

統計学 2020 Lecture 4: ポアソン分布と統計的推測

北門 利英（東京海洋大学海洋生物資源学科）

2020 年 6 月 3 日

Attention:

- 授業 HP に授業に関する情報をアップデートしていきますので参照ください. (URL: <https://toshihidekitakado.github.io/STAT2020/index.html>)
- 今回の授業に関してわからないことがあれば、メールカリアルタイム接続時に遠慮なく質問してください (6 月 3 日は 13:00-14:00 とします).

Point:

- ポアソン分布とは何か説明できるようになること
- ポアソン分布の期待値と分散の式を自身で導けるようにすること
- ポアソン分布の確率分布を利用して様々な計算ができること
- ポアソン分布におけるパラメータの推定値と標準誤差の計算ができること

1 ポアソン分布とは

これまで、確率変数・確率分布には、離散型と連続型の 2 種類があること、期待値や分散などの特性値が確率分布のパラメータで規定されていること、そして今回は離散型確率分布のうち 2 項分布に焦点を当て、その定義と性質の他、パラメータの推定について学びました. 今回は離散型確率分布のうち 2 項分布と並んで最重要であるポアソン分布について学びます.

1.1 ポアソン分布の定義

ポアソン分布は 2 項分布と並んでよく利用される離散型分布のひとつであるが、大きな違いは、2 項分布は N 中に何回というように値の下限 (0) と上限 (N) が決まっているのに対して、ポアソン分布では 0 以上の値をとることは 2 項分布と同様であるが上限が特にない. この違いは、例えば次のような観測手続きの違いを考えるとわかりやすいかもしれない. ある生物のサンプリングにおいて、2 項分布では「サンプリングした N 個体のうち雄は何個体であるか」という内訳の観測となるが、ポアソン分布では「サンプリングされる個体数は何個体であるか」あるいは「サンプリングされる雄個体は何個体であるか」というようにそもそも何個体獲れるのか、というように興味の対象が異なる.

あるいは別の例を考えると、ある浜に定点カメラを設置したとする. この定点カメラの撮影範囲を 24 時間以

内の中に訪れるウミガメの個体数を確率変数と考える場合、その数は0かもしれないし、5, 10, 28, 46のように大きな値をとるかもしれない。もちろん1000個体訪れることなどないと考えれば、それを上限としてもよいかもしれないが、では1001や2000ではだめなのか、等の疑問には答えられない。したがって、このような場合には2項分布のように取りうる値の上限を定めてしまうよりも、上限を与えないで、その代わりに「大きな値は確率的にとりにくい」と仮定する方が論理的にも数学的にも取り扱いやすい。

それではまずポアソン分布を数学的に定義する。

離散型分布 2 [ポアソン分布] 確率変数 Y が確率関数

$$P(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!}, \quad y = 0, 1, 2, \dots \quad (\lambda > 0) \quad (1)$$

をもつとき、 Y はポアソン分布 (Poisson distribution) $Pois(\lambda)$ にしたがうという。

突然 λ (ラムダ) という見慣れない記号が出てきたが、この λ がポアソン分布を規定している重要なパラメータであり、この値によって確率変数 Y の確率分布が変わり、したがって観測される確率変数の実現値のメカニズムをコントロールしていることになる。すなわち、図1のように λ の値が2のように小さい場合には Y の値は0という値も含めて小さな値が出やすく、 λ が5のように2よりも大きいと相対的に大きな値を取りやすい (0 をとる確率もほとんどゼロになる。 $P(Y = 0) = e^{-\lambda}$ であるから、 λ に関して指数的に減少する)。逆に言うと、観測値の値が大きければその背後にある λ の値は大きいのだろう、などと推測をするのが統計的推測であり、それは観測値を基に観測の背後のメカニズムを暴こうという科学的な推論の一つである。

```
par(mfrow = c(1, 2), mar = c(3, 4, 4, 4))
y <- seq(0, 20, 1)
plot(y, dpois(y, 2), "h", ylim = c(0, 0.3), ylab="Prob", main = "(a) lambda=2")
plot(y, dpois(y, 5), "h", ylim = c(0, 0.3), ylab="Prob", main = "(b) lambda=5")
```

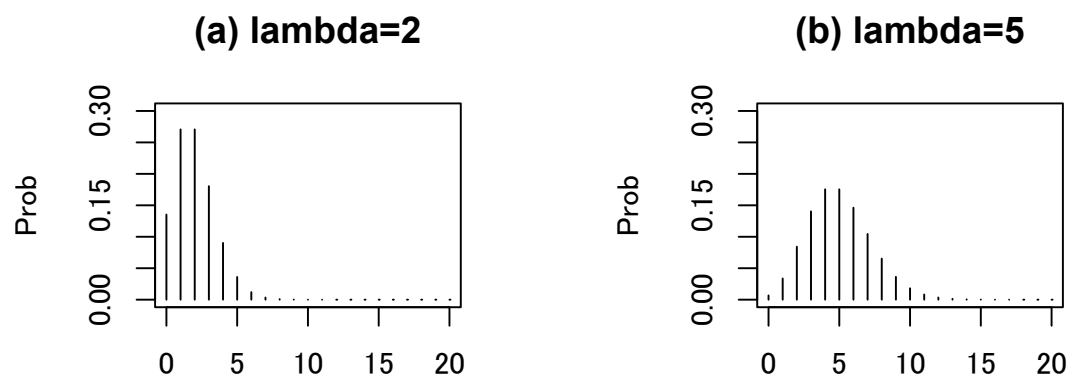


図1 ポアソン分布の確率分布の例 ((a) $\lambda = 2$ および (b) $\lambda = 5$ のとき)

☞ ポアソン分布の起源や導出は大変重要で、その中でも時間や空間分布と絡めた理論的背景は興味深いが、2年生の統計学の授業としてはやや高度となるため、ここでは割愛します。

1.2 ポアソン分布の性質

1.2.1 確率の総和が1であることの証明とマクローリン展開

全事象の確率の総和は1であるから、ポアソン分布の確率 $P(Y = y)$ の0から ∞ までの和は1である。このことは、 e^λ のマクローリン展開 (原点周りの Taylor 展開) を利用して示すことができる。すなわち、

$$e^\lambda = 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \dots = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}$$

$$\sum_{y=0}^{\infty} P(Y = y) = \sum_{y=0}^{\infty} e^{-\lambda} \frac{\lambda^y}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} e^\lambda = 1$$

となる。これと同様の方法で、次の $E[Y] = V[Y] = \lambda$ も導くことができる。

☞ 突然マクローリン展開が出てきて驚いたかもしれませんが、微分積分や解析学の基礎を少し身につけていると、他の分野でも役に立つ一つの例とってください。

1.2.2 期待値

期待値は以下のように求められる。

$$\begin{aligned} E[Y] &= \sum_{y=0}^{\infty} y \cdot P(Y = y) \\ &= \sum_{y=1}^{\infty} y \cdot e^{-\lambda} \frac{\lambda^y}{y!} \quad (y=0 \text{ のとき } y \cdot P(Y = y) = 0 \text{ なので}) \\ &= \lambda \cdot e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!} \\ &= \lambda \cdot e^{-\lambda} \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \quad (z = y-1 \text{ とおくと } z = 0, 1, \dots) \\ &= \lambda \cdot e^{-\lambda} \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \\ &= \lambda \cdot e^{-\lambda} \cdot e^\lambda = \lambda \end{aligned}$$

1.2.3 分散

Lecture 3 の 2 項分布の際にも述べたが、便利な分散の公式があり、

$$\begin{aligned}
V[Y] &= E[(Y - E[Y])^2] \\
&= E[Y^2 - 2YE[Y] + E[Y]^2] \\
&= E[Y^2] - 2E[Y] \cdot E[Y] + E[Y]^2 \\
&= E[Y(Y-1) + Y] - E[Y]^2 \\
&= E[Y(Y-1)] + E[Y] - E[Y]^2
\end{aligned}$$

が成り立つから、 $E[Y]$ の場合とほぼ同様な計算から $E[Y(Y-1)]$ が次のように求まる。

$$\begin{aligned}
E[Y(Y-1)] &= \sum_{y=0}^{\infty} y(y-1) \cdot P(Y=y) \\
&= \sum_{y=2}^{\infty} y(y-1) \cdot e^{-\lambda} \frac{\lambda^y}{y!} \quad (y=0,1 \text{ のとき } y(y-1) \cdot P(Y=y) = 0 \text{ なので}) \\
&= \lambda^2 \cdot e^{-\lambda} \sum_{y=2}^{\infty} \frac{\lambda^{y-2}}{(y-2)!} \\
&= \lambda^2 \cdot e^{-\lambda} \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \quad (z = y-2 \text{ とおくと } z = 0, 1, \dots) \\
&= \lambda^2 \cdot e^{-\lambda} \cdot e^{\lambda} = \lambda^2
\end{aligned}$$

であることが分かり、したがって分散は

$$V[Y] = E[Y(Y-1)] + E[Y] - E[Y]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

となる。したがって、ポアソン分布では期待値と分散が等しく、これはポアソン分布の一つの特徴でもある。

1.2.4 ポアソン分布の再生性

ポアソン分布については以下のような再生性が成り立つ。

定理 1 [ポアソン分布の再生性] 確率変数 Y_i ($i = 1, 2, \dots, m$) が独立に $Po(\lambda_i)$ にしたがうとき、 $T = \sum_{i=1}^m Y_i$

はポアソン分布 $Po(\sum_{i=1}^m \lambda_i)$ にしたがう。

この性質は例えば次のような例を考えるとわかりやすい。 m 箇所の漁場である生物の生息数を観測した。ただし、漁場の面積がそれぞれ異なり、生息数の期待値がその面積に依存すると考えると、漁場 i の生息数を表す確率変数 Y_i は異なるパラメータ λ_i をもつポアソン分布にしたがうと考える。再生性はこの確率分布の合計 $\sum_{i=1}^m Y_i$ もまたポアソン分布にしたがい、そのパラメータも $\sum_{i=1}^m \lambda_i$ となる。証明の方法は、畳み込み計算や積率母関数を用いる方法などいくつかあるが、ここでは割愛する。

1.3 ポアソン分布と 2 項分布の関係

1.3.1 2 項分布の極限分布としてのポアソン分布

ポアソン分布が導出される過程はいくつかあり、したがってこの分布が適合する確率的な現象も多い。例えば、2 項分布 $Bin(n, p)$ において、 p の値が非常に小さい一方で n の値が大きく、しかも $\lambda = np$ の値がそれほど大きな値をとらない場合がある。このとき、2 項分布 $Bin(n, p)$ の極限分布としてポアソン分布 $Po(\lambda)$ が導かれることがわかる。

1.3.2 ポアソン分布に従う確率変数の和を与えた時の条件付分布

定理 2 独立な確率変数 Y_1 と Y_2 がそれぞれ $Po(\lambda_1)$ および $Po(\lambda_2)$ にしたがうとき、 $T = Y_1 + Y_2 = t$ を与えた時の Y_1 の条件付き確率分布は $Bin(t, \theta)$ となる。ただし、 $\theta = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ である。

☞ この定理は、ポアソン分布の性質として非常に重要ですので記載していますが、この授業ではこの後利用しませんし、試験にも出ません。

1.4 ポアソン分布の例

1.4.1 生物の遭遇数

例 1 [到着数の確率分布] 沿岸に設置された定置網にはクジラがごく稀に混入する。1 時間当たりのクジラの到着数の期待値を λ とし、クジラの到着数の観測値を表す確率変数を Y とするとき、 Y の確率分布としてポアソン分布 $Po(\lambda)$ を仮定することができる。また、再生性の性質を使うと、 t 時間当たりのクジラの到着数は $Po(\lambda t)$ で表せる。

1.4.2 個体数区画調査

例 2 [発見ミスがあるときの個体発見数の確率分布] 面積 A の漁場における総個体数を n とし非常に大きい値であるとする。この漁場において面積 a だけで個体数調査を行うとする。調査区画の面積比 $p = a/A$ が漁場面積 A に比べて十分小さい時、その区画に生息する個体数 X の確率分布は近似的にポアソン分布 $Po(\lambda)$ にしたがう (ただし $\lambda = np$)。また、区画内の個体のダイバーによる発見率が独立同一に θ とするとき、ダイバーによる発見個体数 Y は、 $X = x$ が分かっている下では 2 項分布 $Bin(X, \theta)$ に従うが、 $X = x$ は観測できないので、すべての場合を考えて確率を計算すると、ダイバーによる発見数の確率分布は以下のように

$Y \sim Po((1 - \theta)\lambda)$ となる

$$\begin{aligned}
 P(Y = y) &= \sum_{x=0}^n P(X = x) \cdot P(Y = y | X = x) \\
 &= \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} \cdot \binom{x}{y} \theta^y (1 - \theta)^{x-y} \\
 &= e^{-\lambda} \frac{(\theta\lambda)^y}{y!} \sum_{x=y}^{\infty} \frac{\{(1 - \theta)\lambda\}^{x-y}}{(x - y)!} \\
 &= e^{-\theta\lambda} \frac{(\theta\lambda)^y}{y!}
 \end{aligned}$$

2 ポアソン分布における統計的推測

ホエールウォッチングにおける発見回数

いまあなたは xx 湾においてホエールウォッチング業を始めることにしました。事業を始めるにあたって、事前調査として湾内でウォッチング船による 1 時間当たりのクジラへの遭遇数を 5 日間に渡って観察したところ、2, 3, 0, 4, 1 であった。このとき、1 時間当たりの遭遇数がポアソン分布 $Po(\lambda)$ にしたがうとすると、 λ の推定値と推定値の標準誤差を求めなさい。



図2 ある日の広告とアイスランドのホエールウォッチング船

最初にこの問題の観測値に対する仮定を正確に記述すると、以下のよう書ける。

$$Y_1, Y_2, \dots, Y_n \sim (iid) Po(\lambda)$$

☞ *(iid)* は **independently and identically distributed** の略で、独立同一の～分布にしたがうという意味。

次に λ の推定方法としてここでは次の標本平均を用いることにする（この妥当性は以降の授業で、不偏性と分散最小の意味から示す）。

$$\hat{\lambda}(Y_1, Y_2, \dots, Y_n) = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}.$$

表記を簡略化するため、 $\hat{\lambda}(Y) = \hat{\lambda}(Y_1, Y_2, \dots, Y_n)$ と置くことにする。この推定量 $\hat{\lambda}(Y)$ の期待値と分散はそれぞれ

$$E[\hat{\lambda}(Y)] = E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n} \sum_{i=1}^n E[Y_i] = \frac{1}{n} \cdot n\lambda = \lambda$$

$$V[\hat{\lambda}(Y)] = V\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n^2} \sum_{i=1}^n V[Y_i] = \frac{1}{n^2} \cdot n\lambda = \frac{\lambda}{n}$$

となる。したがって、 λ は未知であるが、 $E[\hat{\lambda}(Y)] = \lambda$ がどんな λ に対しても成り立つから、 $\hat{\lambda}(Y)$ を用いることによって偏りなく推定できることが分かる。またその際の散らばりは、 $V[\hat{\lambda}(Y)] = \lambda/n$ で評価でき、観測回数 n が大きいほど散らばりが小さく精度の良い推定が可能なが分かる。推定量の分散の平方根は、「推定量の標準偏差」として定義され、この場合、

$$SD[\hat{\lambda}(Y)] = \sqrt{V[\hat{\lambda}]} = \sqrt{\frac{\lambda}{n}}$$

と表すことができる。

例題のように実際の観測値 $Y_1 = 2, Y_2 = 3, Y_3 = 0, Y_4 = 4, Y_5 = 1$ を得たとき、それを推定量 $\hat{\lambda}(Y)$ に代入したものが推定値であり

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{2+3+0+4+1}{5} = 2 \quad [\text{推定値}]$$

として得られ、さらに $\hat{\lambda}$ を用いて $SD[\hat{\lambda}(Y)]$ を計算した値を標準誤差 (standard error, SE) と呼ぶが、それはこの例では

$$SE = \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{2/5} = 0.632 \quad [\text{推定値の標準誤差}]$$

として求められる。

ホエールウォッチング (続き)

上記の例では、1 時間当たりのクジラ群への遭遇数は期待値 $\hat{\lambda} = 2 (/hour)$ をもつポアソン分布にしたがうと推定された。この情報を基に、あなたは「クジラをウォッチングできる確率は 99.5% 以上！」という謳い文句でセールスを開始したいと考えています。2 時間のホエールウォッチングツアーとすると、このセールストークの通りの確率を満たせるでしょうか？あるいは、この謳い文句を満たすには何時間走行する必要があるでしょうか？

例 1 のクジラの到着数の確率分布と同様に、 t 時間ホエールウォッチングを行ったときの遭遇数を確率変数 T で表すと、

$$T \sim Po(\hat{\lambda}t)$$

となり、「少なくとも 1 回クジラをウォッチングできる確率」は、1 から「1 度もクジラをウォッチングできない確率」を引けばよいので、

$$P(T \geq 1) = 1 - P(T = 0) = 1 - \exp(-\hat{\lambda}t) \geq 0.995$$

を満たす t を求めればよい。したがって

$$\exp(-2t) \leq 0.005 \iff -2t \leq \log(0.005) \iff t \geq 2.65$$

となり、切りよく 3 時間走行すれば謳い文句の通りお客さんを満足させることができる。

3 提出課題

前回の Lecture 3 と今回の Lecture 4 の課題と併せて一つの課題とします。計算式と答えを書いて、必ずまとめて 1 つのファイルで提出してください。ワードで数式の書けない人は、手書きの計算結果を写真に撮ってワードに張り付けて下さっても構いません。答だけの回答は大きく減点します。提出期限は当初 6 月 9 日 23:59 としていましたが、6 月 12 日 23:59 とします。

Lecture04-HW1

あなたはお店を経営しているとし、1 日当たりのお客さんの数は一定でなく、これを確率的に表現するとポアソン分布 $Po(\lambda)$ にしたがうことが分かりました。1 日 3 人以上お客さんが訪れないと経営が成り立ちません。3 人以上お客さんが訪れる確率はいくらでしょうか。ただし、 λ の値は学籍番号の末尾の数字とします。(0 の人は $\lambda=10$ としてください)

Lecture04-HW2

上記において、平日 10 日間のお客さんの数を記録したところ、3, 3, 6, 0, 7, 2, 1, 4, 2, 2 人であったとします。このとき、1 日当たりのお客さんの数の期待値を推定し、その推定値と標準誤差を求めなさい。

4 アンケート

下記フォームにてアンケートに回答してください。出欠確認を兼ねます。<https://forms.gle/GEZLKKEqW7Xjfiax7>