

手動 bot のツイート間隔

hsjoihs

1. 概要

およそ 1 時間おきにツイートする手動 bot を運用していたが、睡眠や体調の都合上実際のツイート間隔には比較的大きなばらつきがあった。しかし、50 件のデータをもとに統計を取りたところ、確率変数「ツイート間隔の対数」が「一様分布に三次関数を適用したもの」として近似できること、その三次関数はツイート間隔を 1 時間の周辺に集中させるような性質を持つことに気づいた。以下、そのことについて解説と考察する。

2. データ

50 件のデータは以下のとおりである。

0:59:42, 6:29:36, 0:54:56, 5:13:22, 1:13:42, 3:24:21, 1:04:07, 0:53:05, 0:59:23, 1:57:27, 1:14:50, 0:45:24, 0:59:06, 1:01:30, 8:54:04, 4:50:01, 0:58:24, 0:51:30, 0:49:54, 1:03:14, 1:28:24, 3:31:16, 3:06:19, 1:20:29, 1:36:08, 11:21:56, 0:41:41, 0:53:58, 2:25:50, 4:24:47, 1:48:54, 0:51:39, 8:33:02, 1:07:54, 1:06:54, 1:08:01, 1:05:13, 3:14:49, 3:01:49, 1:29:07, 3:06:35, 0:59:12, 1:46:46, 14:02:29, 0:52:34, 19:45:05, 1:07:07, 0:51:03, 1:12:10, 1:17:38,

これを 1 時間で割って、ソートすると次のようになる。

0.695, 0.757, 0.832, 0.851, 0.858, 0.861, 0.876, 0.885, 0.899, 0.916, 0.973, 0.985, 0.987, 0.990, 0.995, 1.025, 1.054, 1.069, 1.087, 1.115, 1.119, 1.132, 1.134, 1.203, 1.228, 1.247, 1.294, 1.341, 1.473, 1.485, 1.602, 1.779, 1.815, 1.958, 2.431, 3.030, 3.105, 3.110, 3.247, 3.406, 3.521, 4.413, 4.834, 5.223, 6.493, 8.551, 8.901, 11.366, 14.041, 19.751

ここで、このデータの自然対数が、区間 [0, 1] 上の連続一様分布に三次関数を適用して作られたものであるという仮定をおき、これら 50 個のデータの自然対数にそれぞれ 0.01, 0.03, ...0.99 を対応させ、desmos で curve-fitting をしたところ (<https://www.desmos.com/calculator/rcvlqu>)

jl9i)、求める三次関数は $g(x) = -0.31942 + 1.97413x - 4.66222x^2 + 5.94988x^3$ であることが分かった。なお、 $R^2 = 0.9939$ であった。

3. 三次関数そのものに関する考察

上記の三次関数は $x = 0.26119$ で変曲点を持つ。さらに、 $x = 0.26119$ においてこの三次関数が取る値は -0.01583 であり、非常に 0 に近い。

手動 bot を運用しているときに、「およそ 1 時間おきにツイートする」という意識のもとでツイートを行っていたことを考えると、この結果は自然である。単調増加する三次関数において最も変化率が小さくなるのは変曲点であり、これはつまりツイートの間隔の自然対数が変曲点の近くとなる確率が高いということを意味する。今回はツイート間隔を 1 時間で割ってから自然対数を取っているので、つまりツイート間隔が多くの場合 1 時間程度であるということを意味する。

この考察をもとに、 $g(x)$ の形を $g(x) = a + bx + cx^2 + dx^3$ として 4 つのパラメータで書くのではなく $g(x) = A(x - B)^3 + C(x - B)$ としてパラメータを 3 つにしてみると (<https://www.desmos.com/calculator/nih5fwbofx>)、 $A = 6.17755$, $B = 0.271117$, $C = 0.764937$ と表わせ、 $R^2 = 0.9939$ のままであった。

4. 得られた確率変数に関する考察

この考察で得られた確率変数は、区間 $[0, 1]$ 上の連続一様分布に従う確率変数 K を導入することで $X = e^{A(K-B)^3+C(K-B)}$ と書ける。ただし、 $A > 0$, $C > 0$, $0 < B < 1$ とする。 B は X が 1 以下になる確率を表す。 C は何を表すだろうか？

K が B に非常に近いときは $X \approx e^{C(K-B)}$ と考えることができるため、 $X = 1$ 周りでの累積頻度関数は $y = e^{C(x-B)}$ の逆関数 $y = B + \frac{\ln x}{C}$ で近似でき、故に確率密度関数はその導関数である $f(x) = \frac{1}{Cx}$ と近似できる。したがって、 $x = 1$ 付近での確率密度関数の値の逆数が C であると考えることができる。（後に示すが、実は $x = 1$ においての確率密度関数の値は厳密に $\frac{1}{C}$ である。）

4.1. 台

$A(x - B)^3 + C(x - B)$ も e^x も単調増加するので、台は当然 $[e^{-AB^3-BC}, e^{A(1-B)^3+C(1-B)}]$ で

ある。今回のパラメータの場合、[0.719, 19.100] である。

4.2. 中央値、第1四分位点、第3四分位点

中央値、第1四分位点、第3四分位点は、それぞれ自明に $e^{A\left(\frac{1}{2}-B\right)^3+C\left(\frac{1}{2}-B\right)}$ 、 $e^{A\left(\frac{1}{4}-B\right)^3+C\left(\frac{1}{4}-B\right)}$ 、そして $e^{A\left(\frac{3}{4}-B\right)^3+C\left(\frac{3}{4}-B\right)}$ である。今回の場合、それぞれ 1.283、0.984、2.843 である。

4.3. 期待値

区間 $[-B, 1-B]$ 上の連続一様分布に従う確率変数 L を導入する。

$$E(X) = E(e^{AL^3+CL}) = \int_{-B}^{1-B} e^{Ax^3+Cx} dx$$

厳密解など出るはずもないで、近似する。

$$= \int_{-B}^{1-B} e^{Cx} \sum_{k=0}^{\infty} \frac{(Ax^3)^k}{k!} dx = \sum_{k=0}^{\infty} \frac{A^k}{k!} \int_{-B}^{1-B} e^{Cx} x^{3k} dx$$

ここで $\int e^{Cx} x^{3k} dx = \frac{e^{Cx}}{C^{3k+1}} \sum_{l=0}^{3k} (-1)^{3k-l} (Cx)^l \frac{(3k)!}{l!}$ はい、使い物になりませんね。おしまい。

もうこれシンプソンの公式でいいんじゃないかな、上で第1四分位点とか出したんだし。 $G(x) = e^{A(x-B)^3+C(x-B)}$ とおけば、

$$E(X) = \int_0^1 G(x) dx \approx \frac{1}{4 \cdot 3} (G(0) + 4G(0.25) + 2G(0.5) + 4G(0.75) + G(1))$$

今回のパラメータだと誤差は 10% ぐらい。十分許容範囲では？

5. 累積分布関数と確率密度関数

5.1. 累積分布関数と確率密度関数を求める

$G(x) = e^{A(x-B)^3+C(x-B)} (0 < x < 1)$ の逆関数が累積分布関数である。つまり、 $x = e^{A(y-B)^3+C(y-B)}$ な

ので $\ln x = A(y - B)^3 + C(y - B)$ である。ここで $k = \sqrt{\frac{3A}{4C}}$ とおくと

$$\frac{3k}{C} \ln x = 4k^3(y - B)^3 + 3k(y - B)$$

ここで $k(y - B) = \sinh \lambda$ とおくと $\frac{3k}{C} \ln x = 4 \sinh^3 \lambda + 3 \sinh \lambda = \sinh 3\lambda$ であり、これ
は $\frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3} = \lambda$ と書ける。

ゆえに、累積分布関数は $y = F(x) = B + \frac{1}{k} \sinh \frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3}$ ($G(0) < x < G(1)$) であ
り、確率密度関数はこれを微分して

$$\begin{aligned} f(x) &= \frac{1}{k} \frac{d}{dx} \sinh \frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3} = \frac{1}{k} \cosh \frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3} \cdot \frac{d}{dx} \frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3} \\ &= \frac{1}{3k} \cosh \frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3} \cdot \frac{1}{\sqrt{1 + \left(\frac{3k}{C} \ln x\right)^2}} \cdot \frac{d}{dx} \frac{3k}{C} \ln x \\ &= \frac{1}{Cx} \cosh \frac{\operatorname{arsinh}\left(\frac{3k}{C} \ln x\right)}{3} \cdot \frac{1}{\sqrt{1 + \left(\frac{3k}{C} \ln x\right)^2}} \end{aligned}$$

つまり、 $f(x)$ の式の形の中に露わには B が現れない。(B は定義域を定める役割のみを果たす。)

以後、 $D = \frac{3k}{C}$ なるパラメータを導入することとする。今回の例では $D = 9.652$ である。

5.2. パラメータと関数の分離

以上の結果の一部を <https://www.desmos.com/calculator/lmw7kroctq> にまとめた。図示する

と、 $(G(0), f(G(0)))$ から始まり、 $(1, f(1))$ 付近にあるピークまで下凸で登った後、一気に下凸で下っていく様子が見て取れる。この挙動を捉えたより単純な関数で確率密度関数を近似することを考える。

まず、 $H(u) = \cosh \frac{\text{arsinh}(u)}{3} \cdot \frac{1}{\sqrt{1+u^2}}$ とおくと $f(x) = \frac{H(D \ln x)}{Cx}$ と書ける。 $h(x) = H(D \ln x)$ とおくと h は $x = 1$ で極値 1 を取る。なぜなら、まず $h(1) = \cosh \frac{\text{arsinh}(0)}{3} \cdot \frac{1}{\sqrt{1+0^2}} = 1$ であり、なおかつ

$$\begin{aligned} \frac{dh}{dx} &= \frac{dH}{du} \cdot \frac{d(D \ln x)}{dx} \\ &= \frac{D}{x} \left(\left(\frac{d}{du} \cosh \frac{\text{arsinh} u}{3} \right) \cdot \frac{1}{\sqrt{1+u^2}} + \cosh \frac{\text{arsinh} u}{3} \cdot \frac{d}{du} \left(\frac{1}{\sqrt{1+u^2}} \right) \right) \\ &= \frac{D}{x} \left(\frac{\sinh \left(\frac{1}{3} \text{arsinh} u \right)}{3 \sqrt{u^2 + 1}} \cdot \frac{1}{\sqrt{1+u^2}} + \cosh \frac{\text{arsinh} u}{3} \cdot -\frac{u}{(1+u^2)^{3/2}} \right) \end{aligned}$$

は $u = D \ln x = 0$ で 0 になるからである。

さて、以上より、 $H(u) = \cosh \frac{\text{arsinh}(u)}{3} \cdot \frac{1}{\sqrt{1+u^2}}$ とおくと $f(x) = \frac{H(D \ln x)}{Cx}$ と書け、なおかつ $H(u)$ は $u = 0$ で極値 1 を取ることが分かる。

ここまで結果をまとめ、 B, C, D, H のみで書き直すと、次のようになる。

- $G(x) = e^{\frac{4C^3 D^2}{27}(x-B)^3 + C(x-B)}$ の逆関数 $F(x) = B + \frac{3}{CD} \sinh \frac{\text{arsinh}(D \ln x)}{3}$ ($G(0) < x < G(1)$) が累積分布関数である
- 確率密度関数はその導関数である $f(x) = \frac{H(D \ln x)}{Cx}$ ($G(0) < x < G(1)$)
- ただし、 $H(u) = \cosh \frac{\text{arsinh}(u)}{3} \cdot \frac{1}{\sqrt{1+u^2}}$

- 特に、 $f(1) = \frac{H(0)}{C} = \frac{1}{C}$ であるので、パラメータ C は $\frac{1}{f(1)}$ という意味合いを持った量である。
- D は、大きくすると $G(1)$ を大きくさせ、また $f(x)$ を $x = 1$ の辺りに尖らせる役割を持つ。
- B は、変わらず X が 1 以下になる確率を表す。これは $F(1) = B + \frac{3}{CD} \sinh \frac{\operatorname{arsinh}(0)}{3} = B$ という式でも表現される。

これにより、パラメータの方は説明をつけることができたので、あとはパラメータを含まないこの謎の関数 $H(u)$ の性質を調べればよい。

6. $H(u)$ の性質

6.1. 近似

$\operatorname{arsinh}(u) = \ln(u + \sqrt{u^2 + 1})$ だが、これを $\ln(2\sqrt{u^2 + 1})$ で近似しても $H(u)$ の概形は $u = 0$ の周り以外ではほとんど変わらず、 $u = 0$ の周りでも 2% ほどしか変わらない。

さて、その近似をした $H'(u)$ は $J(v) = v^{2/3} + v^{4/3}$ を用いて $H'(u) = J\left(\frac{1}{2\sqrt{1+u^2}}\right)$ と書ける。何故かと言うと、

$$\begin{aligned} H'(u) &= \frac{\exp\left(\frac{\ln(2\sqrt{u^2 + 1})}{3}\right) + \exp\left(-\frac{\ln(2\sqrt{u^2 + 1})}{3}\right)}{2} \cdot \frac{1}{\sqrt{1+u^2}} \\ &= \left(\exp\left(\frac{\ln(v^{-1})}{3}\right) + \exp\left(-\frac{\ln(v^{-1})}{3}\right) \right) \cdot v \\ &= \left(\exp\left(\ln(v^{-1/3})\right) + \exp\left(\ln(v^{1/3})\right) \right) \cdot v = J(v) \end{aligned}$$

と書けるからである。つまり、

$$H(u) \approx \left(\frac{1}{4(1+u^2)} \right)^{1/3} + \left(\frac{1}{4(1+u^2)} \right)^{2/3}$$

6.2. 別方面

$u \rightarrow \infty$ のとき $\text{arsinh}(u) \approx \ln 2u$ なので、 $\cosh \frac{\text{arsinh}(u)}{3} \approx \frac{u^{1/3}}{2^{2/3}}$ である。故に、 $H(u) \approx (2u)^{-2/3}$ である。したがって、 $(H(u))^3 \approx (2u)^{-2}$ であることが分かる。

0 付近の挙動を調べるため、 $(H(u))^{-3}$ をテイラーフレーム展開すると、

$$(H(u))^{-3} = 1 + \frac{4u^2}{3} + \frac{16u^4}{81} - \frac{128u^6}{2187} + \frac{512u^8}{19683} - \dots$$

とまあ、なんとも興味深い形が現れる。（念のために言っておくと、この次の項でこのパターンは崩れ、2 のべき乗でも 3 のべき乗でもない数が現れる。）

$v = \frac{4u^2}{9}$ とすると、

$$(H(u))^{-3} = 1 + 3v + v^2 - \frac{2}{3}v^3 + \frac{2}{3}v^4 - \dots$$

一方、これが $u \rightarrow \infty$ のときは $(H(u))^{-3} \approx 4u^2 \dots$ とはいえ、これはあくまで漸近的な評価にすぎず、かなり悪い近似である。

もっと精密に評価しよう。 $k = \left(\frac{1}{4(1+u^2)} \right)^{1/3}$ としてやると $H(u) \approx k + k^2$ なので

$$(H(u))^{-3} \approx (k^3 + 3k^4 + 3k^5 + k^6)^{-1}$$

だが、実際に近似してみると k^6 を抜いたほうがむしろ精度が上がる気がする。（最初に「 $\ln(2\sqrt{u^2 + 1})$ で近似」で導入された誤差を打ち消すように働く）

具体的に係数をどうすべきか考えてみよう。

$$(H(u))^{-3} = 1 + 3v + v^2 - \frac{2}{3}v^3 + \frac{2}{3}v^4 - \dots$$

と

$$k^{-3}(H(u))^3 \approx 1 + 3k + 3k^2 + k^3$$

一方 $k = \left(\frac{1}{4+9v}\right)^{1/3}$ より $k^{-3} = 4 + 9v$ である。前述のテイラーライフ展開の逆数を取ると

$$(H(u))^3 = 1 - 3v + 8v^2 - \frac{61}{3}v^3 + \frac{151}{3}v^4 - \dots$$

なので

$$k^{-3}(H(u))^3 = 4 - 3v + 5v^2 - \frac{28}{3}v^3 + \frac{55}{3}v^4 - \dots$$

これが

$$k^{-3}(H(u))^3 \approx a + bk + ck^2 + dk^3$$

となるべく一致してほしい。まずは $d = 0$ で考える。 $v = 0$ としたときの値を比較すると

$$4 = a + b\left(\frac{1}{4}\right)^{1/3} + c\left(\frac{1}{4}\right)^{2/3}$$

導関数の $v = 0$ のときの値を比較すると

$$-3 = b \frac{d(4+9v)^{-1/3}}{dv} + c \frac{d(4+9v)^{-2/3}}{dv} = -\frac{b}{3} \cdot 9 \cdot (4+9v)^{-4/3} - \frac{2c}{3} \cdot 9 \cdot (4+9v)^{-5/3}$$

一旦 $c = 0$ とすると $1 = b \cdot (4)^{-4/3}$ と $4 = a + b\left(\frac{1}{4}\right)^{1/3}$ より $b = 4^{4/3}$, $a = 4 - 4^{4/3}\left(\frac{1}{4}\right)^{1/3} = 0$ 、ふーむ。

グラフ書いてみたんですが、えーなんかめっちゃ合いますね、完璧では。

$$H(u) \approx 4^{4/9}k^{4/3} = \left(\frac{4}{4+9v}\right)^{4/9} = \left(\frac{1}{1+u^2}\right)^{4/9}$$

1 から遠いところではズレないわけではないけどまあ仕方がない。

2 階微分まで使うとか、漸近評価も使うとかすればいいんだろうけど。

7. 近似した $H(u)$ で考察

7.1. 累積分布関数

確率密度関数が $f(x) = \frac{H(D \ln x)}{Cx}$ 、 $H(u) = \left(\frac{1}{1+u^2}\right)^{4/9}$ である場合、

$$F(x') = \int_a^{x'} \frac{H(D \ln x)}{Cx} dx = \int_{D \ln a}^{D \ln x'} \frac{H(q)}{CD} dq = \int_{D \ln a}^{D \ln x'} \frac{(1+q^2)^{-4/9}}{CD} dq$$

... 初等関数で積分できないなこれ。一体なんのための近似だったのやら。

7.2. 期待値

期待値なら x が入るから積分できるのでは？

と一瞬期待してしまったが、 $\ln x$ が入るので普通に無理である。

8. そもそも三次関数を使ったのが良くなかったのでは？
