

統計学 Lecture 3

2項分布と統計的推測

北門 利英（海洋生物資源学科）

Attention: 授業に関してわからないことがあれば、メールか5/27のリアルタイム接続時（12:45-13:45）に遠慮なく質問してください。

POINT:

- 2項分布とは何か説明できるようになること
- 2項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
- 2項分布の確率分布を利用して様々な計算ができること
- 推定量と推定値の違いを理解する事
- 2項分布の推定量の期待値，分散，標準偏差の計算ができること



T. Kitakado

Statistics

Lecture 2 May 20, 2020

生物資源学科の北門です。このビデオは統計学のLecture3の要約版です。最初にこのビデオを見てから、ハンドアウトをご覧いただき、最後に確認のために再度ビデオをご覧いただくと頭の整理になると思います。今回は、2項分布の定義と、2項分布のパラメータの推定について理解してください。それでは始めます。

2項分布とは

Binomial distribution



2項分布とは

確率変数 Y が試行数 N , 試行の成功確率 p ($0 \leq p \leq 1$) をもつ2項分布に従うとは, Y が確率関数

$$P(Y = y) = \binom{N}{y} p^y (1 - p)^{N-y} (y = 0, 1, 2, \dots, N)$$

をもつことをいう。とくに, このことを

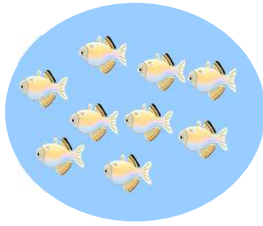
$$Y \sim \text{Bin}(N, p)$$

と略記する。



2項分布の定義についてはハンドアウトに詳しく記載していますのでご覧いただければと思いますが, 簡単な例でいうと, コイン投げのように1回1回の成功の確率が p である試行を N 回繰り返したとき, 成功の回数は確率的にしか予測できませんが, その規則性を数学的に表現したものが2項分布であり, この授業では, 確率変数を Y とおくと, Y チルダ $\text{Bin}(N, p)$ と表します。

例：mtDNA ハプロタイプ観測の確率分布



ハプロタイプ I: AGTTC...

ハプロタイプ II: TGTTC...

母集団中の割合 $p : (1-p)$ 5個体をランダムに
サンプリングサンプリングした5個体のうち、
ハプロタイプ I を持つ個体の数

$$Y \in \{0, 1, 2, 3, 4, 5\}$$



この式の導出について簡単ふれると、たとえば母集団中には遺伝的にハプロタイプIとIIを持った個体しかいないとし、その割合を p および $1-p$ とします。母集団から5個体ランダムにサンプリングしたとき、ハプロタイプIを持つ個体の数は0から5の整数のいずれかとなりますが、どの値をとるかは確率的にしか分かりません。

例：mtDNA ハプロタイプ観測の確率分布

サンプリングした5個体のうち、ハプロタイプIを持つ個体の数

 $Y \in \{0, 1, 2, 3, 4, 5\}$ ○ ハプロタイプI
× ハプロタイプII $Y = 0$ ⇔ × × × × × 1通り $Y = 1$ ⇔
○ × × × ×
× ○ × × ×
× × ○ × ×
× × × ○ ×
× × × × ○ 5通り $Y = 2$ ⇔ $Y = 3$ ⇔ $Y = 4$ ⇔ $Y = 5$ ⇔

$$P(Y = 0) = 1 \cdot (1 - p)^5 = \binom{5}{0} (1 - p)^5$$

$$P(Y = 1) = 5 \cdot p \cdot (1 - p)^4$$

$$P(Y = 2) = 10 \cdot p^2 \cdot (1 - p)^3$$

$$P(Y = 3) = 10 \cdot p^3 \cdot (1 - p)^2$$

$$P(Y = 4) = 5 \cdot p^4 \cdot (1 - p)^1$$

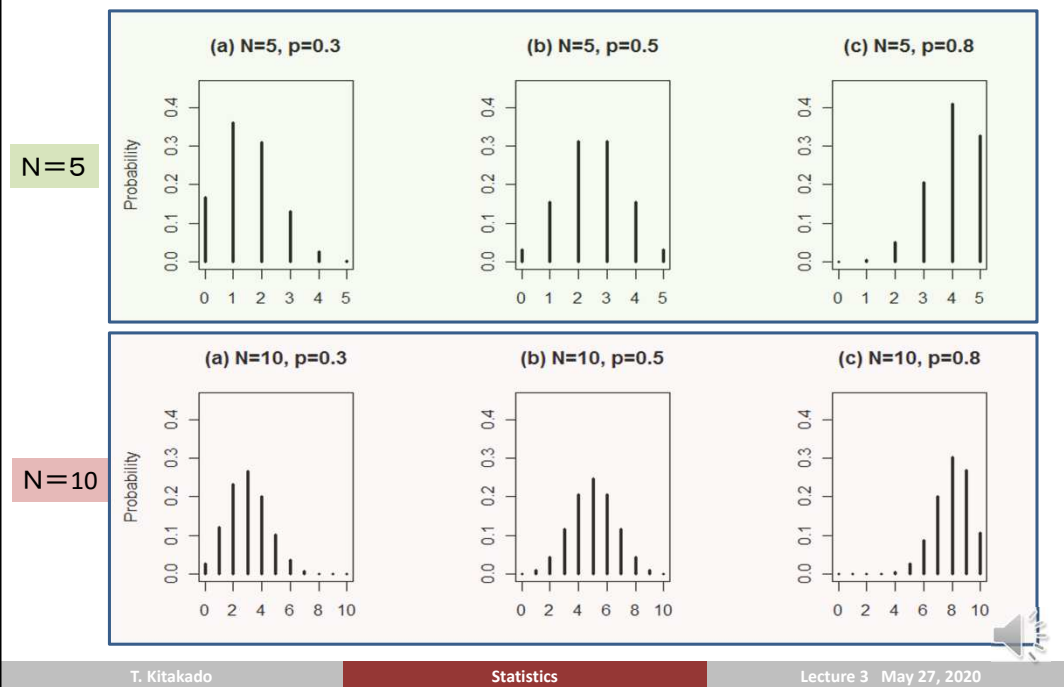
$$P(Y = 5) = 1 \cdot p^5$$

$$P(Y = y) = \binom{5}{y} p^y (1 - p)^{5-y}$$

($y = 0, 1, 2, \dots, 5$)

これを順番に見ていくと、 $Y=0$ すなわち1個体もハプロタイプIを持っていないとは、このように○×であらわすと、×が5個並ぶ確率ですので $1-p$ の5乗となります。また $Y=1$ とは、○が1つ、×が4つ並ぶ組み合わせが5通り、いずれのパターンが生じる確率とも $p(1-p)^4$ で、この5つは同時に起こり得ない排反事象なので確率は5つの和となります。その他の場合も同様で、それを一つにまとめた式が先の2項分布の確率分布の式となります。

確率関数のグラフ



2項分布の確率分布の形状はパラメータ N と p に異なります。ビデオを止めて、じっくりご覧ください。パターンが分かるかと思います。

2項分布の性質 (I)

2項定理 $(a + b)^N = \sum_{y=0}^N \binom{N}{y} a^y b^{N-y}$ より

$$\sum_{y=0}^N P(Y = y) = \sum_{y=0}^N \binom{N}{y} p^y (1-p)^{N-y} = 1$$

$$\begin{aligned} E[Y] &= \sum_{y=0}^N y \binom{N}{y} p^y (1-p)^{N-y} = \sum_{y=1}^N y \frac{N!}{y! (N-y)!} p^y (1-p)^{N-y} \\ &= Np \sum_{y=1}^N \frac{(N-1)!}{(y-1)! (N-y)!} p^{y-1} (1-p)^{N-y} \\ &= Np \sum_{z=0}^{N-1} \frac{(N-1)!}{z! (N-1-z)!} p^z (1-p)^{N-1-z} \\ &= Np(p + 1 - p)^{N-1} = Np \end{aligned}$$



次に2項分布の性質ですが、まず高校時代に習ったはずの2項定理を思い出していただくと、すべての事象の確率の和がちゃんと1になることを示せます。また期待値ですが、期待値の公式に従って計算していくと、階乗をうまく消していき、2項定理が改めて使えるようにすれば、期待値が Np であることを容易に示せます。

2項分布の性質 (2)

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

$$\begin{aligned}
 E[Y(Y-1)] &= \sum_{y=0}^N y(y-1) \binom{N}{y} p^y (1-p)^{N-y} \\
 &= N(N-1)p^2 \sum_{y=2}^N \frac{(N-2)!}{(y-2)!(N-y)!} p^{y-2} (1-p)^{N-y} \\
 &= N(N-1)p^2
 \end{aligned}$$



同様に分散ですが, こちらも詳しくはハンドアウトをよく読んで頂きたいですが, 結果として $Np(1-p)$ となります. 期待値と分散についてはしっかり覚えるとともに, 導出の仕方も理解しておいてください.

2項分布のパラメータの推定



次に2項分布のパラメータの推定に移ります.

視聴率調査 (1)

東京日曜午後6時40分の調査

 p : サザエさんを見ている人の割合

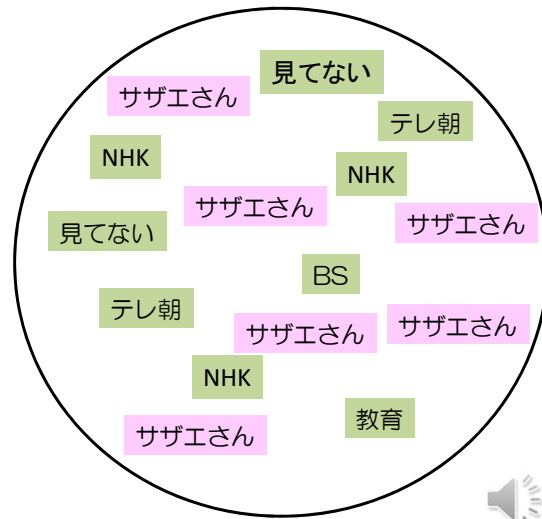
但し、視聴率の推定量の標準偏差を0.01以下にしたい

 $1-p$: サザエさんを見ていない人の割合

何世帯をサンプリングすればよいであろうか?

$$Y \sim \text{Bin}(N, p)$$

$$\hat{p}(Y) = \frac{Y}{N}$$



ここでは、皆さんにはリサーチ会社に働いていると想像していただき、東京の日曜午後6時40分のサザエさんの視聴率を調べる調査の立案を上司に頼まれたとしましょう。ただし、条件として視聴率の推定量の標準偏差を0.01にしなければならない条件があります。さて何世帯調べましょうか？この問題は、典型的な2項分布の問題で、1世帯あたりサザエさんを見ている確率を p 、サンプリングする世帯数を N とし、何世帯サザエさんを見ていたかが確率変数となります。いま、知りたいのは p で、その推定を、 N 世帯のうちサザエさんを見ている世帯数 Y の割合で推定します。このように未知のパラメータを推定する式を推定量と言います。

視聴率調査 (2)

推定量

$$\hat{p}(Y) = \frac{Y}{N}$$

推定量の
期待値

$$E[\hat{p}(Y)] = E\left[\frac{Y}{N}\right] = \frac{1}{N}E[Y] = \frac{1}{N}Np = p$$

推定量の
分散

$$V[\hat{p}(Y)] = V\left[\frac{Y}{N}\right] = \frac{1}{N^2}V[Y] = \frac{1}{N^2}Np(1-p) = \frac{p(1-p)}{N}$$

推定量の
標準偏差

$$SD[\hat{p}(Y)] = \sqrt{\frac{p(1-p)}{N}} \leq 0.01$$

ただし、調査前は
pは未知の値

左辺はp=0.5の時に最大。従って、p=0.5を想定すれば

$$SD[\hat{p}(Y)] = \sqrt{\frac{0.5(1-0.5)}{N}} \leq 0.01 \Leftrightarrow N \geq 2500$$

次に推定量の性質を調べてみましょう。推定量の期待値はp, 分散はp(1-p)/Nとなります。つまり、この推定方法を用いれば平均的にはpを偏りなく推定でき、pの本当の値の周りで散らばるのですが、Nを大きくするとその散らばりも小さく推定精度を上げることができます。また上司の依頼であった推定量の標準偏差は、分散の平方根であり、満たすべき式はこの式となります。ただし、この式にはpとNの2つの変数が入っているため、このままではNを決めることができません。そこで、標準偏差が一番大きくなるワーストケースを考えるとそれはp=0.5のときでその場合はNが2500以上必要となります。

視聴率調査 (3)

しかし視聴率50%は「お化け番組」
もう少し現実な値を考え、高々 $p=0.2$ と想定すれば

$$SD[\hat{p}] = \sqrt{\frac{0.2(1-0.2)}{N}} \leq 0.01 \Leftrightarrow N \geq 1600$$

高々 $p=0.1$ と想定すれば

$$SD[\hat{p}] = \sqrt{\frac{0.1(1-0.1)}{N}} \leq 0.01 \Leftrightarrow N \geq 900$$

このように、必要とされる精度（分散）に応じてサンプル数
に関して決定を下すことができる。



ただ、 $p=0.5$ 、すなわち視聴率が50%のようなお化け番組を想定しているわけではなくサザエさんですので、高々20%を想定すれば1600、10%の場合は900世帯以上と算出することができます。このように、確率分布を想定することで、必要とされる精度（分散）に応じてサンプル数に関して決定を下すことができるわけです。

提出課題（提出に関する説明は次週）

Lecture 3 と次回の Lecture4 の課題と併せて一つの課題とします。必ずまとめて1つのファイルで提出してください。提出期限は6月9日 23:59 です。

Lecture03-HW1

雌雄の比率が 1:1 である生物群集から N 個体のサンプルを取得するとし、雄の個体数を表す確率変数を Y とするとき、確率 $P(Y = y)$ ($y = 0, 1, \dots, N$) の値を求めなさい。ただし、 N は学籍番号の末尾の数字とします。(0 の人は $N=10$ としてください)

Lecture03-HW2

マウスに薬を投与して 80% 以上の割合で効果があつた際に次の実験ステージに進めるとする。50% の確率でしか効かない薬をマウス 10 個体に投与した際に、そのような薬が次のステップに進んでしまう確率はどの程度か求めなさい。

Lecture03-HW3

ある養殖場で生産される魚は 1% の確率で奇形であるとする。この養殖場の魚を 100 個体選んだとき、奇形である個体が 3 個体以下である確率を求めよ。

Lecture03-HW4:

ある生物集団におけるメスの成熟率を推定したい。推定量の標準偏差を 0.05 以下とするためには何個体をサンプリングする必要があるか？



今回は課題を提示します。ただし、提出は次回のLecture4と併せて提出して頂きます。提出方法については次回お伝えします。これで統計学Lecture3のビデオを終わります。