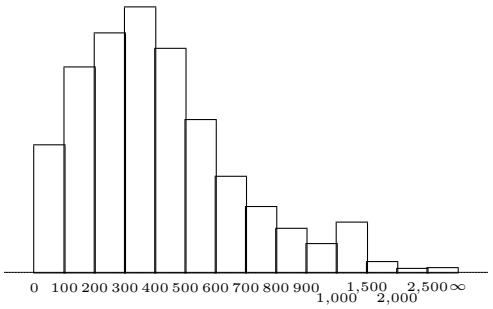
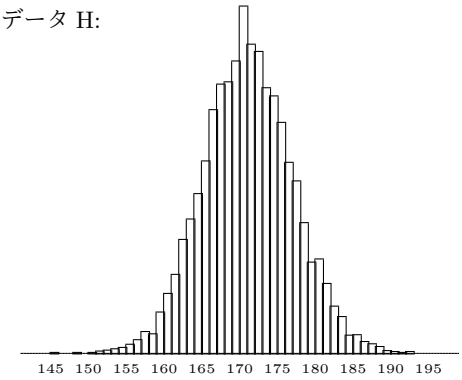
ありがちな分布

これまではデータの整理方法に主眼を置いたので、データの値はだいぶ不自然なものになっていた。しかし、世の中のデータを多く集めた場合は一定の特徴が見られるものである。

データ G:



データ H:



データ G は、2017 年国税庁の調査をもとにした給与階級別の年収状況を、割合（%）で表したヒストグラムである。横軸の単位は（万円）で、階級の範囲は 100 万円、500 万円が混在しているため、グラフの形状としては少し歪（ゆが）んで見えることになる。

階級値をもとに大まかに計算すると、平均値はおよそ 430 万円以上。“以上” というのは、2,500 万円以上の階級値を 3,000 万円で計算したからで、実際は数億円の収入の人たちのせいで平均値は上がるはずだからだ。しかし、最頻値は 300–400 万円であるから、平均値は実態より高めに出ていることになる。

さらに、中央値はおよそ 370 万円。中央値というのは、順位にしてちょうど真ん中に位置する人の年収であるから、半数以上の人の年収は 370 万円以下なのである。また、第一・四分位数はおよそ 220 万円。第一・四分位数というのは、順位にして先頭から（この例では下から）ちょうど 1/4 に位置する人の年収であるから、全体の 25%の人々の年収は 220 万円以下なのだ。ちなみに、第三・四分位数はおよそ 550 万円。すなわち、2017 年の年収が 550 万円以上であったなら、その人の収入は上位 25%の集団に属していたことになる。

データ H は、2017 年度学校保健統計調査をもとにした 17 歳男子の階級別身長を、割合（%）で表したヒストグラムである。横軸の目盛りは階級値で単位は（cm）、階級の幅は 1cm である。階級値 170cm のところに示された割合は、四捨五入して 170cm になる身長の人たちのものと考えてもらってよい。

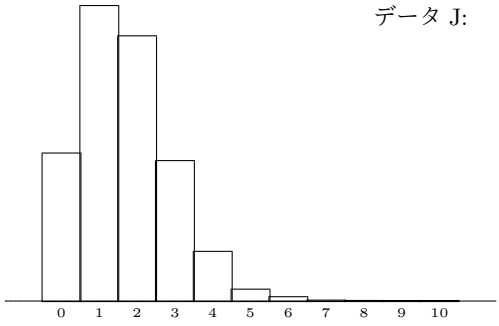
これは、ほぼ左右対称のグラフになっている。階級値をもとに大まかに計算すると、平均値はおよそ 170.6cm で、最頻値・中央値共に 170cm である。

ここに示した 2 種類のグラフは、実はありがちなグラフの形状なのである。とくに、データ G から得られるヒストグラムは 1,000 万円以上の階級幅も 100 万円にすると、右側にもっと長く漸減（ぜんげん）するロングテールが見られる。こういう形状のグラフは案外多いものだ。しかし、いずれにせよ基本的にグラフは最頻値の付近に多くのデータが集まり、離れるにしたがってデータ数が減っていくものである。

二項分布

年収は千円を単位とすると、2,345,678 円は 2,345.678 千円、身長は細かく測れば 169.75cm のように小数点以下までゆうに示せる。そのようなデータを数万人、数十万人集め小さい順に並べたとすれば、ほとんど連続的に変化しているように感じるかもしれない。しかし、それでは扱いが少々難しいと思われるので、データを理論的に整理するにあたって、離散的に変化するものを考えておこう。

数学的に扱いやすいのは確率だろう。サイコロを 10 回振る試行を 1 セットとし、10 回の試行中に 1 の目が何回出たかを調べるものとする。この試行を 2,400 セット行うとすれば、その分布は



のようになる。1 の目が 7 回以上出るところはグラフがないように見えるが、わずかな回数出現している。これは、実際に実験した結果ではなく、確率の計算から求められたものだ。形状はデータ G に近いが、本質はまったくの別物である。

このような分布を**二項分布**という。サイコロを振って 1 の目が出る確率は 1/6 だが、一般に事象 A が起こる確率を p（すると、起こらない確率は $1-p$）のとき、n 回の独立試行中に事象 A が r 回起こる確率は ${}_nC_r p^r (1-p)^{n-r}$ で求めることができる。つまり二項分布は、試行回数 n と事象 A が起こる確率 p で決定づけられる。これを $B(n,\ p)$ で表すことにする[1]。したがって、先のサイコロの話は二項分布 $B\left(10,\ \dfrac{1}{6}\right)$ に従うという。

[1] B は、二項分布（Binomial Distribution）からきている。

ところで証明は省くが、二項分布 $B(n,\ p)$ については、変量の平均 m と標準偏差 s は

$$\boxed{m = np、 \qquad s = \sqrt{np(1-p)}}$$

であることが知られている。このことをサイコロの例 $B\left(10,\ \frac{1}{6}\right)$ に当てはめると、10 回の試行で 1 の目が出る回数の平均は $m = 10 \times \frac{1}{6} \approx 1.67$（回）であり、標準偏差は $s = \sqrt{10 \times \frac{1}{6} \times \frac{5}{6}} \approx 0.952$ である。

サイコロは 6 回振れば、1 の目は 1 回程度は出そうなものだから、10 振ったときの平均値が 1.67 回というのは、当たり前のように思える。また、標準偏差が 0.952 といっても、どういう意味か即座に分かりかねるが、これについては後で理解が深まるだろう。いまは、変量の平均値からの偏差の加重平均が、1 弱であるという認識でしかないと思われる。

二項分布のグラフがデータ G のグラフと異なるのは、そもそも二項分布がヒストグラムで表されるものではないからである。実は、単に棒グラフで十分なのである。それは、確率分布

X	0	1	2	3	4	5	6	7	8	9	10	計
P	$(\frac{1}{6})^0(\frac{5}{6})^{10}$	${}_{10}C_1(\frac{1}{6})(\frac{5}{6})^9$	$\cdots$	${}_{10}C_X(\frac{1}{6})^X(\frac{5}{6})^{10-X}$	$\cdots$	$(\frac{1}{6})^{10}(\frac{5}{6})$	1					

をグラフ化したものになる。表中、X は**確率変数**で、サイコロを 10 回振って 1 の目が X 回出ることを表す。下段の P がそのときの確率である。1 の目は確率 $\frac{1}{6}$ で出る（確率 $\frac{5}{6}$ で出ない）ので、n 回の試行で 1 の目が r 回出る確率が ${}_nC_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{n-r}$ であることは、反復試行の確率で学んだであろう。データ J のグラフは、これをもとに描かれたのであった。

連続的な確率分布

二項分布は飛び飛びの値をとるので**離散的**である。一方、ヒストグラムになるようなデータの分布は、離散的とは言えない。たとえば、年収や身長のデータは 1 円刻み、0.1cm 刻みのように、正しく言えば離散的ではあるが、多くのデータを集めるとひとつながりの様相を呈してくる。厳密には年収や身長の確率変数は連続的に変化していないけれど、**連続的確率変数**と見る方が都合がよい。○以上△未満の範囲をまとめてヒストグラムにするのは、データがおおむね連続的であると考えているからだ。

ここで、離散的な確率変数 X に対する確率を $P(X)$ で表し、$P(X) = {}_{10}C_X(\frac{1}{6})^X(\frac{5}{6})^{10-X}$ などと書くように、a 以上 b 以下の値をとる確率変数 X に対する確率を $P(a \leqq X \leqq b)$ と書くことに

しよう。これまでなら、a 以上 b 未満の範囲を考えることが常だったので、$P(a \leqq X < b)$ と書いてもよいかもしれない。

このとき、サイコロを振る例で $a \leqq X \leqq b$ のような範囲を考えることは適切でないが、年収や身長のデータは X の範囲を考える方が適切である。現実問題としても、年収や身長の分布を連続的な確率分布と捉えてみよう。

確率密度関数

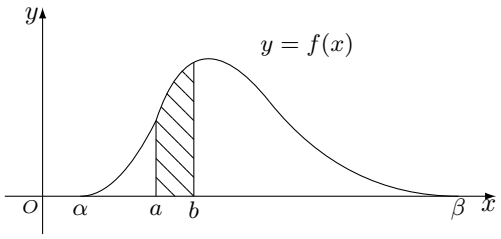
そこで、$\alpha \leqq X \leqq \beta$ を満たす連続的な確率変数 X について、唐突だが次のような性質の関数 $f(x)$ を対応させてみよう。

i) $f(x) \geqq 0$

ii) $P(a \leqq X \leqq b) = \displaystyle\int_a^b f(x)\,dx$

iii) $P(\alpha \leqq X \leqq \beta) = 1$

式だけ見ていると何のことだかよく分からないかもしれないが、積分で関係を与えていることから



のようなグラフにおいて、区間 $[a,\ b]$ の斜線部分の面積を確率 $P(a \leqq X \leqq b)$ の値と定めるということである。iii) は、定義された全区間 $[\alpha \leqq X \leqq \beta]$ の面積が 1 であることを意味するので、確率の合計が 1 であることに対応している。この条件を満たす関数 $f(x)$ を **X の確率密度関数**、曲線 $y = f(x)$ を**分布曲線**と呼ぶ。確率を区間の面積で表したので、ある点 $X = c$ の確率 $P(c)$ は 0 となることに注意されたい。

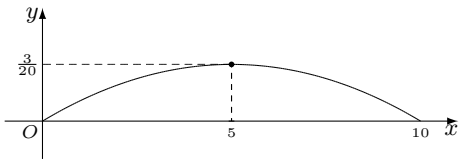
また、$\alpha \leqq X \leqq \beta$ の範囲を前提にしているが、$X \leqq \alpha,\ \beta \leqq X$ の範囲の確率が 0 ならば、実質的に $-\infty < X < \infty$ と同じことであるから、iii) は深く意識しなくてもよいだろう。

* * *

たとえば $f(x)=\dfrac{3}{500}x(10-x)$ $(0 \leqq x \leqq 10)$ は確率密度関数の条件を満たす。実際、定義域で $f(x) \geqq 0$ であり、

$$\int_0^{10} \frac{3}{500}x(10-x)\,dx = \frac{3}{500}\cdot\left\{-\frac{1}{6}(0-10)^3\right\} = 1$$

である。グラフの形状は



となっている。しかし、このような確率分布にしたがうものは現実的にはありそうもなく、特別な状況を創作しなくてはならないかもしれない。▌