平成 19 年 5 月 31 日

プログラム中のある乱数の評価

新潟工科大学 情報電子工学科 竹野茂治

1 はじめに

C 言語の rand() という標準関数は、0 から RAMD_MAX[1] の間の整数の乱数を返す。もちろん、実際にはある規則によって疑似乱数列を生成するだけであるが、これが一様な乱数であると仮定して、次のような問題を考えてみる:

> rand() を用いて、与えられた実数値 r $(0 \leq r \leq 1)$ に対して、確率 r で 1、確率 $1-r$ で 0 を返す関数を作るにはどのようにするのがいいか

容易に想像できるだろうが、if 文を使ってほぼこれを満たす関数を作ることはそれほど難しくはない。例えば C の関数として書けば次のような具合である:

```
int rand_real(double r)
{
    int x;
    double y;

    x = rand();
    y = x/(double)RAND_MAX; /* 0.0 <= y <= 1.0 */
    if (y <= r) return 1;
    else return 0;
}
```

すなわち、x を rand() の値とし、

$$\frac{x}{\texttt{RAND_MAX}} \leq r$$

ならば 1 を、そうでなければ 0 を返すようにするわけである。これは、それなりによさそうであるが、このようなものの妥当性について、少し数学的に検討をしてみたい。

[1] RAMD_MAX は、stdlib.h 等のヘッダーファイルで定義されている値で、例えば 16 進数の 7fffffff などのように定義されている。

2 実際の確率と r とのずれ

以後、簡単のために RAND_MAX を R_M と書くこととし、rand() の値を意味する x は、0 から R_M までの整数値を一様にとる確率変数であるとする。

この場合、$x = j$ $(j = 0, 1, 2, \ldots, R_M)$ である確率はすべて等しいことになるので、

$$\mathrm{Prob}\{x = j\}\ (=「x = j \text{ である確率」})\ = \frac{1}{R_M + 1} \tag{1}$$

となる。

1 節で紹介した方法では、$x \leq R_M \times r$ となるとき 1 となるようにしているが、この確率

$$P_1 = \mathrm{Prob}\{x \leq R_M \times r\}\ (=「x \leq R_M \times r \text{ である確率」})$$

がほぼ r に等しいことが期待される。

以下、まず一般の負でない実数 t に対して、確率

$$P = \mathrm{Prob}\{x \leq t\}$$

を t の式であらわすことを考えてみる。x は実際には整数の値しか取らないので、

$$P = \mathrm{Prob}\{x \leq \lfloor t \rfloor\}$$

に等しい。ここで、$\lfloor t \rfloor$ は、t 以下の最大の整数を表わすものとし[2]、例えば $\lfloor 3.2 \rfloor = 3$, $\lfloor 3 \rfloor = 3$ となる。

よって、(1) より、

$$\begin{aligned} P &= \mathrm{Prob}\{x = 0\} + \mathrm{Prob}\{x = 1\} + \cdots + \mathrm{Prob}\{x = \lfloor t \rfloor\} \\ &= \frac{\lfloor t \rfloor + 1}{R_M + 1} \end{aligned} \tag{2}$$

[2] $\lfloor t \rfloor$ は、C 言語でいう floor(t) に相当する。

となる。この公式 (2) を用いれば P_1 は

$$P_1 = \mathrm{Prob}\{x \leq R_M \times r\} = \frac{\lfloor R_M \times r \rfloor + 1}{R_M + 1} \tag{3}$$

となる。

ところで、$\lfloor t \rfloor$ は定義、あるいは図 1 のグラフより不等式

$$t - 1 < \lfloor t \rfloor \leq t \tag{4}$$

を満たすことがわかる。

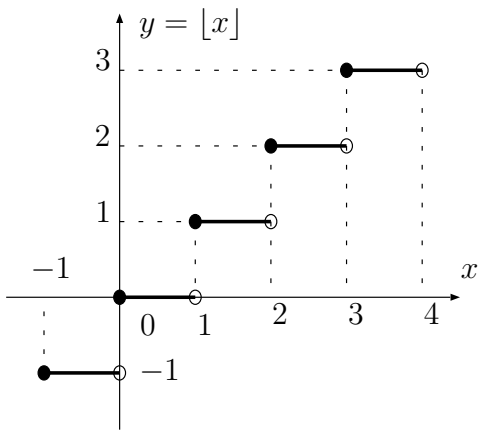

図 1: $\lfloor x \rfloor$ のグラフ

$(3), (4)$ より、P_1 は

$$\frac{R_M r}{R_M + 1} < P_1 \leq \frac{R_M r + 1}{R_M + 1}$$

を満たすので、よって、$P_1 - r$ については

$$-\frac{r}{R_M + 1} < P_1 - r \leq \frac{1 - r}{R_M + 1} \tag{5}$$

が成り立つ。$0 \leq r \leq 1$ だから、R_M が十分大きければ (5) の両端の値は十分 0 に近くなり、よってそれなりに P_1 は r に近いことになる。

3 その他の不等式

2 節の P_1 と r との違いをよりわかりやすく見るために、(3) を r の関数とみて、グラフに書いてみることにする。なお、ここでは簡単のため $R_M = 3$ としてみる。

このとき、P_1 は (3) より、

$$P_1 = \frac{\lfloor 3r \rfloor + 1}{4}$$

となるので、グラフは図 2 のようになる。

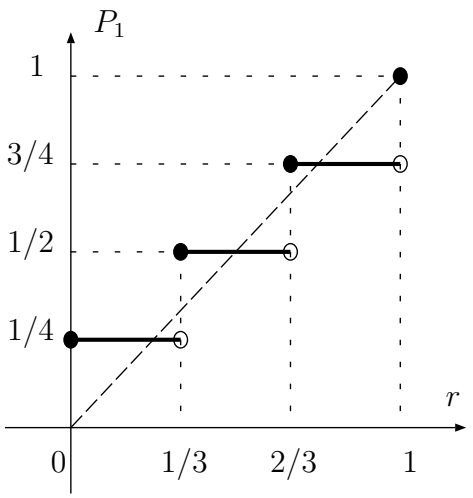

図 2: P_1 と r (破線) のグラフ

このグラフの横軸は 3 等分されているが、縦軸は 4 等分されている。これは、x の値が 0 から R_M までの $(R_M + 1)$ 通りあるのに、P_1 では

$$\frac{x}{R_M} \leq r \tag{6}$$

という不等式を利用していることに由来する。よって、むしろこれに変えて、

$$\frac{x+1}{R_M+1} \leq r \tag{7}$$

や、

$$\frac{x}{R_M+1} \leq r \tag{8}$$

という不等式を利用する方が自然であると思われる。(7) を利用する場合は、この左辺は $(0,1]$ の実数値で、(8) を利用する場合は、この左辺は $[0,1)$ の実数値となる。

これらの不等式を利用した場合の確率をそれぞれ P_2, P_3 とすると、(2) より、

$$\begin{aligned} P_2 &= \mathrm{Prob}\{(x+1)/(R_M+1) \le r\} = \mathrm{Prob}\{x+1 \le (R_M+1)r\} \\ &= \mathrm{Prob}\{x+1 \le \lfloor (R_M+1)r \rfloor\} = \mathrm{Prob}\{x \le \lfloor (R_M+1)r \rfloor - 1\} \\ &= \frac{\lfloor (R_M+1)r \rfloor}{R_M+1}, \end{aligned} \tag{9}$$

$$\begin{aligned} P_3 &= \mathrm{Prob}\{x/(R_M+1) \le r\} = \mathrm{Prob}\{x \le (R_M+1)r\} \\ &= \mathrm{Prob}\{x \le \lfloor (R_M+1)r \rfloor\} \\ &= \frac{\lfloor (R_M+1)r \rfloor + 1}{R_M+1} \end{aligned} \tag{10}$$

となる。

(4) より、P_2, P_3 は、

$$\begin{aligned} \frac{(R_M+1)r-1}{R_M+1} < P_2 \le \frac{(R_M+1)r}{R_M+1} = r, \\ r = \frac{(R_M+1)r}{R_M+1} < P_3 \le \frac{(R_M+1)r+1}{R_M+1} \end{aligned}$$

を満たすので、r との差は、

$$-\frac{1}{R_M+1} < P_2 - r \le 0, \quad 0 < P_3 - r \le \frac{1}{R_M+1} \tag{11}$$

と評価される。

R_M が十分大きければ、(11) から P_2, P_3 も r に十分近いことがわかるが、P_2 は r 以下、P_3 は r 以上であり、(5) と比較してみても、(11) の P_2, P_3 はそれほどよいとも言えないように見える。

P_2, P_3 を r の関数と見ると、$R_M = 3$ の場合は、

$$P_2 = \frac{\lfloor 4r \rfloor}{4}, \quad P_3 = \frac{\lfloor 4r \rfloor + 1}{4}$$

であるので、グラフは図 3 のようになる。

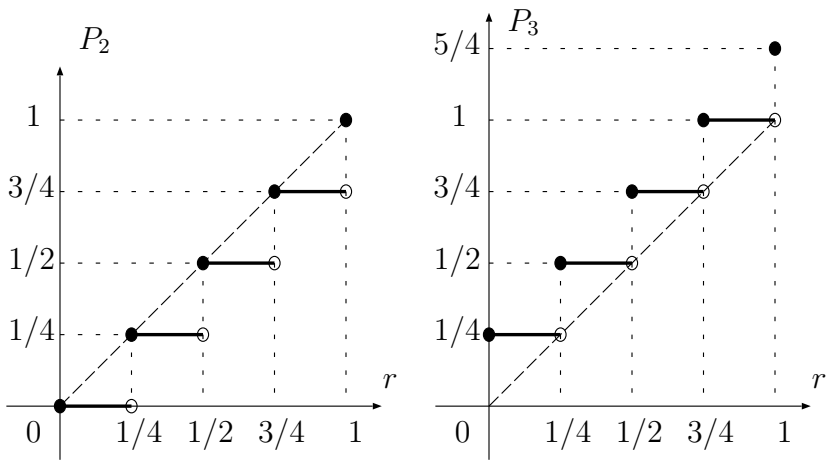

図 3: P_2 と P_3 のグラフ

見てわかる通り、この P_2, P_3 の中間のようなグラフがむしろ r に近いことがわかる。P_2, P_3 の中間のようなグラフとしては、例えば次の 2 種類が考えられる:

$$P_4 = \frac{\lfloor (R_M+1)r \rfloor + 1/2}{R_M+1} \tag{12}$$

$$P_5 = \frac{\lfloor (R_M+1)r + 1/2 \rfloor}{R_M+1} \tag{13}$$

例えば $R_M = 3$ のときは、

$$P_4 = \frac{\lfloor 4r \rfloor}{4} + \frac{1}{8}, \quad P_5 = \frac{\lfloor 4r + 1/2 \rfloor}{4}$$

であり、これらのグラフは図 4 のようになる。

P_4 は P_2 を上に $1/8$ ずらしたもの、P_5 は P_2 を左に $1/8$ ずらしたものとなっている。方向は違うが、いずれも r とのずれは同じ位になっている。

P_4 は、$(9), (10), (12)$ より

$$P_4 = \frac{1}{2}P_2 + \frac{1}{2}P_3$$

なので、例えば 2 回に 1 回 (7) を使い、2 回に 1 回 (8) を使えば、その確率は P_4 となる。

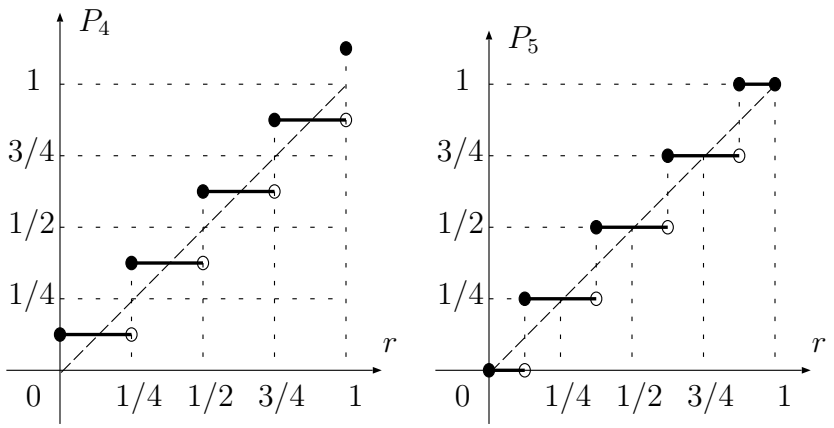

図 4: P_4 と P_5 のグラフ

一方 P_5 は、

$$\begin{aligned} P_5 &= \frac{\lfloor (R_M+1)r + 1/2 \rfloor}{R_M+1} = \mathrm{Prob}\{x \le \lfloor (R_M+1)r - 1/2 \rfloor\} \\ &= \mathrm{Prob}\{x \le (R_M+1)r - 1/2\} \end{aligned}$$

であるので、

$$\frac{x+1/2}{R_M+1} \le r \tag{14}$$

という不等式を用いれば P_5 になる。もちろん、P_4 よりも P_5 の方が実装は容易である。

4 平均自乗誤差

$P_1 \sim P_5$ と r との近さをはかる尺度として、誤差評価でよく用いられる平均自乗誤差を計算してみる。$f(r)$ と $g(r)$ $(a \le r \le b)$ の 平均自乗誤差 (以後 E とする) とは、

$$E = \left\{ \frac{1}{b-a} \int_a^b (f(r) - g(r))^2 dr \right\}^{1/2}$$

で定義されるものであり、関数解析学では L^2 ノルム と呼ばれている[3]。これを $P_1 \sim P_5$ に対して計算してみる。

[3]通常 L^2 ノルムでは $1/(b-a)$ はつけない。

P_j に対する平均自乗誤差を E_j $(1 \leq j \leq 5)$ とすると、グラフから、

$$E_2 = E_3, \quad E_4 = E_5$$

となることが予想されるが、一応すべて計算してみることにする。まずは E_1 から。

$$\begin{aligned}
E_1^2 &= \int_0^1 (r - P_1)^2 dr = \int_0^1 \left(r - \frac{\lfloor R_M r \rfloor + 1}{R_M + 1} \right)^2 dr \\
&= \frac{1}{R_M} \int_0^{R_M} \left(\frac{t}{R_M} - \frac{\lfloor t \rfloor + 1}{R_M + 1} \right)^2 dt \quad (t = R_M r) \\
&= \frac{1}{R_M} \sum_{k=1}^{R_M} \int_{k-1}^{k} \left(\frac{t}{R_M} - \frac{k}{R_M + 1} \right)^2 dt \quad (k-1 \leq t < k \text{ では } \lfloor t \rfloor = k-1) \\
&= \frac{1}{R_M} \sum_{k=1}^{R_M} \left[\frac{R_M}{3} \left(\frac{t}{R_M} - \frac{k}{R_M + 1} \right)^3 \right]_{t=k-1}^{t=k} \\
&= \frac{1}{3} \sum_{k=1}^{R_M} \left\{ \left(\frac{k}{R_M} - \frac{k}{R_M + 1} \right)^3 - \left(\frac{k-1}{R_M} - \frac{k}{R_M + 1} \right)^3 \right\} \\
&= \frac{1}{3} \sum_{k=1}^{R_M} \left\{ \frac{k^3}{R_M^3 (R_M + 1)^3} - \left(\frac{k}{R_M (R_M + 1)} - \frac{1}{R_M} \right)^3 \right\} \\
&= \frac{1}{3R_M^3} \sum_{k=1}^{R_M} \left\{ \frac{k^3}{(R_M + 1)^3} - \left(\frac{k}{R_M + 1} - 1 \right)^3 \right\} \\
&= \frac{1}{3R_M^3} \sum_{k=1}^{R_M} \left\{ \frac{3k^2}{(R_M + 1)^2} - \frac{3k}{R_M + 1} + 1 \right\} \\
&= \frac{1}{3R_M^3} \left\{ \frac{3R_M (R_M + 1)(2R_M + 1)}{6(R_M + 1)^2} - \frac{3R_M (R_M + 1)}{2(R_M + 1)} + R_M \right\} \\
&= \frac{1}{3R_M^2} \left\{ \frac{2R_M + 1}{2(R_M + 1)} - \frac{3}{2} + 1 \right\} = \frac{1}{3R_M^2} \frac{R_M}{2(R_M + 1)} = \frac{1}{6R_M (R_M + 1)}
\end{aligned}$$

よって、E_1 は

$$E_1 = \frac{1}{\sqrt{6R_M (R_M + 1)}}$$

となる。

次に E_2。

$$\begin{aligned} E_2^2 &= \int_0^1 (r-P_2)^2 dr = \int_0^1 \left(r - \frac{\lfloor (R_M+1)r \rfloor}{R_M+1} \right)^2 dr \\ &= \frac{1}{R_M+1} \int_0^{R_M+1} \left(\frac{t}{R_M+1} - \frac{\lfloor t \rfloor}{R_M+1} \right)^2 dt \quad (t=(R_M+1)r) \\ &= \frac{1}{(R_M+1)^3} \sum_{k=1}^{R_M+1} \int_{k-1}^{k} (t-k+1)^2 dt = \frac{1}{(R_M+1)^3} \sum_{k=1}^{R_M+1} \int_0^1 t^2 dt \\ &= \frac{1}{3(R_M+1)^2} \end{aligned}$$

よって、E_2 は

$$E_2 = \frac{1}{\sqrt{3(R_M+1)^2}} = \frac{1}{\sqrt{3}(R_M+1)}$$

となる。E_1 と比較すると、

$$E_1^2 - E_2^2 = \frac{(R_M+1)-2R_M}{6R_M(R_M+1)^2} = -\frac{R_M-1}{6R_M(R_M+1)^2} < 0$$

より、$E_1 < E_2$ である。

E_3 は、

$$\begin{aligned} E_3^2 &= \int_0^1 (r-P_3)^2 dr = \int_0^1 \left(r - \frac{\lfloor (R_M+1)r \rfloor + 1}{R_M+1} \right)^2 dr \\ &= \frac{1}{R_M+1} \int_0^{R_M+1} \left(\frac{t}{R_M+1} - \frac{\lfloor t \rfloor + 1}{R_M+1} \right)^2 dt \\ &= \frac{1}{(R_M+1)^3} \sum_{k=1}^{R_M+1} \int_{k-1}^{k} (t-k)^2 dt = \frac{1}{(R_M+1)^3} \sum_{k=1}^{R_M+1} \int_0^1 (t-1)^2 dt \\ &= \frac{1}{3(R_M+1)^2} = E_2^2 \end{aligned}$$

となるので、確かに $E_3 = E_2$ となっている。

次は E_4。

$$\begin{aligned}
E_4^2 &= \int_0^1 (r-P_4)^2 dr = \int_0^1 \left(r - \frac{\lfloor (R_M+1)r \rfloor + 1/2}{R_M+1} \right)^2 dr \\
&= \frac{1}{(R_M+1)^3} \sum_{k=1}^{R_M+1} \int_{k-1}^{k} \left(t-k+\frac{1}{2} \right)^2 dt = \frac{1}{(R_M+1)^3} \sum_{k=1}^{R_M+1} \int_0^1 \left(t-\frac{1}{2} \right)^2 dt \\
&= \frac{1}{12(R_M+1)^2} = \frac{1}{4} E_2^2
\end{aligned}$$

となるので、$E_4 = E_2/2$ となる。

最後に E_5。

$$\begin{aligned}
E_5^2 &= \int_0^1 (r-P_5)^2 dr = \int_0^1 \left(r - \frac{\lfloor (R_M+1)r + 1/2 \rfloor}{R_M+1} \right)^2 dr \\
&= \frac{1}{R_M+1} \int_{1/2}^{R_M+3/2} \left(\frac{t-1/2}{R_M+1} - \frac{\lfloor t \rfloor}{R_M+1} \right)^2 dt \quad (t = (R_M+1)r + 1/2) \\
&= \frac{1}{(R_M+1)^3} \int_{1/2}^{R_M+3/2} \left(t - \lfloor t \rfloor - \frac{1}{2} \right)^2 dt \\
&= \frac{1}{(R_M+1)^3} \left\{ \int_{1/2}^{1} \left(t-\frac{1}{2} \right)^2 dt + \sum_{k=2}^{R_M+1} \int_{k-1}^{k} \left(t-k+\frac{1}{2} \right)^2 dt \right. \\
&\quad \left. + \int_{R_M+1}^{R_M+3/2} \left(t-R_M-\frac{3}{2} \right)^2 dt \right\} \\
&= \frac{1}{(R_M+1)^3} \left\{ \int_{1/2}^{1} \left(t-\frac{1}{2} \right)^2 dt + \sum_{k=2}^{R_M+1} \int_{0}^{1} \left(t-\frac{1}{2} \right)^2 dt \right. \\
&\quad \left. + \int_{0}^{1/2} \left(t-\frac{1}{2} \right)^2 dt \right\} \\
&= \frac{R_M+1}{(R_M+1)^3} \int_0^1 \left(t-\frac{1}{2} \right)^2 dt = \frac{1}{12(R_M+1)^2} = E_4^2
\end{aligned}$$

となるので、確かに $E_5 = E_4$ となる。

E_5 と E_1 を比較すると、

$$E_1^2 - E_5^2 = \frac{2(R_M+1) - R_M}{12R_M(R_M+1)^2} = \frac{R_M+2}{12R_M(R_M+1)^2} > 0$$

より、$E_1 > E_5$ となる。

よって、結果として、

$$E_2 = E_3 > E_1 > E_4 = E_5$$

となることがわかる。

5 実数軸上の分布

4 節の計算によれば、平均自乗誤差の点では (6) (すなわち P_1) という不等式は (7) (すなわち P_2) や (8) (すなわち P_3) を用いるよりもよく、(14) (すなわち P_5) はそれらよりもさらによいことがわかったが、これは次のように実数軸上の区間 $[0,1]$ 内の分布として考えるとより直感的にわかるだろう。

簡単のため、$R_M = 3$ とすると、P_1 は (6) を用いるということは、$x = 0, 1, 2, 3$ の 4 通りが同様に確からしく起こるのに対して、$y = x/R_M = 0, 1/3, 2/3, 1$ の 4 通りが同様に現れることになり、これで 0 から 1 までの実数の乱数を作っていることになる (図 5)。

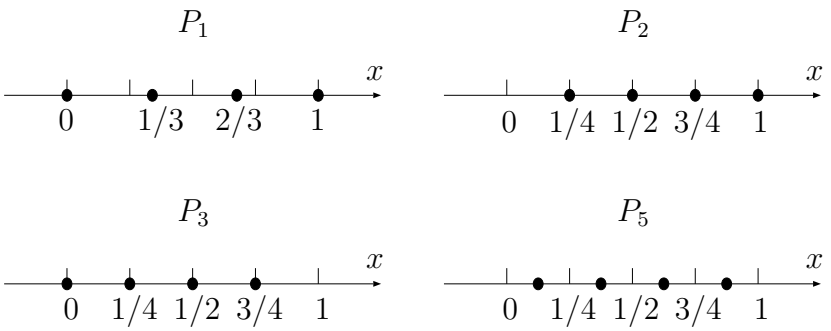

図 5: P_j $(j = 1, 2, 3, 5)$ の分布 $(R_M = 3$ の場合$)$

これに対して、P_2 は、(7) であるから $y = (x+1)/(R_M+1) = 1/4, 1/2, 3/4, 1$ の 4 通り、P_3 は、(8) であるから $y = x/(R_M+1) = 0, 1/4, 1/2, 3/4$ の 4 通りを作っていくことになり、これは、P_1 に比べて区間 $[0,1]$ 全体に均等に分布しているとは言いづらい。

一方 P_5 は、(14) であるから $y = (x+1/2)/(R_M+1) = 0.5/4, 1.5/4, 2.5/4, 3.5/4$ の 4 通りで、より区間 $[0,1]$ に綺麗に均等に分布していることがわかると思う。

6 等号を含めない条件の場合

ここまでは、不等式の条件として

$$P = \mathrm{Prob}\{x \leq t\}$$

を考えたが、これを

$$P' = \mathrm{Prob}\{x < t\}$$

にした考察ももちろん行える。この場合は、

$$\begin{aligned} P' &= \mathrm{Prob}\{x < t\} = 1 - \mathrm{Prob}\{x \geq t\} = 1 - \mathrm{Prob}\{x \geq \lceil t \rceil\} \\ &\quad (\lceil t \rceil \text{ は、} t \text{ 以上の最小の整数}) \\ &= 1 - (\mathrm{Prob}\{x = \lceil t \rceil\} + \mathrm{Prob}\{x = \lceil t \rceil + 1\} + \cdots + \mathrm{Prob}\{x = R_M\}) \\ &= 1 - \frac{R_M - \lceil t \rceil + 1}{R_M + 1} = \frac{\lceil t \rceil}{R_M + 1} \end{aligned}$$

となるので、これを用いれば、上と同じ考察が行える。

この $\lceil x \rceil$[4]のグラフは図 6 のようになるから、$\lfloor x \rfloor$ (図 1 参照) と比較すればわかるが、

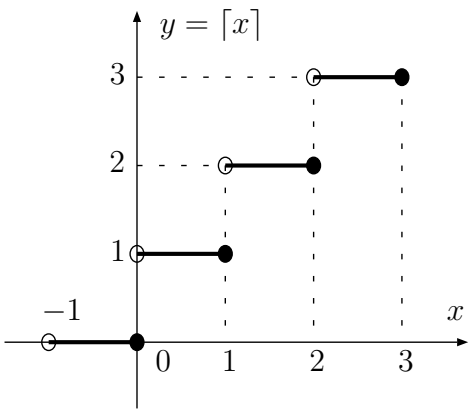

図 6: $\lceil x \rceil$ のグラフ

ほぼ $\lceil x \rceil \approx \lfloor x \rfloor + 1$ のように考えて考察できる。

[4] $\lceil x \rceil$ は C 言語でいう ceil(x)。

7 最後に

実際の C 言語の処理系には、drand48() のように区間 $[0, 1)$ の実数の乱数の値を返す標準関数が用意されていることが多い。よって、実は今回の問題はそれを利用して、

```
int rand_real(double r)
{
    if (drand48() <= r) return 1;
    else return 0;
}
```

とすればいいだけのことなのであるが、C 言語のコードを数学的に評価する、しかも積分によって定量的な評価を行う、といったことの例 (学生向けの例題) として、それなりに意味があるのではないかと思う。