負の二項分布とガンマ分布の関係

プログラマたん bot

2019/7/28

負の二項分布

複数の異なる定義があるので、順に説明する。

Wikipediaに従い、負の二項分布を、「1 回の試行に確率 p で成功するとき、r 回目の失敗までの、成功回数 y の分布」と定義する。この分布は

$$NegativeBinomial(y|r,p) = \binom{y+r-1}{y}(1-p)^r p^y \tag{式 1}$$

これは r 回目に失敗するまでに、y 回成功し $r-1$ 回失敗する試行の組み合わせと、その確率である。

「高校数学の美しい物語」の定義は、成功回数 y を試行回数 $z := y + r$ に置き換えて、

$$\binom{y+r-1}{x} = \binom{y+r-1}{r-1}$$

であることを用い、

$$NegativeBinomial(z|r,p) = \binom{z-1}{r-1}(1-p)^r p^{z-r} \tag{式 2}$$

としたものである。

“Stan Functions Reference” による、別の定義を与える。

$$NegativeBinomial(y|\alpha,\beta) = \binom{y+\alpha-1}{\alpha-1}(\frac{\beta}{1+\beta})^\alpha(\frac{1}{1+\beta})^y \tag{式 3}$$

R の dnbinom 関数の説明によると、

$$NegativeBinomial(y|size,prob) = \frac{\Gamma(y+size)}{\Gamma(size)y!}p^{size}(1-p)^y \tag{式 4}$$

である。これは階乗をガンマ関数におきかえ、式 3 を以下のように置き換えることで得られる。

$$\Gamma(k+1) = k!, size := \alpha, \beta := 1/p - 1 \tag{式 5}$$

ガンマ分布

Wikipediaによると、ガンマ分布の確率密度関数は

$$GammaDistribution(x|shape, rate) = \frac{rate^{shape}}{\Gamma(shape)} x^{shape-1} e^{-rate*x} \quad (式\ 6)$$

である。慣習的に、$\alpha := shape$, $\beta := rate$, $\theta := scale = 1/\beta$ と表記する。平均 $\mathbb{E}[x]$ と分散 $Var[x]$ は、以下の通りである。

$$\mathbb{E}[x] = \frac{\alpha}{\beta} = \alpha\theta, Var[x] = \frac{\alpha}{\beta^2} = \frac{\mathbb{E}[x]}{\beta} = \alpha\theta^2 = \mathbb{E}[x]\theta \quad (式\ 7)$$

平均と分散が既知なら (観測値があれば)、負の二項分布のパラメータ α, β を求めることができる。

$$\beta = \frac{\mathbb{E}[x]}{Var[x]}, \alpha = \mathbb{E}[x]\beta = \frac{\mathbb{E}[x]}{\theta} \quad (式\ 8)$$

ガンマ-ポアソン分布

観測値が過分散であるときに、標本には個体差があり、応答変数は個体差に従った分布である、というモデルを適用することがある。例えば発芽率は、n 個の種の発芽率 $p_i, i \in 1..n$ がある分布に従い、それぞれの種は確率 p_i で発芽する、と仮定することである。詳しくは、“データ解析のための統計モデリング入門一般化線形モデル・階層ベイズモデル・MCMC (確率と情報の科学)”, 久保拓弥著, 岩波書店, 2012/5 を参照すること。

ここで個体差をガンマ分布で表現し、それぞれの個体差を説明変数とするポアソン分布にした値が観測される、というモデルを考える。

$$y \sim PossionDistribution(y), x \sim GammaDistribution(shape, rate) \quad (式\ 9)$$

ガンマ-ポアソン分布を一まとめにして、負の二項分布で表現することができる。証明はStackExchangeにあるが、長いのでここには載せない。

平均と分散

これまでに挙げた文献から、平均と分散を引用する。

NB(α, β) の平均と分散

負の二項分布の定義 式 3 において、平均 $\mathbb{E}[y]$ と分散 $Var[y]$ は、以下の通りである。

$$\mathbb{E}[y] = \frac{\alpha}{\beta}, Var[y] = \frac{\alpha}{\beta^2}(\beta + 1) = \mathbb{E}[y](1 + \frac{1}{\beta}) \tag{式 10}$$

であり、平均と分散が既知なら (観測値があれば)、負の二項分布のパラメータ α, β を求めることができる。

$$\beta = \frac{\mathbb{E}[y]}{Var[y] - \mathbb{E}[y]}, \alpha = \mathbb{E}[y] * \beta \tag{式 11}$$

負の二項分布の代わりに、ガンマ-ポアソン分布 式 9 におけるガンマ分布 式 6 のパラメータで表現することを考える。ポアソン分布 $y \sim poisson(\lambda)$ の平均 $\mathbb{E}[y]$ と分散 $Var[y]$ は

$$\mathbb{E}[y] = \lambda, Var[y] = \lambda \tag{式 12}$$

であり、StackExchangeによると、独立な期待値と分散の加法性 (共分散が 0) から、x がガンマ分布 $Gamma(\alpha_{gamma}, \beta_{gamma})$ に従うとき、

$$\mathbb{E}[y] = \lambda = \mathbb{E}[x], Var[y] = Var[Gamma(\alpha_{gamma}, \beta_{gamma})] + \lambda \tag{式 13}$$

$$\mathbb{E}[x] = \mathbb{E}[Gamma(\alpha_{gamma}, \beta_{gamma})] = \frac{\alpha_{gamma}}{\beta_{gamma}} \tag{式 14}$$

$$Var[x] = \mathbb{E}[x] + Var[Gamma(\alpha_{gamma}, \beta_{gamma}] = \frac{\alpha_{gamma}}{\beta_{gamma}}(1 + \frac{1}{\beta_{gamma}}) = \mathbb{E}[x](1 + \frac{1}{\beta_{gamma}}) \tag{式 15}$$

である。つまり、$\alpha = \alpha_{gamma}, \beta = \beta_{gamma}$ が成り立つ。

NB(size, prob) の平均と分散

負の二項分布の定義 式 4 において、平均 $\mathbb{E}[y]$ と分散 $Var[y]$ は、以下の通りである。

$$\mathbb{E}[y] = size\frac{1 - prob}{prob}, Var[y] = size\frac{1 - prob}{prob^2} = \frac{\mathbb{E}[y]}{prob} \tag{式 16}$$

であり、平均と分散が既知なら (観測値があれば)、負の二項分布のパラメータ $size, prob$ を求めることができる。

$$prob = \frac{\mathbb{E}[y]}{Var[y]}, size = Var[y] * \frac{prob^2}{1 - prob} = \frac{\mathbb{E}[y]^2}{Var[y] - \mathbb{E}[y]} \tag{式 17}$$

負の二項分布の代わりに、ガンマ-ポアソン分布 式 9 におけるガンマ分布 式 6 のパラメータで表現することを考える。先ほどとは逆に、x がガンマ分布 $Gamma(\alpha_{gamma}, \beta_{gamma})$ に従うとしてボトムアップに考える。

$$\beta_{gamma} = \frac{Var(x)}{\mathbb{E}(x)} = \frac{Var(y) - \mathbb{E}(x)}{\mathbb{E}(x)} = \frac{Var(y) - \mathbb{E}(y)}{\mathbb{E}(y)} = \frac{1}{p} - 1 \tag{式 18}$$

$$\alpha_{gamma} = \frac{\mathbb{E}[x]}{\beta} = \frac{\mathbb{E}[y]}{\beta} = size \tag{式 19}$$

つまり、$\alpha_{gamma} = size, \beta_{gamma} = 1/p - 1$ が成り立つ。これは式 3と式 4の関係であった。

再生性

NB(size, prob) の再生性

負の二項分布 $NB(size, prob)$ は $size$ に対して再生性がある。つまり

$$NB(size_A + size_B, prob) = NB(size_A, prob) + NB(size_B, prob) \tag{式 20}$$

$$NB(y|size_A + size_B, prob) = \sum_{i=0}^{y} NB(i|size_A, prob) * NB(y - i|size_B, prob) \tag{式 21}$$

負の二項分布のモーメント母関数

$$M_x(t) = (\frac{p}{1 - (1 - p)e^t})^r \tag{式 22}$$

の定義から明らかと言えるが、ここではガンマ関数の引数が整数に限るつまり階乗に置き換えられるものとして、sum 記号の中を幾何的に解く。

言葉で書くと、r 回目に失敗するまでに、y 回成功するのは、i 回成功してから $y - i$ 成功することと同じ、という意味である。直感的な証明は以下の通りである。式 4 より、

$$NB(y|size_A + size_B, prob) = \sum_{i=0}^{y} \frac{\Gamma(i + size_A)}{\Gamma(size_A)i!} p^{size_A}(1 - p)^i \frac{\Gamma(y - i + size_B)}{\Gamma(size_B)(y - i)!} p^{size_B}(1 - p)^{y-i} \tag{式 23}$$

$$= p^{size_A + size_B}(1 - p)^y \sum_{i=0}^{y} \frac{\Gamma(i + size_A)}{\Gamma(size_A)i!} \frac{\Gamma(y - i + size)}{\Gamma(size)(y - i)!} \tag{式 24}$$

碁盤目の状の街があり、縦 m 区間、横 n 区間あるとする。このとき待ちの左上の端から右下の端まで、右または下方向に移動することを考える。つまり縦、縦、横、縦、... と移動する。このとき左上の端から右下の端までの経路は、$(m+n)!/(m!n!)$ 通りある。

なぜなら縦同士、横同士の移動を移動時の座標に基づいて、縦 1、縦 2、横 1、縦 3、と区別したときの経路数は $(m+n)!$ であるが、実は縦同士、横同士の移動を区別しないので、$m!n!$ で割ると、上記の通りになるからである。これは m+n 個から m 個取り出す組み合わせの数である。

ここで縦 y 区間、横 $size_A + size_B - 1$ 区間ある街を考える。ここで、

1. 縦 i 区間、横 $size_A - 1$ 区間移動して
2. 横に 1 区間移動して
3. 縦 $y - i$ 区間、横 $size_B - 1$ 区間移動する

経路の数を考え、取りうるすべての $i \in 0..y$ について足す。これは i 回成功してから $y - i$ 成功する経路を漏れなく重複なく数えることと同じである。よって、

$$\binom{y + size_A + size_B - 1}{y} = \sum_{i=0}^{y} \binom{y + size_A - 1}{y} \binom{y + size_B - 1}{y} \quad \text{(式 25)}$$

である。

ガンマ分布の再生性

ガンマ分布 $GammaDistribution(\alpha, \beta)$ は $\alpha = shape$ に対して再生性がある。つまり

$$x_A \sim GammaDistribution(\alpha_A, \beta), x_B \sim GammaDistribution(\alpha_B, \beta) \quad \text{(式 26)}$$

なら

$$x_A + x_B \sim GammaDistribution(\alpha_A + \alpha_B, \beta) \quad \text{(式 27)}$$

である。

ガンマ分布のモーメント母関数

$$M_x(t) = (1 + t/\beta)^{\alpha} \quad \text{(式 28)}$$

の定義から明らかと言えるが、証明はこちら。

NB(size, prob) の再生性を確認する

再生性の定義より、

$$\sum_{i=1}^{n} NB(size, prob) = NB(size * n, prob) or \sum_{i=1}^{n} NB(size/n, prob) = NB(sizeprob) \qquad (式\ 29)$$

である。これは n 人がそれぞれ size 回当たるまでくじを引く ($NB(size, prob)$) ときの合計試行回数と、一人で size*n 回当たるまでくじを引く ($NB(size * n, prob)$) ときの試行回数が、同じ分布に従うということである。n 人が同一人物と考えれば、そうなるのが自然である。

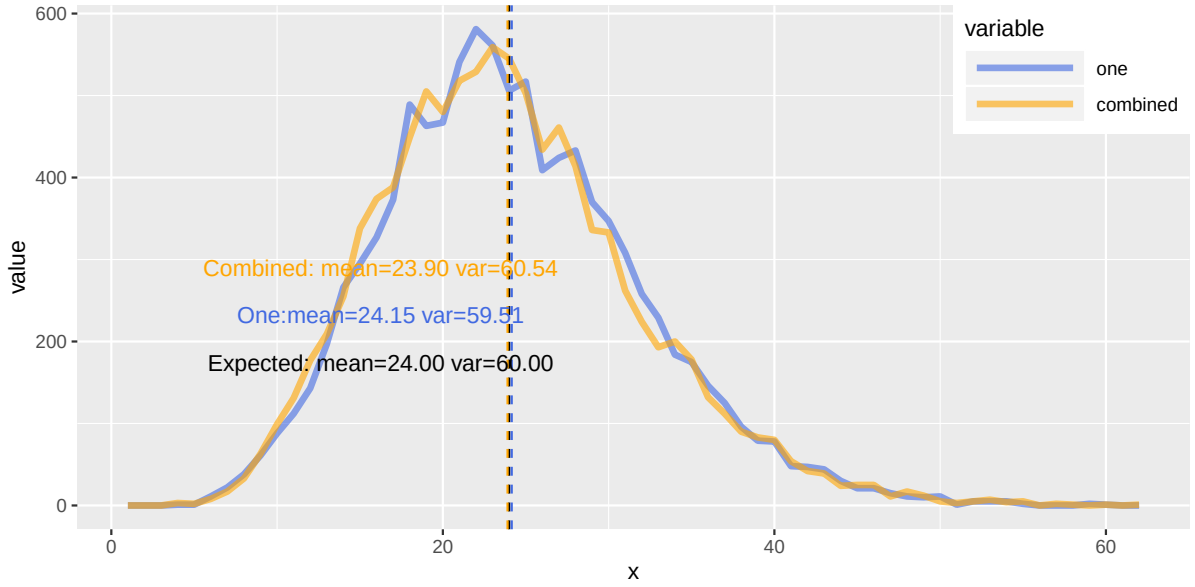

図1　negative binomial(size=16.0, prob=0.4) alpha combined

図3は size=16, prob=0.4 の負の二項分布に基づいて生成した乱数のヒストグラムである。横軸は x、縦軸は各 x の出現回数である。Expected は式 17による平均と分散、one は単一の負の二項分布、combined は size を 16 分割した複数の負の二項分布の和である。

式 7より、ガンマ分布の期待値は $\alpha := shape$ にも $\theta := scale = 1/\beta$ にも比例するが、式 10よりガンマ-ポアソン分布の期待値と分散は α に比例しても、分散は $\theta = 1/\beta$ に比例しない。なので、θ を n 分割した負の二項分布の和は、元の負の二項分布とは異なる (図4)。

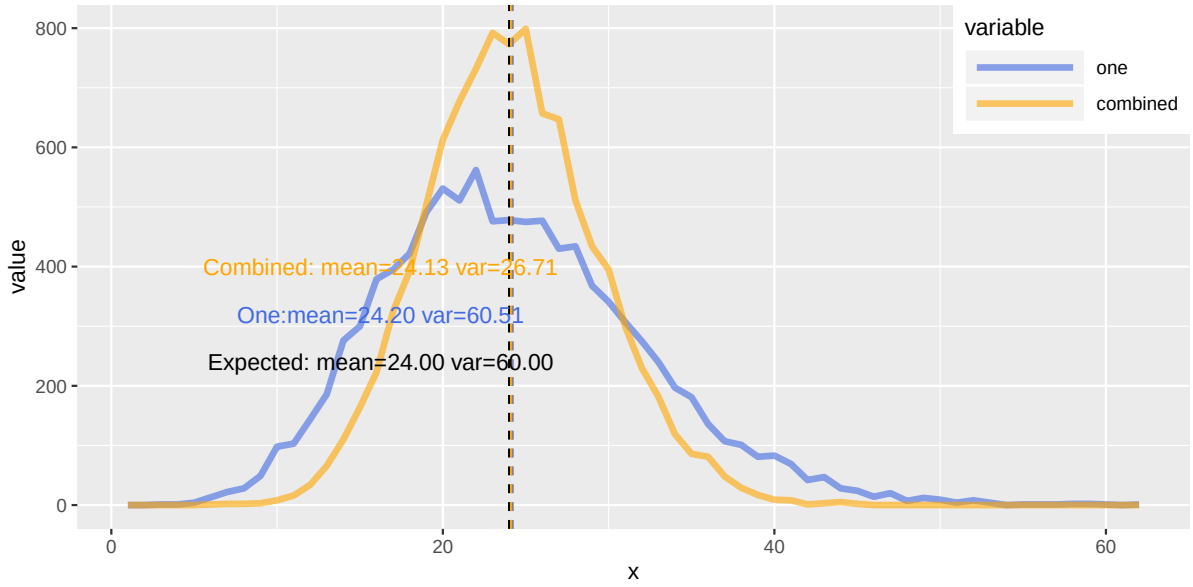

図2　negative binomial(size=16.0, prob=0.4) theta combined

Gamma(α, θ) の再生性を確認する

負の二項分布と同様、$\alpha = size$ の分割は再生性がある。

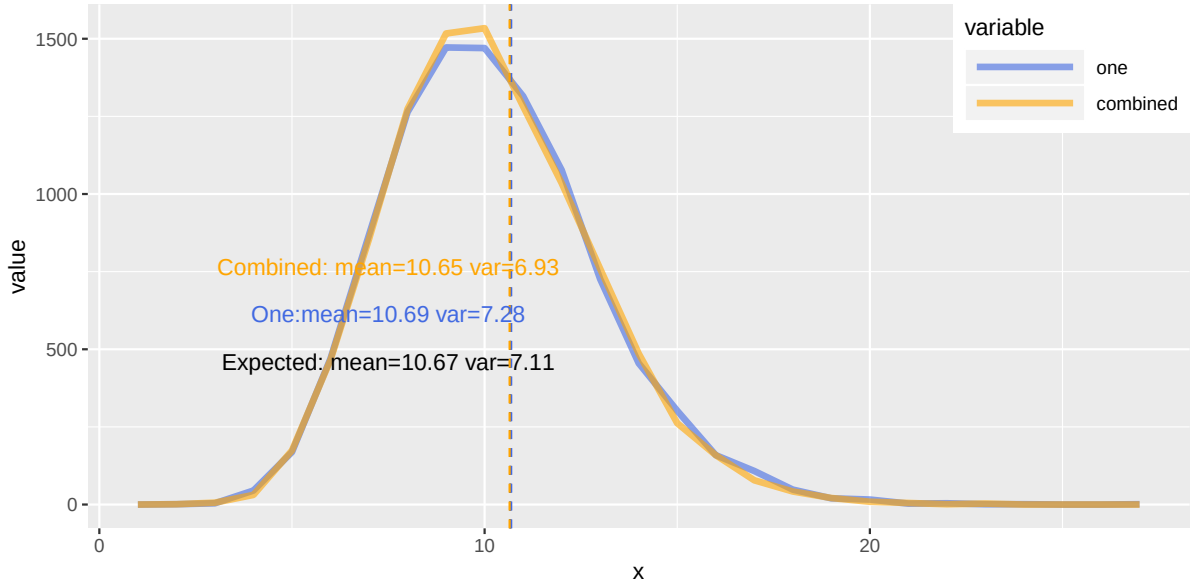

図3　Gamma distribution alpha combined

やはり負の二項分布と同様、θ の分割は再生性がない (期待値は同じだが分散は異なる)。

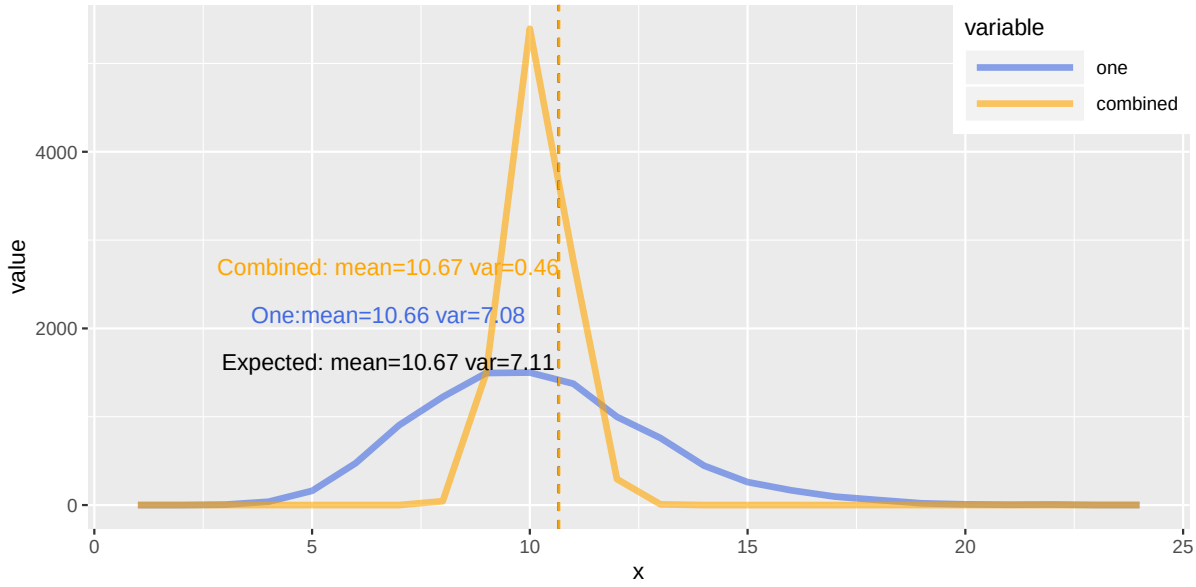

図4　Gamma distribution theta combined