5月

大分学習センターのオープンユニバーシティカフェ

越智　義道　先生（統計科学）

統計学の活用法について

４月のカフェでは、データにはばらつきを含むものがあること、ばらつきとはどういうものかを考えました。そして、何が起きるかわからないけれど、その起こりやすさについては何がしかの知見があることを話し合いました。統計学の手法はそれらをふまえる必要があります。そこで、

５月は、**ばらつきを扱う方法**　について考えてみたいと思っています。

日時　14:00～16:00　**5月15日（水）**

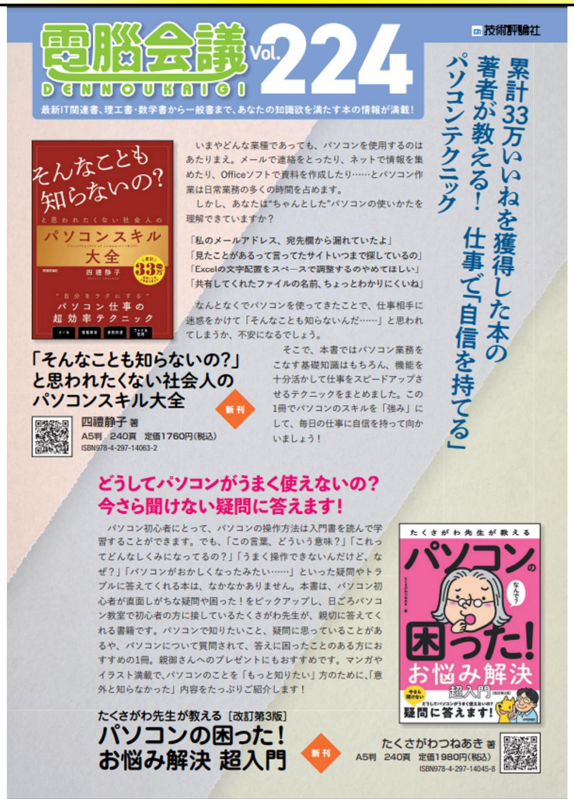
※日程は都合により変更になる場合があります。センターウェブサイトでご確認ください。

情報と計算機

- 情報の処理を支える技術：情報工学
- 計算機の進歩　電話や映像も計算機で
- 計算機でできること

計算機＝Computer＝電脳

PC:　Personal Computer
パソコン



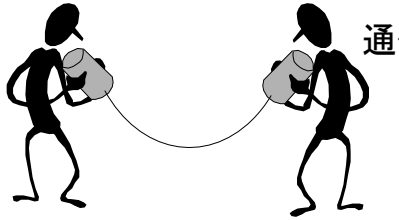
2024.4.13発行

計算機でできること

- 計算（数値計算，数式処理）
- 推論・演繹，翻訳
- 検索（電話帳，時刻表，地図，文献，情報検索）
- 文書処理（ワープロ，ホームページ作成）
- 画像処理（デジカメ，仮想現実VR，画像認識）
- 通信（E-mail，インターネット，動画配信，携帯電話）
- などなど

情報処理

情報検索



通信・コミュニケーション

データの分析

知識・表現・理解・伝達

情報処理

情報とデータ

データ（X01.01.02 ）

情報の表現であって，伝達，解釈又は処理に適するように形式化され，再度情報として解釈できるもの。

データ

事実，事象，事物，過程，着想などの対象物に関して知り得たことの表現であって，伝達，解釈又は処理に適するように形式化され，再度情報として解釈できるもの。

情報（X01.01.01 ）

事実，事象，事物，過程，着想などの対象物に関して知り得たことであって，概念を含み，一定の文脈中で特定の意味をもつもの。

日本工業規格（JIS） 情報処理用語−基本用語より

情報とは

人間が，外部からのさまざまな変化を感覚を通じて受け取ったデータを処理，判断し，人間にとって役に立つものを情報という。

また，現在コンピュータは人間のように，自分で処理，判断することができないので，コンピュータの内部に記録されたものや，通信ネットワークを流れるものはデータである。
このデータを人間が状況に応じて，利用判断することで情報に変わる。

教科「情報」講習テキストより

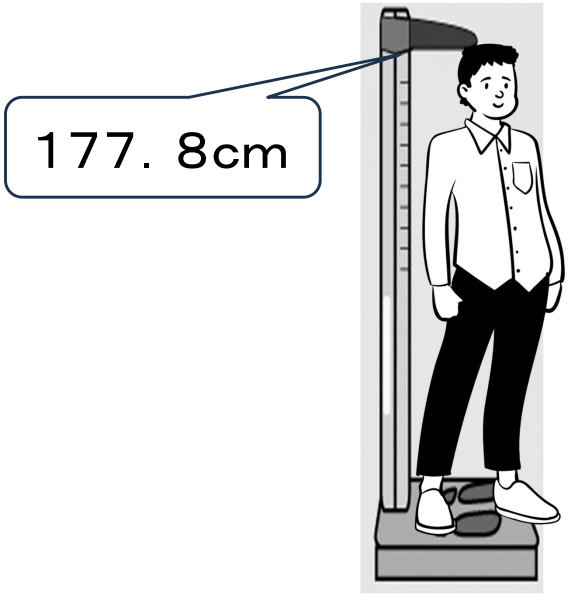
データを処理し，

人間にとって役に立つもの（概念を含み，一定の文脈中で特定の意味をもつもの）

が　情報

そもそもデータとは

事実，事象，事物，過程，着想などの対象物に関して知り得たことの表現であって，伝達，解釈又は処理に適するように形式化され，再度情報として解釈できるもの。



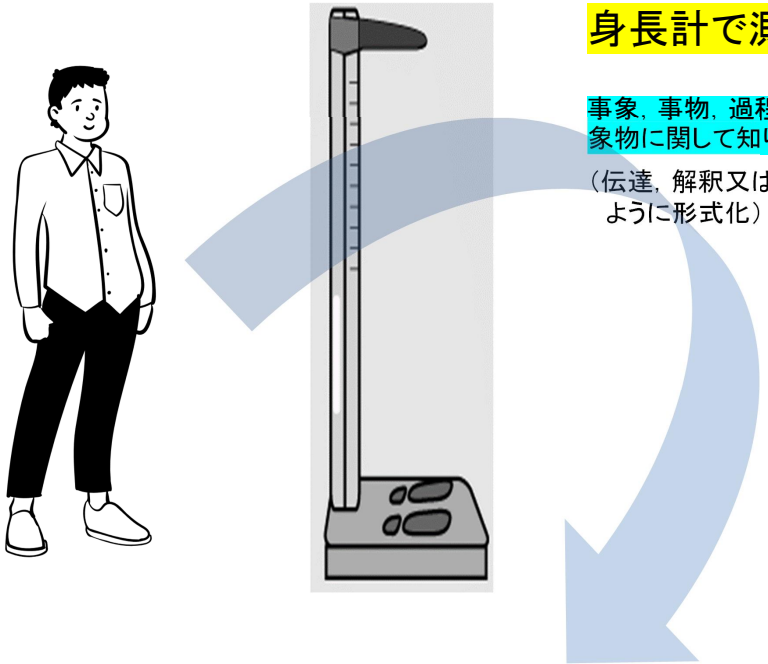
身長計で測る

事象，事物，過程，着想などの対象物に関して知り得たことの表現

（伝達，解釈又は処理に適するように形式化）

日本
177．8cm
アメリカ
5’ 10”
(5 feet 10 inch)
昔の日本
5尺8寸6分7厘

太郎さんの身長：？



身長計で測る

事象，事物，過程，着想などの対象物に関して知り得たことの表現

（伝達，解釈又は処理に適するように形式化）

太郎さんの身長：177．8cm

身長計で計測　（伝達，解釈又は処理に適するように形式化）

177．8cm　174．2cm　185．3cm　178．4cm　176．7cm

この5人の平均身長：178．5cm（178．48cm）

身 長 計 で 計 測 （伝達，解釈又は処理に適するように形式化）

177．8cm　174．2cm　185．3cm　178．4cm　176．7cm

太郎さんの身長：177．8cm

情報処理

データの中から必要な（重要な）情報を取り出す。

データにばらつきを考えなくてもよいとき

検索（Search）　蓄積方法・検索方法（Retrieval）

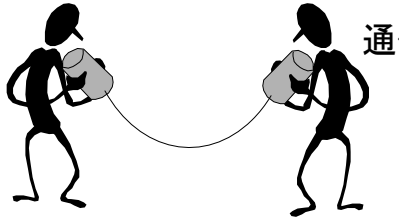
データベース

データにばらつきがあるとき

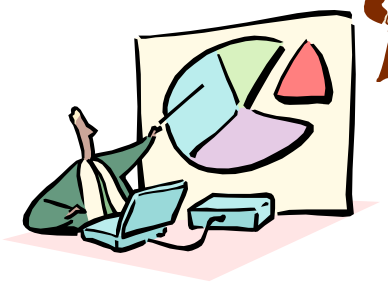
分析・解析（Analysis）　統計的データ解析

統計科学

情報検索



通信・コミュニケーション



データの分析

知識・表現・理解・伝達

統 計 科 学 ・ 統 計 学

- データを入手して必要な情報を抽出する技術
  分析・解析
  統計的データ解析
  ⇓
  抽出された情報は意思決定の基礎となる

- 観測，実験，アンケート　⇒　データ

  – 適切に収集されたデータ
  – 適切に設定された問題認識

  } 実験計画（デザイン）

統計的データ解析のエッセンス

データは　本質的に　重要な要素　と　ばらつき　をもつ。

⇧

これが情報として大切

ばらつき（確率的変動）　を含む

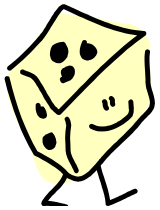
確率的　ランダム　でたらめ

でたらめ　　とはいうけれど

まったく　でたらめ　でもない。

起きやすさ　について　なにがしかの知識がある

（これも重要な情報の一つ）



ばらつきの世界

ばらつきを扱う方法

さいころの

1の目の出る確率が　$\frac{1}{6}$　とは？

さいころを6回振って1の目が1回出ること　　ではない

さいころを60回振って1の目の出る回数は　10回　でもない

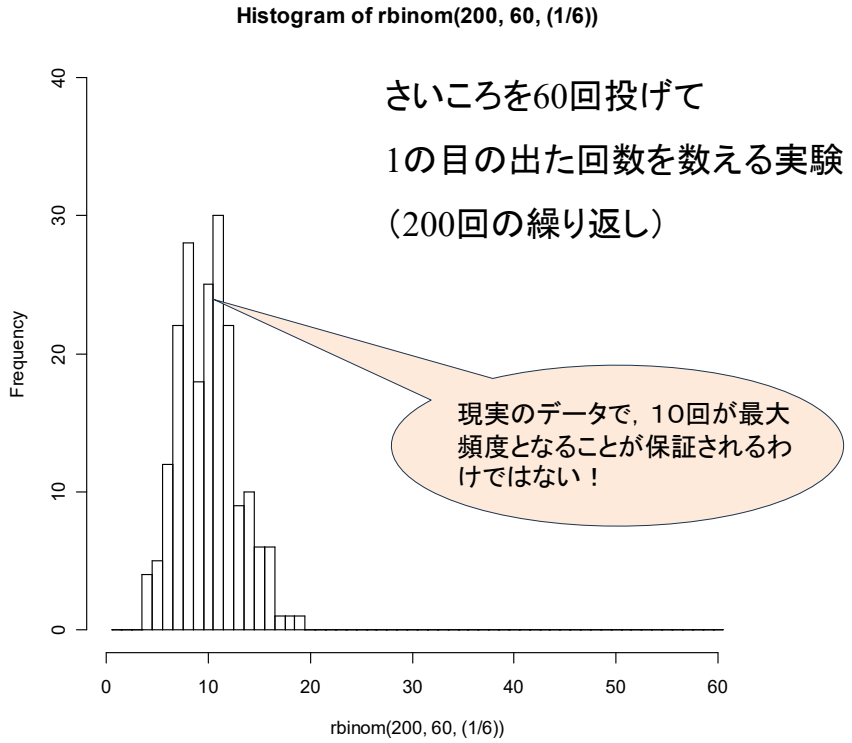
0~60回だけど何になるか　分からん！！　　けれど

60回全部1が出る，なんてほとんどありそうもないね

1回も出ないこと，もあまりなさそうだし….

でも，起きんともかぎらんよ！！

10回ぐらいなら，おきても　よさそう　だよ



さいころを60回投げて

1の目の出た回数を数える実験

（200回の繰り返し）

確率分布

ばらつきを把握するための道具

$$\Pr(X=k) = {}_nC_k\, P^k(1-P)^{n-k} = \frac{n!}{k!(n-k)!}P^k(1-P)^{n-k}$$

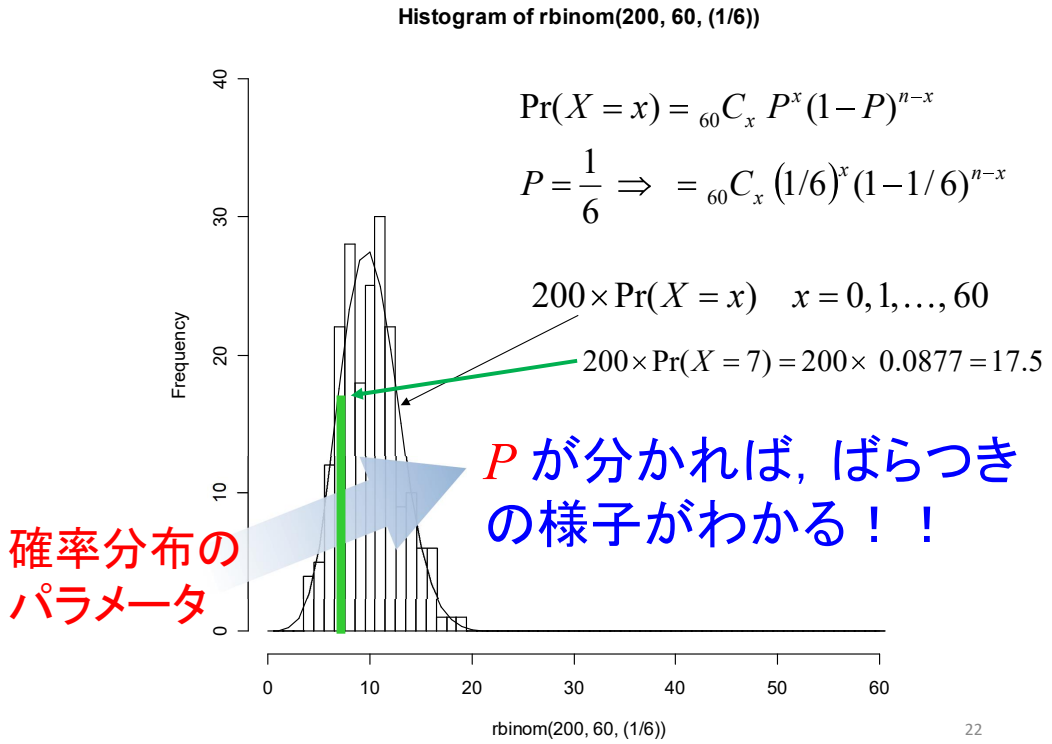
ある事象の生起確率が P であるとき，n 回の試行の後，k 回の事象の生起を確認する確率　（2項確率・分布）

X：確率変数：観測・計測・測定などを表す変数
この例では，ある事象の生起回数を示す変数

さいころの目の 1 が出てくる確率を 1/6 としたとき，60回さいころを振った後に，7回ほど1の目の出てくる確率は，

$$\Pr(X=7) = {}_{60}C_7\left(\frac{1}{6}\right)^7\left(1-\frac{1}{6}\right)^{60-7} = \frac{60!}{7!(60-7)!}\left(\frac{1}{6}\right)^7\left(\frac{5}{6}\right)^{53}$$

$$= 0.08773144$$



$$\Pr(X=x) = {}_{60}C_x\, P^x(1-P)^{n-x}$$

$$P=\frac{1}{6} \Rightarrow = {}_{60}C_x\,(1/6)^x(1-1/6)^{n-x}$$

$$200\times\Pr(X=x) \quad x=0,1,\ldots,60$$

$$200\times\Pr(X=7)=200\times\ 0.0877=17.5$$

確率分布のパラメータ

P が分かれば，ばらつきの様子がわかる！！

いろいろな棒グラフ（ヒストグラム）

- > hist(rbinom(200,60,(1.0/6)),ylim=c(0,40),br=0:60+0.5)
- > lines(0:60, 200*dbinom(0:60,60,(1.0/6)))
- >
- > BH_graphs( )

```
bin_hist<- function(Rep=200,Size=60, Prob=1/6, Count=1){
  Title=paste("Histgram of Binomial Trial",
  "Rep=", Rep,"Size=",Size,"P=", round(Prob,3),"   [", Count,"]");
  hist(rbinom(Rep,Size,Prob),ylim=c(0,40),br=0:Size+0.5, main=Title)
  lines(0:Size, Rep*dbinom(0:Size,Size,Prob) )
}

BH_graphs<-function(Rep=200,Size=60, Prob=1/6, Times=30){
    for(k in 1:Times){
        Sys.sleep(1); bin_hist(Rep, Size, Prob,Count=k)
    }
}
```


したがって　サイコロの目の1の出方について言及するには，その確率 P を知ることが本質的に重要！！

これも？　あれも？

きっとちがうよね。

1/6に近いだろうけど。

でも，やっぱり・・・

ちがうはず。

さいころの目の出かたを知る。

⇔

その目の出る　確率　を知ること。

実際のサイコロでの目の出る確率は？

分からない！！

なんとか知りたい！　　どうする！

データを集めて，調べる。

データからの情報（確率）の抽出（推定・近似）

60回投げて1の目の出た回数が7回だった。

確率＝7/ 60　？？？　　ではなくて，確率≒7/60

なぜなら，　もう一度60回投げたときは・・・

データをとりなおすたびに推定された確率は違っている。

サイコロは同じなのに！！

では，　$\hat{P} = \left(\frac{1\text{の目の出た回数 } x}{\text{サイコロを投げた回数 } n} \right)$

は役に立たない？？？？

ことはない

投げる回数を十分大きくすると，$\hat{P}$ は真の確率に近くなる

大数の法則

```
> rbinom(1,60,(1.0/6))/60
```

[1] 0.1　#60回投げた場合

```
> rbinom(1,200,(1.0/6))/200
```

[1] 0.145　#200回投げた場合

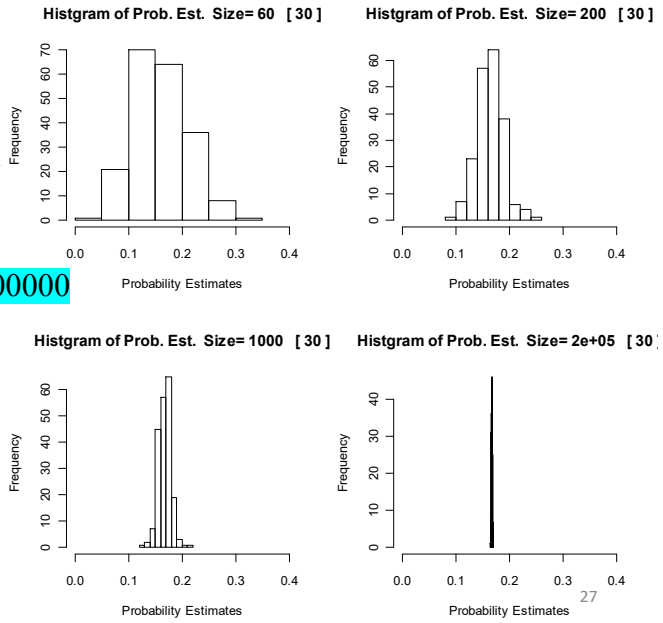
```
> rbinom(1,200000,(1.0/6))/200000
```

[1] 0.168325

#200000回投げた場合

真の値

1／6＝1．6666666・・・



大数の法則

```
> rbinom(1,60,(1.0/6))/60
[1] 0.1
> rbinom(1,200,(1.0/6))/200
[1] 0.145
> rbinom(1,200000,(1.0/6))/200000
[1] 0.168325
```

```
> LL_graphs()
```

```
l_large_num<-function(Sizes=c(60,200,1000,200000),Prob=1/6){
 v<-1:length(Sizes);  for(i in v) v[i]<-rbinom(1,Sizes[i],Prob)/Sizes[i]
 c(Prob,v)
 }

Loop_lln<-function(Sizes=c(60,200,1000,200000),Prob=1/6,Rep=200){
  VV<-l_large_num(Sizes, Prob);
  for(i in 1:Rep-1){VV<-rbind(VV, l_large_num(Sizes, Prob))}
  list(VV=VV,Sizes=Sizes)
 }

LL_graphs<-function(
Sizes=c(60,200,1000,200000),Prob=1/6,Rep=200,Times=1:30){
 par(mfrow=c(2,2));
 for(time in Times){
  Sys.sleep(1);
  V<-Loop_lln(Sizes, Prob, Rep );
  for(i in 1:4) {
    Xlab<-paste("Probability Estimates");
    Title=paste("Histgram of Prob. Est.  Size=",V$Sizes[i],"  [", time,"]");
    hist(V$VV[,i+1],xlim=c(0, 0.4),main=Title,xlab=Xlab);
  }
 }
 par(mfrow=c(1,1));
}
```

ここまで学んできたこと

- データはばらつきを含む。
- ばらつきにもその起き方を扱う手立てがある。それが，確率・確率分布という考え方だった。
- 確率分布を表現するもの，それが本当にほしいもの（情報）である。
- データから計算するものは，その近似である。

５月（５月１５日（水）実施）のカフェでは，４月のカフェの内容を振り返ったのちに，

・まず，前回深堀していなかった，情報とデータという言葉の持つ意味を考えました。
※情報：information(uc)，データ：data/datum(c)

・計算機と情報処理の立ち位置や役割を考え，情報とデータという言葉の違いと関係を，日本工業規格や教科「情報」に関するテキストの記載で確認しました。

・さらに，データを得るということを，身長と身長計をもとにデータと情報の役割を考え，統計的な観点での，データから情報を得るということの意味を話し合いました。

５月（５月１５日（水）実施）のカフェでは（２）

・そのうえで，前回のカフェで議論したばらつきについて話を戻し，データのもつばらつきを（科学的に）扱う方法として，確率や確率分布を表す数式を利用することが考えられていることを紹介しました
（さいころの１の目が出る回数を数える状況を例に，
２項分布という確率分布 を紹介）。

・２項分布の確率関数と2項乱数について，確率分布の数式表現と，実際のデータの現れ方との関係を確認しました。

５月（５月１５日（水）実施）のカフェでは（３）

・確率を数式で表現するときに，ばらつきの様子を特徴づけるために，重要な役割をなしている数式中の変数のことをパラメータと呼ぶことを紹介し，実際の現象から，このパラメータを探すには，どうすればよいかについて考えました。

・さいころを投げる例で，そのパラメータ（１の目のでる確率）の値を近似するために，
１の目の出た回数と投げた回数の比率
を用いると，投げる回数を増やせば，
その比率が真のパラメータの値に近づく性質があること
（対数の法則）を確認しました。

参考：いろいろな（よく使われる）確率分布

離散型の確率分布

この分布のパラメータは　P

1　2項分布　B(n, P)

それぞれが影響を及ぼさない状況で

1 回の試行で事象 A の起こる確率が P である試行を 独立に n 回反復して行う。
このとき事象 A の起こる回数を X とする。

確率変数 X の確率分布は，次の確率関数 $f(k)$ で表される。

$$f(k)(=\Pr(X=k)) = {}_nC_k P^k (1-P)^{n-k} = \frac{n!}{k!(n-k)!} P^k (1-P)^{n-k} \quad (k=0,1,2,\ldots,n)$$

（ただし，$0<P<1$ ）

このとき，X の平均 $E(X)$，分散 $V(X)$ は

$E(X)=nP$，　$V(X)=nP(1-P)$　（従って，X の標準偏差は $\sqrt{nP(1-P)}$ ）

となる。

このとき，確率変数 X は2項分布に従う（あるいは，2項分布を持つ）と言い，
X ～B(n, P) と表す。

離散型の確率分布

この分布のパラメータは　λ

2　ポアソン分布　Po(λ)

確率変数 X の取りえる値が　0, 1, 2, 3, …　（加算無限個・加付番無限個）であり，
この確率変数 X の確率関数 $f(k)$ (k=0, 1, 2, 3, …) が

$$f(k)\ (=\Pr(X=k)) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad (k=0,1,2,\ldots)$$

で与えられるとき（ただし，λ は $\lambda>0$ となる定数），確率変数 X はポアソン分布に従う
（あるいは，ポアソン分布を持つ）と言い，X～Po(λ)と表す。

◎一定期間内に数え上げを行う様な観測に用いられる。
- 1 時間に，ある交差点を通過する自家用車の台数
- 阿蘇山の1ケ月の噴火の回数

など

X の平均: $E(X)=\lambda$，X の分散 $V(X)=\lambda$

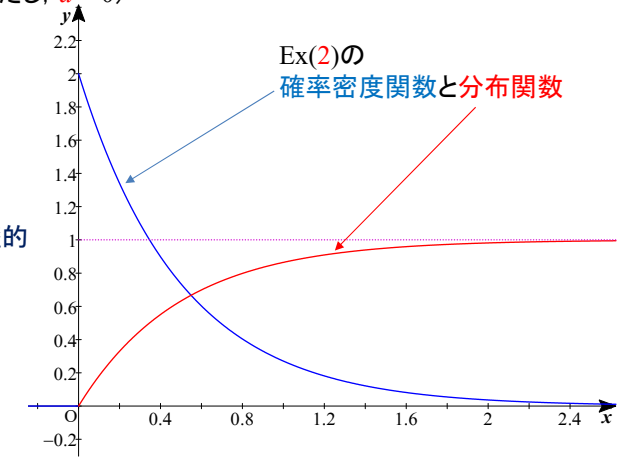
連続型の確率分布

この分布のパラメータは　a

1　指数分布　Ex(a)

確率変数 X の確率密度関数　$f(x)$　が次の式で与えられるとき

$$f(x)=\begin{cases} ae^{-ax} & x\geq 0 \\ 0 & \text{その他} \end{cases} \quad （ただし，a>0）$$



X は指数分布 Ex(a) に従うと言い，

X～Ex(a) と表す。

寿命時間の分析に用いられる基礎的な分布

X の平均 $E(X)$, 分散 $V(X)$ は

$$E(X)=\frac{1}{a},\quad V(X)=\frac{1}{a^2}$$

X の分布関数 $F(x)$ は

$$F(x)\left(=P(X\leq x)\right)=\int_{-\infty}^{x} f(t)dt=\begin{cases} 0 & x<0 \\ \int_0^x ae^{-at}dt=1-e^{-ax} & 0\leq x \end{cases}$$

連続型の確率分布

この分布のパラメータは　μ と σ^2

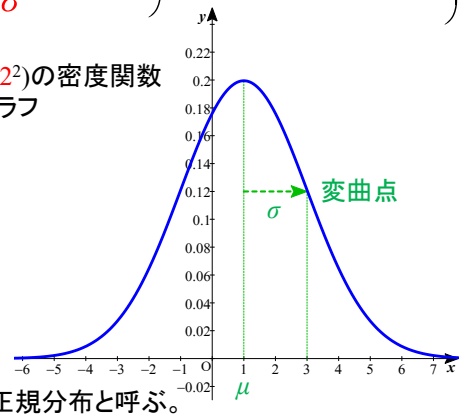
2　正規分布　N(μ, σ^2)

確率変数 X の確率密度関数　$f(x)$　が次の式で与えられるとき

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

$$\left(\text{ただし},-\infty<\mu<\infty,\sigma>0\right)$$

N(1, 2^2)の密度関数のグラフ



X は正規分布 N(μ, σ^2) に従うと言い，

X～ N(μ, σ^2) と表す。

X の平均 $E(X)$, 分散 $V(X)$ は

$$E(X)=\mu,\quad V(X)=\sigma^2$$

特に，μ=0, σ=1 のときの正規分布 N(0,1) を標準正規分布と呼ぶ。

X～ N(μ, σ^2) であるとき，　$Z=\frac{X-\mu}{\sigma}$　とすると　Z～ N(0, 1) となる。

このような確率変数の変換のことを標準化変換と呼ぶ。