第４回　二項分布，ポアソン分布，正規分布

Ａ．代表的な分布

１．離散分布

①　二項分布

大きさ n の標本で，事象Ｅの起こる確率を p とするとき，そのうち x 個にＥが起こる確率は二項分布に従う．

その一般式は，$p(x)={}_nC_x\,p^x(1-p)^{n-x}$

ここで二項係数　${}_nC_x=\dfrac{n!}{x!(n-x)!}$

★　二項分布では母平均 $\mu=np$，母分散 $\sigma^2=np(1-p)$

例　さいころを 10 回振ったときに 1 の出る回数 x の確率分布は二項分布に従う．

$n=10,\ p=\dfrac{1}{6}$ の二項分布になる

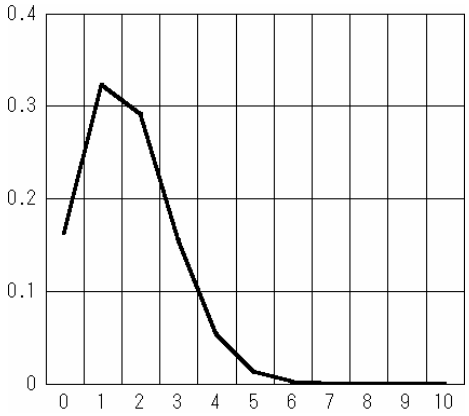
$$p(0)={}_{10}C_0\left(\frac{1}{6}\right)^0\left(1-\frac{1}{6}\right)^{10-0}=0.162$$

$$p(1)={}_{10}C_1\left(\frac{1}{6}\right)^1\left(1-\frac{1}{6}\right)^{10-1}=0.323$$

$$p(2)={}_{10}C_2\left(\frac{1}{6}\right)^2\left(1-\frac{1}{6}\right)^{10-2}=0.291$$

$$p(3)={}_{10}C_3\left(\frac{1}{6}\right)^3\left(1-\frac{1}{6}\right)^{10-3}=0.155$$

$$p(4)={}_{10}C_4\left(\frac{1}{6}\right)^4\left(1-\frac{1}{6}\right)^{10-4}=0.054$$



以下省略

エクセルで計算するときは以下の通りである．

さいころの場合				
1の出る回	確率		一般式から計算すると･･･	エクセルの二項分布関数
0	0.161506		=COMBIN(10,B5)*(1/6)^B5*(5/6)^(10-B5)	=BINOMDIST(B5,10,1/6,FALSE)
1	0.323011		=COMBIN(10,B6)*(1/6)^B6*(5/6)^(10-B6)	=BINOMDIST(B6,10,1/6,FALSE)
2	0.29071		=COMBIN(10,B7)*(1/6)^B7*(5/6)^(10-B7)	=BINOMDIST(B7,10,1/6,FALSE)
3	0.155045		=COMBIN(10,B8)*(1/6)^B8*(5/6)^(10-B8)	=BINOMDIST(B8,10,1/6,FALSE)
4	0.054266		=COMBIN(10,B9)*(1/6)^B9*(5/6)^(10-B9)	=BINOMDIST(B9,10,1/6,FALSE)
5	0.013024		=COMBIN(10,B10)*(1/6)^B10*(5/6)^(10-B10)	=BINOMDIST(B10,10,1/6,FALSE)
6	0.002171		=COMBIN(10,B11)*(1/6)^B11*(5/6)^(10-B11)	=BINOMDIST(B11,10,1/6,FALSE)
7	0.000248		=COMBIN(10,B12)*(1/6)^B12*(5/6)^(10-B12)	=BINOMDIST(B12,10,1/6,FALSE)
8	1.86E-05		=COMBIN(10,B13)*(1/6)^B13*(5/6)^(10-B13)	=BINOMDIST(B13,10,1/6,FALSE)
9	8.27E-07		=COMBIN(10,B14)*(1/6)^B14*(5/6)^(10-B14)	=BINOMDIST(B14,10,1/6,FALSE)
10	1.65E-08		=COMBIN(10,B15)*(1/6)^B15*(5/6)^(10-B15)	=BINOMDIST(B15,10,1/6,FALSE)

練習　A社のチョコレートにはくじが入っていて，当たる確率は 0.15 である．10 個買って１つも当たりが入っていない確率，２つだけ当たりの入っている確率を求めよ．さらに下の表を完成させよ．

当たりの数	エクセルでの計算式	確率
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

② ポアソン分布

ポアソン分布は一定の長さの時間，一定の大きさの空間においてごくまれに起こる事象を表現するときに用いる．二項分布において p をどんどん小さくする一方で，n を無限大にすると得られる．非常に大きな集団においてきわめて起こりにくい事象を対象としたときの分布である．

二項分布と違って，分布の大きさ n は必要ない．例えば，交通事故死はきわめてまれなものである．その対象となる n はしかも何人か決めようがない．運転者や歩行者の数は毎日異なるからである．そういうときにポアソン分布は有効である．

ポアソン分布の一般式：母平均 μ が与えられたとき，事象が x 回出現する確率は

$$p(x)=\frac{\mu^x e^{-\mu}}{x!}$$

e は自然対数の底で，$e=2.7182818\ldots$

★　ポアソン分布では，母平均 μ と母分散 σ^2 は等しい．

例　ある島では毎年，何千羽ものヒナが生まれる．毎年平均 0.5 羽の出現率で黄金色の羽をもつヒナが生まれるという．$\mu=0.5$ であるから，黄金の羽を持つヒナが 0,1,2 羽，出現する確率はそれぞれポアソン分布に従う．（ここでは誕生するヒナの数が何千羽であるかを正確にわかっていなくてもよい．そこが二項分布と違う．）

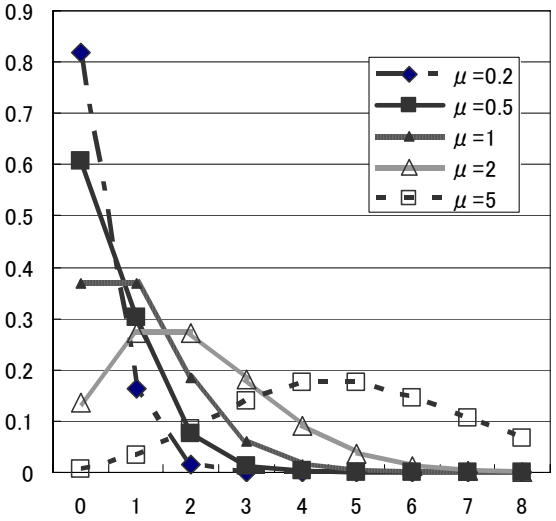
$$p(x)=\frac{\mu^x e^{-\mu}}{x!}=\frac{0.5^x e^{-0.5}}{x!}$$

=0.5^0*EXP(-0.5)/FACT(0)	0.606531
=POISSON(0,0.5,FALSE)	

$x=0$ のとき　$p(0)=\dfrac{0.5^0 e^{-0.5}}{0!}=e^{-0.5}=0.607$

$x=2$ のとき　$p(2)=\dfrac{0.5^2 e^{-0.5}}{2!}=\dfrac{0.5^2 e^{-0.5}}{2\times 1}=0.0758$

★　ポアソン分布では平均（分散も同じ）μ が決まると分布の形が決まる．



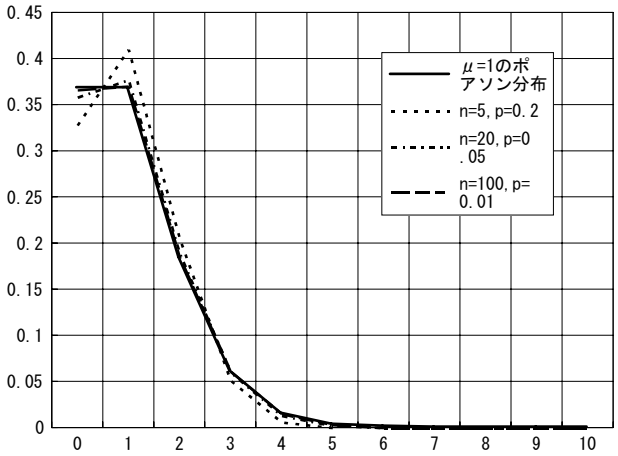
二項分布のうち $n>10,\ p<0.1$ のとき，とくに n が 50 以上，p が 0.1 以下，$np=\mu$ が 0～10 のときは，二項分布をポアソン分布でよく近似できる．二項分布は n が増えると計算が大変であり，ポアソン分布に近似すれば，n は計算上，必要なくなる．

例　ある農園で収穫したトマトでは空洞果の割合が 0.1%である．100 個を箱詰めにすると，空洞果が箱にある個数はどういう確率分布を示すか？

これを二項分布で解こうとすると 100 乗を計算することになり，電卓では面倒である．さらに n と x が大きくなるとエクセルでも計算できないときもある．

平均は二項分布の場合，np であるから，$\mu=np$ としてポアソン分布で近似すると簡単である．

空洞果	二項分布	ポアソン分布
0		
1		
2		



２．連続分布

①　一様分布（矩形分布）　すべてが同じ確率で起こる分布

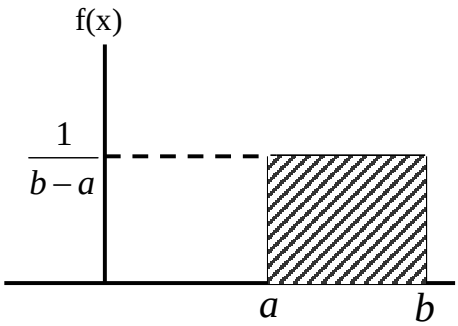


図　一般的な矩形分布

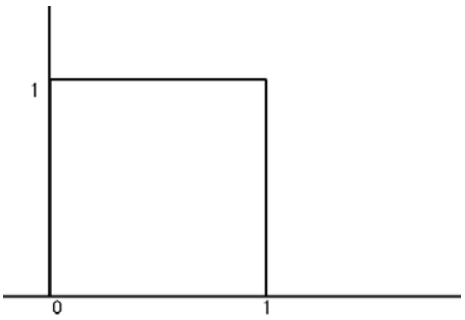
右の確率分布において

確率変数が 0 となる確率は？

確率変数が 1 となる確率は？

確率変数が 0～0.1 となる確率は？

確率変数が 0～1 となる確率は？

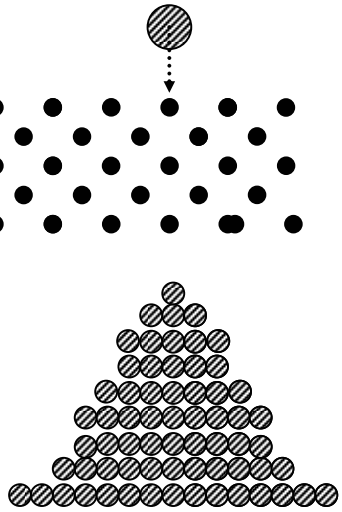


矩形分布の例

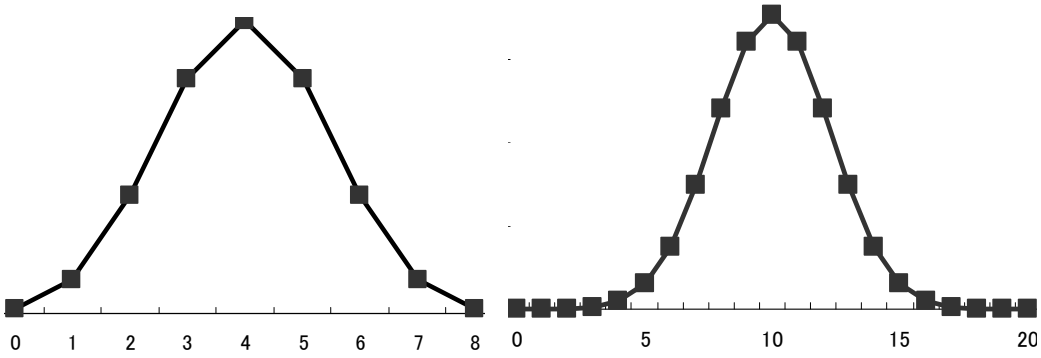
② 正規分布

二項分布で $p=0.5$ としたときに n を無限大にするとえられる．

たくさんのランダムなことが組み合わさった場合に $p=0.5$ の二項分布の極限である正規分布となるとみなすことができる．例えば，　概念的には，パチンコ台での球の分布を見ると釘に当たり右と左に行く確率は同じだとすれば，球の分布は図のようになる．これは二項分布に従う（パスカルの三角形: $n=8$，$p=0.5$ の場合と $n=20$，$p=0.5$ の場合）．



パチンコ台での玉の分布



パスカルの三角形

★ 正規分布に（近似的に）従う事象はたくさんある．

人の身長の分布，犬の体重，卵の重さなどの分布は正規分布に近似できる．

B．正規分布

1．正規分布の特徴

① 母平均 μ と母分散 σ^2 を与えると形が決まる．これを $N(\mu, \sigma^2)$ と書いて表現する．

② 平均 μ を中心にして左右対称である．よって，平均より大きい値あるいは小さい値を取る確率はそれぞれ（　，　）である．

③ 曲線は平均 μ の近傍で高く，両側に行くにしたがって単調に低くなる．

④ 平均 μ は曲線の位置を決める．平均 μ のみ異なる 2 つの曲線は左右に移動させれば重ねることができる（図 2）．

⑤ 標準偏差 σ は曲線の形を決める． σ が大きければ曲線は扁平になる（図 3）．

⑥ (a) $\mu-\sigma$ と $\mu+\sigma$ の間の確率変数を取る確率は約 0.683 である（図 4）．
(b) $\mu-2\sigma$ と $\mu+2\sigma$ の間の確率変数を取る確率は約 0.954 である．
(c) $\mu-3\sigma$ と $\mu+3\sigma$ の間の確率変数を取る確率は約 0.997 である．

⑦ 0.95（95%）の確率で $\mu-1.96\sigma$ と $\mu+1.96\sigma$ の間の確率変数を取る
0.99（99%）の確率で $\mu-2.576\sigma$ と $\mu+2.576\sigma$ の間の確率変数を取る

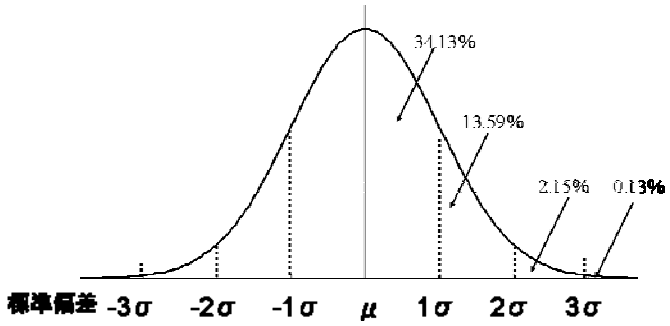


図 1　正規分布

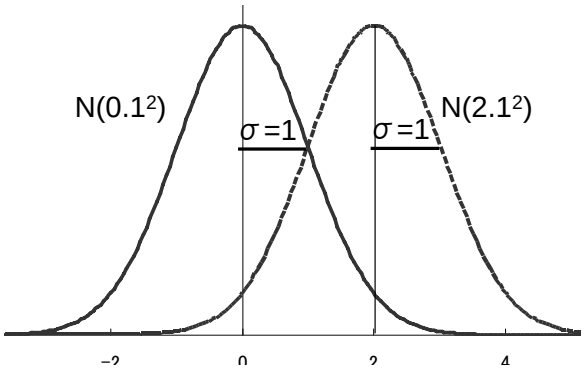


図 2　正規分布 $N(\mu, 1^2)$ の確率密度関数

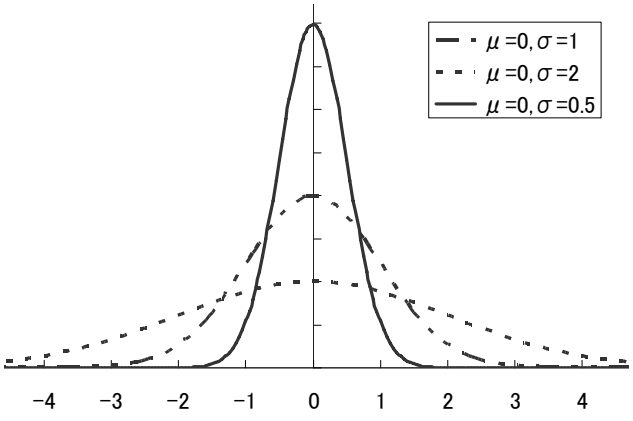


図 3　正規分布 $N(0, \sigma^2)$ の確率密度関

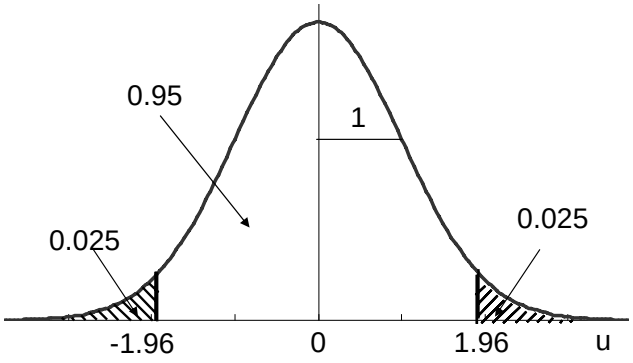
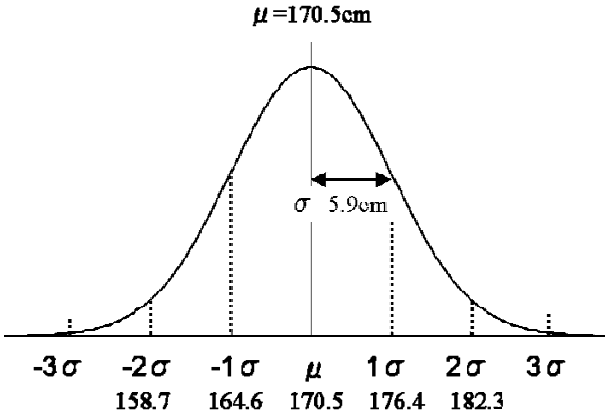


図 4　u の分布，$N(0,1^2)$

正規分布の例：20～24 歳の男性の身長は人間生活工学研究センターの調査（1992-1994）によると平均 170.5cm，標準偏差 5.9cm であった．身長の分布が正規分布するなら，2σ 以上平均より背の高い人，すなわち 182.3cm 以上は全体の 2.28%である．平均から標準偏差以内，すなわち 164.6～176.4cm に全体の約 68%が属する．全体の 95%は 158.9～182.1cm に属する．



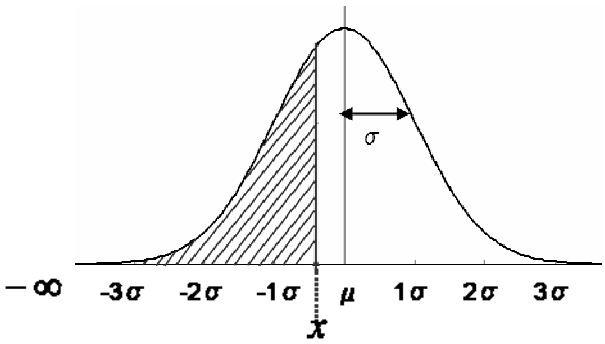
練習：30 歳代の男性の身長の平均は 169.5cm，標準偏差は 5.8cm であった．身長の分布が正規分布するなら，平均から標準偏差以内，すなわち（　　）～（　　）cm に全体の約（　　）%が属する．2σ 以下平均より背の低い人，すなわち（　　）cm 以下は全体の（　　）%である．全体の 95%は（　　）～（　　）cm に属する．

２．正規分布において任意の値と任意の値の間の範囲をとる確率をエクセルから計算する方法

連続分布であるから，正規分布において任意の値を取る確率は 0 である．任意の値と任意の値の間の範囲を取る確率を計算するにはエクセルの関数を利用するのが簡単である．

エクセルの正規分布に関する関数はいくつかある．
今回，利用するのは NORMDIST 関数である．
NORMDIST 関数は平均 μ，標準偏差 σ の正規分布において，$-\infty$（無限大）から x までの値を取る確率を以下のように入力することで計算する．

$= NORMDIST(x, \mu, \sigma, true)$

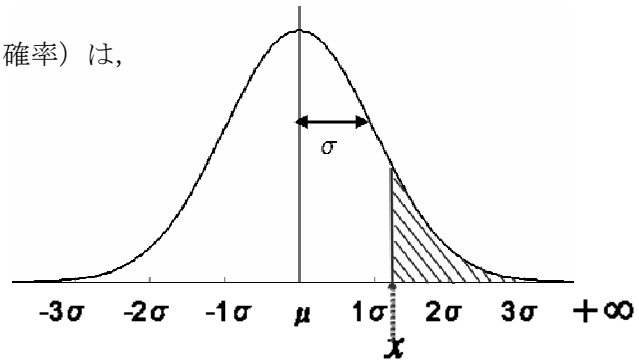


例えば，前述の身長に関する正規分布の例（平均 170.5cm，標準偏差 5.9cm）で，以下 164.6cm 以下の身長の割合は以下の式で求められる．

=NORMDIST(164.6,170.5,5.9,TRUE)	0.158655

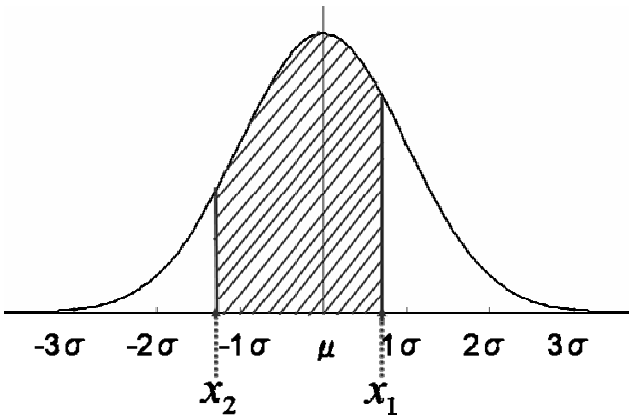
練習　20 歳代の男性の身長の平均は 170.5cm，標準偏差は 5.9cm であった．身長の分布が正規分布するなら，160 cm 以下には全体の約（　　）%が属する．175 cm 以下には全体の約（　　）%が属する．

ある値より大きくなる確率を計算するには，正規分布全体の確率は 1 となることから，下の図のように考えて，1 から下の図の斜線部分の確率を引き算すると
$= NORMDIST(x, \mu, \sigma, true)$ であるから，
すなわち，斜線部に属する確率（x より大きくなる確率）は，
$= 1 - NORMDIST(x, \mu, \sigma, true)$
として，計算する．



練習　20 歳代の男性の身長の平均は 170.5cm，標準偏差は 5.9cm であった．身長の分布が正規分布するなら，173 cm 以上には全体の約（　　）%が属する．　162 cm 以上には全体の約（　　）%が属する．

ある値（x_2）からある値（x_1）をとる確率を計算するには，$-\infty$ から x_1 までを取る確率から $-\infty$ から x_2 までを取る確率の差を取る．すなわち下の図のように計算する．エクセルでは
$= NORMDIST(x_1, \mu, \sigma, true) - NORMDIST(x_2, \mu, \sigma, true)$
として，計算する．



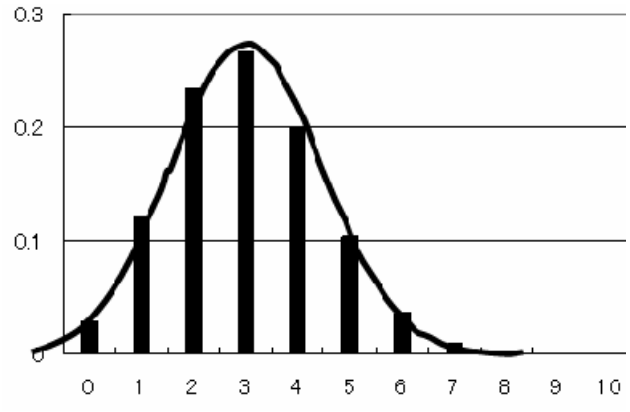
練習　20 歳代の男性の身長の平均は 170.5cm，標準偏差は 5.9cm であった．身長の分布が正規分布するなら，身長が 160～175cm の間にある人は全体の約（　　）%である．

３．二項分布の正規分布への近似

二項分布はある条件を満たせば正規分布に近似できる．

二項分布では平均は np，分散は $np(1-p)$ である． np が 3 以上であれば，この二項分布を $N(np, np(1-p))$ の正規分布で近似でき，5 以上ならよく近似できる．

正規分布は連続分布であり，二項分布は離散分布であるから，例えば二項分布で $x=5$ となる確率を正規分布で求めるには， $4.5<x<5.5$ の確率を求めたらよい．



二項分布（$n=10,\ p=0.3$）の正規分布への近似

例：ジョーカーを抜いたトランプ１組から１枚引く試行を 10 回繰り返す．赤（ハート，クラブ）を何枚引くか．n=10, p=0.5 の二項分布となる．

ジョーカーを抜いたトランプ1組から1枚引く試行を10回繰り返す。
赤(ハート、クラブ)を何枚引くか。n=10, p=0.5の二項分布となる。
N(5, 2.5)の正規分布に近似できる。

出現数	二項分布	正規分布	差
0	0.000977	0.002213	−0.00124
1	0.009766	0.011215	−0.00145
2	0.043945	0.043495	0.00045
3	0.117188	0.114468	0.00272
4	0.205078	0.204524	0.000554
5	0.246094	0.24817	−0.00208
6	0.205078	0.204524	0.000554
7	0.117188	0.114468	0.00272
8	0.043945	0.043495	0.00045
9	0.009766	0.011215	−0.00145
10	0.000977	0.002213	−0.00124

C．宿題

1．第２回の宿題で調べたデータについて①は二項分布，②はポアソン分布で予想される分布とどの程度離れているかを以下の手順で検討せよ．

① それぞれ二項分布，ポアソン分布に従っているとして，確率分布を求めよ．なおポアソン分布の計算で用いる母平均 μ は調査したデータの平均を用いたらよい．

② ①で求めた確率分布のヒストグラムの上に，先週，調べたデータから作ったヒストグラムをトレーシングペーパーなどで書き写したものを，縦軸，横軸の大きさがそろうように重ねて，２つの違いを検討せよ．もし，大きく異なるときはなぜかを考えてみよ．

2．ある分布を正規分布とみなしてよいかを判断するには，厳密にはコルモゴロフ・スミルノフの検定を利用する．第３回の授業の宿題で調べたデータについて，正規分布で予想される分布とどの程度離れているかを以下の手順で検討せよ．

① 調査したデータの標本平均，標本分散をそれぞれ母平均，母分散とした正規分布とすると，理論的には 68.3%の確率変数は $\mu-\sigma \sim \mu+\sigma$ の間にはいる．自分の調べたデータのうち，何%が実際には $\mu-\sigma \sim \mu+\sigma$ の間にはいるかを数えて調べよ．

② さらに $\mu-2\sigma \sim \mu+2\sigma$ ， $\mu-3\sigma \sim \mu+3\sigma$ の間にはいるデータについても数えよ．

③ ①，②の結果から，自分の調査したデータが正規分布に近いかどうかを検討せよ．

3．第３回の授業の宿題で調べたデータについて，そのデータが正規分布で近似できると仮定した場合，第３回の宿題採点表で指定した範囲に属するデータが全体の何パーセントになるかをエクセルの NORMDIST 関数を用いて，計算せよ．

4．次回以降の授業ではあるデータが正規分布あるいは二項分布に基づくと仮定して，統計的に推定あるいは検定を行う．次回の授業では統計的な推定について考える．統計的な推定とは，母集団についてのある数値を知るために，母集団から無作為抽出した標本からデータを集め，そこから母集団についてのある数値を推測することである．統計的な推定の例を以下にいくつかあげる．

① 20 歳代の男女を 100 人，無作為抽出して，1 ヶ月の携帯電話代を調べたところ，平均 5000 円，標準偏差 500 円だった．母集団（20 歳代の男女すべて）の携帯電話代はいくらかを推定したい．

② A 農場は B スーパーにトマトを納入している．B スーパーは仕分けの手間を省くためにトマトの重さの標準偏差を 1g 以内にすることを求めてきた．すべてのトマトの重さを測定できないので，100 個を無作為抽出して標準偏差を調べたところ，0.99g だった．A 農場のトマト全体の重さの標準偏差はいくらと推定できるだろうか？

③ C 林業は D 山を開発しようと考えたが，開発の利益に出費が見合うかわからない．D 山の樹木すべてを調べることは不可能なので，無作為に数地点を選んで樹木の価値を算定し，D 山全体の樹木の価値を推定した．

以上のような例に当てはまる事例をいくつか考えてみよう．