JAPLA 研究会資料　2017/12/9

Ｊで電卓並みの統計ツールを
－統計学はなぜ分かり難いのか－

西川　利男

１．統計学はなぜ分かり難いのか

いまや、統計学は文科、理科を問わず、教養としてより実用のリテラシーとして必要なものである。それにしても、統計学はどうして分かり難いのだろうか。方程式や微分積分などの数学とは、やはり異質なものに感ずる。

データを集めてグラフ化したりする記述統計のあと、それをどう活用するかが、一般の数学と比べて身につかない。確率変数やら、中心極限定理、正規分布、ｔ分布…などいろいろな用語が出てきて、さらに次々と数式による説明でまいってしまう。

Ｊでもっと手軽な「電卓並みの統計ツールを」、という試みを行ってみた。

２．やさしい一つの例題

実は、私の手元にあるシャープの関数電卓 EL-526 では、ちょっとした統計計算もできて、そこに次のような例題がのっていた。

ある試験を受けた。その点数と人数の集計結果は次のようであった。

| 20 1 | 30 3 | 40 5 | 50 8 | 60 13 | 70 10 | 80 7 | 90 3 |

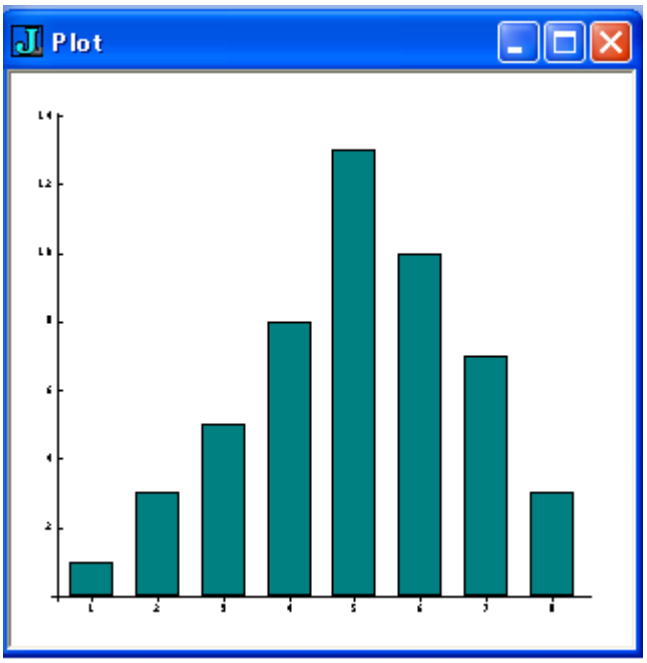
私の得点は 75 点であった。はたして、これはどのあたりの成績になるだろうか。

シャープの電卓では、キーインしただけで、結果がわかるようになっている。

「電卓並みの統計ツールを」とは、このようなものでなければならない。

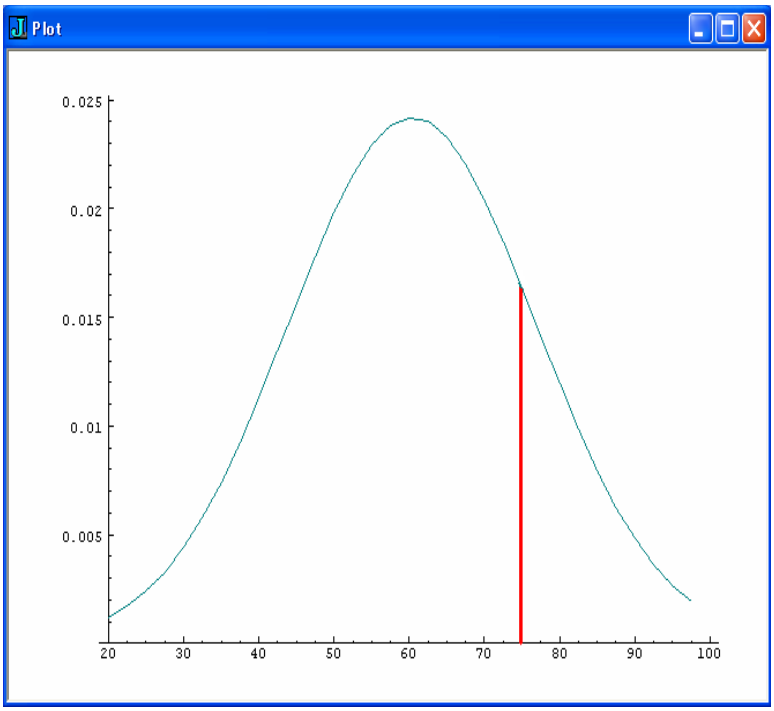
３．１　数式によらず、グラフで考える

まず、ヒストグラムを描いてみる。これは、点数の実際を示したままである。



つぎに、予想される得点の広がりの図を描いて、私の得点、75 点をあてはめて見るとかなり良い成績であることがわかる。

３．２ グラフを描くためのＪのしくみ（Ｊの計算とグラフィック ス）



グラフは、ちょっとした計算で描くためには簡単が必要だが、これは、Ｊに次のように行われる。

まず、最初のデータを次のように名付けて定義する。

```
   DA =: 20 1;30 3;40 5;50 8;60 13;70 7;90 3
   DA
+----+----+----+----+----+----+----+
|20 1|30 3|40 5|50 8|60 13|70 7|90 3|
+----+----+----+----+----+----+----+
```

このように、DA は各点数とその人数をまとめて、ボックスのリストとして定義される。

(1) ヒストグラムのための計算

人数だけをとりだすには次のようにする。

たとえば、ボックスの２番目(0 オリジンなので、実際は３番目)についてやってみる。

```
   >2{DA
40 5
   {: >1{DA
5
```

それぞれのボックスの中について演算を行うには、副詞 L:0 を使った次の機能が有用である。次のように人数だけが取り出される。

```
   Freq =: > {: L:0 DA
   Freq
1 3 5 8 13 10 7 3
```

これを、以上で得た値 Frq をＪのグラフィックツール plot により、図示する。

```
   load 'plot'
 'bar' plot Freq
```

これで、最初のヒストグラムが描かれた。

(2) 分布曲線のための計算

最初にあげたクラスの試験の成績データ DA は標本と呼ばれる。

統計学ではこの標本の値が、十分多数の集合、すなわち母集団から、無作為に得られた値として扱う。そして多くの場合、母集団は正規分布をしているとする。

母集団の平均(μ、ここでは MEAN)と標準偏差(σ、ここでは STDV)とは、Ｊではつぎのようにして計算される。

まず、各点数ごとのグループで、点数に人数を掛けて小合計を出す。ここで活躍するのは、Ｊのボックスの内部ごとに演算する副詞 L:0 である。ボックス内の２つの要素の積を動詞*/により、ボックスのそれぞれに対して行っている。

```
   */ L:0 DA
```

```
+--+--+---+---+---+---+---+---+
|20|90|200|400|780|700|560|270|
+--+--+---+---+---+---+---+---+
```
これらのボックスをほどいて、合計する。つまり、得点の総合計を求める。
```
   +/ > */ L:0 DA
3020
```
一方、人数は、ボックス内の後ろの要素をとって、
```
   (+/ @ ({:)) L:0 DA
+-+-+-+-+--+--+-+-+
|1|3|5|8|13|10|7|3|
+-+-+-+-+--+--+-+-+
```
ボックスを開いて、人数の値を合計する。
```
   +/ > {: L:0 DA
50
```
平均値MEANは
```
   MEAN =: (+/ > */ L:0 DA) % (+/ > {: L:0 DA)
60.4
```
と求められる。

次に偏差を求める。まず、偏差は各点数から－上の値MEANを引く。
```
   MEAN -~ L:0 ({. L:0 DA)
+-----+-----+-----+-----+----+---+----+----+
|_40.4|_30.4|_20.4|_10.4|_0.4|9.6|19.6|29.6|
+-----+-----+-----+-----+----+---+----+----+
```
ボックス内のそれぞれの要素を２乗する。
```
   *: L:0 MEAN -~ L:0 ({. L:0 DA)
+-------+------+------+------+----+-----+------+------+
|1632.16|924.16|416.16|108.16|0.16|92.16|384.16|876.16|
+-------+------+------+------+----+-----+------+------+
```
ボックス内で、各人数を掛けるため、人数を付加する。
```
   (*: L:0 MEAN -~ L:0 ({. L:0 DA)) (,L:0) ({: L:0 DA)
+---------+--------+--------+--------+-------+--------+--------+--------+
|1632.16 1|924.16 3|416.16 5|108.16 8|0.16 13|92.16 10|384.16 7|876.16 3|
+---------+--------+--------+--------+-------+--------+--------+--------+
```
ボックス内で、それぞれの要素の掛け算を行う。
```
   (*/ L:0) (*: L:0 M -~ L:0 ({. L:0 DA)) (,L:0) ({: L:0 DA)
+-------+-------+------+------+----+-----+-------+-------+
|1632.16|2772.48|2080.8|865.28|2.08|921.6|2689.12|2628.48|
+-------+-------+------+------+----+-----+-------+-------+
```
ボックスを開いてリストとし、合計する。
```
   +/ > (*/ L:0) (*: L:0 M -~ L:0 ({. L:0 DA)) (,L:0) ({: L:0 DA)
13592
```
これを総人数で割る。
```
   (+/ > (*/ L:0) (*: L:0 M -~ L:0 ({. L:0 DA)) (,L:0) ({: L:0 DA) )
 % (+/ > {: L:0 DA)
271.84
```
この平方根をとったのが標準偏差STDVである。
```
  STDV =:    %: (+/ > (*/ L:0) (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) (,L:0) ({:
L:0 Sh_DA) ) % (+/ > {: L:0 Sharp_DA)
  STDV
16.4876
```

以上の平均と標準偏差を求める計算をまとめて、プログラムmean_stdとした。
```
mean_std =: 3 : 0
ME =. (+/ > */ L:0 y.) % (+/ > {: L:0 y.)
SD =. %: (+/ > (*/ L:0) (*: L:0 ME -~ L:0 ({. L:0 y.)) (,L:0) ({: L:0 y.) ) %
(+/ > {: L:0 y.)
ME, SD
)
```

```
   mean_std DA
60.4 16.4876
```

平均と標準偏差とが得られれば、その分布の状態はふつうの場合には次の正規分布の式に表すことができる。μ は平均、σ 標準偏差

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

これを元に分布の状態のグラフを描くことができる。

この後も含めて、鈴木義一郎氏の以下の書にはJのプログラム定義が挙げられている。これにより、Jのグラフィックツール plot により描くことができる。

[1] 鈴木義一郎「J言語による統計分析」森北出版、正規分布 p. 45-46

４　得点から成績を予測する

統計学で、通常は正規分布の分布関数表、積分値の表を使ってこのような予測を行うのが普通のやり方である。しかし、Jによれば、このような数表を用いなくとも、積分計算などを含めて、簡単に行うことができる。

つぎの鈴木義一郎氏のJプログラム定義[1]を利用させていただく。

```
NB. 標準正規分布の確率密度関数
Suz_ndens =: 3 : 0
(^ - -: *: y. ) % %: o.2
)
Suz_nden =: 3 : 0
:
s =. %: {: x.
(Suz_ndens(y. - {. x.) % s) % s
)

NB. 0 から y.(右引数)までの標準正規分布の確率密度関数の積分値
Suz_ndfs =: 3 : 0
h =. | y. % 250
(-: h * (Suz_ndens 0) + (Suz_ndens y.) ) + h * +/Suz_ndens (>: i.249) * h
)
Suz_ndf1 =: 0.5&+@(** Suz_ndfs)
Suz_ndf2 =: 0&>. @ -~ & (** Suz_ndfs)
```

つぎのように、関数名を変えた。

```
norm_integ =: Suz_ndf2          NB. renamed from Suzuki's definition
```

まず、簡単に使い方を数値表の値とチェックする。

```
  0 norm_integ 1
0.341344            … 数表の値と一致
   0 norm_integ 2
0.477249            … 数表の値と一致

   1 norm_integ 2
0.135905            ＝0.477249 - 0.341344
```

左、右の引数の値はそのままではなく、平均値、標準偏差を用いて、標準化した値を用いることが必要である。

それでは、シャープの関数電卓EL-526にあるいろいろな問題をやってみよう。

```
NB. 35 点以下  ------------
   (35 - MEAN) % STDV
_1.54055
   (0 - MEAN) % STDV
_3.66337
   _3.66337 norm_integ _1.54055
0.0615893   (6.2 %)

NB. 85 点以上  ------------
   (85 - MEAN) % STDV
1.49203
   0.5 - 0 norm_integ 1.49203
0.0678462   (6.8 %)

NB. 35 点から  75 点まで  ---------------
NB. まず35 点から 平均点(60.4) まで
   ((35 - MEAN) % STDV) norm_integ 0
0.438287
NB. 次に  平均点(60.4) から  75 点まで
   0 norm_integ (75 - MEAN) % STDV
0.31206

NB. 上の計算から、 35 点から  75 点まで
   0.438287 + 0.31206
0.750347
NB. まとめて計算する
   ((35 - MEAN) % STDV) norm_integ ((75 - MEAN) % STDV)
0.750347    (75 %)


NB. 平均点(60.4) から  75 点まで   ---------------
   0 norm_integ (75 - MEAN) % STDV
0.31206      (31 %)

NB. 0点  から  75 点まで   ---------------
   0.5 +  0.31206
0.81206      (81 %)
```

```
NB. 鈴木義一郎「J言語による統計分析」正規分布 p. 45-46 から ================
==============
NB.
NB. 標準正規分布の確率密度関数
Suz_ndens =: 3 : 0
(^ - -: *: y. ) % %: o.2
)
Suz_nden =: 3 : 0
:
s =. %: {: x.
(Suz_ndens(y. - {. x.) % s) % s
)
```

```
norm_dens =: Suz_nden          NB. renamed from Suzuki's definition
NB.    1 4 norm_dens 1 2 3 4                正規分布の確率密度関数の値
NB. 0.199471 0.176033 0.120985 0.0647588      左引数 = 平均, 分散
NB.    0 1 norm_dens 0 0.5 1 2              標準正規分布の密度関数の値
NB. 0.398942 0.352065 0.241971 0.053991


NB. 0 から y.(右引数)までの標準正規分布の確率密度関数の積分値
Suz_ndfs =: 3 : 0
h =. | y. % 250
(-: h * (Suz_ndens 0) + (Suz_ndens y.) ) + h * +/Suz_ndens (>: i.249) * h
)


Suz_ndf1 =: 0.5&+@(** Suz_ndfs)

Suz_ndf2 =: 0&>. @ -~ & (** Suz_ndfs)

norm_integ =: Suz_ndf2          NB. renamed from Suzuki's definition
NB. 0 norm_integ 1 => 0.341344  標準正規分布の(0 から 1) までの積分値 = 数表の
値
NB. 0 norm_integ 2 => 0.477249  標準正規分布の(0 から 2) までの積分値 = 数表の
値
NB. 1 norm_integ 2 => 0.135905  標準正規分布の(1 から 2) までの積分値
```

母集団が正規分布をなしているとして、標本の平均、分散をもとに推測してみると上のグラフになる。その計算はつぎのようになる。

```
   DAT4 =: 20 + 2.5 * i.32

   MEAN =: {. mean_std DA
   MEAN
60.4
   STD =: {: mean_std DA
   STD
16.4876
   VAR =: *: STD
   VAR
271.84

   plot DAT4;((MEAN, VAR) norm_dens  DAT4)
```

```
   load'f:¥j402¥user¥stat_sharp_dentaku.ijs'

   Sharp_DA
+----+----+----+----+-----+-----+----+----+
|20 1|30 3|40 5|50 8|60 13|70 10|80 7|90 3|
+----+----+----+----+-----+-----+----+----+

   */ L:0 Sharp_DA
+--+--+---+---+---+---+---+---+
|20|90|200|400|780|700|560|270|
+--+--+---+---+---+---+---+---+
   (+/ @ ({:)) L:0 Sharp_DA
+-+-+-+-+--+--+-+-+
|1|3|5|8|13|10|7|3|
+-+-+-+-+--+--+-+-+

   +/ > {: L:0 Sharp_DA
50

   +/ > */ L:0 Sharp_DA
3020

   (+/ > */ L:0 Sharp_DA) % (+/ > {: L:0 Sharp_DA)
60.4


   Sh_mean Sharp_DA
60.4

   Sh_DA
+----+----+----+----+-----+-----+----+----+
|20 1|30 3|40 5|50 8|60 13|70 10|80 7|90 3|
+----+----+----+----+-----+-----+----+----+
   Sh_M
60.4


   Sh_M -~ L:0 ({. L:0 Sh_DA)
+-----+-----+-----+-----+----+---+----+----+
|_40.4|_30.4|_20.4|_10.4|_0.4|9.6|19.6|29.6|
+-----+-----+-----+-----+----+---+----+----+

   *: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)
+-------+------+------+------+----+-----+------+------+
|1632.16|924.16|416.16|108.16|0.16|92.16|384.16|876.16|
+-------+------+------+------+----+-----+------+------+
   (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) , ({: L:0 Sh_DA)
+-------+------+------+------+----+-----+------+------+-+-+-+-+--+--+-+-+
|1632.16|924.16|416.16|108.16|0.16|92.16|384.16|876.16|1|3|5|8|13|10|7|3|
+-------+------+------+------+----+-----+------+------+-+-+-+-+--+--+-+-+
   (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) (,L:0) ({: L:0 Sh_DA)
+---------+--------+--------+--------+-------+--------+--------+--------+
|1632.16 1|924.16 3|416.16 5|108.16 8|0.16 13|92.16 10|384.16 7|876.16 3|
+---------+--------+--------+--------+-------+--------+--------+--------+
   (*/ L:0) (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) (,L:0) ({: L:0 Sh_DA)
+-------+-------+------+------+----+-----+-------+-------+
|1632.16|2772.48|2080.8|865.28|2.08|921.6|2689.12|2628.48|
+-------+-------+------+------+----+-----+-------+-------+
```

```
   +/ > (*/ L:0) (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) (,L:0) ({: L:0 Sh_DA)
13592

   (+/ > (*/ L:0) (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) (,L:0) ({: L:0 Sh_DA) )
% (+/ > {: L:0 Sharp_DA)
271.84
   %: (+/ > (*/ L:0) (*: L:0 Sh_M -~ L:0 ({. L:0 Sh_DA)) (,L:0) ({: L:0
Sh_DA) ) % (+/ > {: L:0 Sharp_DA)
16.4876
   load'f:¥j402¥user¥statistics.ijs'
   Sh_S
16.4876


   load'f:¥j402¥user¥statistics.ijs'

   (0 - Sh_M) % Sh_S
_3.66337
   (35 - Sh_M) % Sh_S
_1.54055
   (75 - Sh_M) % Sh_S
0.885516
   (85 - Sh_M) % Sh_S
1.49203
   (100 - Sh_M) % Sh_S
2.40181


   1 norm_integ 2
0.135905
   1.49203 norm_integ 2.40181
0.0596887
   _1.54055 norm_integ 0.885516
0.750347

   Sh_M
60.4
   Sh_S
16.4876

NB. 35 点以下  ------------
   (35 - Sh_M) % Sh_S
_1.54055
   (0 - Sh_M) % Sh_S
_3.66337
   _3.66337 norm_integ _1.54055
0.0615893


NB. 85 点以上  ------------
   (85 - Sh_M) % Sh_S
1.49203
   0.5 - 0 norm_integ 1.49203
0.0678462

NB. 35 点から  75 点まで  ---------------

NB. 35 点から 平均点(60.4) まで
   ((35 - Sh_M) % Sh_S) norm_integ 0
```

```
0.438287
NB. 平均点(60.4) から  75 点まで
   0 norm_integ (75 - Sh_M) % Sh_S
0.31206

NB. 上の計算から  => 35 点から  75 点まで
   0.438287 + 0.31206
0.750347
NB. まとめて計算する
   ((35 - Sh_M) % Sh_S) norm_integ ((75 - Sh_M) % Sh_S)
0.750347
```