バイオスタティスティクス基礎論
第 1 回講義テキスト

岩田洋佳 hiroiwata@g.ecc.u-tokyo.ac.jp

2020/4/17

はじめに

最近では、農学や生命科学の分野において、様々な種類のデータが大量に収集・蓄積されるようになってきています。こうしたデータに潜む未発見の「知」を見逃さずに確実に引き出すためには、研究の目的やデータのもつ性質に適した方法を用いてデータを解析する必要があります。　統計解析には様々な手法がありますが、各手法の特徴を把握し、解析の原理を理解し、得られた解析結果を適切に解釈できるようになるためには、相応の学習を必要とします。また、その学習をより効果的なものにするには、実際のデータを自分で解析してみるという経験も不可欠です。自分で計測したデータを解析してはじめて、講義や参考書で学んだことが明瞭に理解できるようになることは少なくなりません。　本講義は、受講生の皆さんが自らデータ解析を行い、統計解析のスキルを高めていくための「最初の第一歩」を提供することを目的としています。具体的には、今後の研究で必要となると考えられるいくつかの統計手法について、R を使った実践的なデータ解析の方法に重点をおいて解説していきます。本講義の目標は、回帰分析、分散分析、主成分分析などの汎用的な統計解析手法について、それを自分のデータ解析に利用するためのスキルを身につけること、さらには、より発展したデータ解析を行うための足場をかためることです。全 4 回と短い講義ではありますが、統計解析の面白さや巧みさについて興味をもってもらえるように講義を進めていきたいと思っています。

R

R は統計解析のためのフリーソフトウエアです（少しだけ正確にいうと、R とはコンピュータ言語の名称であり、パソコン上にソフトウェアとしてインストールされる R は R 言語を利用するための“環境”となります）。R には数多くの機能が備わっており、その利用場面は、統計解析だけでなく、データの前処理から、データの俯瞰、さらには、論文用のグラフ作成にまで及びます。また、パッケージ（package）として配布されている拡張プログラムをインストールすることで、様々な解析を容易に実行することができます。新しく開発された統計手法が R では比較的早く利用できるようになります。このようなことから、R を使うためのスキルは、農学や生命科学の研究者にとって非常に有用なものとなってきています。　なお、R については、現在、非常に多くの参考書が出版されています。私のおすすめの入門書は、以下の通りです。

- Peter Dalgaard 著、Introductory Statistics with R (Statistics and Computing) Second Edition, Springer, 2008, ISBN: 978-0387790534
- Brian Everitt, Torsten Hothorn 著、An Introduction to Applied Multivariate Analysis with R (Use R!), Springer, 2011, ISBN: 978-1441996497

R を用いた簡単な計算

R では、基本的には、コマンド（命令文）を順次入力しながら対話的に解析を進めていきます（ただし、実際に解析を行う場合は、R スクリプトとして一連のコマンドを先に入力しておき、それを実行する方が部分的修正や履歴の確認ができて便利です）。　ここでは、コマンド入力で簡単な計算を行いながら、R に慣れるところから始めて見ましょう。

R の最も簡単な利用方法は、簡単な算術表現を入力し、その答えを得ることです。例えば、

```
3 + 5 * 3
```

```
## [1] 18
```

得られた結果をもとに次の計算をしたい場合には、次のように値を変数に代入しておきます。

```
x <- 1 + 2
x
```

```
## [1] 3
```

代入しておいた値は、変数名を介して別の計算に用いることができます。

```
x + 5 * x
```

```
## [1] 18
```

関数を用いて様々な計算を行うことができます。

```
abs(x)
```

```
## [1] 3
```

```
sin(x)
```

```
## [1] 0.14112
```

```
atan(x)
```

```
## [1] 1.249046
```

```
log(x)
```

```
## [1] 1.098612
```

```
log10(x)
```

```
## [1] 0.4771213
```

では、少し複雑な計算をしてみましょう。平均 、分散 の正規分布の確率密度関数（図 1）は、

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

ですが、これを R で計算してみましょう。

```
mu <- 3
s2 <- 2
x <- 5
1 / sqrt(2 * pi * s2) * exp(- (x - mu)^2 / (2 * s2))
```

```
## [1] 0.1037769
```

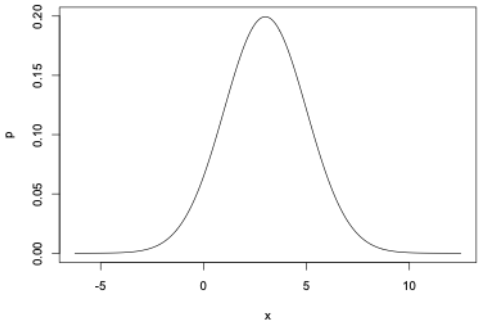


Fig1. 平均3, 分散2 の正規分布

確認のために正規分布の確率密度を計算する関数 dnorm で計算してみると同じ値が得られます。

```
dnorm(x, mu, sqrt(s2))
```

```
## [1] 0.1037769
```

Quiz1

では、ここで練習問題を解いてみましょう。練習問題は、授業中に提示します。

https://www.menti.com/ に移動して、361599 と入力して下さい。その後、ニックネームを登録してクイズが始まるまで待機していて下さい。

ベクトルや行列を用いた計算

R の優れた点のひとつは、ベクトルや行列の演算を非常に簡単に実行できることです。ここでは、ベクトルや行列の演算を用いていくつかの要約等計量を計算してみましょう。

例えば、6 個の数値からなるベクトルを以下のように簡単に作成できます。なお、このデータは、6 品種・系統のイネの籾長を mm 単位で計測したデータです（データの出典は後述します）。

```
length <- c(8.1, 7.7, 8.2, 9.7, 7.1, 7.3)    # mm scale
length
```

```
## [1] 8.1 7.7 8.2 9.7 7.1 7.3
```

同じ品種・系統の籾幅を計測したデータも入力し、籾長と籾幅の比を計算します。

```
width <- c(3.7, 3.0, 2.9, 2.4, 3.3, 2.5)
ratio <- length / width
ratio
```

```
## [1] 2.189189 2.566667 2.827586 4.041667 2.151515 2.920000
```

まず、籾長と籾幅の比の平均を計算してみましょう。母平均の推定値は、

$$\sum_{i=1}^{n} x_i /n$$

として計算できます。ここで、 は i 番目のサンプルの値、n はサンプル数です。

```
sum(ratio)
```

```
## [1] 16.69662
```

```
length(ratio)
```

```
## [1] 6
```

```
sum(ratio) / length(ratio)
```

```
## [1] 2.782771
```

平均は、関数 mean を使って計算できます。

```
mean(ratio)
```

```
## [1] 2.782771
```

次に、分散を計算してみましょう。母分散の推定値は、

$$\sum_{i=1}^{n} (x_i - \overline{x})^2/(n-1)$$

として計算できます。ここで、$\overline{x}$は先ほど計算した平均です。

```
xbar <- mean(ratio)
(ratio - xbar)^2
```

```
## [1] 0.352338947 0.046700930 0.002008434 1.584819189 0.398483500 0.01883189
5
```

```
sum((ratio - xbar)^2)
```

```
## [1] 2.403183
```

```
sum((ratio - xbar)^2) / (length(ratio) - 1)
```

```
## [1] 0.4806366
```

分散は、関数 var を使って計算できます。

```
var(ratio)
```

```
## [1] 0.4806366
```

次に、共分散を計算してみましょう。2 変量 x と y 間の共分散の推定値は、

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})/(n-1)$$

として計算できます。ここで、$\overline{x}$および$\overline{y}$は各変量の平均を表します。

```
xbar <- mean(length)
ybar <- mean(width)
sum((length - xbar) * (width - ybar)) / (length(length) - 1)

## [1] -0.1773333
```

なお、R の関数 cov を使って共分散を計算することもできます。

```
cov(length, width)

## [1] -0.1773333
```

共分散に続いて Pearson の積率相関係数（以下、相関係数）を計算してみましょう。相関係数を式で書くと、

$$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

となります。

```
s12 <- sum((length - xbar) * (width - ybar))
s1 <- sum((length - xbar)^2)
s2 <- sum((width - ybar)^2)
s12 / (sqrt(s1) * sqrt(s2))

## [1] -0.3901388
```

式を見て分かるように、相関係数は共分散を両変数の標準偏差で割ったかたちになっています。実際に計算して確認してみましょう。

```
cov(length, width) / (sd(length) * sd(width))

## [1] -0.3901388
```

相関係数では、両変数の標準偏差で割ることにより基準化してあるために、共分散と異なり、計測値のスケールに影響されずに変数間の関係を把握できます。したがって、異なる尺度（重さと長さなど）で計測された変数間で関係の強さを比較するのに適しています。

なお、R の関数 cor を使って相関係数を計算することもできます。

```
cor(length, width)

## [1] -0.3901388
```

では、行列計算を用いて分散と共分散を計算してみましょう。まずは、length と width を結合して6×2の行列を作成します。

```
x <- cbind(length, width)
x
```

```
##      length width
## [1,]    8.1   3.7
## [2,]    7.7   3.0
## [3,]    8.2   2.9
## [4,]    9.7   2.4
## [5,]    7.1   3.3
## [6,]    7.3   2.5
```

次に、関数 apply を用いて各列の平均を求めます。

```
m <- apply(x, 2, mean)
m

##   length    width
## 8.016667 2.966667
```

求めた列平均を各列から引き算します。

```
z <- sweep(x, 2, m)
z

##           length       width
## [1,]  0.08333333  0.73333333
## [2,] -0.31666667  0.03333333
## [3,]  0.18333333 -0.06666667
## [4,]  1.68333333 -0.56666667
## [5,] -0.91666667  0.33333333
## [6,] -0.71666667 -0.46666667
```

あとは行列の積を用いることで分散と共分散（分散共分散行列）を計算できます。

```
t(z) %*% z / (nrow(z) - 1)

##            length      width
## length  0.8656667 -0.1773333
## width  -0.1773333  0.2386667
```

対角成分が分散、非対角成分が共分散です。

分散共分散行列は関数 cov で計算することができます。

```
cov(x)

##            length      width
## length  0.8656667 -0.1773333
## width  -0.1773333  0.2386667
```

Quiz2

では、ここで練習問題を解いてみましょう。練習問題は、授業中に提示します。

クイズのページを閉じてしまった人は、https://www.menti.com/ に移動して、361599 と入力して下さい。その後、ニックネームを登録してクイズが始まるまで待機していて下さい。

外部データを読み込んで解析する

自分の研究のために R を利用する場合は、表計算ソフト等で整理されたデータを読み込んで解析する場合がほとんどだと思います。ここでは、他のソフトで保存されたデータを R に読み込み解析するための手順を説明します。　なお、ここでは、Zhao ら（2011; Nature Communications 2:467）がイネ遺伝資源を用いたゲノムワイドアソシエーション解析に用いられたデータ（Rice Diversity http://www.ricediversity.org/data/　からダウンロードできる）をデータ例として用います。

csv 形式で保存されたファイルの読み込みには read.csv という関数を用います。

```
pheno <- read.csv("RiceDiversityPheno.csv")
```

読み込んだデータのサイズやデータの一部を確認するには以下のようにします。

```
dim(pheno)

## [1] 413  38

head(pheno[,1:4])

##        HybID NSFTVID Flowering.time.at.Arkansas Flowering.time.at.Faridpur
## 1 081215-A05       1                   75.08333                          64
## 2 081215-A06       3                   89.50000                          66
## 3 081215-A07       4                   94.50000                          67
## 4 081215-A08       5                   87.50000                          70
## 5 090414-A09       6                   89.08333                          73
## 6 090414-A10       7                  105.00000                          NA
```

このデータには、各遺伝資源の由来などが記述されたファイルが別に存在します。ここでは、そのファイルを読み込んで pheno データに結合してみます。まずは、ファイルを読み込みます。

```
line <- read.csv("RiceDiversityLine.csv")
head(line)

##   GSOR.ID        IRGC.ID NSFTV.ID Accession.Name Country.of.origin  Latitu
de
## 1  301001 To be assigned        1       Agostano             Italy 41.8719
40
## 2  301003         117636        3  Ai-Chiao-Hong             China 27.9025
27
## 3  301004         117601        4       NSF-TV 4             India 22.9030
81
## 4  301005         117641        5       NSF-TV 5             India 30.4726
64
## 5  301006         117603        6       ARC 7229             India 22.9030
81
## 6  301007 To be assigned        7          Arias         Indonesia -0.7892
75
##   Longitude Sub.population     PC1     PC2     PC3     PC4
## 1  12.56738            TEJ -0.0486  0.0030  0.0752 -0.0076
## 2 116.87256            IND  0.0672 -0.0733  0.0094 -0.0005
## 3  87.12158            AUS  0.0544  0.0681 -0.0062 -0.0369
## 4  75.34424       AROMATIC -0.0073  0.0224 -0.0121  0.2602
```

```
## 5  87.12158            AUS  0.0509  0.0655 -0.0058 -0.0378
## 6 113.92133            TRJ -0.0293 -0.0027 -0.0677 -0.0085
```

line データの NSFTV.ID と pheno データの NSFTVID が対応しているので、この列の情報をもとに 2 つのデータを結合します。

```
data <- merge(line, pheno, by.x = "NSFTV.ID", by.y = "NSFTVID")
head(data[,1:14])
```

```
##   NSFTV.ID GSOR.ID        IRGC.ID Accession.Name Country.of.origin  Latitu
de
## 1        1  301001 To be assigned       Agostano             Italy 41.8719
40
## 2        3  301003         117636  Ai-Chiao-Hong             China 27.9025
27
## 3        4  301004         117601       NSF-TV 4             India 22.9030
81
## 4        5  301005         117641       NSF-TV 5             India 30.4726
64
## 5        6  301006         117603       ARC 7229             India 22.9030
81
## 6        7  301007 To be assigned          Arias         Indonesia -0.7892
75
##   Longitude Sub.population     PC1     PC2     PC3     PC4      HybID
## 1  12.56738            TEJ -0.0486  0.0030  0.0752 -0.0076 081215-A05
## 2 116.87256            IND  0.0672 -0.0733  0.0094 -0.0005 081215-A06
## 3  87.12158            AUS  0.0544  0.0681 -0.0062 -0.0369 081215-A07
## 4  75.34424       AROMATIC -0.0073  0.0224 -0.0121  0.2602 081215-A08
## 5  87.12158            AUS  0.0509  0.0655 -0.0058 -0.0378 090414-A09
## 6 113.92133            TRJ -0.0293 -0.0027 -0.0677 -0.0085 090414-A10
##   Flowering.time.at.Arkansas
## 1                   75.08333
## 2                   89.50000
## 3                   94.50000
## 4                   87.50000
## 5                   89.08333
## 6                  105.00000
```

Quiz3

では、ここで練習問題を解いてみましょう。練習問題は、授業中に提示します。

クイズのページを閉じてしまった人は、https://www.menti.com/ に移動して、361599 と入力して下さい。

読み込んだデータの解析

計測データセットには、実験上の都合から欠測したデータが含まれる場合が少なくありません。また、数少ない変数について解析をするだけでなく、たくさんの変数についてその分布や変数間の関係をみたい場合が少なくありません。ここでは、先ほど読み込んだデータを用いて、データ解析を行ってみましょう。

では、先ほどと同じようにして、籾長と籾幅の比を計算し、その平均を計算してみましょう。

```
ratio <- data$Seed.length / data$Seed.width
mean(ratio)

## [1] NA
```

すると NA と表示されるだけで平均が計算できません。何故でしょうか。

これは、ratio に欠測値（R では NA として表す）が含まれているためです。

```
ratio[1:14]

##  [1] 2.188254 2.610704 2.814950 4.075973 2.168927 2.905327 3.229055 2.2706
83
##  [9] 2.430681       NA 3.418122 3.061198 4.024255       NA
```

このような場合は、na.rm というオプションを指定して計算します。

```
mean(ratio, na.rm = T)

## [1] 2.752084
```

data 内の全ての変数について平均を求めるには以下のようにします。ここでは、1 番目から 14 番目のデータについて計算します。

```
sapply(data[, 1:14], mean, na.rm = T)

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

##                   NSFTV.ID                   GSOR.ID
##               2.340896e+02              3.016027e+05
##                    IRGC.ID             Accession.Name
##                          NA                         NA
##          Country.of.origin                   Latitude
##                          NA              2.166591e+01
##                  Longitude             Sub.population
##               4.359763e+01                         NA
##                         PC1                        PC2
##              -2.716707e-04             -1.280872e-04
##                         PC3                        PC4
##              -3.072639e-04             -2.106538e-05
##                      HybID Flowering.time.at.Arkansas
##                         NA              8.794439e+01
```

数値データでないデータについては警告メッセージが表示されて計算結果は NA となります。

なお、次のコマンドを用いると、数値（numeric）データについては平均だけでなく、四分位点、最小値、最大値が表示され、因子（factor）データについては、各階級に属するサンプルの数え上げ結果が表示されます。ここでは、1 番目から 14 番目のデータについて計算します。

```
summary(data[,1:14])
```

```
##     NSFTV.ID        GSOR.ID                  IRGC.ID          Accession.Nam
e
##  Min.   :  1.0   Min.   :301001   To be assigned: 58   Azucena      :  2
##  1st Qu.:112.0   1st Qu.:301109   117616        :  2   Carolina Gold:  2
##  Median :217.0   Median :301215   117638        :  2   Moroberekan  :  2
##  Mean   :234.1   Mean   :301603   117756        :  2   N 22         :  2
##  3rd Qu.:322.0   3rd Qu.:301318   117808        :  2   Nipponbare   :  2
##  Max.   :652.0   Max.   :312018   (Other)       :343   1021         :  1
##                  NA's   :4        NA's          :  4   (Other)      :402
##      Country.of.origin    Latitude        Longitude         Sub.population
##  United States: 39     Min.   :-38.42   Min.   :-102.553   ADMIX   :62
##  India        : 34     1st Qu.: 14.06   1st Qu.:  -7.093   AROMATIC:14
##  China        : 31     Median : 23.70   Median :  71.276   AUS     :57
##  Bangladesh   : 27     Mean   : 21.67   Mean   :  43.598   IND     :87
##  Japan        : 19     3rd Qu.: 34.66   3rd Qu.: 113.921   TEJ     :96
##  Taiwan       : 19     Max.   : 55.75   Max.   : 179.414   TRJ     :97
##  (Other)      :244     NA's   :20       NA's   :20
##       PC1                 PC2                  PC3
##  Min.   :-0.0516000   Min.   :-0.0801000   Min.   :-0.0846000
##  1st Qu.:-0.0422000   1st Qu.:-0.0090000   1st Qu.:-0.0332000
##  Median :-0.0326000   Median :-0.0027000   Median : 0.0045000
##  Mean   :-0.0002717   Mean   :-0.0001281   Mean   :-0.0003073
##  3rd Qu.: 0.0599000   3rd Qu.: 0.0023000   3rd Qu.: 0.0266000
##  Max.   : 0.0689000   Max.   : 0.1193000   Max.   : 0.0934000
##
##       PC4                                          HybID
##  Min.   :-4.140e-02   @5206720064940610241040863209221 4:  1
##  1st Qu.:-1.670e-02   @5206720064940610241040863209222 1:  1
##  Median :-9.400e-03   @5206720064940610241040863209222 5:  1
##  Mean   :-2.107e-05   @5206720064940610241040863209222 7:  1
##  3rd Qu.:-5.000e-04   @5206720064940610241040863209223 1:  1
##  Max.   : 2.784e-01   @5206720064940610241040863209223 3:  1
##                       (Other)                           :407
##  Flowering.time.at.Arkansas
##  Min.   : 54.50
##  1st Qu.: 79.75
##  Median : 87.71
##  Mean   : 87.94
```

```
##  3rd Qu.: 96.83
##  Max.   :150.50
##  NA's   :39
```

では、籾長と籾幅で相関係数を計算してみましょう。

```
cor(data$Seed.length, data$Seed.width)
```

```
## [1] NA
```

結果が NA となってしまいます。これは先ほどと同様欠測値によるものです。

欠測値に対する対処の仕方を指定して再度計算してみます。

```
cor(data$Seed.length, data$Seed.width, use = "pair")
```

```
## [1] -0.2837094
```

無事計算されました。

Quiz4

　では、ここで練習問題を解いてみましょう。練習問題は、授業中に提示します。

クイズのページを閉じてしまった人は、https://www.menti.com/ に移動して、361599 と入力して下さい。

データの視覚化

　実際に統計解析を行う前に、データをいろいろな角度から眺めてみることは非常に重要です。例えば、上述した平均や分散といった統計量は要約のための統計量であり、同じような平均や分散をもつ変数であっても、観察値の分布が大きく異なる場合もあります。したがって、まずはデータをじっくり眺めるということが、そのデータのもつ特性を理解するためにも非常に重要です。また、データの視覚化はデータ解析の結果を論文等にまとめる際にも必要です。ここでは、様々なデータ視覚化手法について説明します。

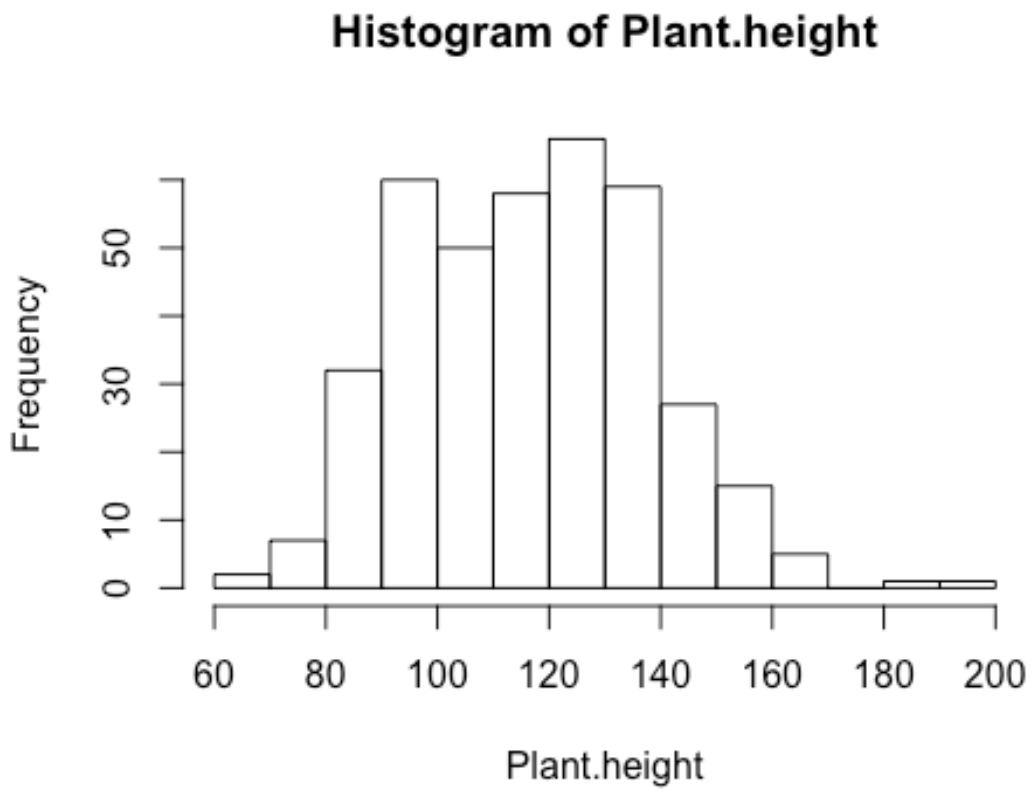
まず、視覚化手法の説明の前に、data 内にあるデータを直接呼び出せるようにしましょう。

```
attach(data)
```

こうしておくことで、例えば、これまで data*Plant.height* と指定していたところを、*data* 無しの Plant.height として入力できるようになります。

では、まずヒストグラムを描いてみましょう。

```
hist(Plant.height)
```



stem-and-leaf プロットを描いてみましょう。

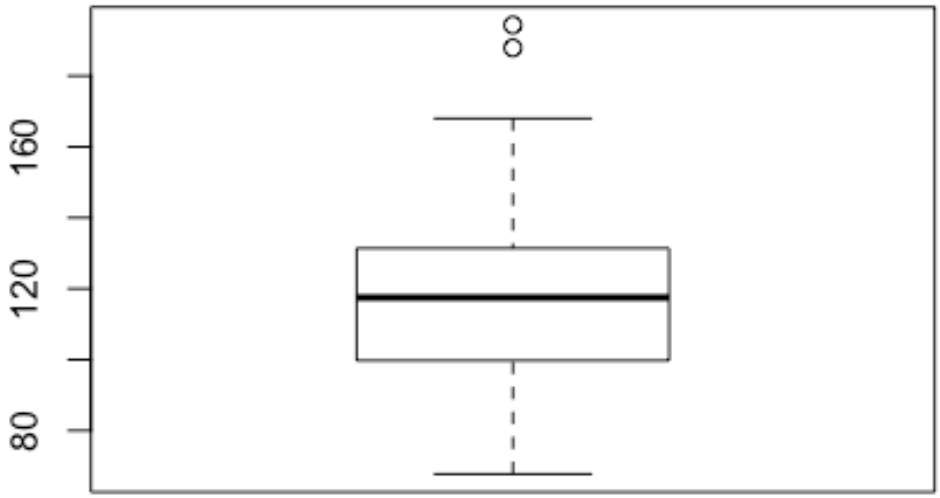
```
stem(Plant.height)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    6 | 8
##    7 | 0155578
##    8 | 012223344566677777888888899999
##    9 | 000111111111112223333344445555556666667777788888889999
##   10 | 00000000111111222222223333334445566666667777777778888899
##   11 | 000011111111222333334444444555555566677778888888899999
##   12 | 00000000111111222223333334444455556666777777777788888888999999999
##   13 | 000000001111111122222222223333444444555556666666777777999
##   14 | 00000111122223334444444566778889
##   15 | 00122223344479
##   16 | 001278
##   17 |
##   18 | 8
##   19 | 4
```

こちらは図ではなくテキスト表示で結果が示されます。

箱ひげ図（box plot）を描いてみましょう。

```
boxplot(Plant.height)
```



次に、いもち病抵抗性（Blast.resistance）についてヒストグラムを描いてみます。

```
hist(Blast.resistance)
```

うまく分布が図示できているように見えますが、実は落とし穴があります。

いもち病抵抗性データは抵抗性の強さを 9 段階（0-9）のスコアで表されています。そこで、まずは、9 段階のどの階級に何品種・系統が含まれているのか集計してみましょう。

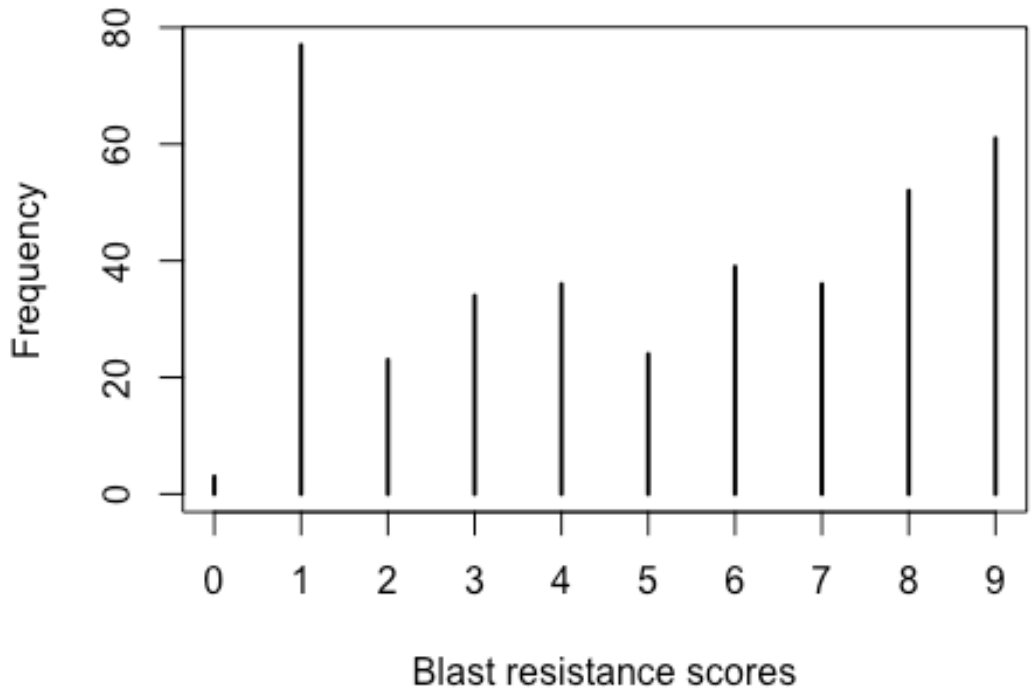
```
t <- table(Blast.resistance)
t

## Blast.resistance
##  0  1  2  3  4  5  6  7  8  9
##  3 77 23 34 36 24 39 36 52 61
```

さきほど描いたヒストグラムでは全階級をうまく表せていなかったことが分かります。

上記のように table 関数を用いて集計されたデータから、棒グラフ（bar plot）を描くことができます。

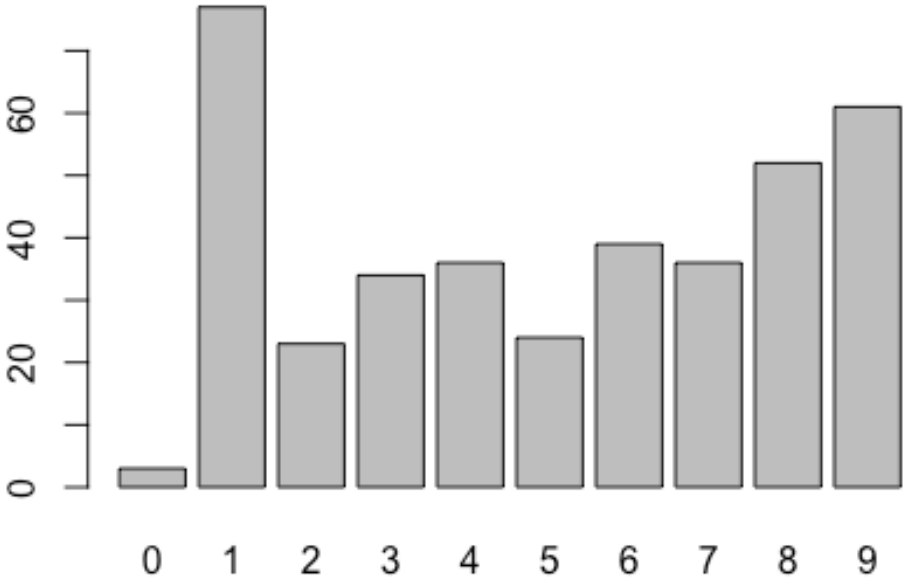
```
plot(t)
```


```
plot(t, xlab = "Blast resistance scores", ylab = "Frequency")
```



棒グラフは barplot 関数を用いて描くこともできます。ただし、上記の棒グラフと少し見た目が異なります。

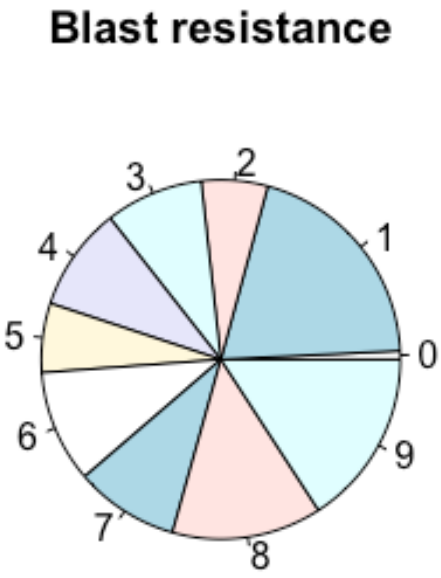
```
barplot(t)
```



円グラフを描くと各スコアの割合を図示できます。

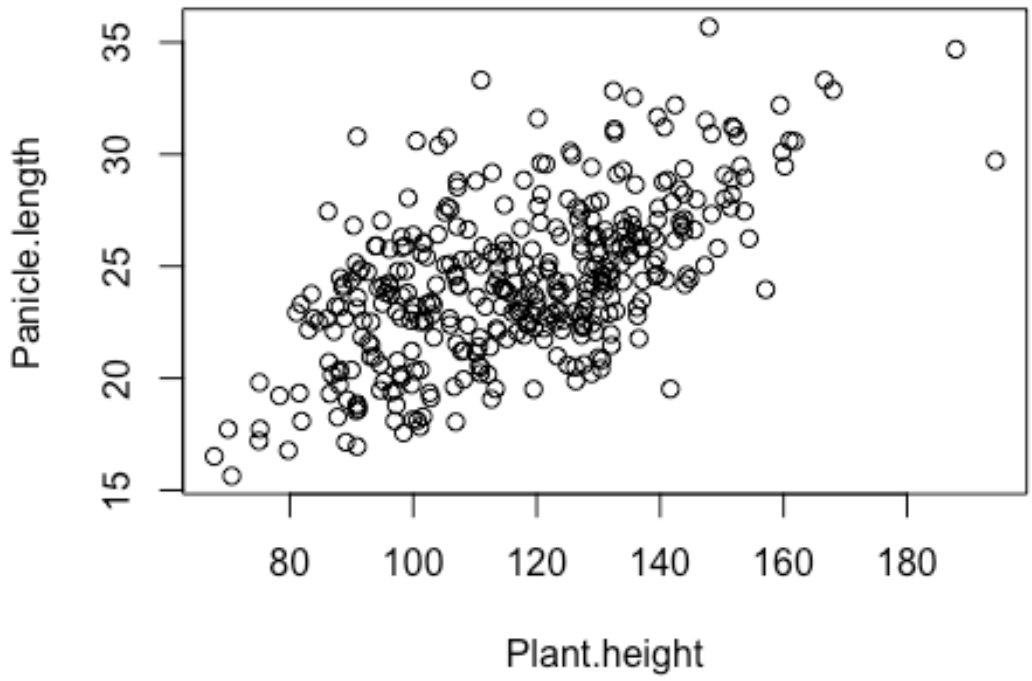
```
pie(t)
```


```
pie(t, main = "Blast resistance")
```



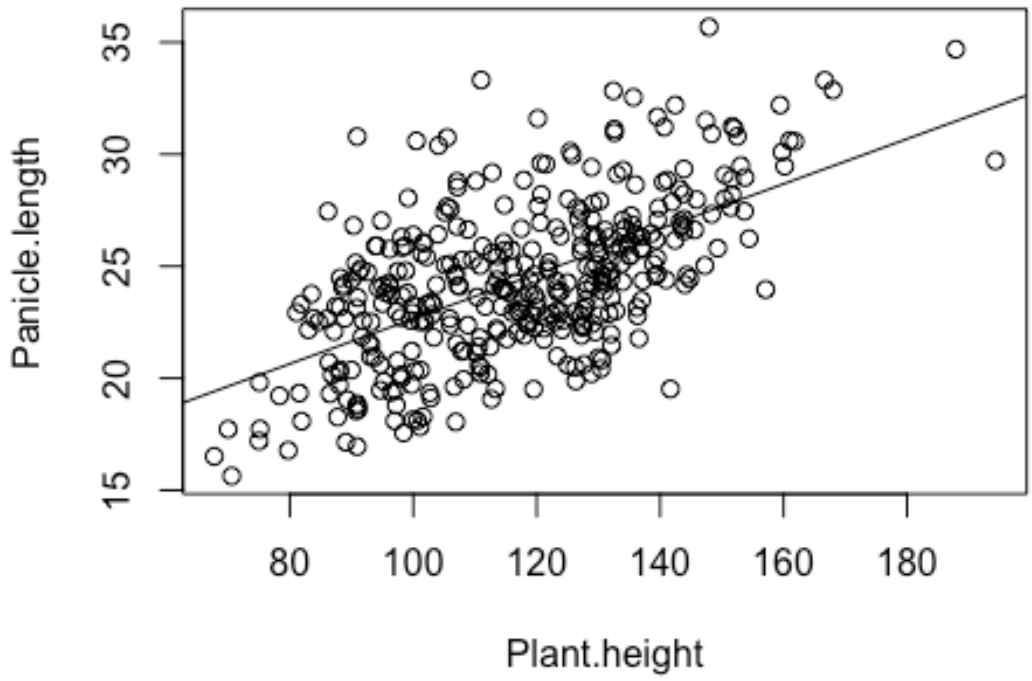
ここからは、2 変数間の関係を見て行きましょう。

```
plot(Plant.height, Panicle.length)
```



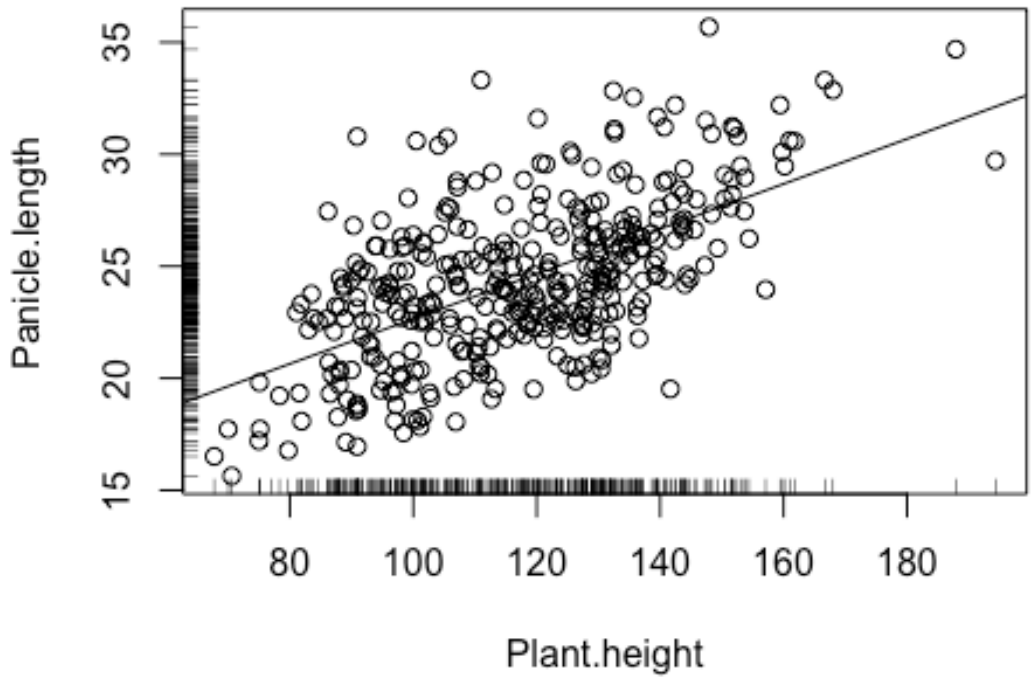
回帰分析により直線をあてはめて重ね描きします。

```
plot(Plant.height, Panicle.length)
abline(lm(Panicle.length ~ Plant.height))
```



ラグ（織物）プロットを重ね描きします。分布の疎密を視覚化するのに便利です。

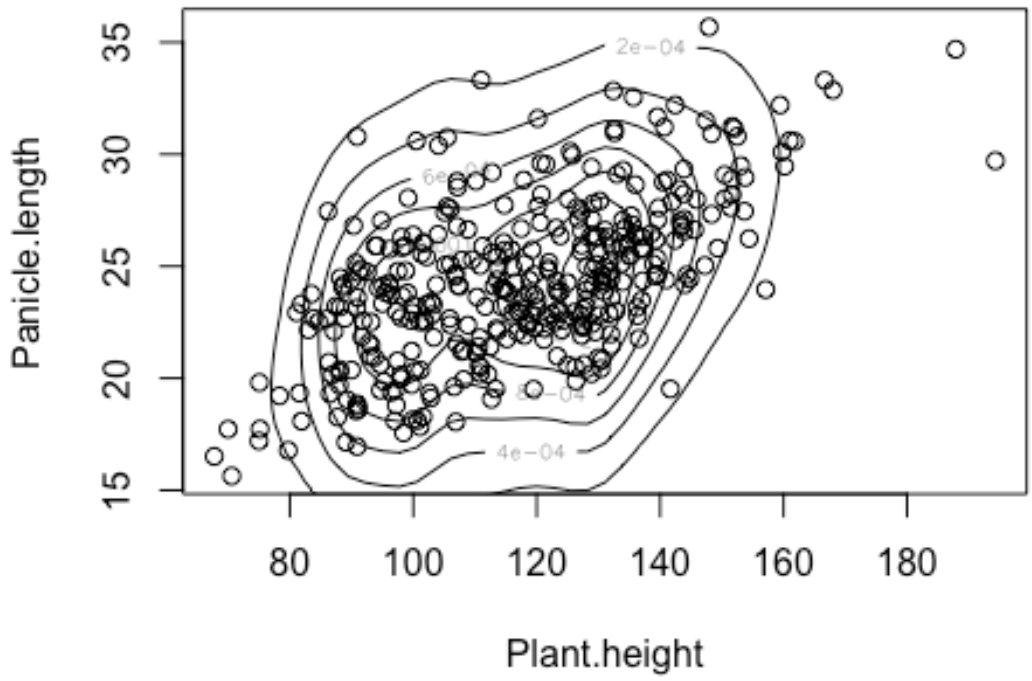
```
plot(Plant.height, Panicle.length)
abline(lm(Panicle.length ~ Plant.height))
rug(Plant.height, side = 1)
rug(Panicle.length, side = 2)
```



2 変数間の関係を、カーネルを用いた平滑化（kernel smoothing）を用いて図示してみましょう。

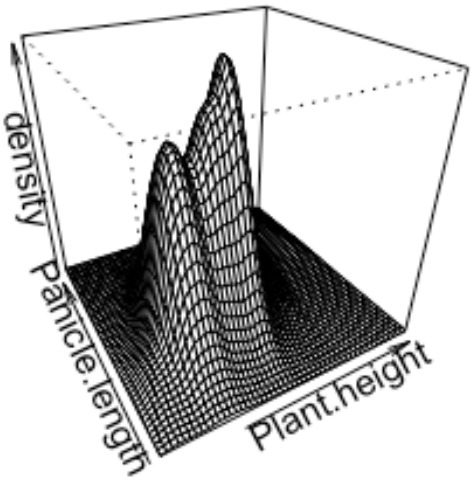
では、実行してみます。

```
library("KernSmooth")

## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009

x <- data.frame(Plant.height, Panicle.length)
x <- na.omit(x)
d <- bkde2D(x, bandwidth = 4)
plot(x)
contour(d$x1, d$x2, d$fhat, add = T)
```



等高線のように表されているのがカーネルで平滑化された点の密度です。 では、この平滑化された密度を 3 次元で表示してみましょう。

```
persp(d$x1, d$x2, d$fhat, xlab = "Plant.height", ylab = "Panicle.length", zlab = "density", theta = -30, phi = 30)
```



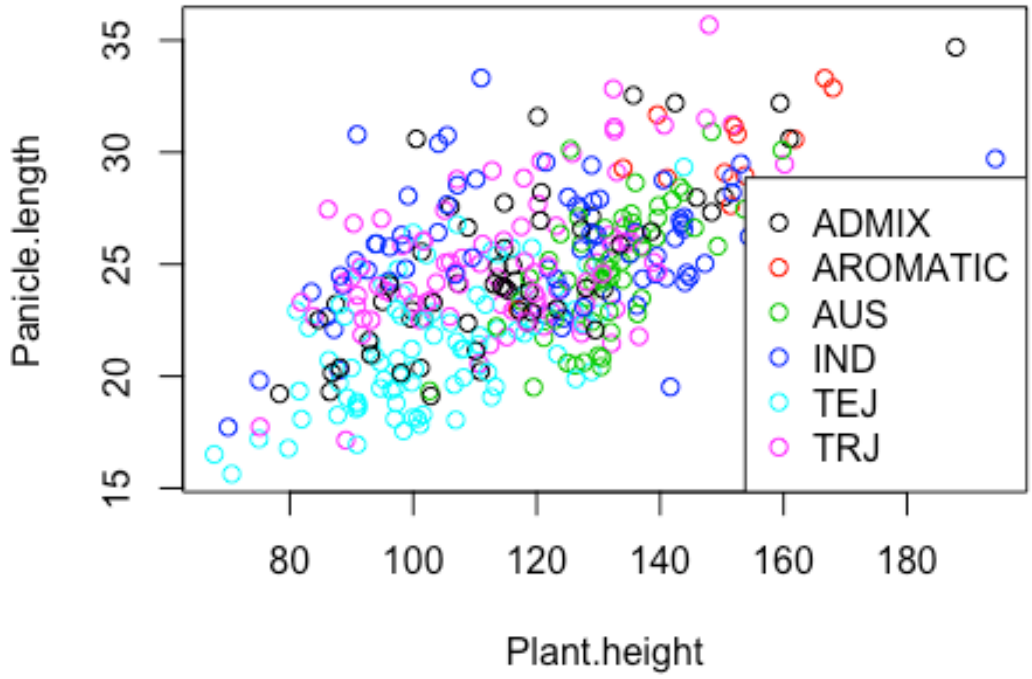
読み込まれた Zhao ら（2011）のデータには、形質データだけでなく、遺伝資源の遺伝的背景に関するデータも含まれています。遺伝的背景と形質の間にどのような関係があるのか、両データを併せて図示して調べてみましょう。

Sub.population という変数は、各遺伝資源の遺伝的背景の違いを表しています。これは Structure 解析（Pritchard et al. 2000, Genetics 155:945)を用いて推定されたものです。では、遺伝的背景と草丈や穂長にどのような関係があるのか視覚化して見てみましょう。

```
pop.id <- as.numeric(Sub.population)
plot(Plant.height, Panicle.length, col = pop.id)
levels(Sub.population)
```

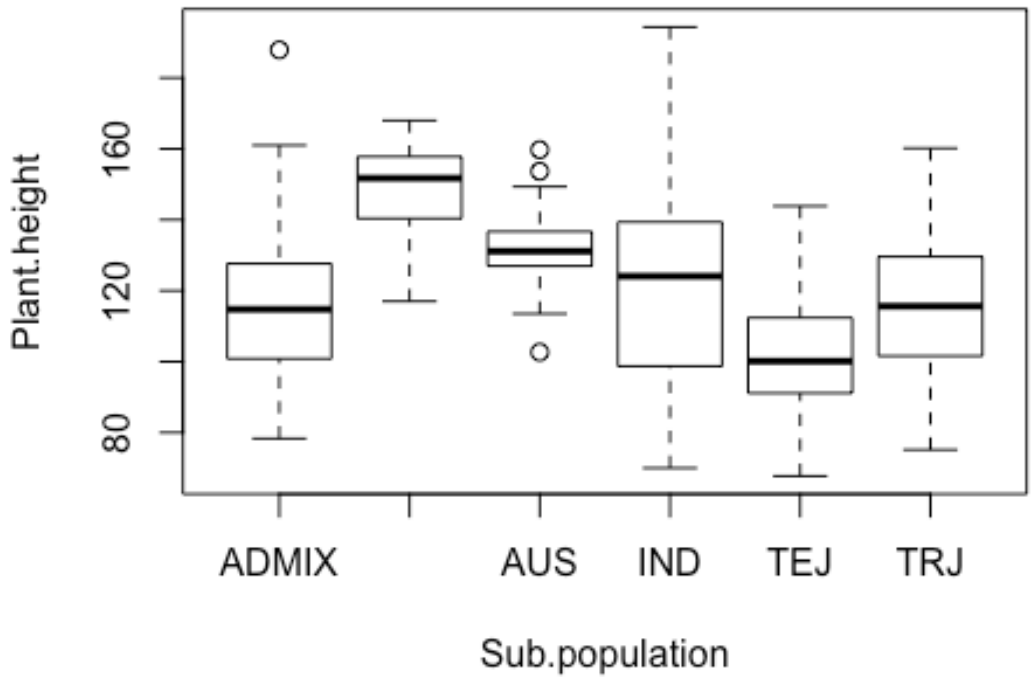
```
## [1] "ADMIX"    "AROMATIC" "AUS"      "IND"      "TEJ"      "TRJ"
```

```
legend("bottomright", levels(Sub.population), col = 1:nlevels(Sub.population),
 pch = 1)
```

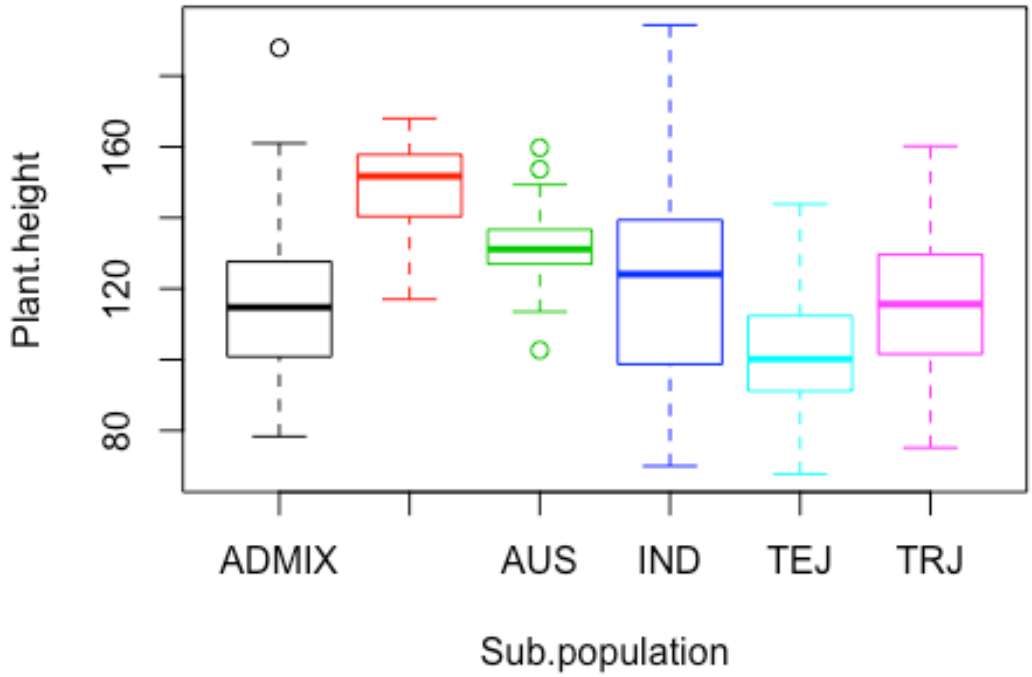


遺伝的背景の違いにより値にどのような違いがあるのかを箱ひげ図で示してみましょう。

```
boxplot(Plant.height ~ Sub.population)
```

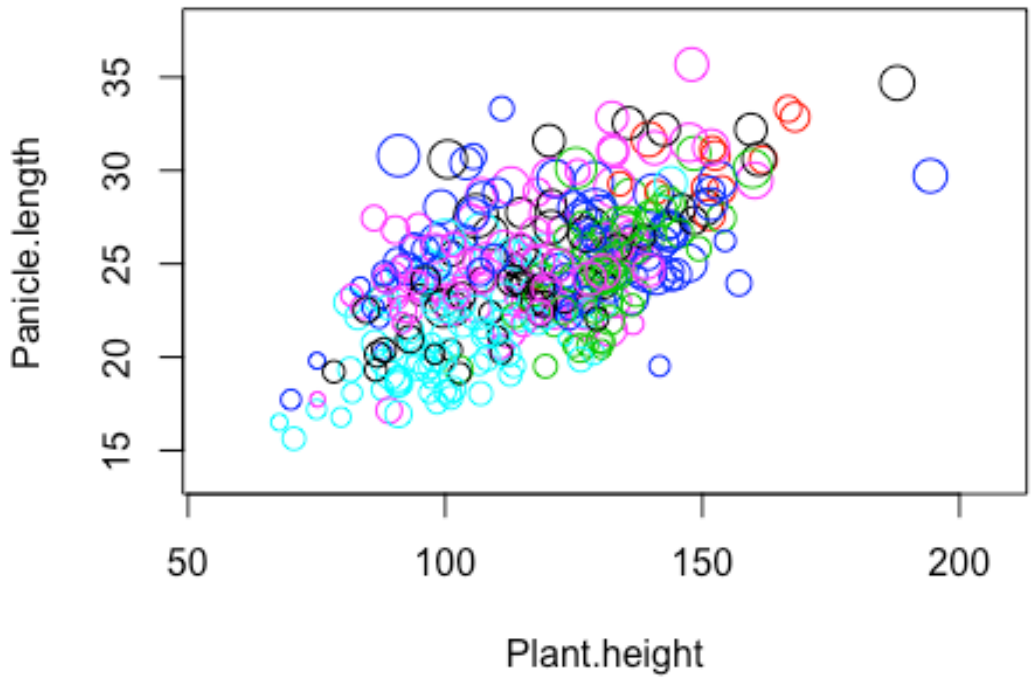



```
boxplot(Plant.height ~ Sub.population, border = 1:nlevels(Sub.population))
```



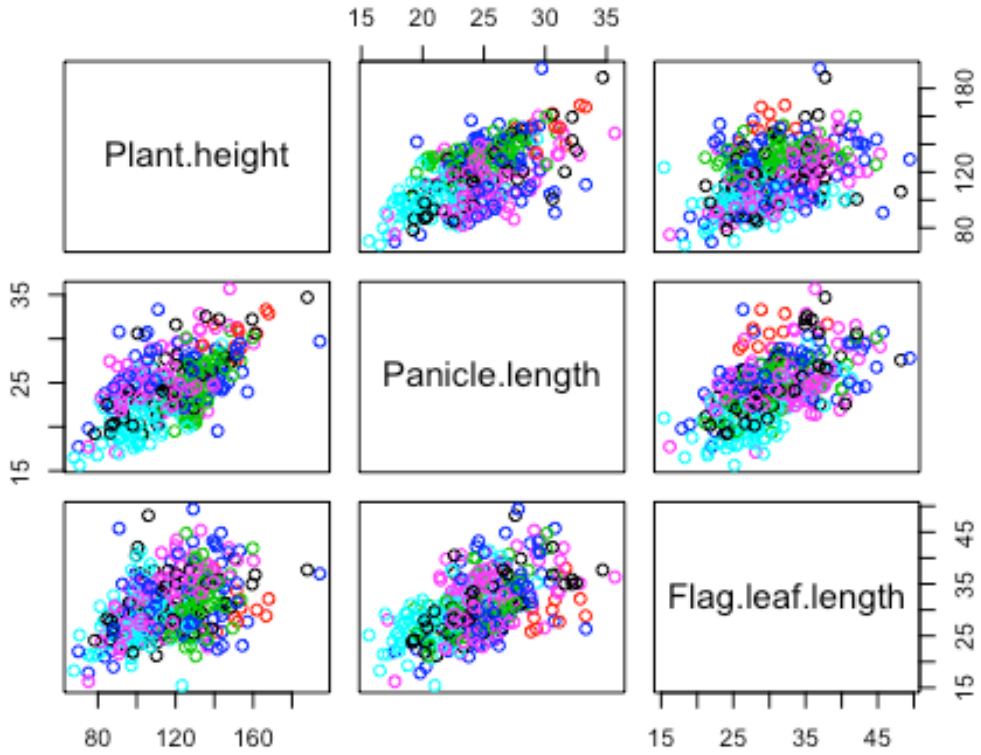
草丈（Plant.height）と穂長（Panicle.length）に加え、止め葉の長さ（Flag.leaf.length）の 3 変数の間の関係がどのようになっているかをバブルプロット（bubble plot）によって確かめてみましょう。ここでは、バブルの大きさが止め葉の長さを表しています。

```
symbols(Plant.height, Panicle.length, circles = Flag.leaf.length, inches = 0.
1, fg = pop.id)
```



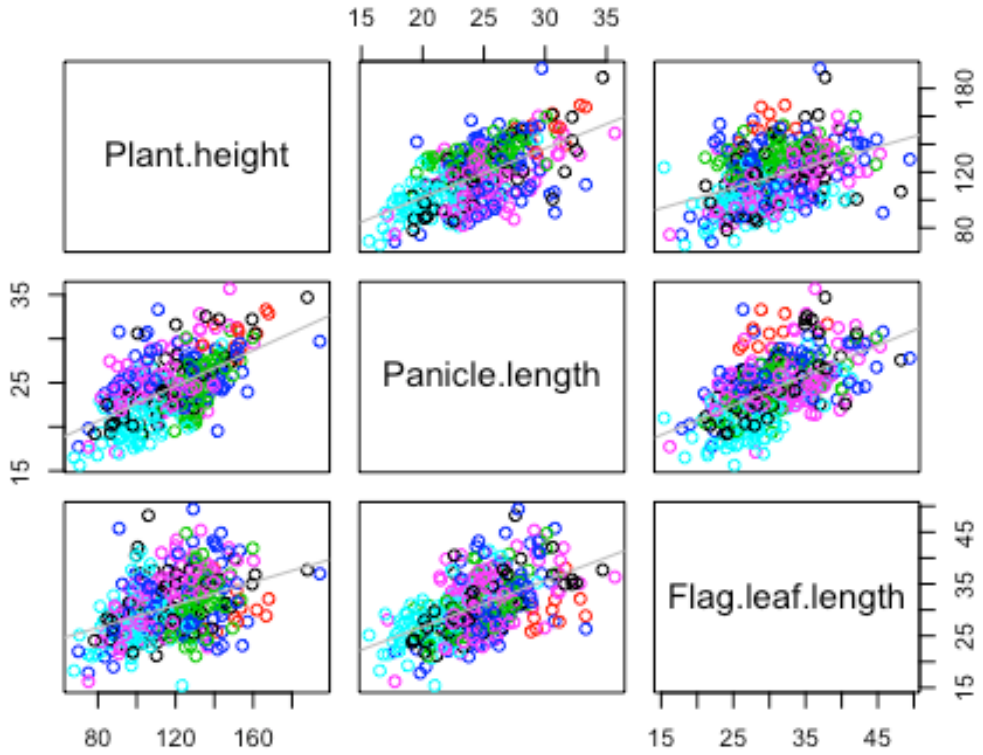
3 変数間の関係を総当たりの散布図で描いてみましょう。

```
x <- data.frame(Plant.height, Panicle.length, Flag.leaf.length)
pairs(x, col = pop.id)
```



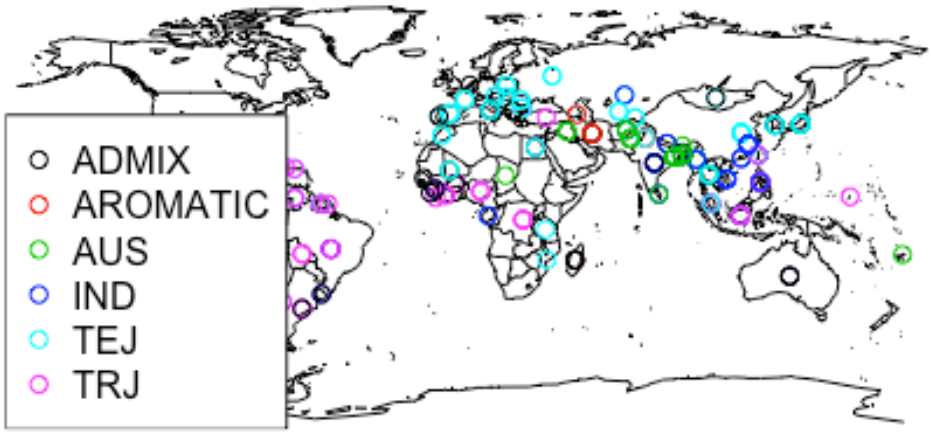
回帰直線を加えた少し複雑な散布図にしてみましょう。

```
pairs(x, panel = function(x, y, ...) {
    points(x, y, ...)
    abline(lm(y ~ x), col = "gray")
}, col = pop.id)
```



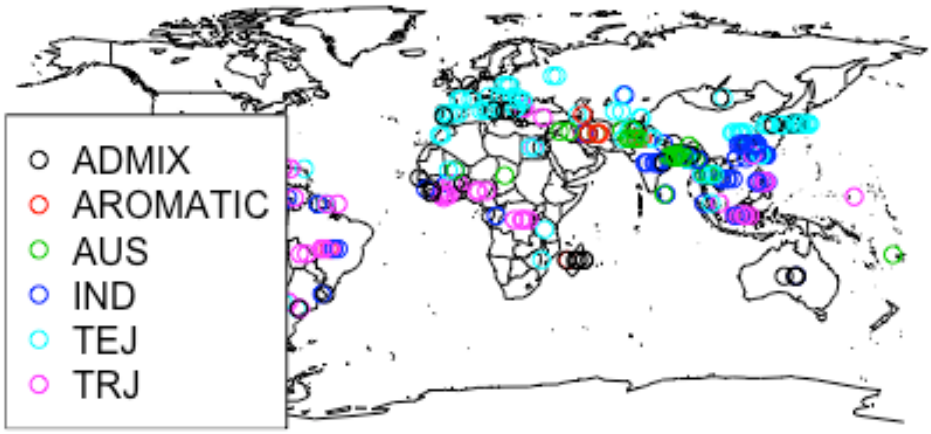
読み込まれているデータには、各遺伝資源の由来している場所の緯度経度のデータも含まれています。そこで、各遺伝資源の由来を世界地図上にマップして確認してみましょう。

```
library(maps)
library(mapdata)
map('worldHires')
points(Longitude, Latitude, col = pop.id)
legend("bottomleft", levels(Sub.population), col = 1:nlevels(Sub.population),
 pch = 1)
```



上のコマンドでは、遺伝資源数よりもずっと少ない数の点しか描かれません。これは、同じ地域からの遺伝資源が互いに重なり合って表示されているためです。重なり合いを防ぐには関数 jitter で重なっている点を少しだけ動かします。

```
map('worldHires')
points(jitter(Longitude, 200), Latitude, col = pop.id)
legend("bottomleft", levels(Sub.population), col = 1:nlevels(Sub.population),
 pch = 1)
```



Quiz5

では、ここで練習問題を解いてみましょう。練習問題は、授業中に提示します。

クイズのページを閉じてしまった人は、https://www.menti.com/ に移動して、361599 と入力して下さい。

図のファイルへの出力

作成した図を論文やプレゼン用資料などを利用するためには、図を PDF ファイルなどに出力できると便利です。ここでは、簡単にその方法を説明します。

先ほど描いた図を map.pdf というファイルに出力してみましょう。

```
pdf("map.pdf")
map('worldHires')
points(jitter(Longitude, 200), Latitude, col = pop.id)
legend("bottomleft", levels(Sub.population), col = 1:nlevels(Sub.population),
 pch = 1)
dev.off()

## quartz_off_screen
##                 2
```

上のコマンドを実行すると map.pdf というファイルが R の作業ディレクトリに出力されます。

関数 pdf では、出力する図のサイズを指定することができます。今回の図のように横長のほうが合っていて、かつ、大きなサイズで出力したほうがよい場合には、サイズを指定して出力したほうがきれいな図が描けます。

```
pdf("map_large.pdf", width = 20, height = 10)
map('worldHires')
points(jitter(Longitude, 200), Latitude, col = pop.id)
legend("bottomleft", levels(Sub.population), col = 1:nlevels(Sub.population),
 pch = 1)
dev.off()

## quartz_off_screen
##                 2
```

なお、複数の図を同じ pdf ファイルに繰り返し出力すると複数ページの pdf ファイルとして保存されます。同種の図を繰り返し大量に出力したい場合には、1 つの pdf ファイルにまとめておく方が便利かもしれません。

インタラクティブな図を描く

パッケージ plotly を使うと、インタラクティブな図を描くことができます。ここでは、先に描いた 3 次元の図を plotly の関数 plot_ly を用いて描いてみましょう。

まずは、データの密度の 3 次元表示を行ってみましょう。

```
require(plotly)
plot_ly(data = data.frame(d), x = d$x1, y = d$x2, z = d$fhat) %>%
add_surface()
```

最後に、3 変量間の関係を 3 次元で眺めてみましょう。

```
df <- data.frame(Sub.population, Plant.height, Panicle.length, Flag.leaf.leng
th)
df <- na.omit(df)
plot_ly(data = df, x = ~Plant.height, y = ~Panicle.length, z = ~Flag.leaf.len
gth, color = ~Sub.population, type = "scatter3d", mode = "markers")
```

Quiz 6

では、ここで練習問題を解いてみましょう。練習問題は、授業中に提示します。

クイズのページを閉じてしまった人は、https://www.menti.com/ に移動して、361599 と入力して下さい。

レポート課題

講義で学んだ様々なデータ視覚化法を用いて形質間の関係や、形質と遺伝的背景間の関係について図を描いてください。また、描いた図から読み取ることができる関係について記述してください。

提出方法：

- レポートは pdf ファイルとして作成し、ITC-LMS から提出する。

- 何らかの問題で ITC-LMS で提出できない場合は、メール添付で report@iu.a.u-tokyo.ac.jp に送る。
- レポートの最初に、所属、学生番号、名前を忘れずに。
- 提出期限は、5 月 1 日