JAPLA 研究会資料 2009/1/24

素人さんの統計学ノートーその１

度数分布と中心極限定理

西川　利男

１．はじめに

統計学について、例えば高校の数学教科書から単元「身近な統計」を見てみよう。「ある学級の通学時間を調べた 5 分から 80 分の 40 個のデータがある。まず、全体のようすがわかり易くなるよう整理してみる。それには 10 分おきの区間に分けて、度数分布表を作って図示する。そして、階級と度数とでできるヒストグラムを眺める。すると、正規分布しているように見られ、平均、標準偏差を求めて、…、さらに中心極限定理を、……」となる。

このようなのが、ふつうの統計学の入門になっているが、これで良いのだろうか。度数分布表を作るには「正」の字で集計して、とあるが、これが案外、面倒である。この程度のデータ数であれば、度数分布表などつくらないでそのままグラフ表示して眺めたらよい。今のコンピュータ時代ではその方がずっと易しい。

われわれが知りたいのは、度数分布表に集約されたデータではなく、生のデータの値の分布、ばらつきなどの情報である。度数分布表は個々の値を無視して情報量を減少させ、ときにはゆがめてしまうこともある。

いくつかの分布をもとに度数分布表を作りグラフ表示を行い、これを比較する、という実験を行なってみた。

２．正規分布の場合

シューハート・チップあるいは日本規格協会チップとして、各種の平均、標準偏差を持つ正規分布の色分けチップが使われ、従来はこれを実際に取り出し標本とする統計実験がなされたが、今ではコンピュータの中で容易に行なうことができる。

例として、日本規格協会の緑チップのデータを GDA とする。度数分布表の作成のための階級値を GDX とする。

```
   GDX
50 51 52 53 54 55 56 57 58 59 60 61
```

階級に対する頻度を計算する関数を freq として定義し、次のように度数分布を求める。なお、J のいろいろな定義は末尾にあげた。

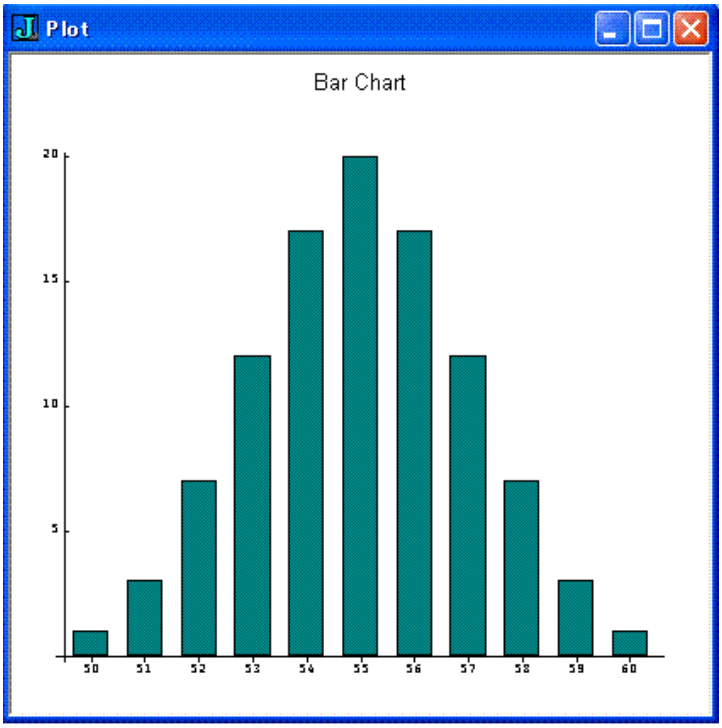
```
   GDX freq GDA
1 3 7 12 17 20 17 12 7 3 1
```

これは緑チップの内容を示すことになる。

次に、ヒストグラムを描く。これには簡単に J の plot ルーチンでもできるが、横軸の目盛りの表示を階級値で示すなどのため、pd 命令を用いた J のプログラム histogr を作った。また関数 xclass は階級値を pd の表示文字列に合わせて、これを histogr の左引数とし、右引数には度数の値をとる。

こうして、ヒストグラムは次のように描かれる。

```
   (xclass GDX) histogr GDX freq GDA
```

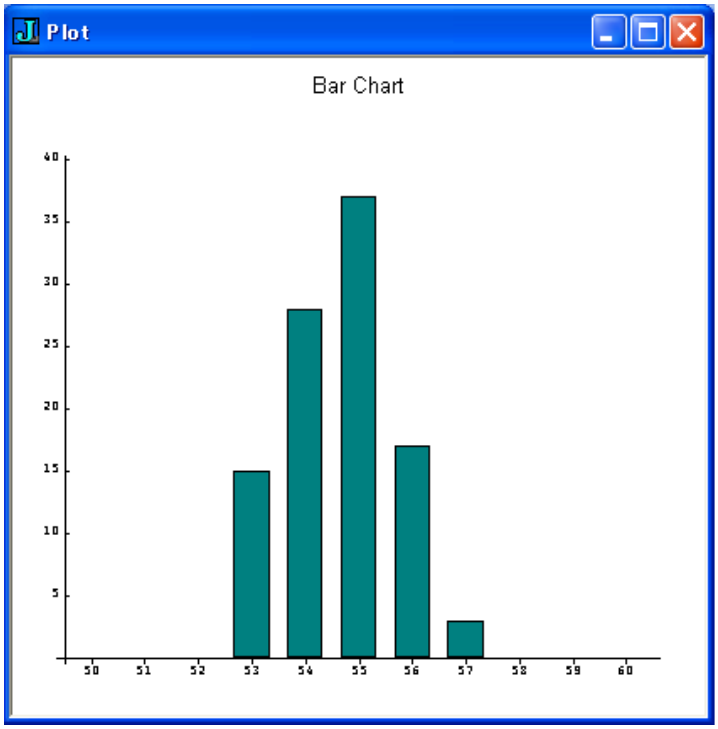


次にこれを母集団として、5つずつ100回、無作為に取り出した標本をMDAとする。

```
$MDA
100 5
```

この各回ごとに標本の平均をとりMDMとする。そしてMDMの分布を見てみよう。

```
MDM =: mean"(1) MDA
(xclass GDX) histogr GDX freq MDM
```



このように標本では平均は同じだが、ばらつきはずっと小さいヒストグラムが得られた。そして正規分布を予想させるもので、これは中心極限定理とよばれている。

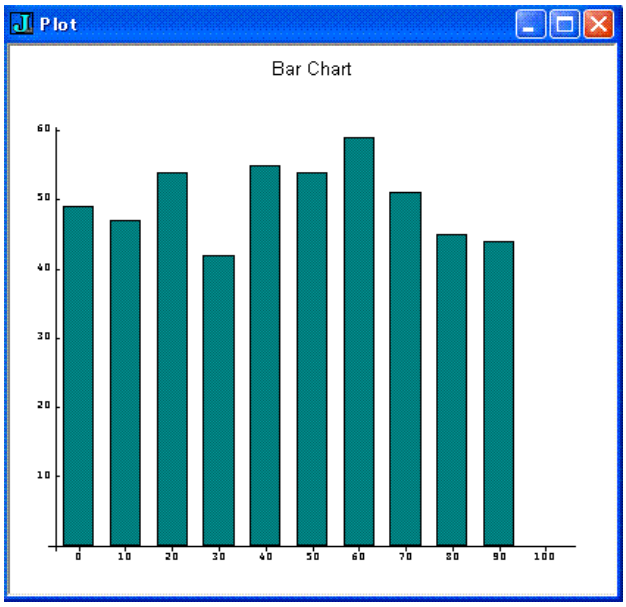
３．一様分布の場合

J の関数 rnd で 0 から 100 の乱数, 500 個を生成させ、これを一様分布(XP)とする。

```
XP =: 100 rnd 500
```

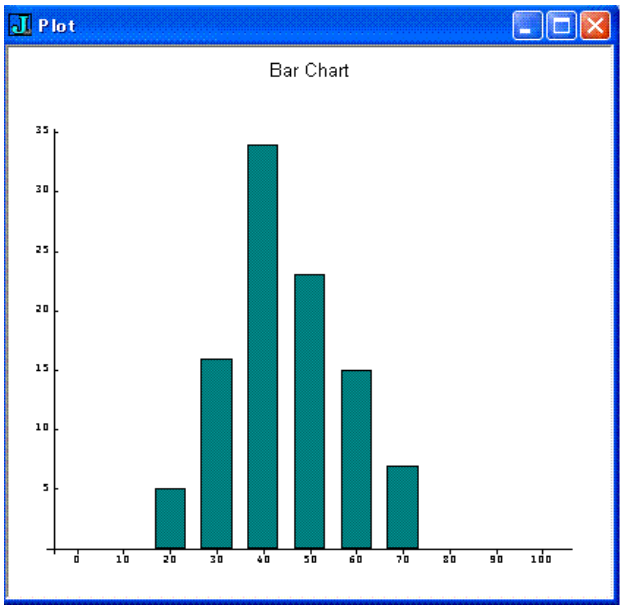
分布のヒストグラムは以下のようになる。なお、A は階級値である。

```
(xclass A) histogr A freq XP
```



前と同様、一様分布から無作為標本を取り出し、その平均の分布をヒストグラムに示す。驚くべきことに一様分布からの標本でも同じ正規分布をなす。

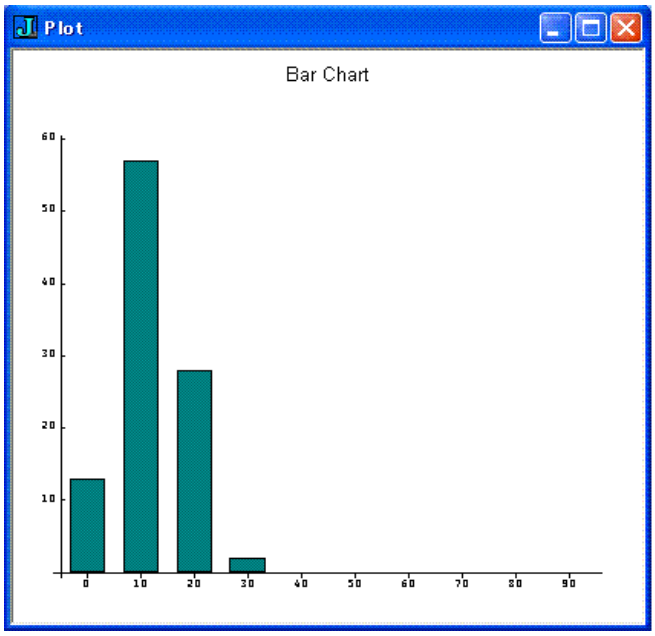
```
XPA =: 100 5$XP
XPM =: mean"(1) XPA
(xclass A) histogr A freq XPM
```



４．指数分布の場合

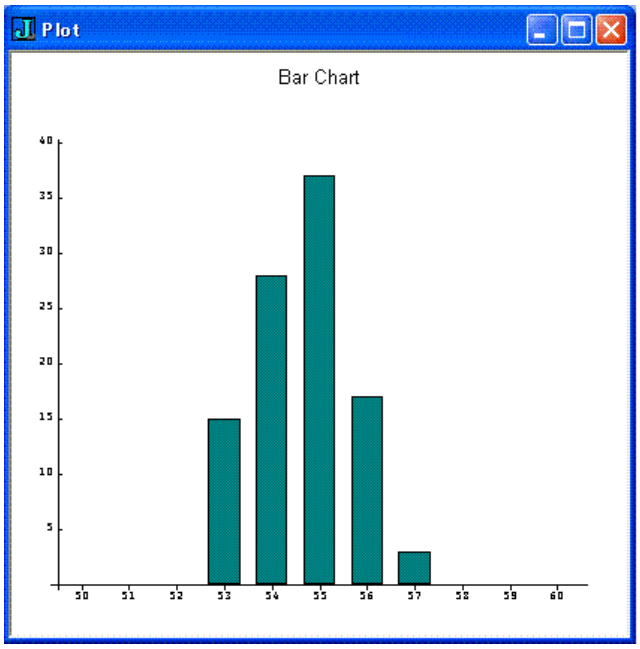
今度は指数分布についてもやってみる。

```
EXPA =: <. 1000 * 0.05 expd i.100
EXPB =: EXPA # i.100
(xclass A) histogr A freq EXPB
```



同様に指数分布から無作為標本を取り出し、その平均の分布を図示する。すると指数分布のときも前と同様に正規分布をなす。

```
EXPP =: 970 rnd 500
EXPD =: 100 5$EXPC
EXPM =: mean"(1) EXPD
(xclass A) histogr A freq EXPM
```



５．度数分布表と中心極限定理

中心極限定理(Central Limit Theorem)とは「正規分布をなす母集団から標本抽出して出来たグループごとの平均はやはり正規分布をなし、ただそのばらつきは小さくなる。正確に言えば平均は変わらず、標準偏差は1/√n（標本の個数をn）になる。そしてこれはどんな母集団であっても変わらない。」となっている。

以上の実験はこれを示すものだが、このような説明のしかたは私には不満である。中心極限定理などとありがたそうな名前がついているが、度数分布表で平均なる操作をすることで値が集中しただけで、その操作の誤差の分布が正規分布と同じ式で広がっているというだけのものである。従って、母集団の分布がどんなものであっても平均操作をすれば、その誤差は正規分布するのは当然のことである。

鈴木義一郎氏もかねて言われるように、ガウスの誤差分布という用語を使った方が良い、という意見に私も同感である。正規という名前がよくない。これ以外はノーマルでなく、アブノーマルなのか。正規とか標準とかという形容詞は軽々しく使うべきではない。権威主義のにおいがする。もっと冷静に見るのがサイエンスである。

正規分布は数学的扱いがすっきりするというだけでの理由からで教科書で取り上げているに過ぎない。正規分布以外の統計データはいくらでもある。実際の統計処理に当たって、まず平均、標準偏差を計算して、… と進める統計学はおかしい。この２つのパラメータだけでは、元のデータの情報量をあえて捨ててしまっている。それよりもグラフに描いて、頭を使ってよく観察するのが統計学の第一歩と私は思う。

Jのプログラムリスト

```
NB. My Statistic Mathematics  中心極限定理

NB. uniform distribution
NB.    y. from rundom values 0 thru x.
NB.    eg. 200 rnd 10 => XDA =: 95 47 54 71 33 97 179 181 12 180
rnd =: 3 : 0
:
?. y. # x.
)

NB. normal distribution
norm =: 3 : 0
:
'm s' =. x.
z =. ^ - -: ((y.-m)%s)^2
NB. z =. z % (s * %: 2p1)
)

NB. exponential distribution
expd =: 3 : 0
1 expd y.
:
lamb =. x.
z =. lamb * ^ - lamb * y.
```

```
)
NB. mean
mean =: +/ % #

NB. standard deviation
std =: %:@(mean@(*:@(- mean)))

NB. maximum and minimum
max =: ({.@¥:){]
min =: ({.@/:){]
NB. Frequecy
NB. XCLASS freq XDATA = extract subgroup from XDATA as class XCLASS
freq =: 3 : 0
:
+/"(1) > ~:/ L:(0) 2 <¥ 0 < x. - / , y.
)

NB. Histogram
NB. (xclass A) histogr A f B
histogr =: 3 : 0
:
require 'plot'
pd 'reset'
pd 'new'
NB. pd 'new 500 500 460 460'
pd 'type bar'
pd 'border 0'
pd 'xlabel ', x.
pd 'title Bar Chart'
pd  y.
pd 'show'
)

xclass =: 3 : 0
; '"', L:0 '" ',~ L:0 ": L:0  <"(0) }: y.
)

NB. 松井進作「おはなし統計的手法」 p.45-47, 日本規格協会(1982).
NB. 緑色チップ
GDX =: 50 + i.12
GDA =: 50, (3#51), (7#52), (12#53), (17#54), (20#55), (17#56), (12#57),
(7#58), (3#59), 60
```