1　問題編

問題です。

「某社の新入社員は 120 人います。
さて、1 年を 365 日としたとき、誕生日が同じ新入社員は何人いると予想できますか？」

もう少し教科書的にいえば、
「番号の振られた箱が n 個、番号の振られた玉が m 個あります。
すべての玉を箱の中に適当に入れるとします。
一つの箱に玉を何個でも入れてかまいません。
では、玉が 1 個だけ入った箱の数の期待値（平均値）は何個でしょうか？」
このとき $n = 365, m = 120$ とすれば、

$$\text{誕生日が同じ新入社員の期待値} = m - \text{誕生日が誰とも同じでない人数の期待値} \tag{1}$$
$$\text{誕生日が誰とも同じでない人数の期待値} = \text{玉が 1 個だけ入った箱の数の期待値} \tag{2}$$

の関係が成り立ちます。

さて、何人いると思われますか?

2 解答編 1 (近似解)

m 個の玉を次のようにグループ分けします。

- 1 個の玉からなるグループを a_1 組 ($a_1 \geq 0$)。
- 2 個の玉からなるグループを a_2 組 ($a_2 \geq 0$)。

  $\vdots$
- r 個の玉からなるグループを a_r 組 ($a_r \geq 0$)。

つまり

$$m = a_1 + 2a_2 + 3a_3 + \cdots + ra_r, \quad (a_i \geq 0) \tag{3}$$

となるように $a_1, a_2, \ldots$ を選びます。以降、この操作を m の分割とよび、$a_1, a_2, \ldots$ を $\{a_i\}_m$ と略記します。

各分割について、$a = a_1 + a_2 + \cdots$ 組のグループをそれぞれ別の箱 (全部で n 個) に入れる組み合わせ $P_n(\{a_i\}_m)$ は次のようになります。

$$\begin{aligned}
P_n(\{a_i\}_m) = {}& {}_nC_{a_1} \prod_{i_1=0}^{a_1-1} {}_{m-i_1}C_1 \\
& \times {}_{n-a_1}C_{a_2} \prod_{i_2=0}^{a_2-1} {}_{m-a_1-2i_2}C_2 \\
& \times {}_{n-a_1-a_2}C_{a_3} \prod_{i_3=0}^{a_3-1} {}_{m-a_1-2a_2-3i_3}C_3 \\
& \quad \vdots \\
& \times {}_{n-a_1-a_2-\cdots-a_{r-1}}C_{a_r} \prod_{i_r=0}^{a_r-1} {}_{m-a_1-2a_2-\cdots-(r-1)a_{r-1}-ri_r}C_r \qquad (4) \\
= {}& \frac{m!n!}{(n-a)!} \prod_{k=1}^{r} \frac{1}{a_k!(k!)^{a_k}} \qquad (5)
\end{aligned}$$

$i > r$ に対して $a_i = 0$ であるから $r \to \infty$ としてもこの式は成り立ちます。

従って、m の分割 $\{a_i\}_m$ を n 個の箱に入れる組み合わせは

$$P_n(\{a_i\}_m) = \frac{m!n!}{(n-a)!} \prod_{k=1}^{\infty} \frac{1}{a_k!(k!)^{a_k}} \tag{6}$$

と表せることが分かります。

よって、求める玉が 1 個だけ入った箱の数の期待値 $\langle a_1 \rangle_{n,m}$ は

$$\langle a_1 \rangle_{n,m} = \frac{L_{n,m}^{(1)}}{n^m} \tag{7}$$

$$L_{n,m}^{(1)} = \sum_{\{a_i\}_m} a_1 P_n(\{a_i\}_m) \tag{8}$$

を計算すれば得られます。ここで $\sum_{\{a_i\}_m}$ は m の分割に関する和 ($m = a_1 + 2a_2 + 3a_3 + \cdots, (a_i \geq 0)$ を満たす $a_1, a_2, \ldots$ に関する和) を表します。

さて次に実際に $L^{(1)}_{n,m}$ を計算をしたいわけですが、まずは計算機を使って $m = a_1 + 2a_2 + 3a_3 + \cdots, (a_i \geq 0)$ を満たす $a_1, a_2, \ldots$ の組み合わせを生成して、式 (8) の和を実行する方法について考えましょう。

しかし、この方法には問題があります。m の分割の仕方は膨大な数になるからです。例えば $m = 120$ の場合、1844349560 通りの分割の仕方があります。そのため各分割について $P_n(\{a_i\}_m)$ を求める計算が 1 秒間に 10 万回計算できるとしても、$\langle a_1 \rangle_{n,m}$ を求めるのに 5 時間もかかることになります。従って、この方法では、式 (8) の和に強く寄与する項を優先的に和をとるなどして、うまく近似解を求めるのがいいと考えられます。

式 (6) を見ると、k の値が大きいほど $(k!)^{-a_k}$ の効果で $P_n(\{a_i\}_m)$ の値が非常に小さくなることが読み取れます。従って、a_1 の値が大きい分割ほど $a_i, (i > 1)$ の値が小さくなるため、式 (8) の和により寄与することが予想できます。(ただし、a_1 の値が小さい分割ほど分割の仕方が多いため、塵も積もればの要領で結果的に大きな数になる可能性があります。今回はいくつか実験してみて予想通りの結果が得られたのでよしとしました。)

この考察をもとに、a_1 の値が大きい分割から順に和をとって $\langle a_1 \rangle_{n,m}$ を計算するプログラムが

- `http://sonoisa.github.io/misc/birthday/birthday.c`

です。ちゃちゃっとプログラムを書いて計算したかったので、多倍長演算に Mathematica 4.2 の MathKernel を使用しています。(面倒にはなりますが、もちろん C 言語の多倍長演算ライブラリでも代替可能です。)

実行結果は

- `hhttp://sonoisa.github.io/misc/birthday/L1sequence.txt`

です。$120, 118, \ldots$ が a_1 の値、sum とラベルされた値が $L^{(1)}_{n,m}$ の値です。予想通り収束していく様子が見られます。

有効桁 4 桁として、$a_1 = 60$ (計算時間は iBook G4 800MHz で約 1 時間半) で打ち切った $L^{(1)}_{n,m}$ の値

2585449862832371164981251820404082605635333416346864947950293037282755
8276979842885127836327724944830615068666224887998064112620889373561559
7153487833994083684628617523795551821349594477711910929603986937298249
4484917730882656173656442269582364671423894980680527756459013065105293
118747610972160000000000000000

を採用すると、箱が $n = 365$ 個、玉が $m = 120$ 個のとき、玉が 1 個だけ入った箱の数の期待値として

$$\langle a_1 \rangle_{n,m} = \frac{L^{(1)}_{n,m}}{n^m} \simeq 86.57 \tag{9}$$

が得られます。よって式 (1), (2) より、

$$\text{誕生日が同じ新入社員の期待値} \simeq 33.43\ \text{人} \tag{10}$$

となります。

つまり「新入社員が 120 人いるとき、誕生日が同じ新入社員は約 33 人いる」と予想できます。率にして全体の約 1/4 が誰かと誕生日が一緒だということになります。意外と多いですね。

3　解答編 2 (厳密解)

勘のいい人は $P_n(\{a_i\}_m)$ (式 (6)) を見た瞬間に指数関数の Tayler 展開

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{11}$$

の係数に似ていることに気がつかれたと思います。この視点で手を少し動かしてみると、次の面白い関数 (母関数) を発見できます。

$$G(x,y) = e^{ye^x} \tag{12}$$

この関数と $P_n(\{a_i\}_m)$ の関係をみてみましょう。

$$G(x,y) = e^{ye^x} = e^{y\sum_{k=0}^{\infty}\frac{x^k}{k!}} = \prod_{k=0}^{\infty} e^{y\frac{x^k}{k!}} \tag{13}$$

$$= \prod_{k=0}^{\infty} \sum_{a_k=0}^{\infty} \frac{1}{a_k!(k!)^{a_k}} y^{a_k} x^{ka_k} \tag{14}$$

$$= \left(\sum_{b=0}^{\infty} \frac{y^b}{b!}\right) \prod_{k=1}^{\infty} \sum_{a_k=0}^{\infty} \frac{1}{a_k!(k!)^{a_k}} y^{a_k} x^{ka_k} \tag{15}$$

$$= \sum_{b=0}^{\infty} \sum_{a_1=0}^{\infty} \sum_{a_2=0}^{\infty} \cdots \frac{1}{b!} \left(\prod_{k=1}^{\infty} \frac{1}{a_k!(k!)^{a_k}}\right) y^{a+b} x^m \tag{16}$$

ここで $a = \sum_{k=0}^{\infty} a_k = a_1 + a_2 + a_3 + \cdots, m = \sum_{k=0}^{\infty} ka_k = a_1 + 2a_2 + 3a_3 + \cdots$ です。従って、$y^n x^m$ の項は

$$\sum_{a_1=0}^{\infty} \sum_{a_2=0}^{\infty} \cdots \frac{1}{(n-a)!} \left(\prod_{k=1}^{\infty} \frac{1}{a_k!(k!)^{a_k}}\right) y^n x^m \tag{17}$$

と表せます。ただし、$a_1, a_2, \ldots$ は $a \leq n$ と $m = \sum_{k=0}^{\infty} ka_k$ を満たす範囲を動くとします。この式と式 (6) を比較することにより

$$G(x,y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \sum_{\{a_i\}_m} \frac{P_n(\{a_i\}_m)}{n!m!} y^n x^m \tag{18}$$

という関係が得られます。一方で、Tayler 展開公式より

$$G(x,y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{n!m!} \left(\frac{\partial^m}{\partial x^m} \frac{\partial^n}{\partial y^n} G(0,0)\right) y^n x^m \tag{19}$$

$$= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{n^m}{n!m!} y^n x^m \tag{20}$$

がわかります。この式 (20) と式 (18) の係数を比較すれば

$$\sum_{\{a_i\}_m} P_n(\{a_i\}_m) = \sum_{\{a_i\}_m} \frac{m!n!}{(n-a)!} \left(\prod_{k=1}^{\infty} \frac{1}{a_k!(k!)^{a_k}} \right) = n^m \tag{21}$$

という関係式が導けます。これは m 個の玉を

- 1 個の玉からなるグループを a_1 組 ($a_1 \geq 0$)。
- 2 個の玉からなるグループを a_2 組 ($a_2 \geq 0$)。
  ⋮

と分割したものをそれぞれ別の箱 (全部で n 個) に入れる組み合わせの総数は、当然ながら、n^m 通り (m 個の玉を重複を許して n 個の箱に入れる組み合わせ) になることを示しています。

さて、次に $G(x,y)$ を手本に本題の

$$\frac{L^{(1)}_{n,m}}{n!m!} = \sum_{\{a_i\}_m} \frac{a_1}{(n-a)!} \left(\prod_{k=1}^{\infty} \frac{1}{a_k!(k!)^{a_k}} \right) \tag{22}$$

を係数に持つ母関数を作りましょう。式 (14) の $k=1$ の項をみると、y^{a_1} を y について微分して y を掛ければ、欲しい a_1 という係数が y^{a_1} の肩から降りてくることがわかります。この $k=1$ の項は e^{yx} の展開から得られたため (式 (13) 参照)

$$G^{(1)}(x,y) = z \frac{\partial}{\partial z} e^{y(e^x - x) + zx} \Big|_{z=y} \tag{23}$$

$$= xye^{ye^x} = xyG(x,y) \tag{24}$$

が求める母関数であるとわかります。つまり

$$G^{(1)}(x,y) = xyG(x,y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{L^{(1)}_{n,m}}{n!m!} y^n x^m \tag{25}$$

です。従って、式 (20) より

$$L^{(1)}_{n,m} = m!n! \frac{(n-1)^{m-1}}{(m-1)!(n-1)!} = mn(n-1)^{m-1} \tag{26}$$

となり、$\langle a_1 \rangle_{n,m}$ の厳密解は

$$\langle a_1 \rangle_{n,m} = \frac{L^{(1)}_{n,m}}{n^m} = m \left(\frac{n-1}{n} \right)^{m-1} \tag{27}$$

であることがわかります。同様の方法で標準偏差を求めることもできます。

$$\sigma^{(1)}_{n,m} = \sqrt{\langle a_1^2 \rangle_{n,m} - \langle a_1 \rangle^2_{n,m}} \tag{28}$$

$$= \sqrt{m(m-1) \left(\frac{n-1}{n} \right) \left(\frac{n-2}{n} \right)^{m-2} + m \left(\frac{n-1}{n} \right)^{m-1} - m^2 \left(\frac{n-1}{n} \right)^{2(m-1)}} \tag{29}$$

これらの式に $n = 365, m = 120$ を代入すると

$$\langle a_1 \rangle_{n,m} = 120 \left(\frac{364}{365} \right)^{119} \simeq 86.5755 \tag{30}$$

$$\sigma^{(1)}_{n,m} \simeq 6.15626 \tag{31}$$

が得られ、近似解 (9) が十分よい値を与えていることがわかります。

この母関数を使用した方法には様々な応用が考えられます。例えば

$$G^{(k)}(x,y) = z\frac{\partial}{\partial z}e^{y(e^x-\frac{x^k}{k!})+z\frac{x^k}{k!}}\bigg|_{z=y} = y\frac{x^k}{k!}e^{ye^x} \tag{32}$$

という母関数からは $\langle a_k \rangle_{n,m}$ の値を導くことができます。標準偏差についても同様に計算でき、表 1 のようになります。従って、それぞれ 95% の確率 (2 シグマ) で

- 誕生日が誰とも同じでない人が 86.6 ± 12.3 人
- 誕生日が誰か 1 人と同じの人が 28.3 ± 12.0 人
- 誕生日が誰か 2 人と同じの人が 4.59 ± 7.06 人
- 誕生日が誰か 3 人と同じの人が 0.492 ± 2.78 人
- 誕生日が誰か 4 人と同じの人が 0.0392 ± 0.884 人

  ⋮

程度いると推定できます。(k の値が大きいものは正規分布で近似するのはよくないようですが。)

	期待値	標準偏差
a_1	86.5755	6.15626
a_2	14.1518	2.98888
a_3	1.52922	1.17678
a_4	0.122884	0.348123
a_5	0.00783215	0.0884387
a_6	0.000412407	0.0203068
⋮	⋮	⋮

表 1: $n = 365, m = 120$ のときの期待値と標準偏差

また、

$$H^{(k)}(x,y) = e^{y(e^x-\frac{x^k}{k!})+z\frac{x^k}{k!}} \tag{33}$$

という母関数の $z^{a_k}y^{n-a_k}x^m$ 項の係数に注目すれば、図 1 のような a_k の確率分布を求めることができます。このように、母関数を用いる手法は他にも様々な量の計算に応用できると考えられます。

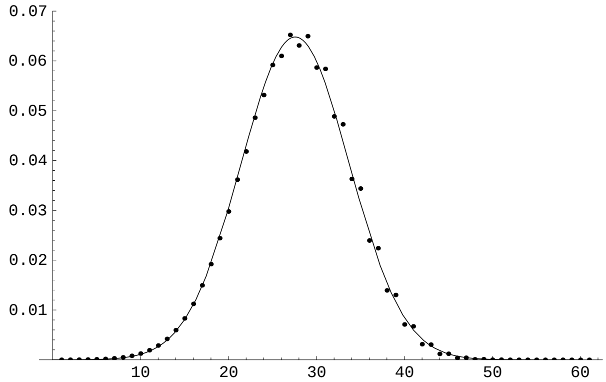

図 1: $n = 365, m = 120$ のときの a_1 の確率分布と正規分布によるフィット ($a_1 = 59$ が原点)