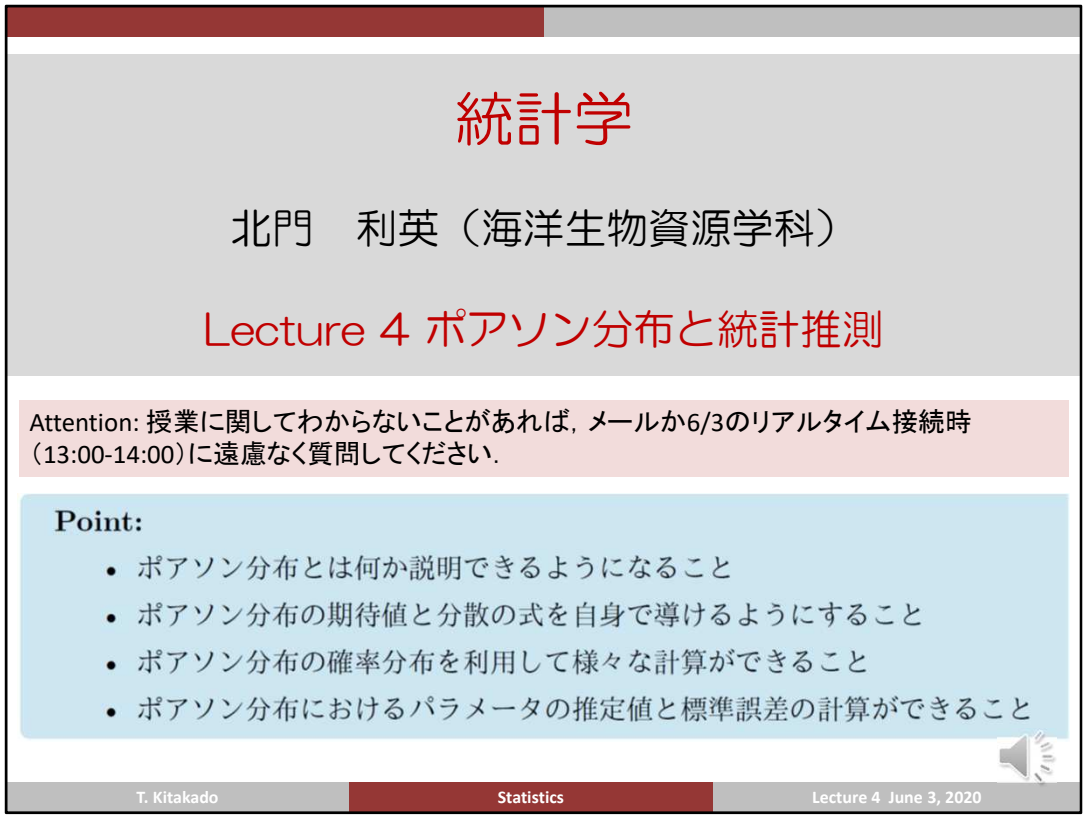


生物資源学科の北門です．このビデオは統計学のLecture4の要約版です．最初にこのビデオを見てから，ハンドアウトをご覧いただき，最後に確認のために再度ビデオをご覧いただくと頭の整理になると思います．今回は，ポアソン分布の定義と，ポアソン分布のパラメータの推定について理解してください．それでは始めます．

今回の例題

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き，λ の推定値と推定値の標準誤差を求めなさい．

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き，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？

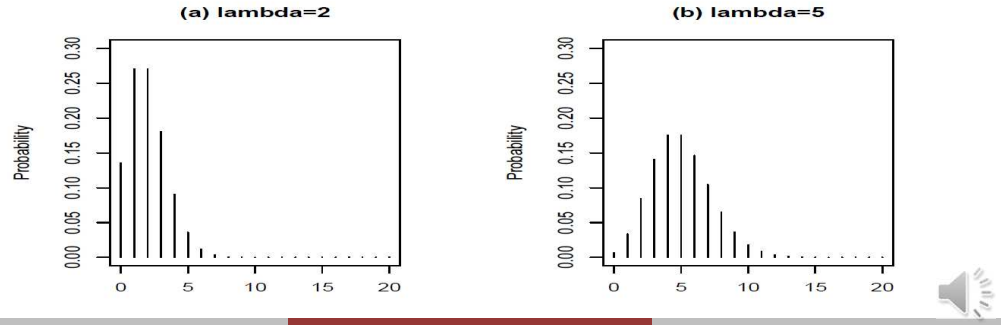
今日は最初に，取り組むべき課題について説明したいと思います．いま仮想的に，あなたはある湾においてホエールウォッチング業を始めることにしました．事業を始めるにあたって，まず事前調査として1時間船で走行する実験を5日間にわたって行たところ，初日は2回，二日目は3回，という風に結果を得ました．このような変動をポアソン分布に従うと想定し，パラメータであるラムダの推定値と標準誤差を求めたい，という問題です．さらに，あなたは広告に，クジラをウォッチングできる確率は99.5パーセント以上，という謳い文句を書きたいとします．そのためには，この実験結果を基にするとき，何時間走行する必要があるでしょうか？というとても現実的な問題です．すごく臨場感のある設定ではないでしょうか？では，まず今回の主題でもあるポアソン分布から説明していきたいと思います．

ポアソン分布とは

確率変数 Y が期待値 λ (λ>0) をもつポアソン分布に従うとは，Y が確率分布

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

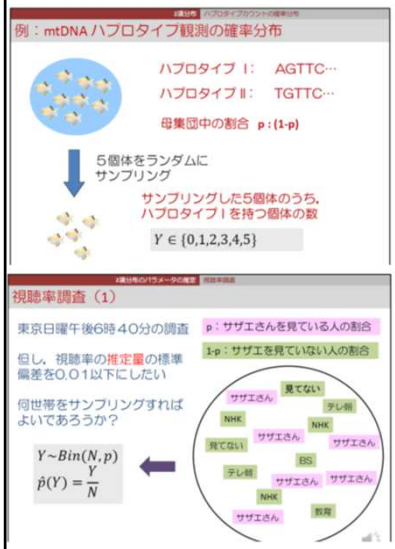
をもつことをいい，これを $Y \sim Po(\lambda)$ と略記する



そのポアソン分布ですが，これも2項分布と同様に離散型確率分布の一種で，非常に応用範囲の広い重要な確率分布の一つです．ポアソン分布にはパラメータが一つあり，ここではそれをラムダとおきますが，そのラムダは期待値としての意味をもちます．つまり確率変数Yが期待値ラムダをもつポアソン分布にしたがうとは，その確率分布がこのような式で表せることを意味します．見慣れない式に見えますが，ある意味とてもきれいな式です．それはあとでわかります． この確率分布で定義される確率ですが，たとえばラムダが2おときには，Y=0というような値をとる一方で，6よりも大きな値はあまりとりません．一方でラムダ＝5とすると，Y=4,5あたりをもっとも取りやすくなりますが10以上の値もしばしばとります．このように確率分布の様子はラムダがコントロールしていることが分かります．逆に，小さなYの値しか観測しなければラムダの値は小さいのだな，というような推測もできるわけです．

2項分布とポアソン分布の違い

2項分布



ポアソン分布

- あるアワビ漁場で区画サンプリングを行ったときの個体数の確率分布
- 単位時間あたりに，刺網に到達する魚の数の確率分布
- ミンククジラの10分間当たりの浮上回数の確率分布
- ホエールウォッチング中1時間におけるクジラへの遭遇数の確率分布
- ．．．

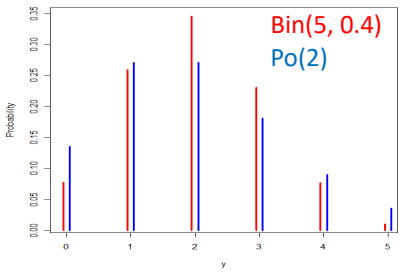
2項分布とポアソンの違いは

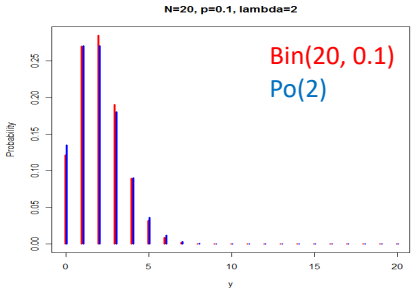
2項分布	Y = 0, 1, 2, …, N	上限有
ポアソン分布	Y = 0, 1, 2, 3, ……	上限無

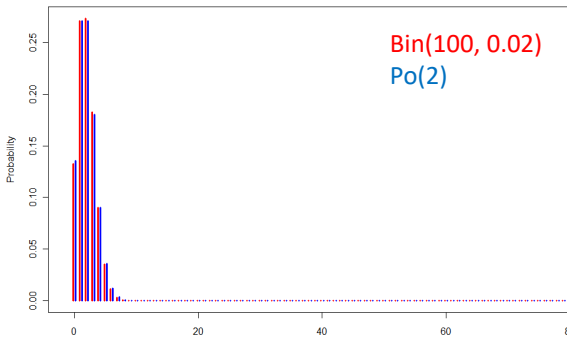
ところでポアソン分布と2項分布の違いですが，例えば2項分布の場合には，先週のハプロタイプや視聴率調査のように，N個体あるいはN世帯のうちいくつというように，取りうる範囲が0～Nと範囲が決まっています．一方でポアソン分布の場合には，ここに挙げている例のように，，，，，特に上限が定まらない定める必要がない，という違いがあります．その他，理論的背景や導出の違いもあるのですが，まずはその点を抑えてください．

2項分布のポアソン分布近似

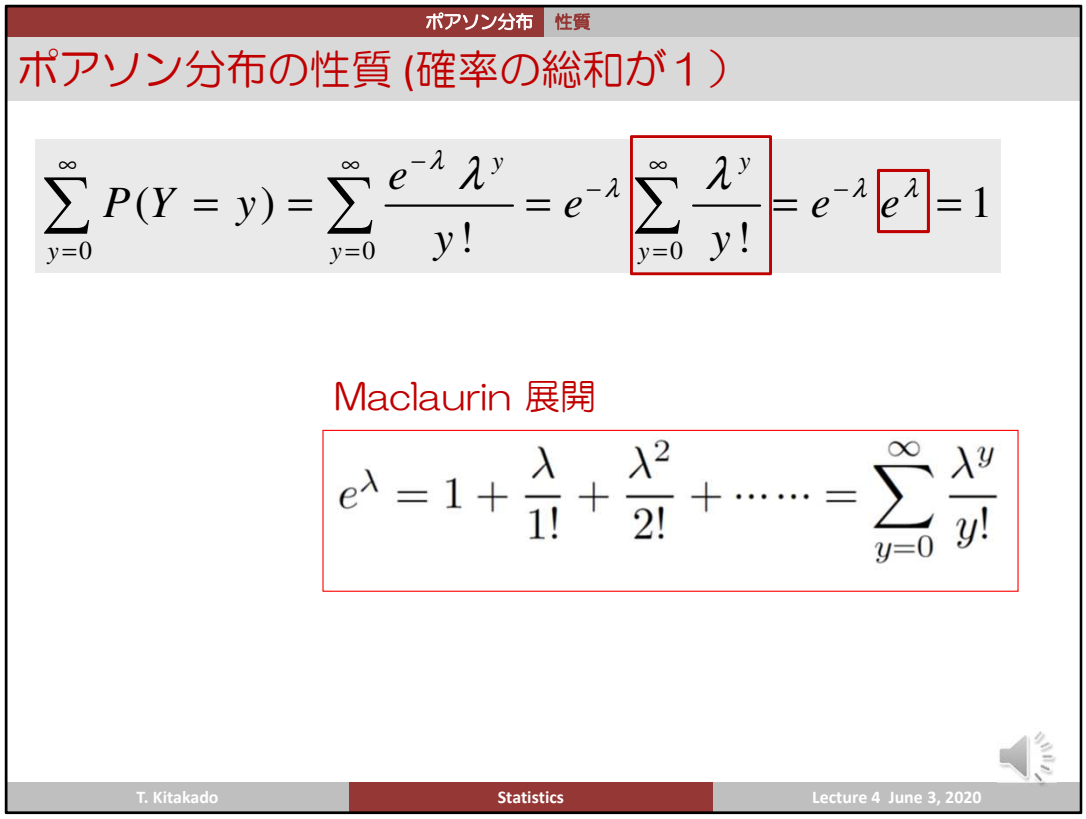
N → ∞, p → 0 (Np → λ) の時, Bin(N,p) は Po(λ) に近づく







ただ，2つの確率分布は全く無関係ではなく，たとえば2項分布においてNが大きくpが小さいという状況では，2項分布の極限分布がポアソン分布となるいう事実もあります．ここに示した図のように，N=5程度では2項分布とポアソンに多少差が見られますが，N=100になるとほとんど差がなくなります．



次に確率分布の性質ですが，その前に，重要なこととして，確率の総和が1になることを確認したいと思います．その際，皆さんは覚えておられるかどうかわかりませんが，1年生の時にテイラー展開とかマクローリン展開とかを習ったはずです．この展開の直接的な意味は，何回でも微分可能な滑らかなか関数は，多項式で近似できる，ということです．ここでは，指数関数がこのような形で近似できるできることを利用しますが，実際，ポアソン分布の確率の式を足し合わせると，ここにマクローリン展開の結果が出てきますから，これをexp(lambda)と置き換えれば，合計が1であることが証明できます．これと全く同じ方法で，期待値と分散という特性値も計算可能です．なお，テーラー展開，マクローリン展開は微分積分のどの教科書にも出ていますが，参考までに付録としてこのビデオの最後にスライドを付けておきます．興味のある方はご覧ください．

ポアソン分布の期待値

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

その期待値ですが，定義に従って値と確率をかけてすべて足す，わけですが，前回の2項分布と同様な計算で，期待値がラムダであることが示せます．

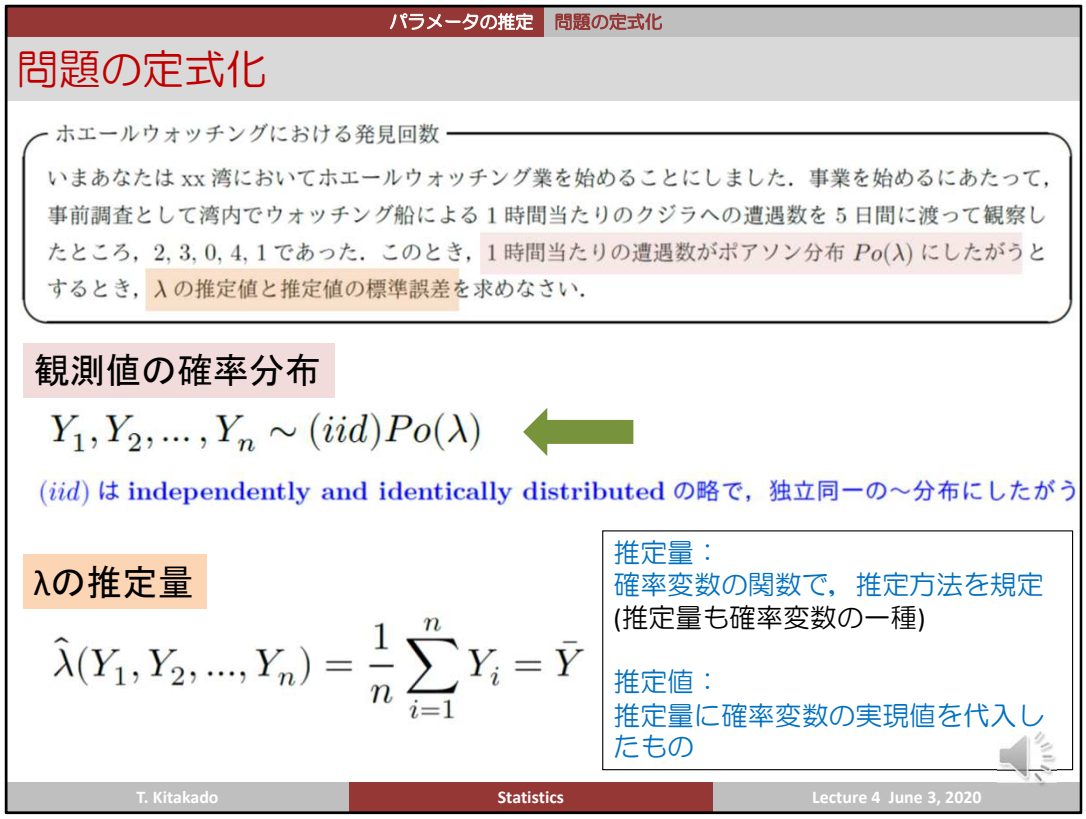
ポアソン分布の分散

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

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

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

分散についても同様で，離散型分布の場合は，分散を計算する際に，この式を利用すると計算が格段に楽になりますが，ここでも，階乗をうまく消して，マクローリン展開を使える形にすれば，分散が期待値と同じラムダとなることが示せます．



では次にホエールウォッチングの問題に戻りたいと思います．1時間走行を5回行って，クジラのへの遭遇回数のデータを得ています．これを一般的にこのような表記で表したいと思います．ここでiidとは独立かつ同一の確率分布に従うという意味で，この授業でこのあともほぼ毎回でててくる表記になりますから必ず覚えてほしいですが，同じ条件での観測を複数回繰り返す際に利用します．いま例がまさにそれに当てはまるわけですが，この観測データを基に未知のパラメータ ラムダを推定します．ここでは観測値の平均値を利用することとします．この式はまだ実際の値を入れる前の式で推定の仕方を規定している関数となります．このような式を推定量ということは前回も説明しましたが，確率変数の関数ですから期待値と分散が計算できます．これとよく似ていますが，この推定量に実現値（実測値ともいう）を代入して具体的な値として得られるものを推定値と言います．ちょっとややこしいと思われるかもしれませんが，あとで少し整理します．

パラメータの推定量の性質

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

λの推定量の期待値

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

λの推定量の分散

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

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

最初に推定量の性質を調べると，λハットの期待値は，定数倍が期待値の記号のそとにでて，そて足し算の期待値は期待値の足し算，したがって，λハットの期待値はラムダに．したがって偏っていないことが分かります．また分散については定数倍は分散の記号の外に出るときは2乗されて出ますし，いま確率変数が独立ですので，足し算の分散は分散の足し算になりますから，λハットの分散はnぶんのλとなり，nが大きいほど，すなわち観測回数が多いほど分散を小さくでき，散らばりの小さいよい推定が期待されます．それから分散の平方根を標準偏差と呼びますが，ここではこののように求まります．ここで使った期待値と分散の性質はLecture2の資料で述べています．忘れた人は確認してください．

推定値と標準誤差

実測値

$$Y_1 = 2, Y_2 = 3, Y_3 = 0, Y_4 = 4, Y_5 = 1$$

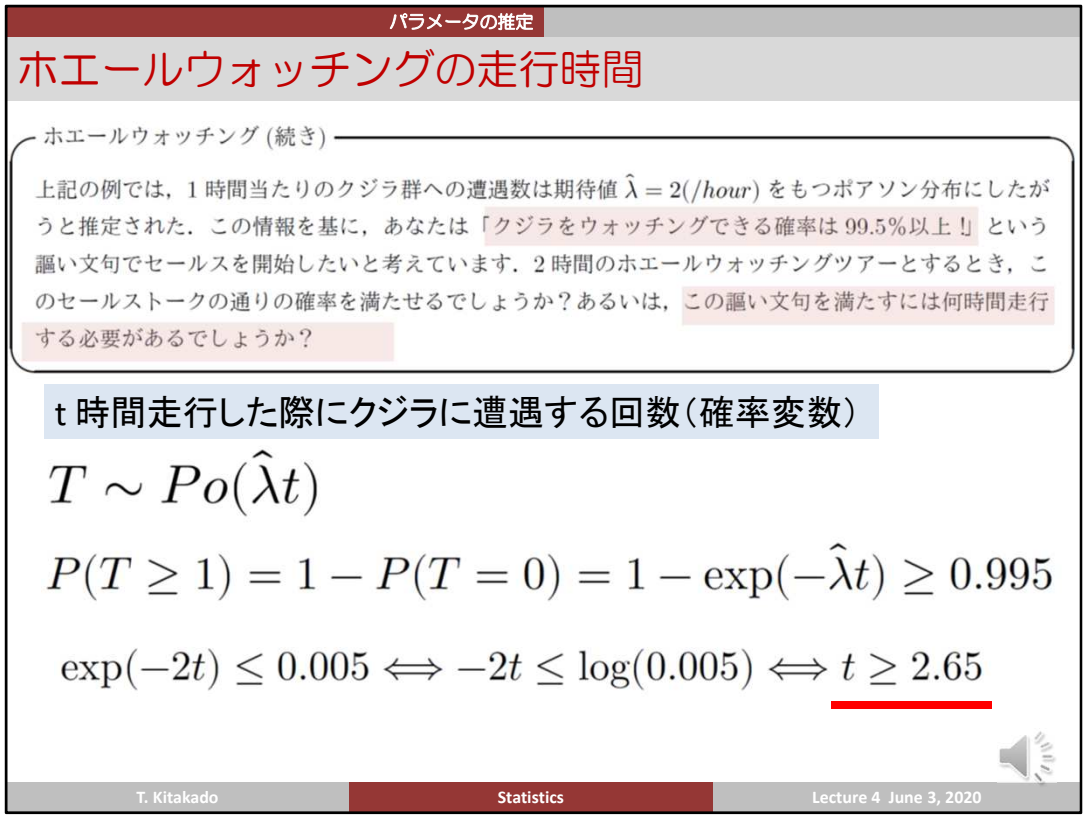
λの推定値

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{2+3+0+4+1}{5} = 2$$

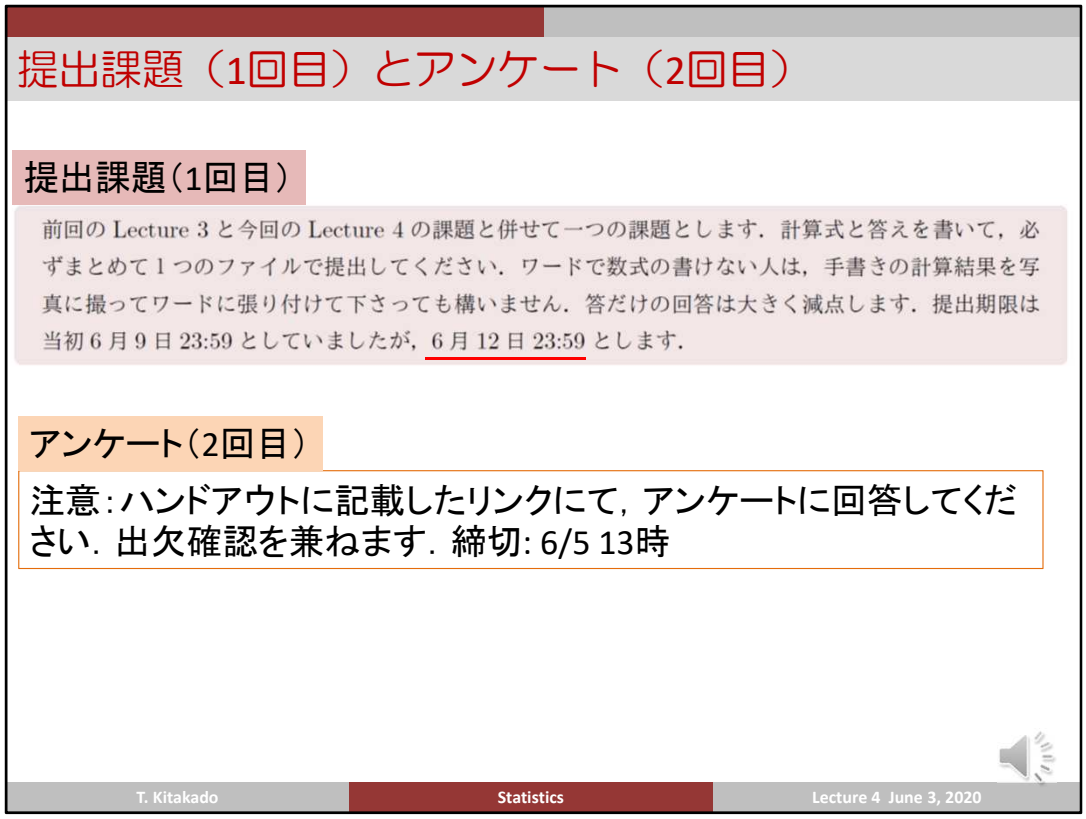
λの推定値の標準誤差(standard error, SE)

$$SE = \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{2/5} = 0.632$$

さて，推定量の性質が分かりましたので，今度は実際に推定値を計算してみましょう．実測値はこの通りで，したがってこの平均値がラムダの推定値λハットとなります．また標準誤差とは，ラム推定量の標準偏差の値に推定値を代入して求めますこの場合は，このように計算できます．これで課題の答えが求まりました．



また，ホエールウォッチング業を誠実に行うために，ウォッチングできる確率を99.5%以上にする船の走行時間ですが，いま単位時間当たりの遭遇回数の期待値が求まりました．したがってハンドアウト記載した再生性の性質を利用すると，t時間走行した時の遭遇回数Tの確率分布はこのようなポアソン分布に従うと考えられます．少なくとも1回遭遇する確率を求めればよいですから，それは1から一度も遭遇しない確率を引けばよいく，式を変形してtが2.65時間以上の時に条件を満たすことが分かります．つまり2時間ではだめですね．このように，実践的な場面でも統計学を用いて問題を解決することができます．



今回は宿題が出ています．前回のLecture3の4問と今回の2問について，一つのファイルに回答を保存して，学務システムに提出してください．また，2回目のアンケートを実施します．出欠も兼ねますので，期日までに必ず回答ください．

補足：Taylor 展開

定理 1.12 (Taylor の定理) 関数 $y = f(x)$ が $x = a$ を含む開区間で n 回微分可能ならば,

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + R_n, \\ R_n &= \frac{f^{(n)}(a+\theta(x-a))}{n!}(x-a)^n, \quad 0 < \theta < 1 \end{aligned}$$

をみたす θ が少なくとも 1 つ存在する．また，関数 $y = f(x)$ が何回でも微分可能であり，$R_n \to 0$ $(n \to \infty)$ となるならば，$f(x)$ は無限級数

$$f(x) = f(a) + \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

で表すことができる．これを関数 $f(x)$ の $x = a$ の周りでの Taylor（テイラー）展開あるいは Taylor 級数とよぶ.

「農学・水産学系学生のための数理科学入門」
第1章　数学 （北門著） より抜粋

補足： Maclaurin 展開

定理 1.12 において，$a = 0$ とおくと $f(x)$ は

$$f(x) = f(0) + \sum_{n=1}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

と表すことができる．このような Taylor 展開の特別な形を関数 $f(x)$ の Maclaurin（マクローリン）展開あるいは Maclaurin 級数とよぶ．このように，任意の高次微分可能な関数は，Taylor 展開や Maclaurin 展開を利用して多項式で近似できることが分かる．

例 1.3 次の級数展開が成り立つ．

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots \quad (-\infty < x < \infty),$$
$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots + (-1)^{n-1}\frac{x^n}{n} + \cdots \quad (-1 < x \leq 1),$$
$$\sin x = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots \quad (-\infty < x < \infty).$$

補足： Maclaurin 展開による関数の近似

図 4 は関数 $f(x) = e^x$ および $f(x) = \sin x$ の Maclaurin 展開による近似の様子を表している．図中の n は展開の次数を表している．いずれも近似に用いる多項式の次数 n が上がると近似の精度が上がることを示している．また，$x = 0$ の周りの展開なので，その付近での近似精度が良好である様子がわかる．

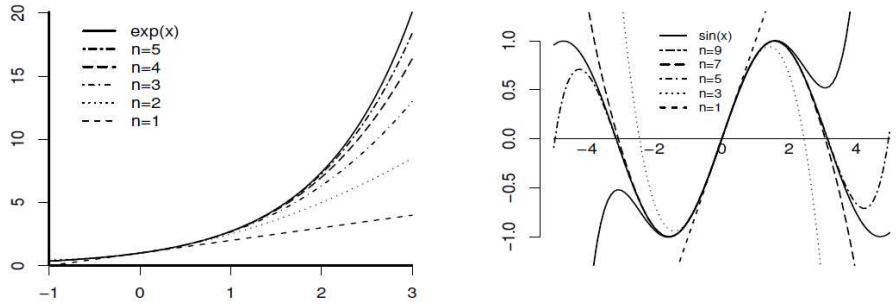


図 4. 関数 $f(x) = e^x$（左）および $f(x) = \sin(x)$（右）の Maclaurin 展開