

# 統計学 2020 Lecture 1: イントロダクション

北門 利英（東京海洋大学海洋生物資源学科）

2020 年 5 月 13 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>)
- 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（水曜 3 限）に遠慮なく質問してください。

## Point:

- Lecture 1 はイントロですので気楽に読んでください。用語や式はこれからの授業で少しずつ説明していきます（今回は宿題や提出課題はありません）。

## 1 はじめに

### 1.1 この授業のねらい

我々の日常の出来事には多くの変動や不確実性を伴います。その中で、観察による客観的情報を基にして、様々な判断をする必要があります。海洋大の学生である皆さんは将来、生物実験や生態調査を行ったり、資源調査による観察情報や漁業情報を集めたりする機会があるでしょう。そのとき、自分で取得したデータの解析を自らの手で行う必要があります。そこでこの講義では、統計学の考え方を理解し、実践で利用できるよう基礎を身につけることを目標とします。生物学、生態学、水産学などの農学系分野での適用を念頭におき、講義と練習課題により授業を進めます。統計ソフト R を用いた演習も交えて、生物や海洋のデータだけでなく身近なデータの解析を通じて統計学の考え方を学んでいきます。

### 1.2 学習・教育到達目標

- 不確実性の科学かつ不確実性の下での科学である統計学の考え方を理解できるようになる
- 統計学で用いる確率分布を実際の現象と関連付け、適切な確率分布を選択できるようになる
- 統計学の基礎的な推測方法（推定、区間推定、仮設検定）について説明できるようになる
- 基礎的な統計解析を統計ソフトを用いて解析できるようになる

## 2 統計学の役割

統計学と言っても、扱うデータや対象は様々な種類があります。この授業では、限られた時間の中で基礎的なことしか扱えませんが、基礎なくして発展はありません。以下は、統計学で扱う幾つかの例です。この他にも多くのことに適用できますが、まずは代表的な例を挙げます。

- 母集団の様子をサンプリング（一部を調べて全体の様子を知ること）を通して調べる
- 野外調査やアンケート調査の設計を行う
- 確率的な現象を数学的に表現する（統計的モデリング）
- 未知の量について推定を行い、その推定精度についても付与する（統計的推定）
- 仮説など主張したいことについて、観測データに基づき客観的な方法で推論する（統計的検定）
- データの視覚化と仮説・関係性探索（探索的データ解析）
- 現象や観測値の因果関係を知る（回帰分析）
- 高次元のデータから情報を集約する（次元縮小）
- 時間的に変化する対象の構造を追う（時系列解析）
- 空間的な特性を追う（空間解析、時空間解析）
- ...

## 3 この授業の内容で扱う例題（いくつか抜粋）

幾つかの例題を通して、統計学がどのような場面で利用できるかを見ていきます。詳細は今後の授業中に説明する予定です（今回はあくまで頭出しとして）。

### 3.1 例題 1: 魚の体長の母集団の推測

例題 1: 魚の体長の母集団の推測

養殖場で商用に飼育している魚種があり、体長 30cm 以上が商品サイズとして考えられているとする。飼育場の母集団における 30cm 以上の個体数の割合を知りたいとする。どうすればそれを知ることができるだろうか？魚はすべて同じ年齢とする。

東京海洋大学の大泉ステーションではニジマスを飼育している。冬に孵化した卵の状態から 1 年半経過した夏には、体長が 20 センチ近くまで成長する。成長の様子には個体差があり、体長の大きな個体もいれば、小さい個体もある。もちろん平均値に近い体長をもつ個体の頻度が高い。母集団の体長組成を知るには魚をサンプリングし、体長を測定したのち、ヒストグラムを描けば体長の組成が表現できる。さらに確率分布と呼ばれる式を利用すれば、数学的にその組成を表現することも可能である。

以下は令和元年の実データの度数分布表とヒストグラムである。非常にきれいなベル型の図である。

表1 令和元年の大泉ニジマス体長の度数分布表

| LenBin  | LenClass | LenNumber |
|---------|----------|-----------|
| (10,11] | 10.5     | 0         |
| (11,12] | 11.5     | 0         |
| (12,13] | 12.5     | 0         |
| (13,14] | 13.5     | 0         |
| (14,15] | 14.5     | 2         |
| (15,16] | 15.5     | 12        |
| (16,17] | 16.5     | 27        |
| (17,18] | 17.5     | 56        |
| (18,19] | 18.5     | 98        |
| (19,20] | 19.5     | 151       |
| (20,21] | 20.5     | 189       |
| (21,22] | 21.5     | 189       |
| (22,23] | 22.5     | 138       |
| (23,24] | 23.5     | 81        |
| (24,25] | 24.5     | 45        |
| (25,26] | 25.5     | 8         |
| (26,27] | 26.5     | 5         |
| (27,28] | 27.5     | 2         |
| (28,29] | 28.5     | 0         |
| (29,30] | 29.5     | 0         |

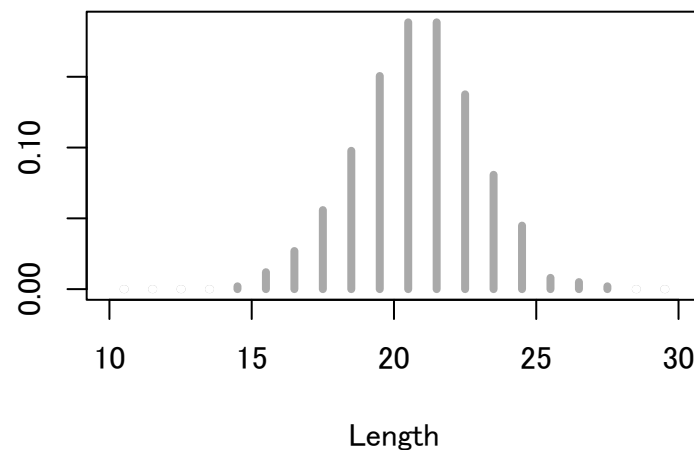


図1 大泉ニジマス体長測定結果 (R1)

これを数学的に表すための代表的な例（確率分布）として、次の正規分布 (normal distribution) がある。

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty$$

詳しくはこれからの授業で説明するのでここでは杓子定規的な説明にとどめるが、このような正規分布を  $N(\mu, \sigma^2)$  と略記し、 $\mu$  は期待値、 $\sigma^2$  は分散を表す。また関数  $f(y)$  は確率密度関数と呼ばれ、その概形は図1の通りである。上記のヒストグラムの概形と大変よく似ていることが分かる。正規分布のパラメータ  $\mu, \sigma^2$  は観測データから推定することができる。このような方法を統計的推定と呼ぶ。また、パラメータ  $\mu, \sigma^2$  の値が

分かれば、例題に答える確率  $Pr(Y \geq 30)$  の計算できる。

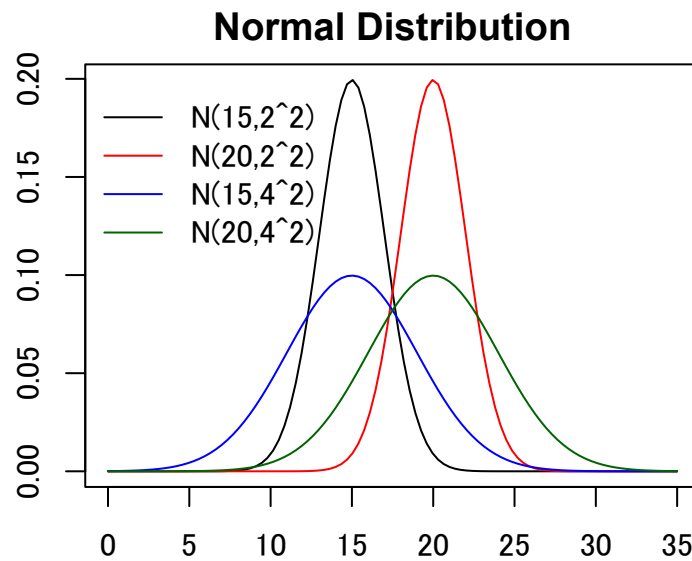


図2 正規分布の概形

なお、実際の野生生物のサンプリングでは、複数の年齢の個体が混ざった状態で観察されることも多い。大泉のニジマスの場合でも H30 年度のサンプルには 1.5 歳と 2.5 歳の個体が混ざっていたが統計的手法を用いればこれらを分解することも可能である。

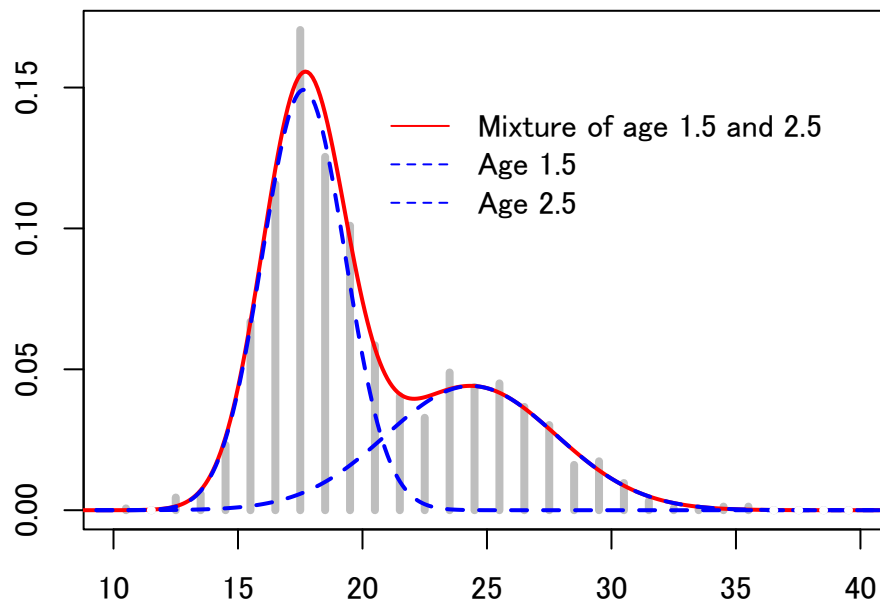


図3 大泉ニジマス体長測定結果 H30

### 3.2 例題 2: 貝の個体密度の推定

#### 例題 2: 貝の個体密度の推定

ある磯においてアワビの生息数を調べるためにサンプリングを計画したい。調査となる対称海域とサンプリングユニットを下図のように定める。また、この海域内に個体が一様にランダムに分布していると想定する。このとき、個体密度の推定誤差を変動係数 20 パーセント以下にするためには、何か所サンプリングする必要があるか？

一般に調査するエリアが広い場合、一部を調べて全体の様子を把握するサンプリングが行われる。ここでは簡略化のために同じサイズのサンプリングユニットが 100 個あるとし、そのうちいくつかをランダムに抽出し、各ユニットにおける個体数を観測することで個体密度を推定するものとする。ここで大切になるのが推定精度で、その尺度としてしばしば変動係数  $CV(= \text{標準偏差}/\text{平均}, \text{ここでは推定値の標準偏差}/\text{推定値})$  を用いる。

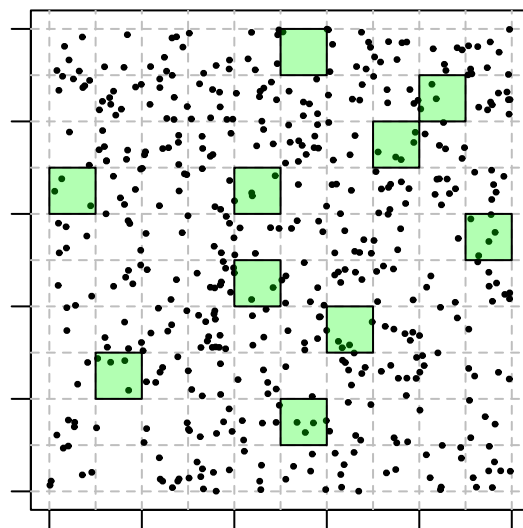


図4 サンプリングエリアとサンプリングユニット。黒丸はアワビ個体

| Estimate | CV    |
|----------|-------|
| 3.5      | 0.109 |

個体密度とサンプリングユニット数に対する推定精度（変動係数）を図 5 に示したが、多くのサンプリングユニットの個体数を観測した方が推定精度がよくなることが分かる。事前に経験的に密度が 2 以上であろう、と想定される場合に、サンプリングユニット数は 10 程度でよいことが分かる。このような推定精度の事前評価は、調査の計画に有効である。

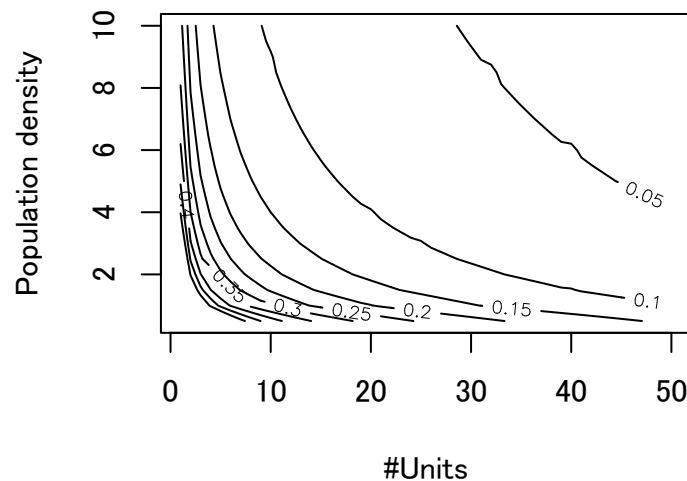


図5 個体密度とサンプリングユニット数に対する推定精度 (変動係数)

### 3.3 例題 3: データの視覚化 (魚の成長パラメータ)

例題 3: データの視覚化 (魚の成長パラメータ)

魚の大きさと成長速度に関係があるだろうか？大きく成長する魚ほど成長速度は相対的に速いのか，それとも遅いのか？この疑問に答えるべく，いくつかの魚種を抽出して調べて視覚的に表現したい。

人間と同じように魚も年齢を追うごとに体長が増加する．この様子を数学的に示す代表的なモデルが von Bertalanffy 成長式で，以下のような式で表される．

$$L_t = L_{\infty}(1 - e^{-K(t-t_0)})$$

ここで，それぞれの記号は以下に表す通り．

- $L_t$  : 年齢  $t$  における体長
- $L_{\infty}$  : 極限体長
- $K$  : 成長係数
- $t_0$  : 体長が 0 となる仮想年齢

ここでは魚の大きさの指標として極限体長  $L_{\infty}$  を，そして成長速度として成長係数  $K$  を考える．下図は成長式の概形である（いずれも  $t_0 = -0.2$ ）．

ここでは FishBase データ（魚種ごとに生物学的情報などを集約した Website）を利用した成長パラメータの可視化の例を示す．この例だけ，参考までに R コードを示しておく．

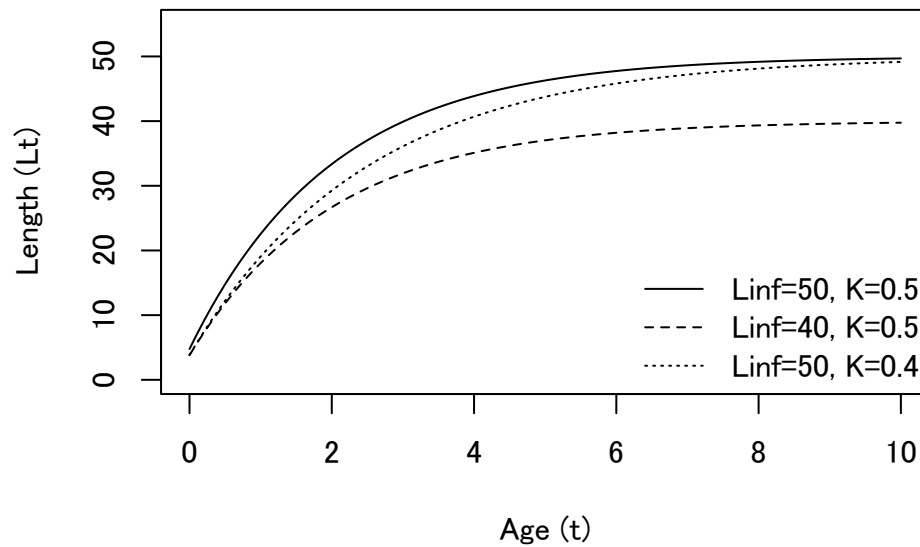


図6 成長式の概形

```
library(rfishbase)
library(kableExtra)
library(tidyverse)

# マグロ類の学名を抽出
fish.name <- unique(common_to_sci("tuna")$Species)
# ハゼを除く (futuna が混じっている)
fish.name <- fish.name[-c(18,19)]

Data <- popgrowth(
  species_list=fish.name,
  fields=c("Species", "Loo", "K", "to", "M", "Locality")
) %>% as.data.frame()

write.csv(Data, file="growth.tuna.csv")
Data <- read.csv("growth.tuna.csv") # 情報を数値データに手早く変換するために
Data %>% filter(Species=="Thunnus alalunga") %>%
  kable("latex", booktabs=T, caption="Growth parameters for albacore") %>%
  kable_styling(font_size=7, latex_options=c("striped", "hold_position"))
```

表2 Growth parameters for albacore

| X  | Species          | Loo    | K     | to      | M    | Locality                                       |
|----|------------------|--------|-------|---------|------|------------------------------------------------|
| 1  | Thunnus alalunga | 141.70 | 0.090 | NA      | NA   | New zealand waters                             |
| 2  | Thunnus alalunga | 121.00 | 0.134 | -1.9220 | NA   | New Zealand waters                             |
| 3  | Thunnus alalunga | 104.00 | 0.340 | -0.5300 | NA   | central and western North Pacific              |
| 4  | Thunnus alalunga | 114.00 | 0.253 | -1.0100 | NA   | central and western North Pacific              |
| 5  | Thunnus alalunga | 100.00 | 0.200 | NA      | NA   | South Pacific                                  |
| 6  | Thunnus alalunga | 124.00 | 0.164 | -2.2390 | NA   | North Pacific                                  |
| 7  | Thunnus alalunga | 94.70  | 0.258 | -1.3540 | NA   | Aegean and Ionian Seas                         |
| 8  | Thunnus alalunga | 125.00 | 0.230 | -0.9900 | NA   | Indian Ocean                                   |
| 9  | Thunnus alalunga | 147.00 | 0.130 | NA      | NA   | Indian Ocean                                   |
| 10 | Thunnus alalunga | 102.00 | 0.280 | NA      | NA   | South Pacific                                  |
| 11 | Thunnus alalunga | 128.00 | 0.160 | -0.9000 | NA   | Indian Ocean                                   |
| 12 | Thunnus alalunga | 127.00 | 0.120 | -2.3100 | NA   | South Pacific                                  |
| 13 | Thunnus alalunga | 142.00 | 0.150 | -0.6700 | NA   | Indian Oean                                    |
| 14 | Thunnus alalunga | 164.00 | 0.100 | -2.0700 | NA   | South Atlantic                                 |
| 15 | Thunnus alalunga | 192.00 | 0.060 | -3.3000 | NA   | South Pacific                                  |
| 16 | Thunnus alalunga | 86.00  | 0.370 | -0.7600 | NA   | Aegean Sea                                     |
| 17 | Thunnus alalunga | 124.74 | 0.230 | -0.9892 | NA   | North and South Atlantic                       |
| 18 | Thunnus alalunga | 142.28 | 0.145 | -0.6740 | NA   | South Atlantic                                 |
| 19 | Thunnus alalunga | 90.50  | 0.314 | -1.0120 | NA   | Taranto G. & N. Aegean Sea                     |
| 20 | Thunnus alalunga | 78.20  | 0.540 | -0.3750 | NA   | Taranto G. & N. Aegean Sea                     |
| 21 | Thunnus alalunga | 109.00 | 0.233 | -2.1400 | NA   | off California                                 |
| 22 | Thunnus alalunga | 118.00 | 0.262 | -1.3100 | NA   | Eastern Atlantic (off Brittany?)               |
| 23 | Thunnus alalunga | 134.00 | 0.183 | -0.3500 | 0.20 | Bay of Biscay                                  |
| 24 | Thunnus alalunga | 135.00 | 0.138 | NA      | NA   | Gulf of Gascogne (45<b0>30' N, 6<b0> W)        |
| 25 | Thunnus alalunga | 135.00 | 0.190 | NA      | NA   | Tropical Atlantic                              |
| 26 | Thunnus alalunga | 104.10 | 0.139 | -4.8800 | NA   | N. Ionian & S. Adriatic Sea                    |
| 27 | Thunnus alalunga | 145.00 | 0.159 | -3.0600 | NA   | North Pacific Current 28-40<b0>N               |
| 28 | Thunnus alalunga | 150.00 | 0.119 | NA      | NA   | Bay of Biscay and off Ireland, 50<b0>N 10<b0>W |
| 29 | Thunnus alalunga | 151.00 | 0.118 | NA      | NA   | NA                                             |
| 30 | Thunnus alalunga | 189.00 | 0.093 | NA      | NA   | Central North Pacific                          |
| 31 | Thunnus alalunga | 134.00 | 0.172 | -1.8700 | 0.22 | North Pacific 40<b0>N 150<b0>W                 |
| 32 | Thunnus alalunga | 143.00 | 0.183 | NA      | 0.20 | Eastern Atlantic                               |
| 33 | Thunnus alalunga | 104.00 | 0.274 | NA      | NA   | Catabro-Galicia Region                         |
| 34 | Thunnus alalunga | 140.00 | 0.141 | NA      | 0.23 | Central Atlantic                               |
| 35 | Thunnus alalunga | 146.00 | 0.150 | -0.8600 | NA   | NA                                             |
| 36 | Thunnus alalunga | 141.00 | 0.146 | NA      | NA   | NA                                             |
| 37 | Thunnus alalunga | 140.00 | 0.130 | NA      | NA   | Bay of Biscay                                  |
| 38 | Thunnus alalunga | 163.00 | 0.100 | NA      | NA   | NA                                             |
| 39 | Thunnus alalunga | 98.60  | 0.406 | -0.6260 | NA   | Tyrrhenian Sea                                 |
| 40 | Thunnus alalunga | 113.00 | 0.370 | 0.1400  | NA   | pacific                                        |

まず、極限体長と成長係数を魚種を跨いでモノトーンでプロットしてみた。二つの層があるように見える。また、極限体長と成長係数に負の相関があるようにも見える。

```
tmp <-  
  Data %>% ggplot(aes(x=Loo, y=K)) + geom_point() +  
  xlab("Asymptotic length (Linf)") + ylab("Growth coefficient (K)") +  
  labs(title="Relationship of growth parameters (Linf and K)")  
tmp
```

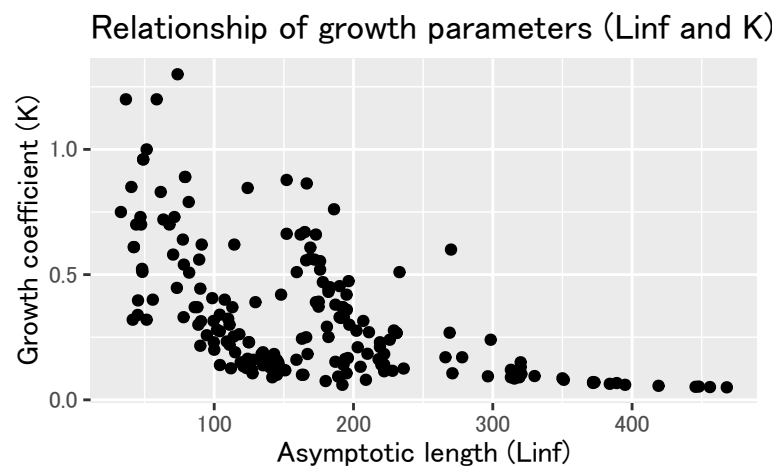
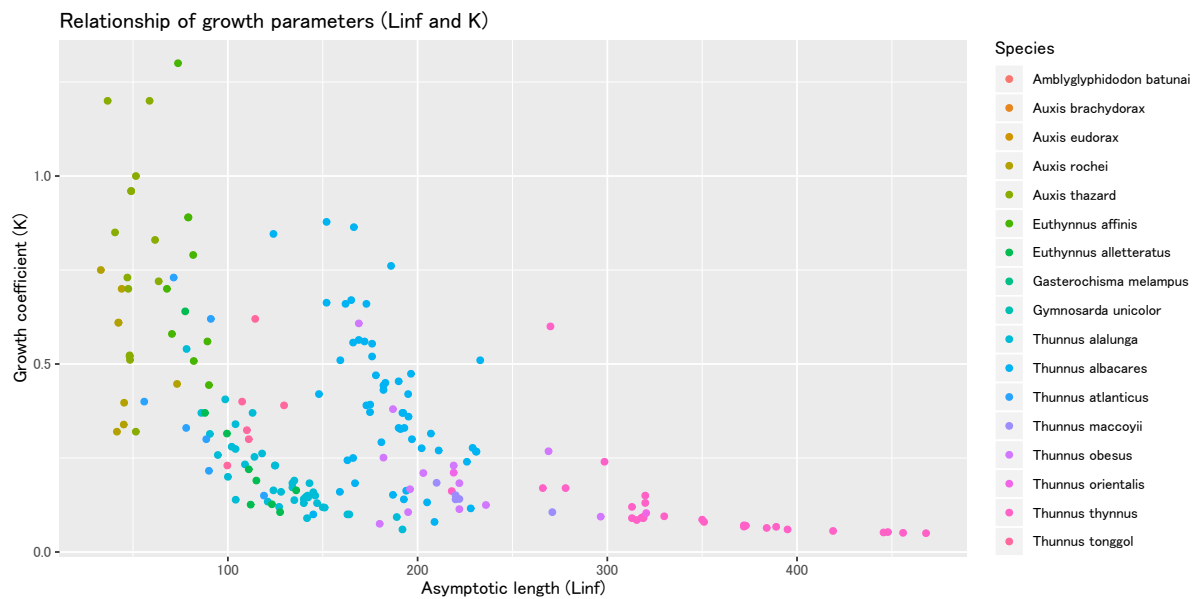


図7 成長パラメータのプロット (Linf,K)

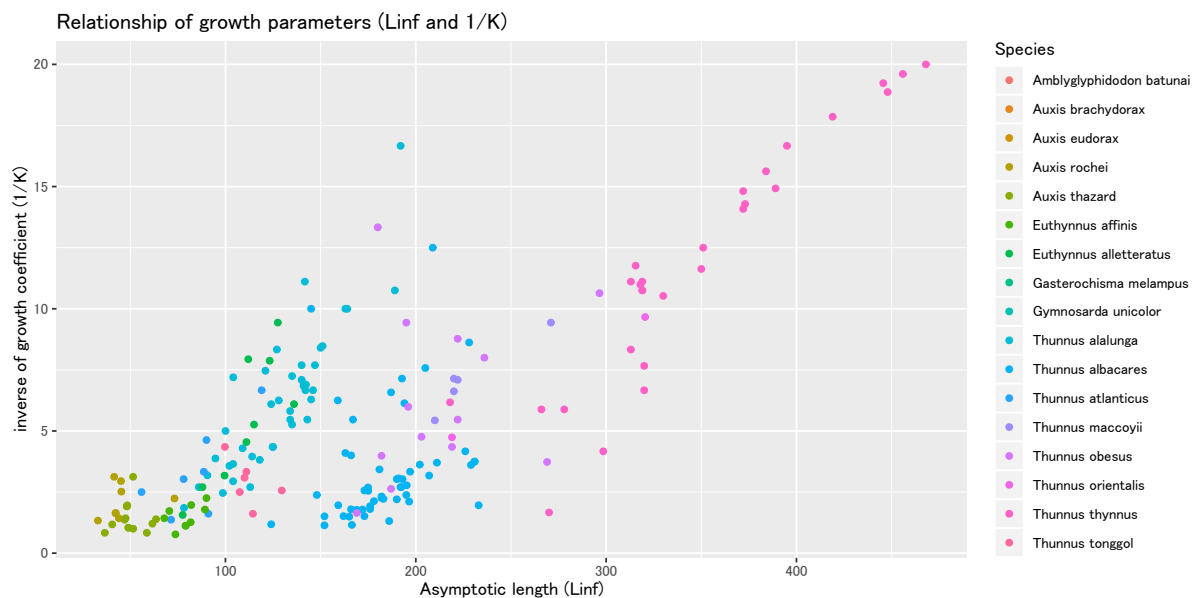
次にこれを魚種別にカラーでプロットしてみると関係がだんだん明らかになってくる。なお、同じ魚種で複数の点がプロットされているのは、複数の独立した研究があることによる。

```
tmp <-  
  Data %>% ggplot(aes(x=Loo, y=K, color=Species)) + geom_point() +  
  xlab("Asymptotic length (Linf)") + ylab("Growth coefficient (K)") +  
  labs(title="Relationship of growth parameters (Linf and K)")  
tmp
```



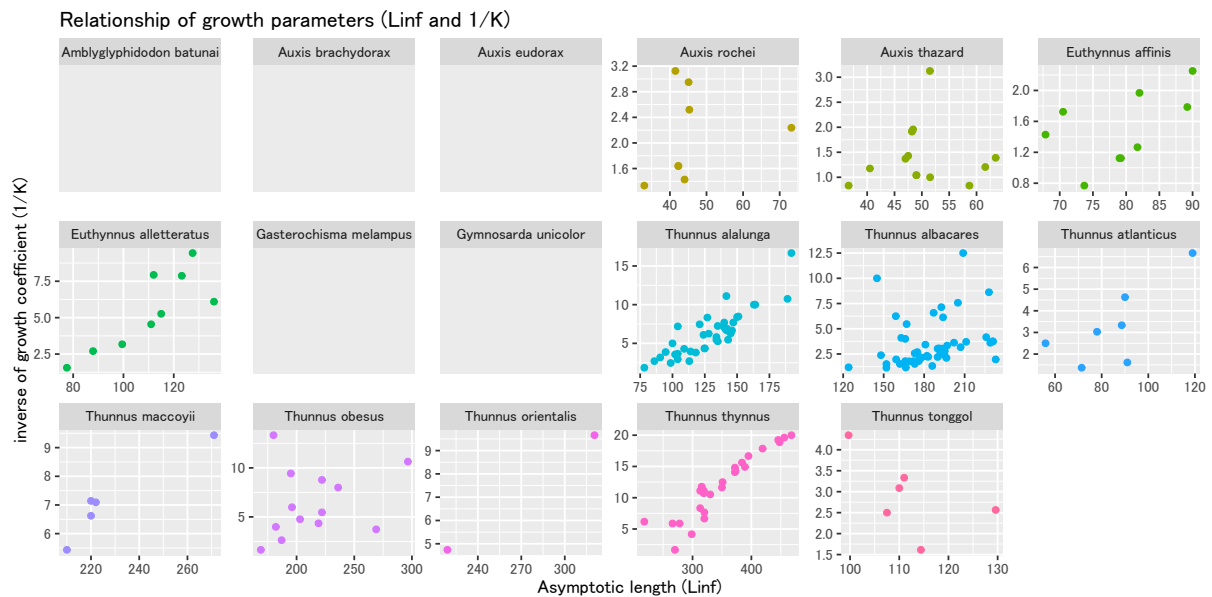
今度は「極限体長」と「成長係数の逆数」の関係をみると、各層で綺麗な直線関係が示唆される。

```
tmp <-
  Data %>% ggplot(aes(x=Loo, y=1/K, color=Species)) + geom_point() +
  xlab("Asymptotic length (Linf)") + ylab("inverse of growth coefficient (1/K)") +
  labs(title="Relationship of growth parameters (Linf and 1/K)")
tmp
```



さらに魚種毎に分割したければ、以下のように示せる。

```
tmp + facet_wrap(~Species, ncol=6, scales="free") + guides(color="none")
```



ちょっとした数行の R コードで可視化が可能となる．コーディングの詳細は授業の後半戦で紹介予定である．

## 4 統計データや解析結果に騙されない

有名なシンプソンのパラドックスについて，視聴教材をご覧ください．

## 5 最後に

この授業では，上記以外にもなるべく多くの事例を取り扱っていきたくと考えている．

- 「保護区の検証～効果はあったか？」
- 「ある生物資源は環境インパクトのためにメス化が進行？」
- 「ある生物集団では遺伝的多様性が失われてきている？」
- 「飼料Aと飼料Bではどちらが成長効率がよい？」
- 「喫煙は心拍数を高める？」
- 「LED 集魚灯は従来のものと比較して効果的？」
- 「〇〇海域に生息する ×× クジラは，むかし枯渇の危機に瀕していたか？いまは回復？」
- 「日本沿岸と沖合に回遊してくる ×× 種は同じ遺伝的に同じ集団（系群）か？」
- 「陸上 100 メートル走においてピストルへの反応速度は選手で異なるか？」
- 「プロ野球のボールは以前と比較して飛ぶようになったか？」
- 「サッカーではホームチームが勝つ確率が高い？」