（自然対数の底）　平均

$1/\sqrt{2\pi\sigma 2}$ で全面積を1に正規化

もとの指数関数 $f(x)=e^{-x^2}$

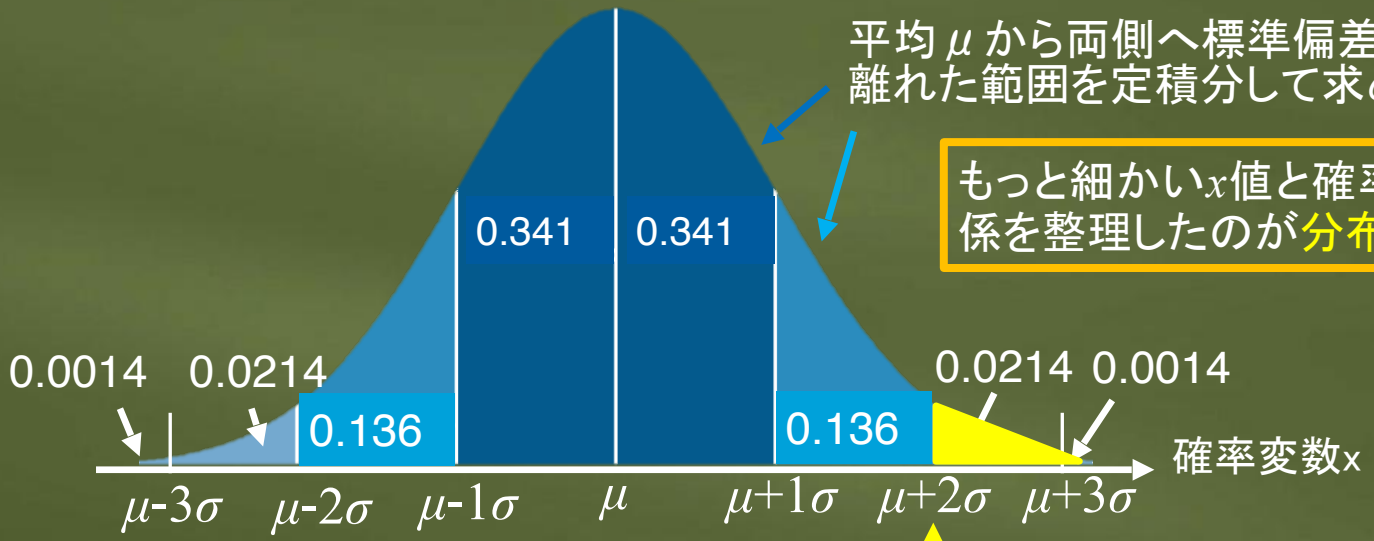
分布の横位置（平均）と横幅（分散）を変更できるよう

母数：μ，σ^2

$f(x)=e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

2.3　正規分布の便利な性質

全体の平均とバラツキが既知なら，ある対象データの値が全体のどのような位置にあるかわかる（左右対称で階乗計算がないので簡単）



例題　Excel関数を使った正規分布の確率計算

Excelの関数を使って，ある対象が持つ“44”という値が正規分布の下側や上側から何％のところに位置するのかを求めてみましょう。なお，集団の平均μは40，標準偏差σは2とします。

答え：正規分布の確率密度を求めるExcel関数は，

Excel関数=NORM.DIST(x，平均，標準偏差，関数形式)

で，最後の関数形式には確率密度（分布の高さを）を知りたい場合FALSE，累積確率（分布の面積）を知りたい場合TRUEを入れる。

よって，=NORM.DIST(44，40，2，TRUE)と入力すればよい。

0.97725が返って来るので，“44”は下から97.7％，1から0.97725値を引くと0.02275なので上から2.28％のところに位置することがわかる。

2.4　標準化と偏差値

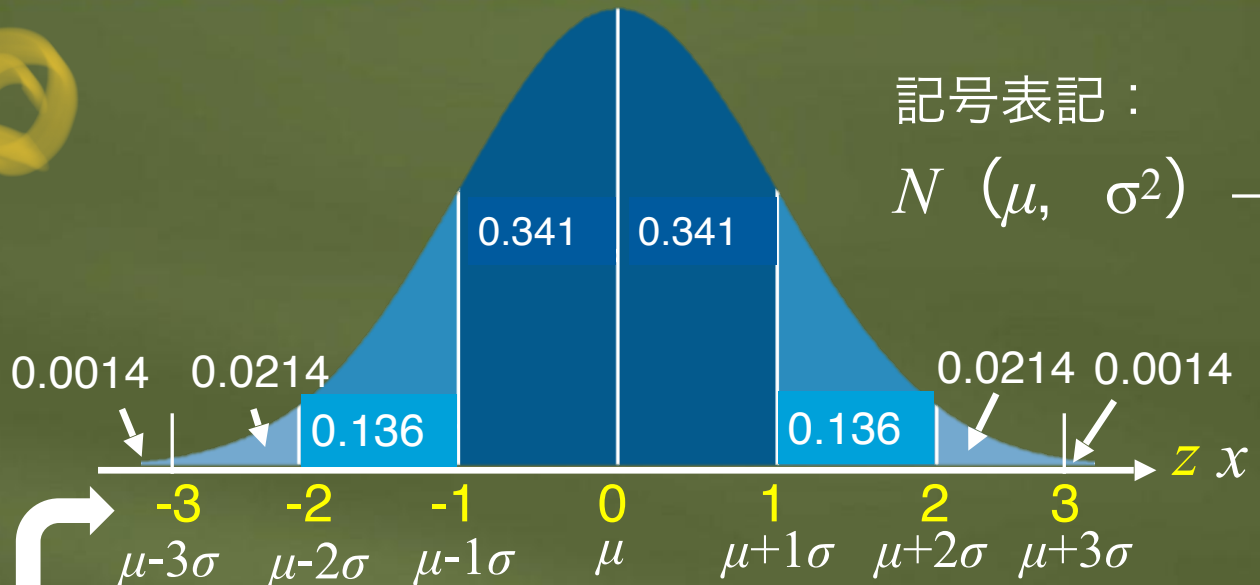
- 確率変数の値と確率との細かな対応関係を示した分布表があれば定積分が不要となる（PCも昔はなかった）
- 問題点：正規分布は母数（平均と分散）によって形状が変化するため，対象ごとに異なる表が必要になる
- 解決策：平均と分散を統一すれば，どのような対象にも使える分布表が得られる
- 標準化：平均μを0，標準偏差σを1（分散も1）に変換

標準化変量

$$z_i = \frac{\text{データ}i\text{の値} - \text{平均}}{\text{標準偏差}} = \frac{x_i - \mu}{\sigma}$$

データの数（i=1～n）だけ変換する

標準正規分布（z分布）



記号表記：

$N(\mu,\ \sigma^2) \rightarrow N(0,\ 1)$

確率変数の値が単純になる

標準正規分布の確率密度関数

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

μやσ²が無くなったため，このz分布で分布表を作成すれば，（異なる平均やバラツキを持つ）どのような対象でも使える

例題　キュウリ収量の標準化

ポット番号	栽培法A	栽培法B
1	3,063	3,157
2	2,275	2,707
3	2,089	3,270
4	2,855	3,181
5	2,836	3,633
6	3,219	3,404
7	2,817	2,219
8	2,136	2,730
9	2,540	3,408
10	2,263	3,203
11	2,140	2,938
12	1,757	3,286
13	2,499	2,920
14	2,093	3,332
15	2,073	3,478
平均	2,443.7	3,124.4
標準偏差	413.5	353.0

栽培法別に標準化

=(値−平均)/標準偏差

Excel関数=STANDARDIZE(x,　平均,　標準偏差)

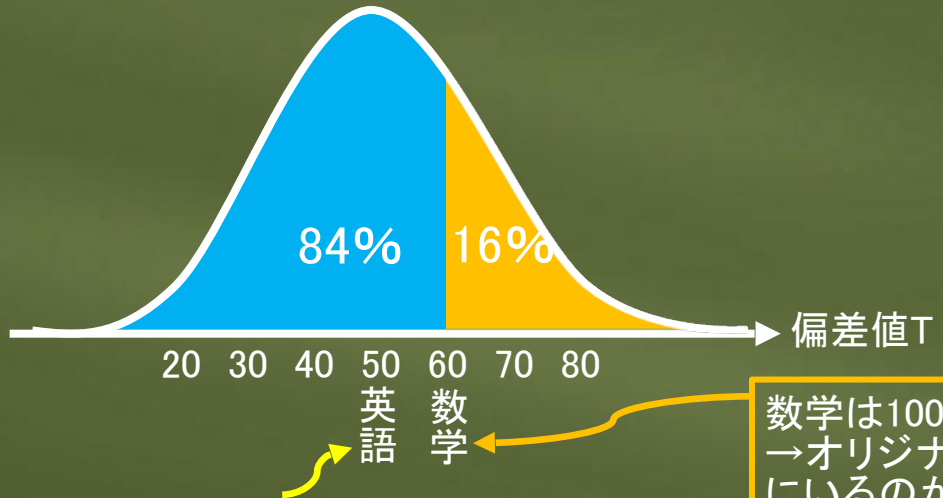
ポット番号	栽培法A	栽培法B
1	1.50	0.09
2	−0.41	−1.18
3	−0.86	0.41
4	0.99	0.16
5	0.95	1.44
6	1.87	0.79
7	0.90	−2.56
8	−0.74	−1.12
9	0.23	0.80
10	−0.44	0.22
11	−0.73	−0.53
12	−1.66	0.46
13	0.13	−0.58
14	−0.85	0.59
15	−0.90	1.00
平均	0.00	0.00
標準偏差	1.00	1.00

平均が0，標準偏差(分散も)が1になっている

偏差値も標準化変量の仲間

偏差値　T_i ＝標準化変量の10倍+50点 $\dfrac{10\times(x_i-\mu)}{\sigma}+50$

試験は0～100点のため，平均を50点，標準偏差を10点の標準化変量に変換することで直感的にグループ内の位置を捉えやすくしている



英語の点数の方が数学より高くても，平均やバラツキ次第では，数学より出来が悪いこともある

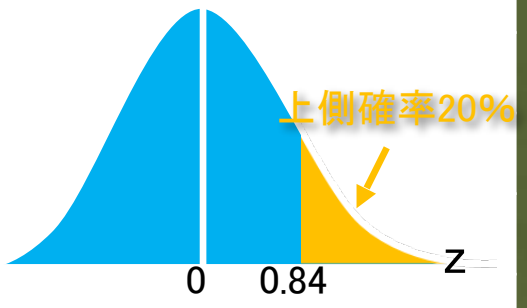
数学は100名中16位ぐらい
→オリジナルの点数では，集団のどの辺りにいるのか不明だが，偏差値が計算できれば容易にわかる

標準正規分布表の読み方（付録I）

z=○.○○（1の位と小数点第1位までは表側，小数点第2位は表頭）

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897
0.3	0.3821	0.37							
0.4	0.3446	0.3409	0.3372	0.3336	0.3300				
0.5	0.3085	0.3050	0.3015	0.2981	0.2946				
0.6	0.2743	0.2709	0.2676	0.2643	0.2611				
0.7	0.2420	0.2389	0.2358	0.2327	0.2297				
0.8	0.2119	0.2090	0.2061	0.2033	0.2005				

第1列と第1行がzの値を，表中の値はそれに対応する上側確率を表している

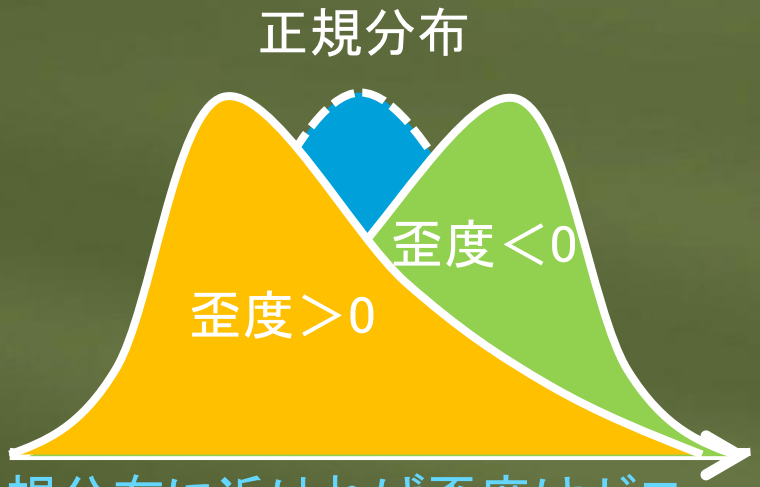


任意のzに対応する下側からの累積確率　Excel関数=NORM.S.DIST(z, FALSE)

下側からの任意の累積確率に対応するz　Excel関数=NORM.S.INV(確率)

2.5　正規分布に近い？離れている？
歪度（わいど）

- データの分布のゆがみ（非対称）の程度を示す統計量



歪度 $S_w = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$

Excel関数=SKEW.P

手元のデータの確認なので標本の平均や標準偏差

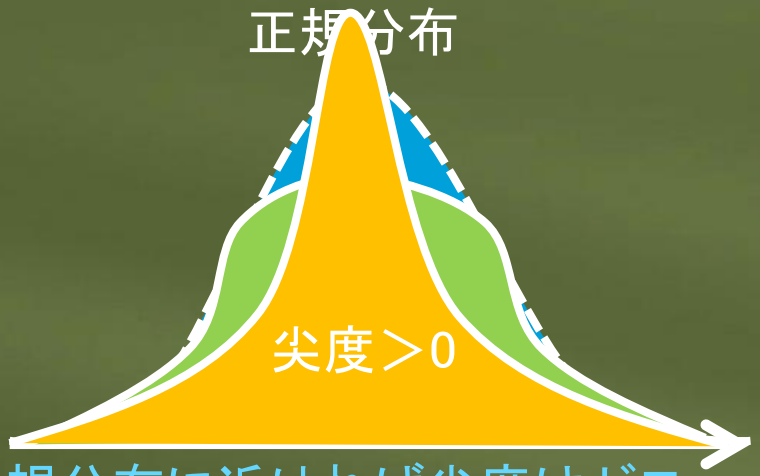
正規分布に近ければ歪度はゼロ

大きい値が相対的に多いと歪度は負の値

小さい値が多かったり，とても大きい値があると歪度は正に大きくなる

尖度（せんど）

- データの分布のとんがり度を示す統計量



$$\text{尖度}\quad S_k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$$

Excel関数=KURT

注：上式とは少し内容が異なる

正規分布に近ければ尖度はゼロ

正規分布よりも平たく鈍角に分布していると尖度は負の値

平均に近い値が多かったり，とても大きな値があると正に大きくなる

ポアソン分布

- 二項分布はカウントデータに適しているが，値の上限は試行数のnであるため，試行の多い事象は苦手
- 正規分布には上限はないが，連続型で負値も無限に取り得るのでカウントデータに適さない

→試行の大きいカウントデータには，ポアソン分布が適している

特徴：試行nが大きく，その成功確率pが小さい場合，二項分布の期待値（ある事象が起きる平均回数）n×pは一定と考えられるため，np=λ（ﾗﾑﾀﾞ）とおけば，一定の時間内に平均でλ回発生する事象がx回発生すると考えられる確率を，次のように単純な式で表せる

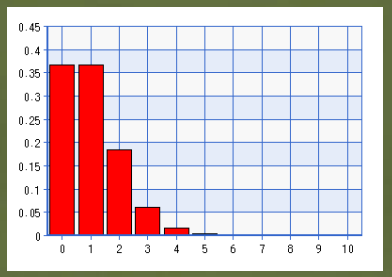
ポアソン分布の確率質量関数 $P(x) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$ ← 母数がλだけで単純

ポアソン分布の例

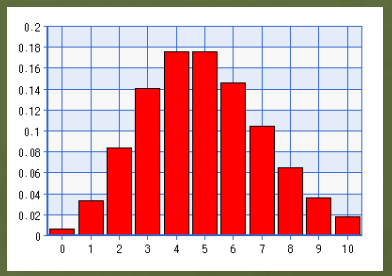
母数はλのみなので，P(λ)と表記

平均も分散もλ
（標準偏差は√λ）

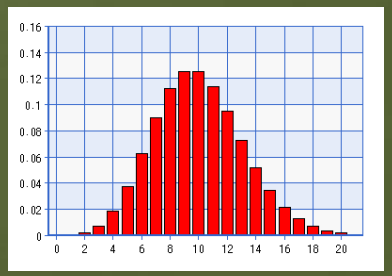
λ大きくなると正規分布に近づく



λ=1



λ=5



λ=10

Excel関数=POISSON.DIST(x, λ, 累積確率ならばTRUE/確率ならばFALSE)

ある交差点で起こる死亡事故の数
（交通量は多くても滅多に起きない）

ある工場で発生する不良品の数
（製造数はとても多いが滅多に発生しない）

事例　地震の起こる確率を
ポアソン分布で予測

- 2006年からの5年間でM7以上の地震は7回発生している（1.4回/年）。今日から3日の間にM7以上の地震が1回だけ起きる確率は？
- 解：1日で1回の試行とするとn=3，地震の確率pは1.4/365で0.0038となるので，平均となるλは，その積である0.0115です。よって，1回だけ地震（x=1）が起こる確率P(1)は，以下のように1.137%となります。

本当はもっと試行が多い場合に適している

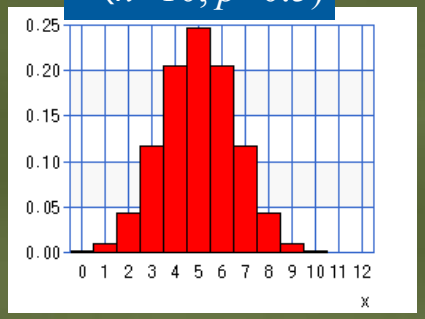
$$P(1) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-0.0115}0.0115^{1}}{1!} = 0.01137$$

Excel関数=POISSON.DIST(1,3*1.4/365,FALSE)

本章で学んだ確率分布との関係

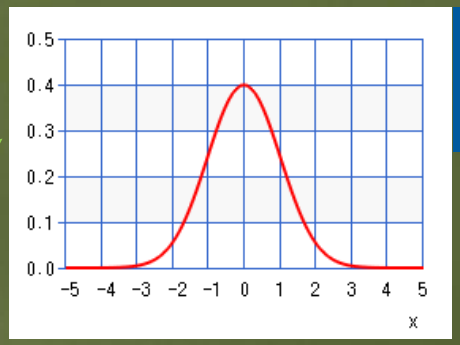
二項分布

(n=10, p=0.5)



非負で上限あり

nが大
pは普通



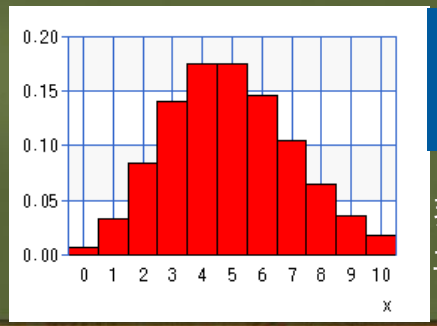
正規分布

(μ=0, σ^2=1)

連続型で
上限も下限もなし

nが大
pが小

$\lambda(=np)$が大



ポアソン分布

(λ=np=5)

非負でほぼ上限なし
平均と分散が同じ

n:ベルヌーイ試行の回数
p:個別の試行で成功する
（事象が起きる）確率

3つの確率分布の使い分け方

・離散したカウントデータには…

二項分布：試行数nが小さく，変数xの上限値が決まっており，事象が起きる確率pがそれほど大きくないとき

例：樹木10本に処置を施し，効果のあった本数を数える実験の繰り返し

ポアソン分布：試行数nが大きく（100以上），変数値の上限が決まっておらず，確率pが小さい（0.05以下）ときで，平均と分散がほぼ同じ

例：1週間に沢山生産する製品中の不良品数を，工場別に数えたデータ

・連続した値のデータには…

正規分布：試行数（データ数）nが大きく，変数の上限値や下限値が決まっておらず，それほどpが小さくないとき

例：動植物の長さや重さなどは非負だが，nが大きければOKとする

以上で第２章は終了です。