統計学 2020 Lecture 5: これまでの用語整理など

北門　利英（東京海洋大学海洋生物資源学科）

2020 年 6 月 10 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)
- 今回およびこれまでの授業に関してわからないことがあれば，メールかリアルタイム接続時に遠慮なく質問してください（6 月 10 日は 13:00-14:00 とします）．
- このハンドアウトで用語確認を行い，Lecture 2-4 の内容を再度復習し，前回提示した課題に取り組み，必ず期日までに提出してください．
- 分からないことをそのままにしないで，毎回しっかり確認してください．

Point: 次の用語をしっかり理解をする事（今回は離散型の確率分布に絞ります）．)

- 確率変数と確率分布
- 期待値，分散，標準偏差
- 2 項分布の定義と性質，そして確率の計算ができること
- ポアソン分布の定義と性質，そして確率の計算ができること
- パラメータの推定量と推定値の違いを区別すること
- パラメータの推定量の性質として，その期待と標準偏差が計算できること

以下は前回授業の補足です．

補足 1：$\binom{N}{y} = \frac{N!}{y!(N-y)!}$ で組み合わせを表します．高校では “C” を使って表したと思いますが，大学以降の授業では $\binom{N}{y}$ を用いることが多く，ここでもそのように表記します．

補足 2：$\exp(x) = e^x$ の意味です．

補足 3: $Y \sim xxx$ は確率変数 Y が xxxx 分布にしたがう，という意味です．

1 Lecture 2-4 の用語まとめ

用語	意味	例 1. サイコロの例	例 2. 雄雌ンプリング
確率変数 (random variable)	実験や試行の結果どのような値が観測されるかは事前には確率的にしかわからない変数. よく X とか Y などの大文字で表す.	サイコロを 1 回ふった時に出る目の数を Y とする場合など.	N 匹サンプリングしたときに含まれる個体雄の数を Y とすうる場合など.
標本空間 (sample space)	確率変数の取りうる値の範囲.	$\{1, 2, 3, 4, 5, 6\}$	$\{0, 1, \dots, N\}$
確率分布 (probability distribution)	確率変数のとりうる値との，それが生じる確率の関係を規定するもの．確率変数のとる値に対する確率的規則性あるいはメカニズムを数学的に表現したもの. 離散型の場合には $P(Y = y)$ と表すことが多い.	$P(Y = y) = 1/6$ $(y = 1, 2, ..., 6)$	$P(Y = y) = \binom{N}{y} p^y (1-p)^{N-y}$ $(y = 0, 1, \dots, N)$|| $< br/ >$ $(expectation)$| = _{y=0}^∞y P(Y=y)$.
分散 (variance)	確率変数が期待値に対してどれくらい散らばるかを表す指標で理論的な値（実際の観測値の分散とは異なる）. $V[Y] = \sum_{y=0}^{\infty}(y - E[Y])^2 P(Y = y)$.	$V[Y] = \sum_{y=1}^{6}(y - E[Y])^2 P(Y = y) = 35/12$	$V[Y] = \sum_{y=0}^{N}(y - E[Y])^2 P(Y = y) = Np(1-p)$
標準偏差 (standard deviation)	分散の平方根. $SD[Y] = \sqrt{V[Y]}$.	$SD[Y] = \sqrt{35/12}$	$SD[Y] = \sqrt{Np(1-p)}$

E[Y]

用語	意味	例 1. サイコロの例	例 2. 雄雌ンプリング
実現値，観測値 (realization,observation)	確率変数の実際の値（具体的な数字として観測した，あるいは現れたもの．確率変数は，どんな値をとるかわからない，ある意味，値の器あるいは数学的な想像のもの）で記号で表すときには小文字を使って確率変数と区別する．	仮に 5 回ふったとき，たとえば 3, 2, 6, 2, 5 などの値．	仮に $N = 10$ の実験を $n = 3$ 回行ったとしたとき，3, 8, 7 などの値．
標本平均 (sample mean)	確率変数を複数観測する場合，その観測値の算術平均値．n 個観測する場合，それを一種の確率変数とみなしたいときには $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ で，その実現値とみなすときには $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ などと表す．	上の場合，$\bar{y} = \sum_{i=1}^{n} y_i = 3$	上の場合，$\bar{y} = \sum_{i=1}^{n} y_i = 6$
標本 (不偏) 平均 (sample variance)	確率変数を複数観測する場合，その観測値の分散．n 個観測する場合，それを一種の確率変数とみなしたいときには $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ で，その実現値とみなすときには $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n} y_i$ などと表す．	上の場合，$s^2 = 3.3$	上の場合，$s^2 = 7$
離散型確率変数 (discrete random variable) と離散型確率分布 (discrete distribution)	確率変数の取りうる値が，0, 1, 2,...のように数直線上の決まった値しかとらない確率変数のこと．またその確率変数に対応する確率分布．	離散型である．特に離散型一様分布という（この授業では覚えなくてよいです）	離散型である．特に 2 項分布という．

用語	意味	例 1. サイコロの例	例 2. 雄雌ンプリング
2 項分布 (binomial distribution)	成功か失敗，雄か雌，病気にかかっているか否か，など 2 つの属性があるときに，N 回中（あるいは N 個体中など）Y 回成功（あるいは Y 個体雄など）というような実験や観測結果を表す場合に利用. $Y \sim Bin(N, p)$ のとき，確率分布は $P(Y = y) = \binom{N}{y} p^y (1-p)^{N-y}$. 期待値や分散は最右列の例 2 を参照のこと.	偶数と奇数のように 2 つの属性を考える場合は N 回サイロころを振ったとき偶数が出る回数	まさに 2 項分布の例
ポアソン分布（Poisson distribution)	生物の発見数や，ある一定時間に起こる地震の回数など，0 を含めた生起回数の観測結果を表す場合に利用. $Y \sim Po(\lambda)$ のとき，確率分布は $P(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!}$, 期待値も分散も λ.	ポアソン分布では扱えない現象	ポアソン分布では扱えない現象
$Y_1, ..., Y_n \sim (iid)xxx$	n 個の確率変数 $Y_1, ..., Y_n$ は，独立で，かつ同一の xxx 分布に従う		
再生性	確率変数の和もまた同じ確率分布に従うこと		
推定量 (estimator)	パラメータを推定するための方式（確率変数の式で表される）.		
推定値 (estimate)	推定量に具体的な数値を代入したもので，推定値も数値.		

離散型分布の場合(一般に)

確率変数 Y
確率分布 $P(Y=y)$

確率変数と確率分布の組み合わせが決まったら，その確率変数の確率分布の性質として，次の特性値が決まる

期待値 $E[Y]$
分散 $V[Y]$
標準偏差 $SD[Y]=\sqrt{V[Y]}$

2項分布の場合

$Y\sim Bin(N,p)$

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

ただし，$\binom{N}{y}=\frac{N!}{y!(N-y)!}$

期待値 $E[Y]=Np$
分散 $V[Y]=Np(1-p)$
標準偏差 $SD[Y]=\sqrt{Np(1-p)}$

(例)25%が成熟している個体群から，N=5個体サンプリングした時，2個体が成熟個体である確率は？

$Y\sim Bin(5,1/4)$

$$P(Y=2)=\binom{5}{2}\left(\frac{1}{4}\right)^2\left(1-\frac{1}{4}\right)^{5-2}$$
$$=\frac{270}{4^5}=\frac{135}{512}$$

ポアソン分布の場合

$Y\sim Po(\lambda)$

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

期待値 $E[Y]=\lambda$
分散 $V[Y]=\lambda$
標準偏差 $SD[Y]=\sqrt{\lambda}$

(例)1区画当たりのアワビの個体数の期待値が2であるとき，ランダムに選んだ2区画には1個体も存在しない確率は？

$Y_1,Y_2\sim(iid)Po(2)$

iidは「独立かつ同一の確率分布に従う」という意味

$$P(Y_1=0,Y_2=0)$$
$$=P(Y_1=0)P(Y_2=0)$$
$$=e^{-2}\frac{2^0}{0!}e^{-2}\frac{2^0}{0!}=e^{-4}=0.0183$$

2 確率変数と確率分布

確率変数は X や Y と表すことが多く，どちらを用いても構いませんが，今回は Y で統一して書くことにする．

2.1 期待値の定義と性質

まずは，期待値の定義を確認しよう．

定義 1 **[確率変数の期待値]** 確率変数 Y の期待値は次のように定義される．

$$E[Y] = \sum_{y=0}^{\infty} yP(Y=y)$$

ここで，期待値の線形性を考える．まず定数 a, b に対して

$$E[a+bY] = \sum_{y=0}^{\infty}(a+by)P(Y=y) = a\sum_{y=0}^{\infty}P(Y=y) + b\sum_{y=0}^{\infty}yP(Y=y) = a+bE[Y]$$

が成り立つ．同様にして，確率変数 Y_1, Y_2 と a_1, a_2 に対して，$a_1Y_1 + a_2Y_2$ の期待値は

$$E[a_1Y_1 + a_2Y_2] = a_1E[Y_1] + a_2E[Y_2]$$

が成り立つ．このように，「和の期待値」は「期待値の和」となることが分かる．たとえば，サイコロを 2 回振って，1 回目は出た目の 1000 倍，2 回目は出た目の 2000 倍のお金が当たるくじがあるとする．このくじの期待値は $1000Y_1 + 2000Y_2$ であるが，これはサイコロを 1 つ振ったときの出る目の期待値は 3.5 だから 1 つ目のサイコロを振ったときの期待値は \$1000 ×3.5=3500 円，2 つ目では $2000 \times 3.5 = 7000$ 円なので，合計 10500 円という計算をしていることと同じである．

2.2 分散の定義とその性質

定義 2 **[確率変数の分散]** 確率変数 X の**分散 (variance)** は次のように定義される．

$$V[Y] = \sum_{y=0}^{\infty}(y-E[Y])^2P(Y=y)$$

分散は言いかえれば「確率変数と期待値の差の 2 乗」の期待値である．

ここで，分散については以下が成り立つ．

$$V[a+bY] = \sum_{y=0}^{\infty}(a+by-E[a+bY])^2P(Y=y) = \sum_{y=0}^{\infty}b^2(y-E[Y])^2P(Y=y) = b^2\sum_{y=0}^{\infty}(y-E[Y])^2P(Y=y) = V[Y]$$

また，確率変数 Y_1, Y_2 が独立ならば，a_1, a_2 に対して次の式が成り立つ．

$$V[a_1Y_1 + a_2Y_2] = a_1^2V[Y_1] + a_2^2V[Y_2]$$

3 2 項分布とは

成功か失敗の試行を N 回行うとする．各回の試行は独立で，成功の確率は試行を通して同一とする．ここでの成功/失敗は，例えば母集団から個体を選んだときの雄/雌，成熟/未成熟，ウイルス感染/非感染と読み替えても同様である．この成功回数を確率変数とし Y とおくと，Y は 2 項分布にしたがう．

離散型分布 1 [2 項分布] 確率変数 Y が確率関数

$$P(Y=y)=\binom{N}{y}p^y(1-p)^{N-y},\quad y=0,1,...,N\quad(0\le p\le 1) \tag{1}$$

をもつとき，Y は 2 項分布 (binomial distribution) にしたがうという．この 2 項分布は $Bin(N,p)$ で表すこととする．

3.1 2 項分布の期待値と分散

期待値と分散は以下のように求められる（標準偏差は分散の平方根）．詳しい計算過程は Lecture 3 の資料で確認してください．

$$\begin{aligned}
E[Y] &= \sum_{y=0}^{N} yP(Y=y)=\sum_{y=0}^{N} y\binom{N}{y}p^y(1-p)^{N-y}=Np.\\
E[Y(Y-1)] &= \sum_{y=0}^{N} y(y-1)P(Y=y)=\sum_{y=0}^{N} y(y-1)\binom{N}{y}p^y(1-p)^{N-y}=N(N-1)p^2\\
V[Y] &= E[Y(Y-1)]+E[Y]-E[Y]^2=N(N-1)p^2+Np-(Np)^2=Np(1-p)
\end{aligned}$$

3.2 2 項分布の再生性

定理 1 [2 項分布の再生性] 確率変数 Y_i $(i=1,2,...,m)$ が独立同一に $Bin(n_i,p)$ にしたがうとき，$T=\sum_{i=1}^{m}Y_i$ は 2 項分布 $Bin(N,p)$ にしたがう．ただし，$N=\sum_{i=1}^{m}n_i$ である．

3.3 2 項分布のパラメータの推定

2 項分布にしたがう確率変数を観測し，パラメータ p を推定したい．

$$Y\sim Bin(N,p)$$

いま p が未知パラメータ値であるから，その推定の方法として次の方式を考える．

$$\hat{p}(Y)=\frac{Y}{N}\quad[p\text{ の推定量}]$$

この式では，N 世帯のサンプリングをするので N は固定されているのに対し，Y は確率変数でありどのような値が観測されるか分からないが，その値に関わらず p の推定のための計算方法を予め規定している，ことを意味している．このような式のことを**推定量 (estimator)** という．

推定量 $\widehat{p}(Y) = \frac{Y}{N}$ の期待値と分散は

$$E[\widehat{p}(Y)] = E\left[\frac{Y}{N}\right] = \frac{1}{N}\ E[Y] = \frac{1}{N}\ Np = p \quad [p \text{ の推定量の期待値}]$$

$$V[\widehat{p}(Y)] = V\left[\frac{Y}{N}\right] = \frac{1}{N^2}\ E[Y] = \frac{1}{N^2}\ Np(1-p) = \frac{p(1-p)}{N} \quad [p \text{ の推定量の分散}]$$

推定量 $\widehat{p}(Y) = \frac{Y}{N}$ の標準偏差 (standard deviation)，分散の平方根だから次のように求まる．

$$SD[\widehat{p}(Y)] = \sqrt{V[\widehat{p}(Y)]} = \sqrt{\frac{p(1-p)}{N}}$$

$\widehat{p}(Y)$ に実際の観測値を代入したものが推定値であり，さらに $\widehat{p}$ を $SD[\widehat{p}(Y)]$ に代入して計算した値を**標準誤差 (standard error, SE)** と呼ぶ．

3.4 演習

問題 1 ある魚種のある漁場には，本種の A 個体群と B 個体群が回遊してくる．この漁場における A 個体群の個体の割合を p とする．この漁場で 20 個体のサンプリングを行ったところ 6 個体が A 個体群の個体であることが分かった．このとき，p の推定値と標準誤差を求めなさい．この漁場から 10 個体サンプリングした時，A 個体群の個体が期待値以上含まれている確率を求めなさい．

解答例 観測値が次の確率分布から得られたと仮定する．

$$Y \sim Bin(20, p)$$

未知のパラメータ λ を推定するために，次の推定量を用いる．

$$\widehat{p}(Y) = \frac{Y}{20} \quad [p \text{ の推定量}]$$

$Y = 5$ であるから，これを $\widehat{p}(Y)$ に代入して

$$\widehat{p} = \frac{6}{20} = 0.3 \quad [p \text{ の推定値}]$$

また標準誤差は推定量の標準偏差にパラメータの推定値を代入した値だから，次の通り．

$$SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{20}} = \sqrt{\frac{0.3 \cdot 0.7}{20}} = 0.102 \quad [p \text{ の推定値の標準誤差}]$$

10 個体サンプリングした時の A 個体群の個体数の期待値は $N\widehat{p} = 10 \cdot 0.3 = 3$ であるから，求める確率は以下の通り．

$$\begin{aligned} P(Y \geq 3) &= 1 - P(Y=0) - P(Y=1) - P(Y=2) \\ &= 1 - \binom{10}{0}0.3^0 0.7^{10} - \binom{10}{1}0.3^1 0.7^9 - \binom{10}{2}0.3^2 0.7^8 = 0.383 \\ &= 1 - 0.0282 - 0.1211 - 0.2335 = 0.6172. \end{aligned}$$

4 ポアソン分布とは

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．

離散型分布 2 [ポアソン分布] 確率変数 Y が確率関数

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

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

($Po(\lambda)$ あるいは $Pois(\lambda)$ と書いたりしますが，この授業では前者を使うことにします．)

4.1 期待値と分散

期待値と分散は以下のように求められる（標準偏差は分散の平方根）．計算過程は Lecture 4 の資料で確認してください．

$$\begin{aligned}
E[Y] &= \sum_{y=0}^{\infty} yP(Y=y)=\sum_{y=0}^{\infty} ye^{-\lambda}\frac{\lambda^y}{y!}=\lambda\\
E[Y(Y-1)] &= \sum_{y=0}^{\infty} y(y-1)P(Y=y)=\sum_{y=0}^{\infty} y(y-1)e^{-\lambda}\frac{\lambda^y}{y!}=\lambda^2\\
V[Y] &= E[Y(Y-1)]+E[Y]-E[Y]^2=\lambda^2+\lambda-\lambda^2=\lambda
\end{aligned}$$

4.2 ポアソン分布の再生性

定理 2 [ポアソン分布の再生性] 確率変数 Y_i $(i=1,2,...,m)$ 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．

4.3 ポアソン分布のパラメータの推定

ポアソン分布にしたがう n 個の観測値を用いてパラメータ λ を推定する．

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

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．

$$\hat{\lambda}(Y)=\frac{1}{n}\sum_{i=1}^{n}Y_i=\bar{Y}.\quad[\lambda\text{ の推定量}]$$

ここで $Y = (Y_1, Y_2, ..., Y_n)$ とする．この推定量 $\hat{\lambda}(Y)$ の期待値と分散はそれぞれ次のように求められる．

$$E[\hat{\lambda}(Y)] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n}\cdot n\lambda = \lambda \quad [\lambda \text{ の推定量の期待値}]$$

$$V[\hat{\lambda}(Y)] = V\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[Y_i] = \frac{1}{n^2}\cdot n\lambda = \frac{\lambda}{n} \quad [\lambda \text{ の推定量の分散}]$$

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ことが分かる．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，この場合次のように表せる．

$$SD[\hat{\lambda}(Y)] = \sqrt{V[\hat{\lambda}]} = \sqrt{\frac{\lambda}{n}} \quad [\lambda \text{ の推定量の標準偏差}]$$

$\hat{\lambda}(Y)$ に実際の観測値を代入したものが推定値であり，さらに $\hat{\lambda}$ を $SD[\hat{\lambda}(Y)]$ に代入して計算した値を**標準誤差 (standard error, SE)** と呼ぶ

4.4 演習

問題 1 あなたは XX 岬のサケ定置網に侵入してくるゼニガタアザラシの個体数を毎日 1 時間水中カメラで観測したところ，10 日間の記録として，0, 3, 2, 1, 0, 0, 3, 4, 2, 1 という結果を得た．1 時間当たりの侵入個体数の期待値を λ とし，その λ の推定値と標準誤差を求めなさい．また，1 時間当たり 1 個体以上が侵入する確率はいくらか．

解答例 観測値が次の確率分布から得られたと仮定する．

$$Y_1, Y_2, \dots, Y_{10} \sim (iid)Po(\lambda)$$

未知のパラメータ λ を推定するために，次の推定量を用いる．

$$\hat{\lambda}(Y) = \frac{1}{10}\sum_{i=1}^{10} Y_i = \bar{Y} \quad [\lambda \text{ の推定量}]$$

$Y = (Y_1, Y_2, ..., Y_{10}) = (0, 3, 2, 1, 0, 0, 3, 4, 2, 1)$ であるから，これを $\hat{\lambda}(Y)$ に代入して

$$\hat{\lambda} = \frac{1}{10}\sum_{i=1}^{10} y_i = 1.6 \quad [\lambda \text{ の推定値}]$$

また標準誤差は推定量の標準偏差にパラメータの推定値を代入した値だから，次の通り．

$$SE = \sqrt{\frac{\hat{\lambda}}{10}} = \sqrt{0.16} = 0.4 \quad [\lambda \text{ の推定値の標準誤差}]$$

1 時間当たり 1 個体以上が侵入する確率は以下のように求められる．

$$P(Y \geq 1) = 1 - P(Y \leq 1) = 1 - P(Y = 0) = 1 - e^{-1.6}\frac{1.6^0}{0!} = 0.798$$