Your 1st step or descriptive statistics 第一歩、記述統計

Introduction 前書き

In our unit many data records (real and/or simulational) are obtained.

うちのラボでは、たくさんのデータレコード(実データかもしれないし、シミュレーションデータかもしれません)を扱います。)

They may have complex structure.

それらは複雑な構造をしているかもしれません。

Eventually various things have to be done.

最終的には、その複雑な構造を知るためにいろいろなことをやることになります。

However, the very basic thing has to be done always for everything.

しかしながら、最初に、すべて対象にやることは基礎的な同じ作業です。

Please make sure the following steps are your routines.

以下のステップをルーチン化してください。

Steps to be done　行うべきステップ

Step 1: to know your variables　変数を知る

- You have to write out “No. samples n x No. variables”.　サンプル数 x 変数の数、は基本です。書き出しましょう
- Group the variables into two; (i) categorical or (ii)　continuous (quantitative) 変数をカテゴリカルか連続(量的)かに二分しましょう

Step 2: to know distribution of each variables それぞれの変数の分布を見ましょう

- Draw histograms. Draw for all continuous variables. hist(X)　すべての変数の分布ヒストグラムにして、眺めることが必要です
- If you find the number of variables = the number of histograms to be drawn, you may sort every variables and draw cumulative distribution curves of the variables in one panel together. A thousand histrograms are too many. A hundred histograms are not too many. matplot(apply(X,2,sort),type=“l”)　変数の数が多すぎて、ヒストグラムの数が多くなりすぎてしまい、全部を眺めることが無理としましょう(１０００を越えたら無理です。１００は可能です)。そんな場合は、各変数をソートしてプロットします。たくさんの変数のソート後カーブを１パネルに描くと用が足ります
- Make a count table for all categorical variables. The bar plots representing fractions may be also good for this task.　すべてのカテゴリカル変数ではカウントテーブルを作ります。カテゴリの割合を表す棒グラフを描くのもよいです

Step 3: to know relation between two vriables 二変数ペアの関係を知る

- All pairs are your targets. すべての変数ペアは興味の対象です
- In the case of categorical variable vs. continuous variable: Histograms per categories https://stats.biopapyrus.jp/r/ggplot/geom_histogram.html (https://stats.biopapyrus.jp/r/ggplot/geom_histogram.html) . カテゴリカル変数 vs. 連続変数の場合には、カテゴリ別ヒストグラムを描きます
- In the case of categorical variable vs. categorical variable: 2D-contingency table. table(X,Y) カテゴリカル変数 vs. カテゴリカル変数の場合には、二元分割表を作ります
- In the case of continuous variable vs. continuous variable: coplot. plot(as.data.frame(X)) 連続変数 vs. 連続変数の場合には、コプロットします

Step 4: to know correlation pattern amang all variables すべての変数ペアの相関の全体像を知る

- Calculate correlation coefficient for all variable pairs and visualize them in a square. 全変数ペアの相関係数を計算して、正方形で描図します
- To know the whole picture, you need something visualizing all. cormat(cor(X)); image(cor.mat) 全体像を知るには、全変数を１つの絵にする必要があります

Step 5: to know difference among samples (not variables)　サンプル間の違いも知る

- Make a distance matrix among samples to know dissimilarity. dist.out <- dist(X) サンプルペアの違いの評価のために距離行列を作りましょう
- Apply clustering method on distance matrix. plot(hclust(dist.out)) 距離行列はクラスタリングして図示しましょう
- Make an inner product matrix as well to know similarlity. ip.mat <- X%*% t(X) サンプルペアの似ている程度の評価のために内積の行列を作りましょう
- Eigen decomposition the inner product matrix and plot its eigen values. eigen.out <- eigen(ip.mat); plot(eigen.out[[1]]) 内積行列は固有値分解して、その固有値をプロットしましょう