データを科学するための統計解析の基礎

**By　tonkyo\_Vc**

# はじめに

しかし、その中身を理解できるとか、使いこなせる人が世の中にどれだけいるかというと、どうもあまりたくさんいるわけではないようです。実際のところ世の中企業もアカデミアもデータ解析にはいろいろ苦労していて、いわゆる「データサイエンティスト」なる人物が実際は絵に描いた餅であったり口裂け女の如き都市伝説的存在であったりするのではないかとまことしやかに囁かれているという噂も後を絶ちません(あくまで誇張表現ですので真に受けないように)。

勿論世の中にはれっきとしたデータ解析を専門業務とする人、データ解析のスキルが極めて高い人も存在しています。一例しか挙げませんがネットを見れば「六本木で働くデータサイエンティストのブログ(<http://tjo.hatenablog.com/>)」のTJOさんとかそのような人は結構いますし、不肖私の職場にも実際に極めてデータサイエンティストとしてのスキルの高い人は複数います。

本屋に行ってみると、最近「統計」に関する本がたくさん並んでいます。数学関係のコーナーなどでは所狭しと「サルでもわかる統計学」とか「数式なしでも完全マスター統計解析」とかいった名前の本が占拠しています(あ、名前はテキトーな思い付きですw)。こんなに統計の本が人気、ということはたぶん、今の世の中統計が強く求められているのでしょう。加えて近頃、データサイエンスとかビッグデータとか、人工知能とか強化学習、深層学習といった言葉が流行っています。そういったことに関する本も山ほど出ていますので、それを目にした方もたくさんいることかと思います。実際、大量のデータや情報が世に氾濫し、何が正しい情報で何が誤った情報か識別することが難しくなっていますし、そういった錯綜した情報を整理して正しい情報を把握するために統計とかデータサイエンスいうのは有効なツールとして人気があるのでしょう。

しかし、その中身を理解できるとか、使いこなせる人が世の中にどれだけいるかというと、どうもあまりたくさんいるわけではないようです。実際のところ世の中企業もアカデミアもデータ解析にはいろいろ苦労していて、いわゆる「データサイエンティスト」なる人物が実際は絵に描いた餅であったり口裂け女の如き都市伝説的存在であったりするのではないかとまことしやかに囁かれているという噂も後を絶ちません(あくまで誇張表現ですので真に受けないように)。

勿論世の中にはれっきとしたデータ解析を専門業務とする人、データ解析のスキルが極めて高い人も存在しています。一例しか挙げませんがネットを見れば「六本木で働くデータサイエンティストのブログ(<http://tjo.hatenablog.com/>)」のTJOさんとかそのような人は結構いますし、不肖私の職場にも実際に極めてデータサイエンティストとしてのスキルの高い人は複数います。

ということで、統計は世の中では理系文系問わず、数学の中では最も広く使われている分野の一つと言えると思いますし、機械学習とかディープラーニングなどもそのユーザは非常に裾野が広がっています。ところがその割には、小学校~大学の算数・数学の授業の中で確率とか統計の部分(特に統計の部分)は結構ないがしろにされているようにも思えます。実際、大学の入学試験要項などには「数学B（確率. 分布と統計的な推測を除く）」とかいうことが書かれていたりします。なので、一応カリキュラムには入っているものの、実質的には授業とかでもスルーされている場合が非常に多いというのが現状です。たぶん、これは統計というものが他の数学と違ってバシッと答えが決まるとかそういうことがない上に、推論の形式が極めて回りくどいという考え方の面倒くささに加え、特に計算機を持ち込めない(持っていても普通の電卓では役に立たない場合が多いのですが)入学試験などでは正規分布表などを配らないとできないなどといった操作上の面倒くささ(PC持ち込み可にしたらできるのですが、それやると他のテストが実質的に無意味になりますよね^^)のために、入試の対象から不幸にも外されてしまっているためなのではないかと思っています。受験で必要ない、ということになると皆さん勉強をやる気もなくなり(私もそうでしたw)、結果として非常に「難しい」ものというイメージが不当にも発生してしまっているのではないでしょうか。

因みに私はといえば学生時代の専門もその後の専門も博士号自体も全く統計と関係なさそうな分野(実験生物学)で、統計とかコンピュータなぞそれこそよくわけのわからぬ修羅の世界というイメージのまま過ごしてきました。ところがある日突然当時の研究所長に呼ばれ、

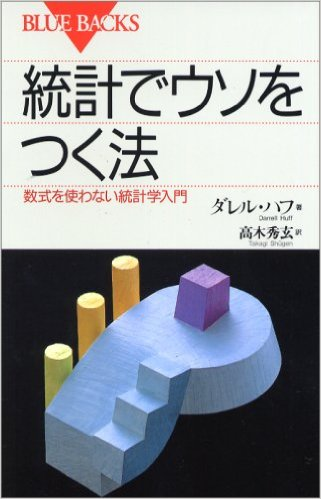
所長「お前これから統計屋になれや」(後から考えるとそれが管理職に昇格するための条件だった様です)

私「え″～マヂすか？」(なんか裏に条件あるっぽいし断ることは無理だな、とうすうす感じつつ)

という流れで統計屋に流れ着いた中年のおっさんです。とはいえ、実際やりだすと統計解析というのは案外面白いもので、仕事とは別にいろいろ自分で手を動かして見様見真似で解析のようなものをやってみたりとしているうちになんだか沼に入ってしまったという感じですね。

同時に、学生の時の統計の授業などで躓くポイントというのがいくつかありますが、なぜ躓くのかというのも何となくわかってきたような気もします。凡そ大学の教科書とか授業というのは学問的な厳密性を重んじたり(少なくとも私が受けた大学の授業はそうでした)、天下り式に授業を行ったりというものが多く、それはそれで重要なことではあるのですが、逆に直感的に「わかる」ことを歓迎していないのではないかと受け取れるふしもあり、その結果その分野へアクセスする人を減少させているのだとすればちょっと残念なことでもあるかと思います。

また、そうやって大学の授業を受けて単位を取った人()は数多いるにもかかわらず、世の中では専門家においてさえも統計の誤用があとを絶たないという現状もありますし、私のような凡人では想像すら及ばぬような超解釈を行うデムパな方も結構な頻度で遭遇します。ググるとそういった例は山ほど出てくるのですが、例えば、「生物医学研究文献の誰でも見つけられる20の統計学的誤り」(<http://www.ronbun.jp/point/follow_jp.html>)や、「ダメな統計学」といった本が出版されているという事実からも(<http://id.fnshr.info/2017/01/20/sdw-trans-publ/>)、その惨状は垣間見ることができます。また、世の中には統計を駆使してガセネタ、というか事実としては間違っていないけれどデータの誤った解釈を流通させるという悪徳商売が横行していたりもします。それどころか、こんな本も出回っていることを考えるときちんと正しい統計の知識を持っていることはいい加減な情報に振り回されないためにも重要だと思います。だからといって逆にこれを使って世の中にいい加減な情報を流そうとはゆめゆめ思わないでくださいね。



この様な事実を考えると理論の面からも厳密によく理解しているというのは非常に大事なのですが、逆にあまり教条主義的に求めた結果統計に関して理解しようという気が失せてしまうという側面も同時にあるとは思います。ですので、きちんとした学習は世の中に教科書として定評のある書籍が沢山ありますので、そういったものを活用していただくことにして、ここでは主に実際のデータをどのように解剖して必要な情報を適切に取り出すかというプロセスの理路を追うことを通じて、皆様に統計解析の考え方、適切な使い方をざっくりと理解していただき、更にはいわゆる「データサイエンス」へつながるような部分の要点についてまとめていこうと考えています。

なお、統計解析を行うためにはある程度コンピュータの解析ソフトが使えることは必要になると思います。例えば世の中には以下のようなソフトウエアがありますが、それぞれ特徴があります(私の個人的評価ですが)。

Excel(有料、もっとも最初からPCに入っている場合も多いですね)：

〇　操作は簡単

〇　データ加工はやりやすい

×　解析方法が限定されている

×　トレーサビリティー(データ解析プロセスの再現性)が取りにくい

×　関数がアレ

SPSSとかJMPとか(有料)：

〇　操作はまあまあ簡単

〇　解析方法も比較的豊富

〇　トレーサビリティーも比較的取りやすい

×　サイズの大きいデータやデータ加工には不向き

GraphPad Prism(有料)：

〇　操作は簡単、グラフ作成が簡単

〇　解析方法も比較的豊富

〇　トレーサビリティーも比較的取りやすい

×　サイズの大きいデータやデータ加工には不向き

×　グラフ作成に特化、解析メインの人には不向き

SAS(有料)：

〇　医学・薬学系の(かつての？)ゴールドスタンダード

〇　解析方法は豊富

〇　データ加工も容易でサイズの大きいデータも扱える

〇　トレーサビリティーが取りやすい

×　プログラミングができないと厳しい

×　高すぎる(＞1000万円)ので個人で購入するのは絶対無理

R(無料)：

〇　SASなどの有料高性能ソフトにとって代わって今は広くデータサイエンスのデファクトスタンダードに

〇　解析方法は極めて豊富、なにより無料！

〇　データ加工も容易でサイズの大きいデータも扱える

〇　トレーサビリティーが取りやすい

×　プログラミングができないと厳しい

PythonとかJuliaとか(無料)：

〇　統計だけでなく様々な数学的解析にも利用可能、徐々にデータサイエンス業界でのデファクトスタンダードになりつつある

〇　解析方法は極めて豊富、なにより無料！

〇　データ加工も容易でサイズの大きいデータも扱える

〇　トレーサビリティーが取りやすい

×　プログラミングができないと基本手も足も出ない

ということで、私がすすめるのは無料のRかPython、Juliaですが、Rは統計解析・機械学習に特化していて関連図書も多数あること(Pythonに関する書籍も多数あるが、統計解析を中心に論じたものは思いのほか少ない、Juliaに関する日本語書籍は今のところ1冊しかなさそう)、統計解析に関する関数が豊富であることもあり、一番とっつきやすいものかと思います（なお、Juliaは個人的にはまだ機能の拡張性という面で若干未完成な気がします）。ですので、ここではRを使って解析の実例を示しながら話を進めていこうと思います。なお、Rのインストールについてもググると情報はたくさん出てきますが、通常のRのソフトウエアを動かすといまいち見にくいとか使いにくいという面がありますのでRをインストールしたら同時にRstudioという統合開発環境（IDE）もインストールしてそこで使うことをお勧めします(こちらも無料です)。インストール方法については三重大学の奥村先生のページ(<https://oku.edu.mie-u.ac.jp/~okumura/stat/R-win.html>　)などを参考にしていただければと思います。なお、使い方についてもここではスクリプトは示しますが、特に細かい説明は行いません。これについても上記HPやその他にもかなり親切に使い方を説明してくれているサイトがあります(日本語のサイトのみ挙げています)。

名古屋市大のページ　<https://sites.google.com/site/webtextofr/home>

東工大間瀬先生のページ　<http://www.is.titech.ac.jp/is-wiki/?maselab%2FR>

R-Tips　 <http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html>

なお、Rの使い方、Rを使った統計解析とかデータサイエンス関連の書籍は無数に出ていますので、本屋で斜め読みして相性のよさそうなものを買って2～3冊持っておくこともよいと思います(それぞれ重点を置いている場所が異なっていたりするため)。

# 1．データの特徴を探る

## 平均と分散

さて、ここにデータがあります。あるテストをA組とB組の生徒に対して実施した時の点数一覧です。ここから何が言えるか考えてみましょう。まず最初に思いつくのは人によって点数が異なるということですね。また、A組とB組でどっちができが良いのか、といったことも気になる点かと思います。とはいえこのままだと数字の羅列でこれくらいの数ならまだ我慢できるレベルですが、これが何千人とか何百クラスとかの数字が並んでいたら何が何だか分からなくなるのは想像に難くありません。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **出席番号** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| A組 | 48 | 60 | 39 | 53 | 54 | 61 | 56 | 50 | 40 | 48 | 55 | 44 | 45 | 30 | 47 | 56 | 50 | 70 | 57 | 63 |
| B組 | 50 | 64 | 36 | 53 | 54 | 66 | 60 | 51 | 40 | 50 | 58 | 43 | 48 | 35 | 50 | 58 | 52 | 70 | 63 | 69 |

ということで、まずはこのデータを少し見やすくしてみようと思います。そのためにデータをRに打ち込んでみましょう。

rm(list=ls(all=TRUE)) #ワークスペースを初期化  
data<- data.frame("A"=c(48,60,39,53,54,61,56,50,40,  
 48,55,44,45,30,47,56,50,70,57,63),  
 "B"=c(50,64,36,53,54,66,60,51,40,  
 50,58,43,48,35,50,58,52,70,63,69))

まず、データを見やすくするためには度数分布表とかヒストグラムというものがあります。度数分布表は点数の範囲、例えば50～59点の間に何人、というような形で示した表です。度数分布表はこんな形で関数を作って出力することができます。

度数分布表の作成

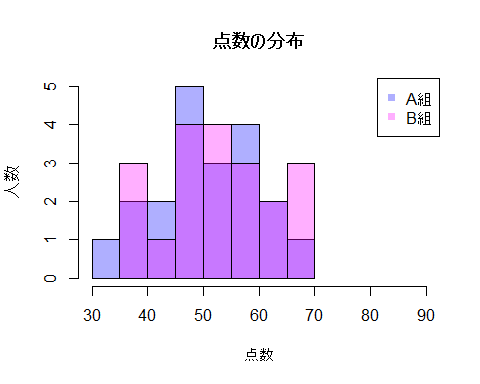
#度数をカウントする関数  
tab.count<- function(x){  
 s<- list()  
 s[[1]]<- x<30  
 s[[2]]<- x>=30&x<40  
 s[[3]]<- x>=40&x<50  
 s[[4]]<- x>=50&x<60  
 s[[5]]<- x>=60&x<70  
 s[[6]]<- x>=70  
 count<- rep(0,length=6)  
 for (i in 1:6){  
 count[i]<- length(which(s[[i]]))  
 }  
 return(count)  
}  
out<- apply(data,2,"tab.count")  
counttable<- data.frame("点数"=c("~29","30~39","40~49","50~59","60~69","70~79"),  
 out)  
print(counttable)

## 点数 A B  
## 1 ~29 0 0  
## 2 30~39 2 2  
## 3 40~49 6 3  
## 4 50~59 8 9  
## 5 60~69 3 5  
## 6 70~79 1 1

次にデータを見る方法としては、ヒストグラムと呼ばれる度数分布表を棒グラフであらわしたものがあります。例えばA組、B組で重ね合わせて出力するようにするとこんな感じになります。

ヒストグラムの出力

hist(data$A,xlim=c(30,90),col="#0000ff50", breaks = 12,  
 main="点数の分布",xlab="点数",ylab="人数")  
hist(data$B,xlim=c(30,90),col="#ff00ff50",add=TRUE,breaks=12,  
 main="",xlab="",ylab="")  
labels <- c("A組","B組")  
legend("topright",legend=labels,col=c("#0000ff50","#ff00ff50"),pch=15,cex=1)



どうでしょう。こうすることによって最初の表の持つ情報がだいぶわかりやすくなってきましたね。ただ、これだと視覚的にはだいぶいい感じにはなってきましたけど、この先データを扱うという意味ではもうちょっと工夫が必要です。そのためにはデータをもっとシンプルな情報に圧縮する必要があります。この圧縮された情報を一般に統計量と呼び、その代表的なものが平均であったり分散であったり標準偏差であったりします。平均は数式としては

と表される統計量で、データの数値の総和をデータ個数で割ったもの、分散(標本分散)は

と表され、データがどれだけばらついているかを示すものになります。悪名高い偏差値で有名な標準偏差は

という統計量です(標本標準偏差)。さて、それぞれRで計算するとこんな感じで求めることができます。因みにRには標本分散、標本標準偏差を計算する関数がないので、これも自作で標本分散の関数を作って計算してみます。

平均、分散、標準偏差の計算

apply(data,2,"mean") #標本平均

## A B   
## 51.3 53.5

#標本分散を算出する関数(Rにはもともとない)  
var.sample<- function(x){  
 x.ave<- mean(x)  
 var.sample<- mean((x-x.ave)^2)  
 return(var.sample)  
}  
  
apply(data,2,"var.sample") #標本分散

## A B   
## 81.31 98.45

sqrt(apply(data,2,"var.sample")) #標本標準偏差

## A B   
## 9.017206 9.922197

こんな結果が出てきます。平均も分散も標準偏差もA組＜B組ですね。つまり、このことは平均点はA組よりもB組の方が2点ほど高い、データのばらつきもA組よりB組の方が大きい、ということを示しています。

## 共分散と相関係数

#rm(list=ls(all=TRUE))#初期化  
data("women")#womenデータの呼び出し  
print(women)#データの出力

## height weight  
## 1 58 115  
## 2 59 117  
## 3 60 120  
## 4 61 123  
## 5 62 126  
## 6 63 129  
## 7 64 132  
## 8 65 135  
## 9 66 139  
## 10 67 142  
## 11 68 146  
## 12 69 150  
## 13 70 154  
## 14 71 159  
## 15 72 164

#定義に従い、共分散を計算する  
t<- vector(length=nrow(women))  
for (i in 1:nrow(women)){  
 t[i]<- (women$height[i]-mean(women$height))\*  
 (women$weight[i]-mean(women$weight))  
}  
cov<- mean(t)  
print(cov)

## [1] 64.4

#相関係数の計算  
rho<- mean(t)/sqrt(var.sample(women$height)\*var.sample(women$weight))  
print(rho)

## [1] 0.9954948

ただこれだけだと統計としてはまだその一番おいしい部分を味わったことになりません。このような有限母集団を対象とした統計(記述統計)はかなり昔に成立していたのですが、統計の醍醐味はこれを背後にある無限母集団を想定し、「Aという属性を持つ集団とBという属性を持つ集団の間に違いがあるかどうか」といったことを理論的に考察することができるというところにあります。そのためにはちょっと遠回りになりますが、次のセクションでは確率と確率分布、そして確率分布の持つ性質というものに触れておこうと思います。若干面倒くさい数式が出てきますが、それに慣れておくといろいろ応用が利くものであるらしいです。

# 2．確率と確率分布

ということで確率についてざっくりと確認をしていきましょう。ところが、確率をまじめに定義しようとすると結構面倒くさいのです。現代では公理的確率論から(抽象的な)確率空間というものを定義してそこから始めるというのが一般的です。で、それを確立したのがコルモゴロフというおっさんでして、「確率論の基礎概念」という本にまとめられています。ちょっと引用してみましょう。

「要素の集合をとし、の部分集合を要素とする集合族をとする。を根源事象といい、を標本空間(根源事象の空間)、の要素を確率事象(または単に事象)という。

1. は集合体である
2. の各集合に、非負の実数が定められている。この数を事象Aの確率という。
3. とBが共通の要素を持たないとき

**公理Ⅰ～Ⅳを満たす3つの組を確率空間という。」**

更に、連続性の公理

1. 「の事象の減少列について、ならば、が成り立つ」

**公理Ⅰ～Ⅴを満たすものを確率空間（公理Ⅰ～Ⅳを満たす3つの組を広義の確率空間）と呼ぶ**というのが現代の確率の定義です。これではわけわかりませんね。ごく平たく言えば、事象に対して確率を対応させるような関係が成立し、その確率の最小値は0であるような集合によって満たされているような空間を確率空間と呼ぼうというものです。まあ、これは「確率」自体に何か経験的な意味(つまり我々が通常確率という言葉でイメージするような実体的なもの)を与えずにまずこういう性質を有するものを確率と定義し、結果として我々が通常使っている確率というものはこの性質を満たしていますよね、という極めて現代の数学に特有の回りくどい形式をとっています。この話を真面目に論じると(そもそも私にかような能力はありませんが)それこそ教科書何冊にもわたる壮大なお話になってしまうのでそこはバッサリとカットして、実際に必要な確率の性質に移りましょう。

## 確率の基本的な性質

基本定理：確率P()は次を満たす

1. 空事象に対して
2. が互いに排反なら　が成立
3. 任意の事象Aに対して
4. 任意の事象AとBに対して

なお、は和事象(or)、は積事象(and)、は余事象(not A)をそれぞれ意味する。

独立の定義：

AとBがを満たすとき、AとBは独立であるという。

条件付き確率の定義：

である事象Aと事象Bに対してAが起きているという条件のもとでBが起こる確率について、これをAのもとでの(given)Bの条件付き確率といい、  
と定義する。

この条件付き確率から次のベイズの定理が導かれます。

ベイズの定理：

がを満たす互いに排反な事象であるなら、任意の事象Bについて

が成立する。

ベイズの定理をもとに展開される統計学はベイズ統計学といい、現在統計学は伝統的(?)な推測統計学とベイズ統計学の2種類に大別されているということもできます。本論ではベイズ統計についても少々扱う予定ですが、基本的には伝統的なものを扱います(医学・生物系ではまだまだ伝統的な統計学が主流であることもあり)。

## 確率分布

さて、確率と一言でまとめていますが、実際には大きく2つ(離散型と連続型)に分けれらます。先の公理的定義②にあったように、事象の種類がそれに対応します。例えばコインを投げると、表か裏いずれかが上になる(立つ、というのもあるとかいうツッコミはしないでね)、という様に互いに排反のとびとびの値をとるような場合と、例えば身長の様に値が連続的に、原理的には場合によっては無限大の事象が発生し得るような場合があります(まあ現実には存在しえませんが、どこかで上限値として打ち切るというのはそれはそれで微妙な問題を含みますので、無限小～無限大で考える、あるいは0～無限大で考えるというのがいろいろ実用的に都合が良いということです)。前者の裏表の様な事象を離散型確率変数(裏表は確率の世界では変数です!)と、後者の身長の様な変数(これは直感的に変数であると理解できますね)を連続型確率変数と呼びます。

離散型確率変数に対しては、各事象(しつこいようだけど変数です)に対応する確率を表す関数を

と定義し、確率変数Xがxの値をとる確率を示します。つまり、X＝{表、裏}でx＝表、x＝裏になる確率が存在し、コインは投げたら表か裏かのいずれに「必ず」なる、ということからXの取り得る値に対して

が成立し、それ以外のxについてはとなります(しつこいけどコインが立つという場合はなしということで)。このとき、このをXの確率分布と呼びます。

また、連続型確率変数については、ある特定の取り得る値に確率を与えても、それは実質的にゼロに等しいのであまり意味がない場合がほとんどです。そこで、a,bを定数(a>b)として

という積分の形であらわされます。

このをXの確率密度関数といい、

になります。

さて、統計ではこの確率密度関数の他に、もう一つ累積分布関数というものを考えることが多いです。それがどのように使われるかはおいおい話を進めるにつれてわかってくると思いますので、とりあえずここでは定義だけしておきましょう。Xと確率変数とし、

と定義される関数をXの累積分布関数と呼びます。離散型分布の場合、必ずしも変数間に順序関係を導入できないものも多いですが(裏と表、どっちが大きいんだよ！的な)、順序関係がはっきりしているものに使うということでそこはあまり深く考えないでください。Xが離散型の場合、

となり、連続型の場合は

となります。これより、が連続である点で

が成立します。

さて、前節で示したような確率の諸性質は確率分布、確率密度関数に対しても当然のように成立するわけです。その前に、複数の確率分布についての同時分布というものを考えてみましょう。

離散型確率変数X,Yに関して、

をX,Yの同時確率関数、連続型確率変数X,Yに対して

をX,Yの同時確率密度関数と呼びます。これらも基本的に可能な全事象の確率の和は1になります。

そして、同じように

をX,Yの同時分布関数、この時のXの分布だけを見た関数

をXの周辺分布関数と呼びます。

さて、確率変数X,Yが独立のときは、

が成立します(必要十分条件)。さらに、条件付き確率については

と定義されます。

## 期待値とかモーメントとか

さて、コインを投げた時に表が出たら100円もらえ、裏が出たらお金をもらえない、という状況を考えてみましょう。1回コインを投げた時にもらえる金額は、コインがきちんとしたものであるならば

円

と期待できることになります。これを期待値と呼び、一般的に離散型確率変数についてはを適当な関数として(この場合は*g*(表)=100、*g*(裏)=0です)

と表されます。連続型確率変数に対しては

と表されます。期待値には以下の様な性質があります。これはなんとなく直感的にわかりますね。

1. X,Yが独立なら

ここで、自然数mを考え、とした場合、この期待値を確率変数Xのm次のモーメントと呼びます。これが何の意味があるかって?もちろん意味はあります。これは分布のパラメータを出す際になかなか便利なものなのです。ただ、実際にこれを定義に従って計算するとなかなか面倒くさいので、更にXが非負の整数値である離散型確率変数の場合は確率母関数

を、Xが連続値の確率変数である場合にはモーメント母関数

というものを考えます。t=0の近傍でこれをm階微分してみましょう。あら不思議(不思議じゃないですけど)、

となり、Xのm次モーメントが出てきます。なんだかおいしそうな感じしますよね?ただ、この場合掛け算が出てくるので、更にこの対数を取ったもの

をキュムラント母関数といい、ここでの例示は省略しますが、これは特に独立な二つ以上の確率変数X,Y,…の和であらわされる確率変数のモーメントを求めるときなどに威力を発揮します。対数を使う利点は掛け算をもうちょい簡単に足し算で表現することができる、というところにありますので、これも分布の特性値(パラメータ)を求めるときに用いられます。

## 分布のパラメータ(もう一度平均とか分散とか、更には相関とか)

さて、前節で確率分布の特徴を示す変数(これもまた確率変数だったりするのですが)を導くための数学的準備ができました。

まず前節の期待値の定義で出てきた(離散型の場合)、(連続型の場合)をそれぞれ*xk, x*とすれば、それが平均になります。つまり、

　(離散型確率分布の場合)　、　(連続型確率分布の場合)

がそれぞれ平均の定義になります。これは確率変数Xの１次のモーメントにほかなりません。

さて、平均は以下のような性質を持ちます。

なお、は独立であることを意味します。

ここで定義されるとして確率変数を考えてみます。当然、この一次モーメント　です。次にこの２次のモーメントを考えると、

離散型の場合、、連続型の場合となります。これが分散の定義になります。さて、分散は以下のような性質を持っています。

分散をとして、その正の値の平方根***σ***が標準偏差になります。

更に、分布の特徴を表すものでより高次のモーメントによってあらわされるものとして、歪度、尖度というものがあります。

歪度は、

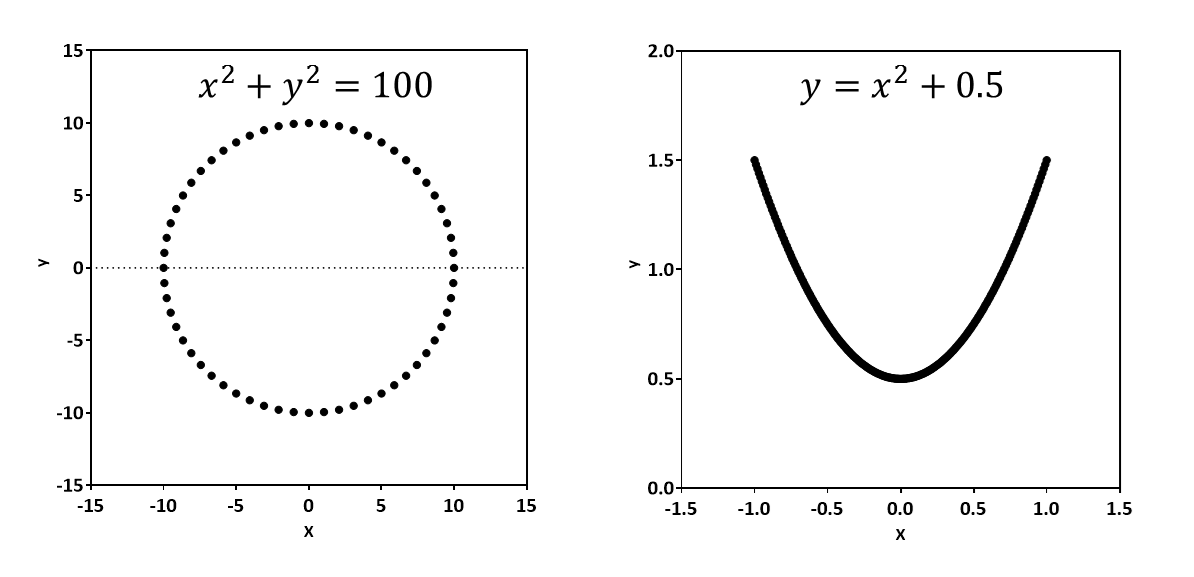
で、分布がどれくらい左右対称から歪んでいるかを示すもので、これが正の値の場合標準正規分布(平均0、分散12の正規分布)左側(負の値側)に分布のピークが寄っている、０のときに左右対称、負の値のときに右側にピークが寄っていることを意味します。これはXの3次のモーメントを標準偏差で規格化したことになります。そして尖度は、

で、なんで-３なんていう項が突然乱入してくるのか訝しむむきもあろうかとは思いますが、この値が正のとき、正規分布に比較して分布の裾が短くなり、ピーク形状が鋭くなる、負の場合に裾が広くなり(ロングテールになり)、ピーク形状が緩やかになります。これはXの４次のモーメントを同じように規格化し、3だけずらした、ということになります。

さて、今度は同時分布に関する特性を考えてみましょう。2つの確率変数X,Y(この場合独立かどうかは仮定しません)に対して共分散を以下のように定義します。

この値はXとYの間に正の相関関係（Xが大きくなるとYも大きくなる、Xが小さくなるとYも小さくなる）があるときに正の値を取り、負の相関関係があるとき（Xが大きくなるとYは小さくなる）に負の値を取ります。先ほどの平均の性質であったようにX,Yが独立、つまり相関関係がないときはが成立しますのでこの値は0になります。で、この共分散をX,Yの標準偏差で規格化した値

を相関係数と定義します。なお、です。ところで、ρ=0は相関関係がないことの必要十分条件ではありません。例えば、下図のような場合、ρ=0ですがx,yの間に相関関係がないということはできませんよね。あくまで相関係数は直線的なX,Yの比例関係を表すものであるということは注意しておく必要があります。例えば以下のグラフのような分布の場合、相関係数は0になりますが、xとyの間が無関係ということはないですよね。なので、データを見るときは単純に数字で表される統計量だけでなく、グラフなどを描いてみて情報を得るというのは重要だったりします。



## 確率分布の性質

Xを非負の値をとる確率変数とすると、任意のに対して、

が成立します。これをマルコフの不等式と言います。

次に、Xの平均E(X)をµ、分散V(X)をσ2とすると、

が成立します。これをチェビシェフの不等式と呼び、これは確率変数が期待値から遠く離れた値を取る確率の上限を定めるもので、例えばとしてみると、平均から 2×標準偏差以上離れた値は全体の 1/4 を超えることはない、ということを意味することになります。

更に、マルコフの不等式とチェビシェフの不等式から大数の法則が導かれます。これは、確率変数が独立に同一の分布に従い、平均=*μ*で、分散が存在するとき、

が成立するというもので、ある同じ確率分布に従うデータが沢山あれば、その平均値は確率分布の期待値に収束するということで、これが成立するからこそ「とりあえずたくさんデータ取れば真の平均値はわかるでっしゃろ」ということが成立します。なお、後述するように平均も分散も存在しないような確率分布というものがあります(例としてコーシー分布というものがあります)。この場合はこの法則が成立しません(右辺の値が存在しないので当たり前ですね)。

いよいよ確率とか統計の世界で最も有名で重要な定理(まあ有名だからといって必ずしも重要であるとは限りませんが、これは大事な定理です)の中心極限定理の登場です。母集団の平均：、母集団の分散：に従うn個の標本： 、があるとき、

母集団の分布がどんな分布であっても、nが十分大きいとはに漸近するという性質があります。なお、は平均0、分散の正規分布を意味します(正規分布については次章で説明します)。これは母集団の分布がどんな分布であっても、その誤差の分布はデータ数を増やせば近似的に正規分布に従うということになりますが、さらっと書いていますが結構驚くべき性質です。この定理があるからこそ正規分布は確率分布の王座の地位を占めているといっても過言ではないと思います。

大数の法則は極限として期待値が現れ、中心極限定理はその期待値がどの程度の速さで収束するのかの情報を与えてくれます。

# 3．統計でよく使う様々な確率分布

さて、確率とか確率分布を考えるうえで必要な情報を前章で大体整理できたと思います。今度は具体的な確率分布の特徴について順番に概観していくことにしましょう。基本的には統計解析でよく使うものはそんなに種類が多いわけではないので、これらを中心に説明をしていきたいと思います。中でも重要（と個人的に思っている）なのが正規分布、*t*分布、*Χ*２（カイ２乗）分布、*F*分布という正規分布をもとに構成される連続型確率分布で、これらは検定とか推定の中で重要な役割を果たしています。なお、それぞれの確率分布がどのような形状をとるかは、[確率分布いろいろ](http://statdist.ksmzn.com/)のサイトが便利です。ここはちょっと面倒くさい数式の羅列になってしまいますが、ちょっと我慢してくださいね💜。

## 離散型確率分布

離散型の確率変数の基本は「あるかないか」だといってもよいでしょう。要するに１ビットの(0,1)の情報です。これはコンピュータの情報の取り扱い方の基本でもありますし、2進法の表記方法でもおなじみです。

### 離散型の一様分布

「ある」が1,2,…nという数値に対応していると考えたとき、

となる分布を一様分布と呼びます。例えば、サイコロが1,2,3,4,5,6のそれぞれの目を出す確率はいずれも1/6ですね。この時、分布の期待値は、分散はになります。

### ベルヌーイ分布

「ある→1、ない→0」を対応させて考えていったとき、最もシンプルなものは1回の行為で1がどれくらいの確率で出てくるか、ということになると思います。これを表すものがベルヌーイ分布というやつで、

という形であらわされます。これ自体はそんなに出てくるものではありませんが(統計で1回の試行というのはまずありえないわけで)、ベルヌーイ分布の期待値=*p*、分散=はすぐに求めることができます(やってみてね♡)。なお、確率母関数はです。

### 2項分布

ベルヌーイ分布はコインを1回投げた時の確率だと考えられます。これを複数回繰り返したときの分布が2項分布であると考えることができます。2項分布の分布関数は、以下の式であらわされます。

なお、2項分布の期待値=*np*、分散=*np(1-p)*です。2項分布には再生性という特徴があります。これは、*X1,X2*が独立でかつとすると、であるというもので、2項分布に従う2つの確率変数の和もまた2項分布になるというものです。確率母関数はで、ベルヌーイ分布の確率母関数のn乗になっています。

### ポアソン分布

nが十分大きい時、2項分布は正規分布に漸近します。さて、今度はnpを一定に保ったまま、n→∞、p→0にした時のことを考えてみましょう。これは一定時間内にある事象が何回か起きる、という様な現象を考えるとわかりやすいかと思います。これがポアソン分布で、分布の関数は以下のようになります。

2項分布とはだいぶ趣の異なる式でなじみにくいかと思いますが、まあそこは我慢してください(というか私にはいかんともしがたいので(笑))。ポアソン分布の期待値、分散共にλ(=*np*)になります(*1-p*→1になるので、これは2項分布の期待値、分散の極限を考えるとすぐにわかるかと思います)。なお、確率母関数はとなります。

ポアソン分布についても2項分布と同様に再生性を示し、*X1,X2*が独立でかつとすると、という特徴があります。

### 超幾何分布

さて、ちょっとおどろおどろしい名前の分布ですが、そんなにえげつないものではありません。2項分布のようなケースで、袋の中にある赤白の玉の合計個数を*N*として、そのうち赤玉が*M*個あるような場合を考えてみましょう。このとき*n*回(*n*<*N*)の非復元抽出(1回出した球は袋に戻さないような抽出方法、逆に戻す場合は復元抽出といいます)試行を行うとして、*x*回赤玉が出てくるような確率を考えよう、というのが問題になります。まず、非復元抽出の場合の数は各玉に色とは別に通し番号がついている、と考えるとNCn通りありますね(以下、NCnはと表記します)。一方、赤玉の個数がxであるとすると、以下の様な関係が成立します。

ただし、

この時、赤玉の個数Xを確率変数と考えると、分布関数は以下のようになります。

超幾何分布の期待値=、分散はとなります(これも計算は自分でやってみてね♡)。なお、超幾何分布には確率母関数は存在しません。

### 負の2項分布

これも何だかよくわからない日本語ですが、まあこけおどしです。コイン投げにおいてr回表が出るまでの間に裏が出た回数Xの分布は以下の様に表されます、

あれれ、2項分布で表が*r*回出るときに*x*回裏が出るのならになってないとおかしくないか?という気がしませんか?ちょっと考えてみましょう。2項分布のとき、*x+r*回目の試行の結果は表であろうが裏であろうがかまいませんよね。ところが、この場合では*x+r*回目の試行の結果は話の流れ上表、しかありえませんね。コインが表でも裏でもよいのは*x+r-1*回目までですので、結果としてという係数になります。負の2項分布の期待値=、分散=となります。確率母関数はとなります。

## 連続型の確率分布

### 連続型の一様分布

となる定数としたとき、

となる分布を連続一様分布と呼びます。これは一般的にと表します。期待値は、分散はになります。

### 正規分布

いよいよ確率分布の真打登場です。いわゆる「釣鐘型」の分布として代表的なものですが、確率分布関数は案外ややこしいものです。

正規分布(と表記)の期待値=*μ*、分散はσ2になり、モーメント母関数はです。特に平均=0、分散=1の正規分布は標準正規分布()と呼び、これはそのあと推定や検定で出てくるさまざまな分布と深い関係を持っています。なお、正規分布も再生性を有していて、*X1,X2*が独立でかつのとき、となります。また、2項分布のnが十分大きい時、B(n,p)は正規分布で近似されるという性質があります。

### ガンマ分布

非負の値をとる確率変数Xが

という確率密度関数を持つものをガンマ分布(と表記)と呼びます。なお、α、βは正定数で、

はガンマ関数と呼ばれる関数です。ガンマ関数はで、正整数nに対してという性質を示します。つまり、階乗を正の実数に拡張したものがガンマ関数です。また、という興味深い性質を持っています。なお、ガンマ分布の期待値=、分散=、モーメント母関数はです。

### ベータ分布

確率変数X(0<X<1)が次の確率密度関数を持つような分布をベータ分布(と表記)と呼びます。

ここで、

はベータ関数といい、というガンマ関数との関連性という意味で興味深い性質を持っています。なお、ベータ分布の期待値＝、分散=です。なお、モーメント母関数は存在しますが、

となかなかに複雑な関数であまり実用性の高いものではないとのことです(というか私もこれ書くためにググって初めて知りましたw)。ガンマ分布と似ているからベータ分布も…と単純にいかないところが数学の面白いところでもありますが同時に厄介な部分ですね。

### カイ2乗分布

さて、ここからは先ほど正規分布のところで仄めかした正規分布から導かれるさまざまな分布の話です。標準正規分布に従うn個の確率変数の2乗の和はどんな分布をしているか、と考えてみましょう。この分布の確率密度関数はこんな感じになります。

この分布を自由度nのカイ2乗分布と呼び、と表記します。は、標準正規分布に従うn個の確率変数の標本分散×標本個数がどんな感じで分布しているかということを示す分布ということになります。期待値はn、分散は2nとなり(シンプルですねぇ)、モーメント母関数は

なお、という興味深い関係があります。

### *t*分布

確率変数Z、Wがそれぞれ標準正規分布、に従うとき、確率変数

が従う分布を自由度nのt分布と呼び、と表記します。の確率密度関数は

と、結構ややこしいものです。特にはコーシー分布と呼ばれ、平均も分散も持たない分布として知られており、逆には正規分布になります。大体n=100くらいで正規分布として扱ってほぼ問題ないと言われています。なお、期待値はn>1のときに0(n=1のときはコーシー分布で存在しない)、分散はn>2のときに存在してになります。モーメント母関数は存在しません。

### *F*分布

U,Vが独立でそれぞれ、に従うとき、確率変数の従う分布というものを考えてみます。これが自由度(m,n)のF分布で確率密度関数は

というt分布よりも更に壮絶な式になります。なお、期待値はn>2のときのみ存在して、分散はn>4のときのみ存在してとこちらもなかなか面倒くさい式になっています。なお、F分布もモーメント母関数は存在しません。

## 実際に確率分布を体感してみよう

さて、結構エグい数式が延々と続いてくらくらしてきましたね。数式を自分でいじってみるというのはこういうものの勉強では効果的と言えば効果的なのですが、見てわかる通り高校の数学ではいじるための道具としては不足で大学の授業でやる微積分ががっつりできないと手も足も出なさそうな感じですし、必ずしも統計を実用的に使うという場合にいわゆる数理統計学を十分理解していることが大前提になるかというとそうでもありません。むしろ、こういった理論的なものだけでは不十分で、それだけではデータ解析をするのには不足しているのだという考え方もあります(これについては実験医学」2014年2月号に書かれた三中先生の記事[https://www.yodosha.co.jp/jikkenigaku/statistics\_pitfall/ pitfall\_1.html](https://www.yodosha.co.jp/jikkenigaku/statistics_pitfall/%20pitfall_1.html)が参考になります)。

ということで、Rを使ったシミュレーションで視覚的に確率分布を感じてみましょう。まず、離散型の確率分布を見てみようと思います。代表例として2項分布とポアソン分布を取り上げたいと思います。まずは2項分布で、1回の試行での事象の発生確率が0.1,0.3,0.5,0.7のとき、試行を20回行ったときにN回事象が発生する確率をプロットしてみましょう。p=0.1であれば一番発生確率が高いのは20×0.1=2回ということになります。

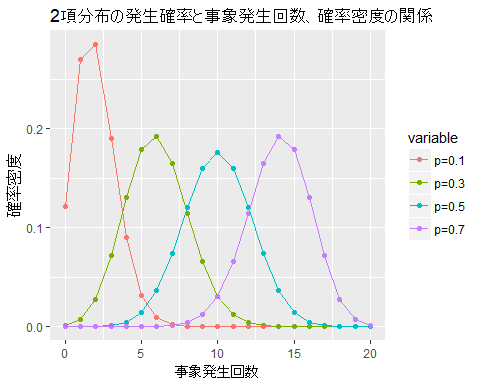
rm(list=ls(all=TRUE))  
#2項分布  
p<- c(0.1,0.3,0.5,0.7) #2項分布の確率  
x<- seq(0,20,by=1) #事象の発生回数のベクトル  
y<- matrix(nrow=length(x),ncol=length(p)) #発生確率の格納先  
cn<- vector() #カラム名  
  
#2項分布の確率の計算  
for (i in 1:length(p)){  
 y[,i]<- dbinom(x,20,p[i])  
 cn[i]<- paste("p=",p[i],sep="")  
}  
y<- data.frame(x,y) #出力のデータフレーム化  
colnames(y)[2:5]<- cn #カラム名指定  
  
#プロット作成  
library(reshape2)

## Warning: package 'reshape2' was built under R version 3.4.3

library(ggplot2)

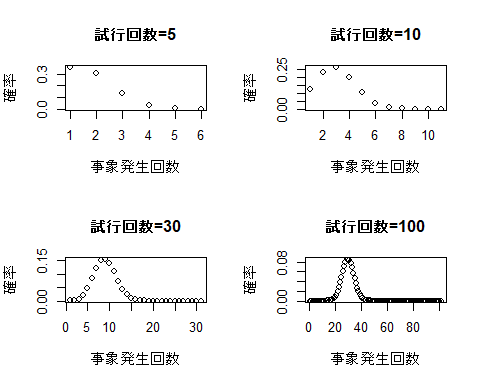
## Warning: package 'ggplot2' was built under R version 3.4.3

y<- melt(y,id.vars="x")  
pl.binom<- ggplot(data=y,aes(x=x,y=value,col=variable))+  
 geom\_point()+  
 geom\_line()+  
 xlab("事象発生回数")+  
 ylab("確率密度")+  
 ggtitle("2項分布の発生確率と事象発生回数、確率密度の関係")  
print(pl.binom)



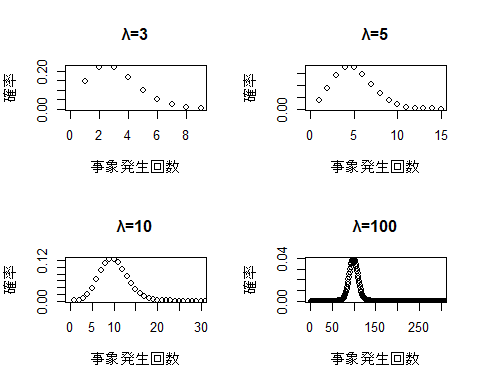
さて、2項分布において今度は事象の発生確率が同じという条件で試行回数をどんどん増やしていくとどうなるかを見てみようと思います。試行回数が増えていくに従い、左右対称な釣鐘型の分布に近づいていくことがわかるかと思います。

#2項分布の試行回数を増やすと正規分布に近づくことを確認  
rm(list=ls(all=TRUE))  
p<- 0.3#事象の発生確率  
r<- c(5,10,30,100)#試行回数  
par(mfrow=c(2,2))  
for (i in 1:4){  
 x<- seq(0:r[i])  
 y<- dbinom(x,r[i],p)  
 plot(x,y,type="p",main=paste("試行回数=",r[i],sep=""),  
 xlab="事象発生回数",ylab="確率")  
}



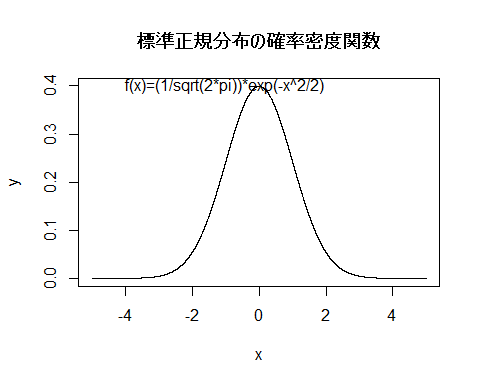
次にポアソン分布で見てみましょう。λ(試行回数×確率)が小さいときは2項分布の試行回数が小さいときと同じく、左右非対称な分布ですがλが増えていくに従い、こちらも左右対称な釣鐘型の分布に近づいていくことがわかるかと思います。

#ポアソン分布の様子を見る  
rm(list=ls(all=TRUE))  
lambda<- c(3,5,10,100)  
par(mfrow=c(2,2))  
for (i in 1:4){  
 x<- seq(0:(lambda[i]\*5))  
 y<- dpois(x,lambda[i])  
 plot(x,y,type="p",main=paste("λ=",lambda[i],sep=""),  
 xlab="事象発生回数",ylab="確率",xlim=c(0,(lambda[i]\*3)))  
}



次は連続型の確率分布です。標準正規分布のプロットをまず見てみましょう。きれいな対称的な釣鐘型を示しますね。

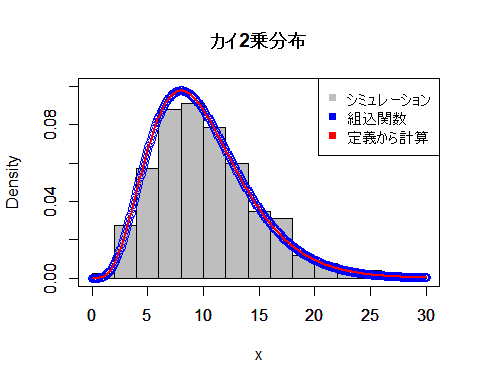
#標準正規分布のプロット作成  
rm(list=ls(all=TRUE))  
x<- seq(-5,5,by=0.001)  
y<- dnorm(x)  
plot(x,y,type="l",main="標準正規分布の確率密度関数")  
text(x=-1,y=0.4,labels="f(x)=(1/sqrt(2\*pi))\*exp(-x^2/2)")



さて、正規分布からカイ2乗分布を導いてみましょう。標準正規分布に従う10個のデータをサンプリングするということを1000回繰り返した時の、それぞれのデータの2乗の和がどのように分布するかをヒストグラムで見てみましょう。これは、「標準正規分布に従う10個のデータの分散×10という値がどのような分布を示すか」ということに等しくなります。これと、Rにあらかじめ組み込まれている関数で計算したカイ2乗の確率分布関数、および先に示した

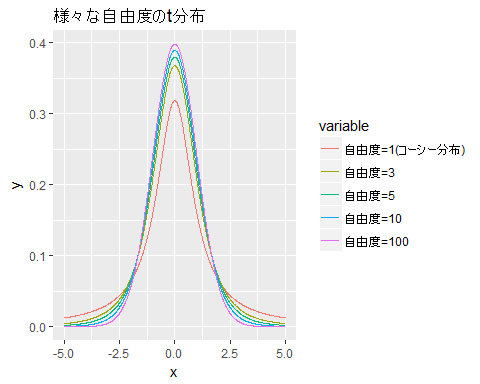
から導かれる値をそれぞれプロットしてみましょう。いずれも同じになることが確認できます。

#正規分布から導かれる確率密度関数  
#カイ2乗分布を正規分布から導く  
rm(list=ls(all=TRUE))  
set.seed(2017)  
A<- matrix(rnorm(10000),nrow=1000,ncol=10)#正規分布に従う10個のサンプルを考える➡これが1000セットあるとする  
A2<- A^2#正規分布の偏差の2乗  
W<- apply(A2,1,"sum")#偏差の2乗和の分布(カイ2乗統計量)  
  
#カイ2乗統計量の分布のヒストグラム(頻度→確率に変換してある)  
hist(W,freq=FALSE,xlab="x",xlim=c(0,30),ylim=c(0,0.1),col="grey",main="")  
  
#組み込み関数からの自由度10のカイ2乗分布の計算  
x.chisq<- seq(0,30,by=0.1)  
y.chisq<- dchisq(x.chisq,10)#自由度10のカイ2乗分布  
  
#組み込み関数からのカイ2乗分布のプロット(ヒストグラムに上書き)  
par(new=TRUE)  
plot(x.chisq,y.chisq,type="p",col="blue",cex=1.2,xlab="",ylab="",xlim=c(0,30),ylim=c(0,0.1))  
  
#定義に従ったカイ2乗分布の作成  
y.chisq.def<- 0.5^(10/2)\*x.chisq^(10/2-1)\*exp(-x.chisq/2)/gamma(10/2)  
  
#定義に従ったカイ2乗分布が組み込み関数と一致することを確認する  
par(new=TRUE)  
plot(x.chisq,y.chisq.def,type="l",col="red",xlab="",ylab="",  
 lwd=2,xlim=c(0,30),ylim=c(0,0.1),main="カイ2乗分布")#定義に従ったカイ2乗分布関数のプロット  
labels <- c("シミュレーション","組込関数","定義から計算")  
legend("topright",legend=labels,col=c("grey","blue","red"),pch=15,cex=1)



最後にt分布をみてみましょう。自由度1(コーシー分布；平均も分散もない)～自由度100まで見ていくと、だんだんピークが高くなり、分布の裾が狭くなってくることがわかります。ただ、コーシー分布もこう見てみるとほかの自由度のt分布と同じ釣鐘型なのですが、これだけ似ていても平均も分散も存在しないというのは感覚的にはちょっと驚きですよね。でも、こうしたちょっとの違いが結果に大きく影響するというのも数学の面白いところであると個人的には思っています。

#t分布  
rm(list=ls(all=TRUE))  
n<- c(1,3,5,10,100)#自由度  
x<- seq(-5,5,by=0.01)#xの値  
y<- matrix(nrow=length(x),ncol=5)#格納先  
for (i in 1:5){  
 y[,i]<- dt(x,n[i])  
}  
  
#t分布のプロット作成  
cn<- c("自由度=1(コーシー分布)","自由度=3","自由度=5",  
 "自由度=10","自由度=100")  
y<- data.frame(x,y)  
colnames(y)[2:6]<- cn  
library(reshape2)  
library(ggplot2)  
y<- melt(y,id.vars="x")  
pl.tdist<- ggplot(data=y,aes(x=x,y=value,col=variable))+  
 geom\_line()+  
 ggtitle("様々な自由度のt分布")+  
 ylab("y")  
print(pl.tdist)



# 4．統計的に推定をするということ

## 母集団と標本集団

さて、長々と確率の話をしてしまいました。本題に戻りましょう。最初に出したテストの得点表のA組、B組の行と全く同じデータです。今度はこのデータが日本全国のある学年の学生から40人無作為に抽出してテストを実施した時の得点一覧であったとします。さて、この40人のデータから何が言えるでしょうか。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 得点  一覧 | 48 | 60 | 39 | 53 | 54 | 61 | 56 | 50 | 40 | 48 | 55 | 44 | 45 | 30 | 47 | 56 | 50 | 70 | 57 | 63 |
| 50 | 64 | 36 | 53 | 54 | 66 | 60 | 51 | 40 | 50 | 58 | 43 | 48 | 35 | 50 | 58 | 52 | 70 | 63 | 69 |

まず、さっきと同じようにデータの平均とか分散、標準偏差を出してみましょう。

rm(list=ls(all=TRUE)) #ワークスペースを初期化  
x<- c(48,60,39,53,54,61,56,50,40,48,  
 55,44,45,30,47,56,50,70,57,63,  
 50,64,36,53,54,66,60,51,40,50,  
 58,43,48,35,50,58,52,70,63,69)  
#標本平均  
mean(x)

## [1] 52.4

#標本分散と標本標準偏差  
#標本分散を算出する関数(Rにはもともとない)  
var.sample<- function(x){  
 x.ave<- mean(x)  
 var.sample<- mean((x-x.ave)^2)  
 return(var.sample)  
}  
#標本分散  
var.sample(x)

## [1] 91.09

#標本標準偏差  
sqrt(var.sample(x))

## [1] 9.544108

さて、同じことを改めて別の40人を無作為に抽出してデータを取ってみるということを繰り返すことを考えてみましょう。このとき、全く同じ点数の分布で同じ平均、同じ分散、同じ標準偏差が得られるでしょうか。普通に考えたら全く同じということはなさそうですよね。というか、まったく同じ結果が得られる方がほぼ奇跡であるといった方がよさそうですね。

では、「日本全国のある学年の学生」全員にこのテストを実施した時というのを考えてみましょう。とてもやる気は起きませんね。大体そんなことやっているうちに時がたってしまって全員に実施することそのものが無理かもしれません。

ところで、最初の40人のデータというのはどういった性質のものだったでしょうか。「日本全国のある学年の学生から40人無作為に抽出してテストを実施した時の得点一覧」ですね。ということは、これは何らかの形で「日本全国のある学年の学生」全体の情報を反映しているものです。統計解析は、限られたサンプル(40人)のデータからそのデータの持つ属性(この場合は「日本全国のある学年の学生」)を有する集団全体の挙動(この場合は得点分布)を見るための解析です。このとき、「日本全国のある学年の学生」を母集団と呼び、その集団での得点の平均とか分散を母数(母パラメータとも呼びます)と呼びます。一方、40人の集団を標本集団と呼びます。言い換えると、標本集団のデータから母数を推定するというのが統計的な推定というものになります。ところで、「日本全国のある学年の学生」というのは現在の時点であれば有限ですが、例えば過去とか未来の集団も含めて考えると有限とは言えなくなります。一般的には「〇〇という属性を有している集団」というのは必ずしもリアルに現在存在しているもの全体に限らない無限の大きさの母集団を想定しています。

## 推定量

さて、ではこの40人のデータから日本全国のある学年の学生全体にこのテストを実施した時の平均点を推定してみましょう。日本全国のある学年の学生全体にこのテストを実施した時の平均点の真の値(=m)というものは実在するはずです。でも、実際にそれを調べるわけではない(この場合は超がんばれば知ることができるかもしれませんが、実際の世の中には現実的に母集団を使って知ることは到底できないような情報というのもたくさんあります)ところで、どうやってmの真の値を知ることができるでしょうか。例えば、以下のような知り方があるかと思います。

1. データなんて知らん。俺は俺の感性を信じる。平均点は55点になるはずだ。
2. やっぱデータには忠実に向き合わないと。平均点は得点の総和/40の52.4点だ。
3. 標本の平均が母集団の平均と一致するなんてことまずないだろ、なんか別の方法あるんじゃね？

上の①～③、どの方法が最適なのでしょうか。普通に考えて①はあまりなさそうだなという気はしますが、②でよいのでしょうか。しかし、③を主張している人が言っていることにも一理ありそうです。とはいえ、③は「なんか別の方法」と言っていますが、それが明らかにされていなければ①と大差ないですね。

実は、推定値の出し方には複数あり、ぶっちゃけて言えば①だってありはありなのです。ただ、推定値の「良さ」というものを考えないといけません。この場合の「良さ」というのは真の値からのずれの大きさです。「真の値がわかんねーのにそのずれもよさもねーだろw」という人もいるかもしれませんが、こう考えてみたらどうでしょう。

ある確率密度関数に独立に従うn個の確率変数ベクトルに対してパラメータ*θ*の推定量は***X***の関数で表すことができます。とりあえずそれをとしましょう(θの上についている記号はハットと呼びます)。

サンプルの値()に対して定まる値に対して、リスク関数

というものを考えてみましょう。これだとの値が真のパラメータに等しい場合は関数の値が0になり、それ以外の場合は正の値になり真の値と隔たるにつれて大きくなっていくということで、推定の良さはの値がどれだけ小さいかで評価することができます。この場合、パラメータ*θ*の任意の値に対してリスク関数の値が一様に小さくなるような推定量が理想的ですが、世の中そんなに甘くはなく、そういった性質を満たす推定量は一般には存在しません。例えば、①のようなものであっても、m=55であるならリスク関数の値は0になって間違いなく最小になりますが、m≠55の場合最小とは言えない場合があります。たまたまデータの標本平均であれば、リスク関数の値はそっちの方が最小になります。

そこで、①のような考え方はそもそもデータに基づかない俺様ルールですので採用しないとし、②③の様なデータに基づくものを俎板の上に乗せるようにしましょう。

### 不偏推定量

このとき、推定量が持っていると嬉しい(いろいろと取扱いに便利な)性質があります。その一つに不偏性というものがあります。これは、任意の*θ*に対してとなる様な性質で、そのような性質を満たす推定量を不偏推定量と呼びます。仮に1回のデータの平均はmからずれていたとしても、先に示した中心極限定理より、サンプルデータの平均値の平均はmに漸近するという性質を有しています。従って、②の考え方は不偏推定の性質を満たしているものです。定式化すれば、平均mの不偏推定量をとしたとき、

となり、平均値の不偏推定量は結局標本平均に等しいということになります。

では、分散の不偏推定量はどうでしょう。標本の分散でよいのでしょうか、実は、そうではないんですね。ここが統計解析を勉強するときに躓きやすい最大のポイントの一つだと思っていますが、標本分散

は不偏推定量にならず、分散の不偏推定量(不偏分散)は

です。基本的に背後に母集団を想定している統計解析ではこの不偏分散を分散の推定量として用います。なぜ、nで割るのではなく、n-1で割らないといけないのでしょうか？

式中に現れるは標本平均ですが、n個のサンプルから標本集団の平均を決めてしまうと、(n-1)番目のサンプルまでは値を任意に取ることができますが、n番目のサンプルはすでに決まった平均に合わせるために値は決まってしまいます。つまり、データがn個あっても、平均が決まっているという束縛がある以上、n個のデータのうち好き勝手に値が動かせるサンプルの数はn-1になります(これを自由度と呼びます)。これが、不偏分散の分母がn-1になる理由です(これでわかりますかね？)。実際、Rでは分散を計算する組み込み関数varは不偏分散を返す仕様になっています。実際に計算してみると先に計算した標本分散、標本標準偏差より少し大きな値になっていることが確認できます。

#不偏推定量の算出  
#不偏平均(≠標本平均)  
mean(x)

## [1] 52.4

#不偏分散  
var(x)

## [1] 93.42564

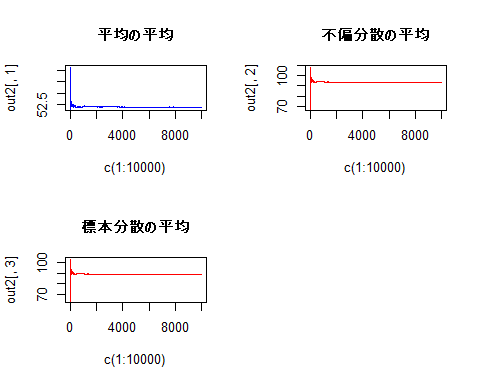
#不偏標準偏差  
sqrt(var(x))

## [1] 9.665694

さて、不偏性の意味ですが、もうちょっと例を出して考えてみましょう。不偏性とはとなる様な性質だと言いました。このデータからさらに適当にいくつかのデータを抽出して(不偏)平均とか(不偏)分散を求める操作を繰り返していくと、多分どこかの値に収束します。そこで、40個のデータを「母集団」だと考え、そこから20個ランダムに抽出し、そのデータの平均、不偏分散、標本分散を算出するという操作を繰り返し、その平均がどのような値に収束するかを見てみましょう。最初は平均、分散ともにその平均はガタついていますが、だんだん落ち着いてきて1万回繰り返した時の平均と不偏分散の値は40個のデータをすべて用いたときの平均値と不偏分散とほぼ等しくなっています。一方、標本分散の値は88.76と40個のデータを用いて算出した標本分散の91.06という値よりいくぶんか小さくなっています。つまり、標本分散はという不偏性の条件を満たしていない、ということになります。

#ランダムにデータをサンプリングしてその収束値を見る  
a<- set.seed(1234) #乱数シードの生成  
ind<- c(1:40) #サンプル見出し  
out<- matrix(nrow=10000,ncol=3)  
for (i in 1:10000){  
 samp<- sample(ind,20) #抽出するサンプルの番号  
 x.samp<- x[samp]  
 out[i,1]<- mean(x.samp)  
 out[i,2]<- var(x.samp)  
 out[i,3]<- var.sample(x.samp)  
}  
#平均、分散の不偏推定量の平均値  
out2<- matrix(nrow=10000,ncol=3)  
for (i in 1:10000){  
 out2[i,1]<- mean(out[1:i,1])  
 out2[i,2]<- mean(out[1:i,2])  
 out2[i,3]<- mean(out[1:i,3])  
}  
par(mfrow=c(2,2))  
plot(x=c(1:10000),y=out2[,1],col="blue",type="l",main="平均の平均")  
plot(x=c(1:10000),y=out2[,2],col="red",type="l",main="不偏分散の平均")  
plot(x=c(1:10000),y=out2[,3],col="red",type="l",main="標本分散の平均")  
print(out2[10000,])

## [1] 52.39988 93.43310 88.76145



### 最尤推定量

さて、②の考え方で推定量を出すというところまではこれでよいかと思いますが、③はどうでしょう。逆転の発想で、確率分布の密度関数に対して、これまで*x*の関数と考えていたものをという*θ*の関数で表されると考えてみましょう。このとき、*θ*を動かしたときにが最大になる様な(つまり、あるデータxに対して最も確率が高くなるような*θ*を当てたときがそうなります) *θ*を求めるということを考えます。この時、計算を楽にするために対数を取るのが一般的で、を尤度(ゆうど)、を対数尤度と呼びます。このとき、対数尤度が最大になる様な*θ*では

が成立します。これを尤度方程式と呼び、これを満たす*θ*を最尤推定量と呼びます。なお、一般的には対数尤度はその負の値(負の対数尤度)を用いますが、これはその導出過程に関係しています(カルバック-ライブラー情報量というものがあり、そこからの導出過程で負の対数尤度が出てくるのですが、この話はいささか専門的すぎるので割愛します)。不偏推定量が分布のパラメータに関して平均、分散くらいしか算出できないのに対して、最尤推定量は様々なパラメータの推定に使えること、尤度を数式で表すことさえできればあとは計算問題(コンピュータを使って解ける)になる、という点がメリットです。

点数の分布が正規分布に従っていると考えます。この時、n個のそれぞれ独立に正規分布に従っているサンプルの値が()となる確率(＝尤度)は

で表されます。ただ、これだと式が掛け算になっていて大変に面倒くさそうです。そこで、この式の対数(＝対数尤度)を取ってみましょう。

こうすると、n個の項それぞれを足し算すれば計算できるのでだいぶ楽になりそうです。更に、これを変形すると

となります。これをそれぞれmとで偏微分すると、

となります。従って、mの最尤推定量として

が、更にはこの結果を下の式に代入するとの最尤推定量として

がそれぞれ得られます。さて、ここで平均の最尤推定量＝不偏推定量ですが、分散の最尤推定量＝標本分散≠不偏分散であることに注意しましょう。

この様に数式にするとなんだか最尤推定量を求めるのは面倒くさそうな雰囲気が漂っていますが、実はそんなことはありません。例えばこのブログに書かれているスクリプトを使って計算してみましょう。

#最尤推定量の算出  
#正規分布の尤度を算出する関数  
LL.norm <- function(x){  
 return(function(par){  
 myu<- par[1]  
 sigma2<- par[2]  
 - length(x) / 2 \* log(sigma2) - 1 / 2 \* sum((x - myu)^2) / sigma2  
 })  
}  
#初期値を(50,100)として上記尤度関数の最適化を行う  
res<- optim(par=c(50,100), fn = LL.norm(x), control = list(fnscale = -1))  
res$par[1] #平均の最尤推定量

## [1] 52.39811

res$par[2] #分散の最尤推定量

## [1] 91.08502

sqrt(res$par[2]) #標準偏差の最尤推定量

## [1] 9.543847

ちょっと小数点以下の桁数が違いますが、平均は標本平均=不偏平均＝最尤推定平均、分散は標本分散＝最尤推定分散になりますね。

### 最小2乗法

さて、ここまでは一つの分布に独立に従ういくつかの確率変数からその分布のパラメータに従う例を考えてきました。