雑感　ヒストグラムは怖い
—スタージェスの公式—

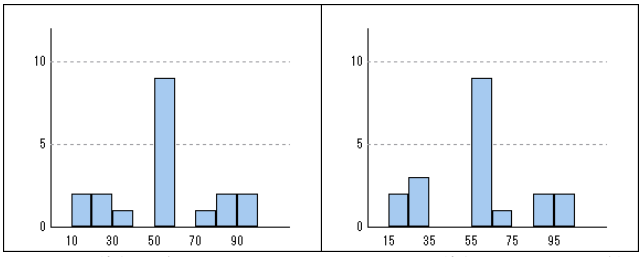
■　「ヒストグラムは」とタイトルしたが，ヒストグラムは度数分布表をグラフ化したものであるから，「度数分布表は」と言い換えても同じである．

■　次のデータ A がある．

A：15,15,25,25,30,55,55,55,55,55,55,55,55,55,55,70,85,85,95,95

このデータから，階級幅を 10 として，①のヒストグラムができる．

①　②

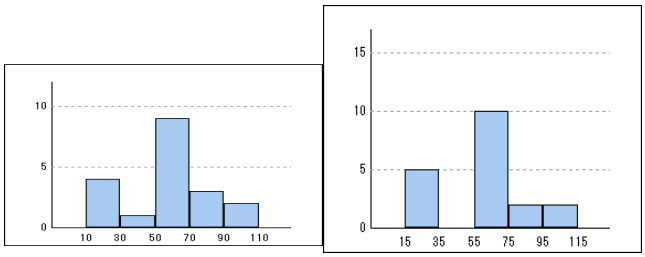


しかし，階級の幅を同じ 10 にとっても，階級の区切りの値を変えると②のヒストグラムができる．

同じデータであるが，ヒストグラムの様子は微妙に違う．

階級の幅を変えて 20 にすると，さらに様相は異なる．③，④は階級の幅を 20 として，階級の区切りを変えたものである．

③　④



このように 4 つ作ってみると，同じデータから作ったヒストグラムなのだろうかと思うくらいに，様子が異なる．

■　ヒストグラムを作るとき，階級の幅をいくつにすべきか（階級数をいくつにすべきか）悩むことが多い．

最近は，「データ数が n である場合，階級数 k は $1+\log_2 n$ 程度にするのがよい」という，スタージェス（Sturges）の公式がよく使われるらしい．

この「公式」の根拠はあるのだろうか？　それとも経験則から導き出された式なのであろうかとずっと思っていた．

最近，「The problem with Sturges' rule for constructing histograms」（Rob J Hyndman 1995）という論文を見つけて，スタージェスのアイデアを知ることができた．

論文の，この公式にかかわる部分はたった数行に過ぎない．

Herbert Sturges (1926) considered an idealised frequency histogram with k bins where the ith bin count is the binomial coefficient $\binom{k-1}{i}$, $i=0, 1, \cdots, k-1$. As k increases, this ideal frequency histogram approaches the shape of a normal density. The total sample size is

$$n=\sum_{i=0}^{k-1}\binom{k-1}{i}=(1+1)^{k-1}=2^{k-1}$$

by the binomial expansion. So the number of classes to choose when constructing a histogram from normal data is

$$k=1+\log_2 n\,.$$

This is Sturges' rule.

■　つまり，こうである．

データ数が多いとき，2 項分布が正規分布に近似されるということをベースにして，階級数が k で，階級 i ($i=0, 1, \cdots, k-1$) の度数が 2 項係数 ${}_{k-1}C_i$ に等しい度数分布を考えると，データ数の合計 n は

$n=\sum_{i=0}^{k-1}{}_{k-1}C_i=(1+1)^{k-1}=2^{k-1}$ に等しい．

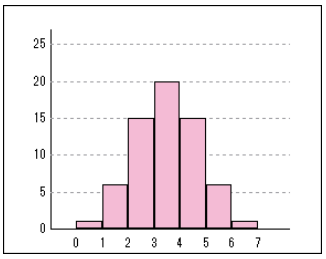
したがって，$k=1+\log_2 n$ である．

度数をデータ数の合計 n で割った「確率分布」でいえば，このデータの分布が 2 項分布 $B\left(n, \frac{1}{2}\right)$ に従うものとしている．

換言すれば，データが 2 項分布 $B\left(n, \frac{1}{2}\right)$ に従うとし，度数分布表を考えたとき，度数の最小値が 1 となるような階級数 k を考えると，$k=1+\log_2 n$ であるということでもある．

■　具体的には，7 個の階級からなる右のような度数分布表を考えると，度数の合計は $2^6=64$ になり，度数の最小値は 1 である．このとき，階級数は $1+\log_2 64=1+6=7$ になっている．ヒストグラムは，2 項係数の対称性から左右対称の「釣り鐘型」になる．

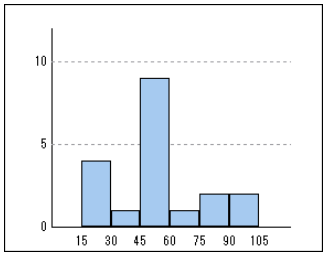
階級値	度数
0	${}_6C_0=1$
1	${}_6C_1=6$
2	${}_6C_2=15$
3	${}_6C_3=20$
4	${}_6C_4=15$
5	${}_6C_5=6$
6	${}_6C_6=1$



■　このスタージェスの公式が本当に有効かどうかは不明であるし，先述したように，ヒストグラムは階級の区切りをどこからにするかで変わってくるのである．

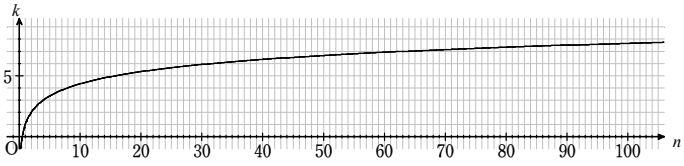
先のデータ A はデータ数が 19 であるから，スタージェスの公式によれば，$1+\log_2 19=1+4.24\cdots \fallingdotseq 5$ より，階級数は 5～6 程度が適切であるということになる．

5 としたヒストグラムは前掲の通りだが，6 としたものを右に載せた．しかし，階級が 15～30，30～45，…というのは余り実用的ではないような気がする．



■　このように，ヒストグラムは作り方次第でずいぶん様相が変わるものだということは，肝に銘じておきたい．したがって，ヒストグラムから「分布は○○である」というのは危険である．

ただ，ここで扱ったデータのようにデータ数が少ない場合にこういった問題が顕在化し，一定数多ければそれほど問題にならないのかも知れない．$k=1+\log_2 n$ からすれば，n が一定大きければ，データ数 n の増加に対して，階級数の増え方は緩やかである．



■　なお，拙 HP「雑感 52　箱ひげ図の能力」で，「データの様子を見るとき，ヒストグラムは最(？)強の表現手段であると思われる」と書いた．そこで取り上げたデータは整数の離散データで，階級幅 1 のヒストグラムであるため，データの様子が正確にヒストグラムに現れていることから，本稿とは矛盾しない．

2013 年 10 月 20 日