2023年9月13日
SASユーザー総会2023

SASによる散布図行列の実装

〇徳田芳稀
（エイツーヘルスケア株式会社）

Scatter Matrix Implementation with SAS

Yoshiki Tokuda
A2 Healthcare Corporation

E-mail: tokuda-y@a2healthcare.com

要旨 :

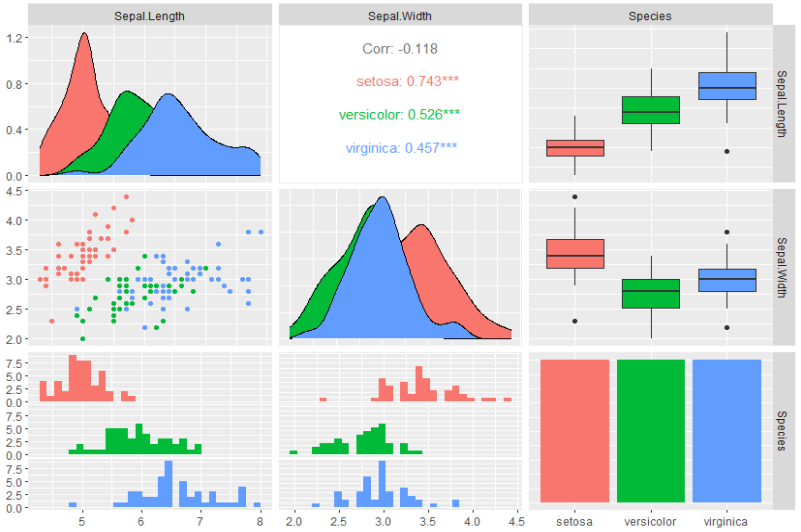
SASではSGSCATTERプロシジャにより散布図行列を作成可能だが、表示方法に限界がある。本発表では、GTL (Graph Template Language)による散布図行列の作成方法を示す。

キーワード : 散布図行列, Graph Template Language (GTL)

背景・目的

- 散布図行列とは？
  複数の変数を持つデータについて、散布図等を行列形式で表示することにより、
  複数の変数間の関連を一度に可視化できる図

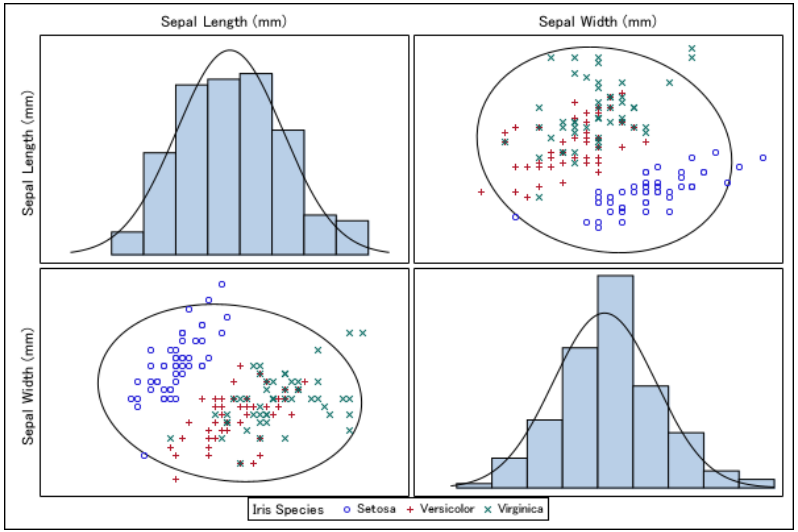
- 実務上の経験と課題①
  - 多様な情報を含む散布図行列を作成する案件が少しずつ増えている
  - 右図はRのGGallyパッケージにより少ないプログラムで作成可能であるが、SAS環境でも実装できるようにしたい
  - 散布図行列自体は、SASのSGSCATTERプロシジャ等により作成可能



背景・目的

- 実務上の経験と課題②
  - SGSCATTERプロシジャの記載例
    (SAS Version 9.4を使用、以降同様)

```
proc sgscatter data=sashelp.iris;
  matrix SepalLength SepalWidth /
         ellipse=(type=predicted)
         diagonal=(histogram normal)
         group=species;
run;
```



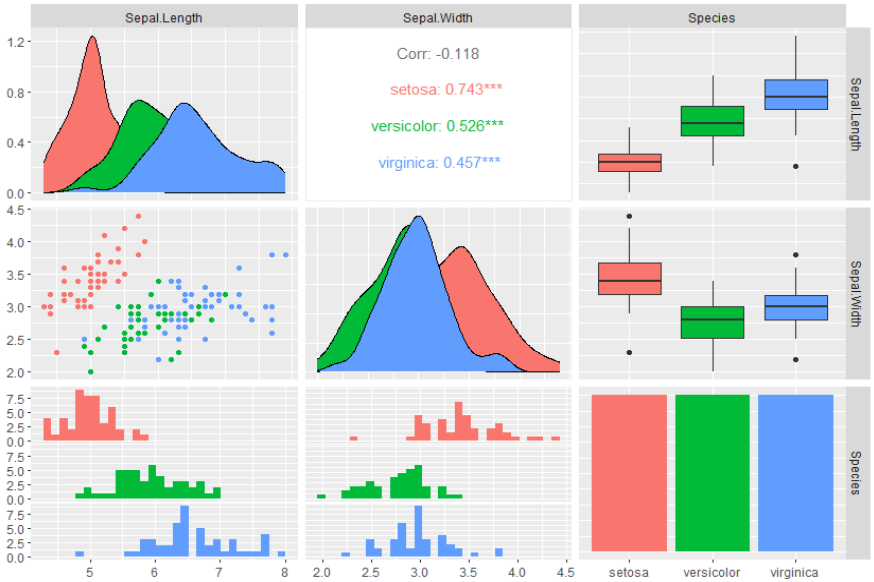
  - 上記以外には回帰直線(REG)・局所回帰(LOESS)・スプライン回帰曲線(PBSPLINE)を描画可能であるが、箱ひげ図等のプロットや文字列の描画ができない

→**Graph Template Language**を使ってみよう！

背景・目的

本発表では、SASのGraph Template Language(以下GTL)を用いて、柔軟に且つ多様な情報を含めた散布図行列の作成方法を示す

Rによる出力



SASによる出力

?

目次

目次

- 使用するデータセット
- GTLの基礎
- GTLによる散布図行列の作成方針
- 出力結果・出力に使用したプログラムと課題
- プログラムの簡略化について
- まとめ

使用するデータセット

- Sashelp.iris: アヤメのデータ

	Species	SepalLength	SepalWidth	PetalLength	PetalWidth
1	Setosa	50	33	14	2
2	Setosa	46	34	14	3
3	Setosa	46	36	10	2
4	Setosa	51	33	17	5
5	Setosa	55	35	13	2
6	Setosa	48	31	16	2
7	Setosa	52	34	14	2
8	Setosa	49	36	14	1
9	Setosa	44	32	13	2
10	Setosa	50	35	16	6

- データの概要

<変数>

・Sepal Length　：がくの長さ(mm)

・Sepal Width　：がくの幅(mm)

・Petal Length　：花弁の長さ(mm)

・Petal Width　：花弁の幅(mm)

・Species　：種類 (Setosa, Versicolor, Virginica)

https://dataaspirant.com/svm-classifier-implemenation-python-scikit-learn/

<概要>

・各Species50レコード、計150レコード

・SAS以外にRやPythonでテストデータとして利用可能

・機械学習の分類問題(判別分析やSVM)の説明によく使用される

目次

GTLの基礎

GTL: Output Delivery System(ODS)の拡張機能で高度なグラフを作成可能

<構文>

```
proc template;
  define statgraph テンプレート名;
    begingraph / オプション;
      layout レイアウトの種類;
        プロットの種類・文字列等;
      endlayout;
    endgraph;
  end;
run;
```

Templateプロシジャ: テンプレート部分を作成

- layoutステートメント: レイアウトの指定
- plot ステートメント: 描画するプロットの指定
- textステートメント: 記載する文字列の指定
- etc…

```
proc sgrender data=XXXX
  template=テンプレート名;
run;
```

SGRENDERプロシジャ:
テンプレートを読み込んで、プロット等を作成

GTLの基礎 (layoutステートメント)

<構文>

```
proc template;
  define statgraph テンプレート名;
    begingraph / オプション;
      layout レイアウトの種類;
        プロットの種類・文字列等;
      endlayout;
    endgraph;
  end;
run;

proc sgrender data=XXXX
  template=テンプレート名;
run;
```

ステートメント	内容
layout overlay	最も汎用され、複数のプロットを重ね書き出来る
layout lattice	複数のプロットをm×nの行列のように配置できる

その他のステートメント

- layout globallegend
- layout datapanel
- etc…

詳細はSAS Helpを参照

GTLの基礎 (plotステートメント)

<構文>

```
proc template;
  define statgraph テンプレート名;
    begingraph / オプション;
      layout レイアウトの種類;
        プロットの種類・文字列等;
      endlayout;
    endgraph;
  end;
run;

proc sgrender data=XXXX
  template=テンプレート名;
run;
```

プロット	構文
楕円	Ellipse x=XXX y=XXX / group=XXX Type=mean or predicted…;
箱ひげ図	Boxplot x=XXX y=XXX / group=XXX display=standard meanattrs=(symbol=diamond)…;
散布図	Scatterplot x=XXX y=XXX / group=XXX;
密度推定曲線	Densityplot 変数 / group=XXX normal() or kernel(weightfunction=)…;
ヒストグラム	Historgram 変数 / …;
棒グラフ	Barchart category=XXX response=XXX / STAT=freq or mean or …

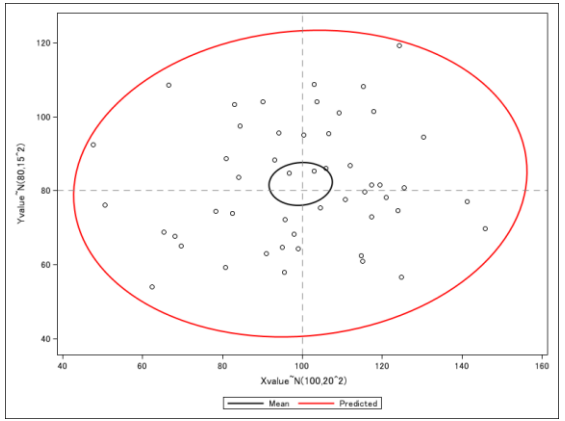
※赤字はデフォルト

GTLの基礎 (補足)

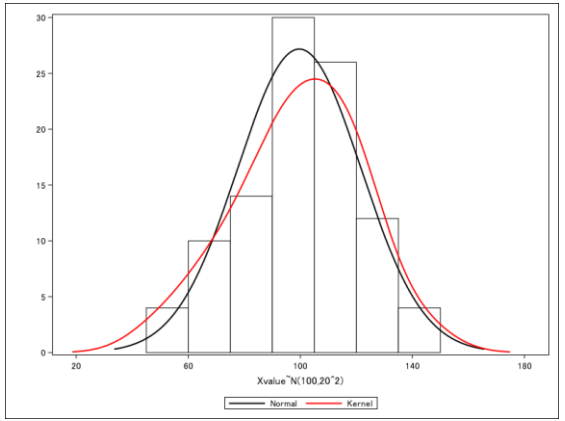
- ellipseステートメント
  - mean: (黒線)
    平均の信頼楕円
  - predicted: (赤線)
    新しい観測値に対する予測楕円

- densityplotステートメント
  - normal(): (黒線)
    標本平均・SDに基づく正規分布の
    確率密度関数(mu, sigmaの指定も可能)
  - Kernel(weightfunction=): (赤線)
    カーネル密度推定に基づく確率密度関数
    (weightfunction=normal/quadratic/triangular)

X軸の値: $N(100, 20^2)$, Y軸の値: $N(80, 15^2)$



X軸の値: $N(100, 20^2)$



GTLの基礎 (実装例:プログラム)

<散布図+信頼楕円>

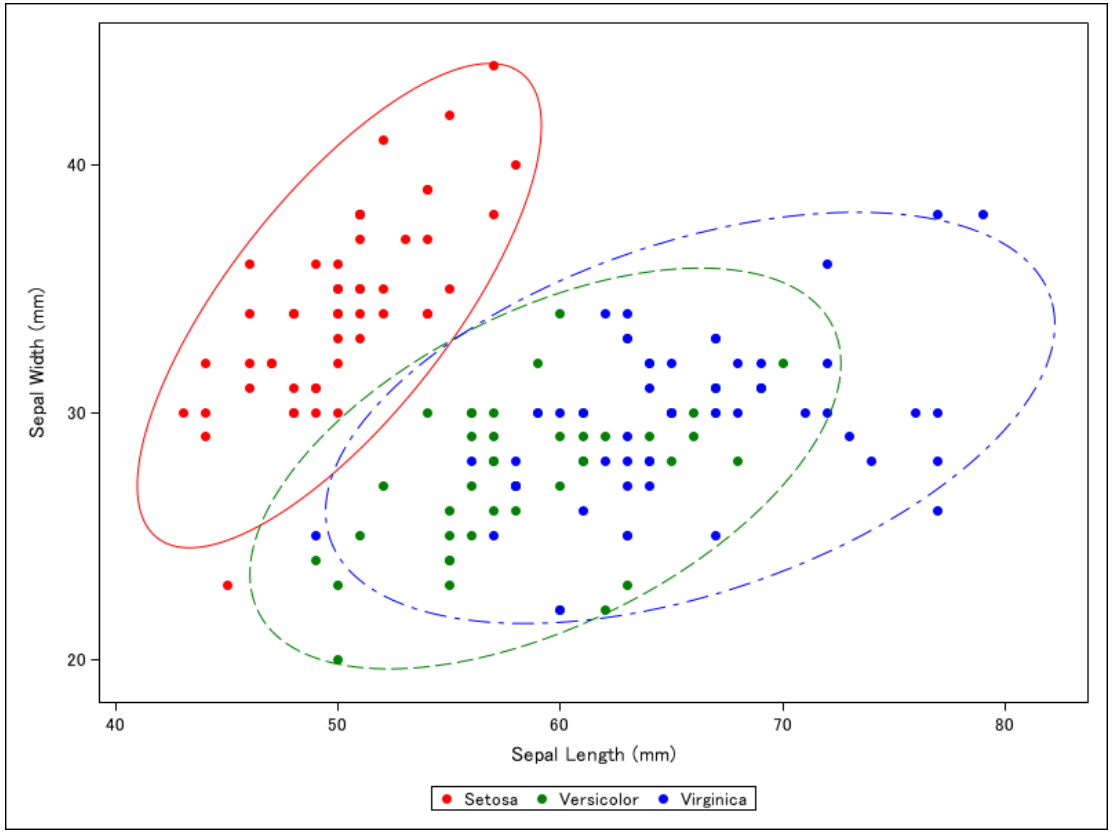
```
proc template ;
 define statgraph SCATTER_ELLIPSE;
  begingraph /
  datacontrastcolors=(red green blue)
  datasymbols=(circlefilled);
   layout overlay /
    xaxisopts=(label="Sepal Length (mm)")
    yaxisopts=(label="Sepal Width (mm)");
    scatterplot x=SepalLength y=SepalWidth / group=species ;
    ellipse x=SepalLength y=SepalWidth / group=species type=predicted ;
   endlayout;
  endgraph;
 end;
run;

proc sgrender data=sashelp.iris template=SCATTER_ELLIPSE;
run;
```

散布図

信頼楕円

GTLの基礎 (実装例:出力結果)

<散布図+信頼楕円>



GTLの基礎 (layout lattice)

<構文>

```
proc template;
  define statgraph matrix;
    begingraph;
      layout lattice / rows=2 columns=3
                order=rowmajor/columnmajor;
        プロットの種類①;
        ・・・
        プロットの種類⑥;
      endlayout;
    endgraph;
  end;
run;
```

layout overlay等で出力するプロットを指定

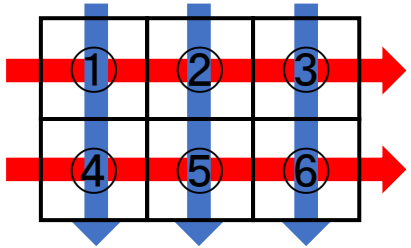
GTLの基礎 (layout lattice)

<構文>

```
proc template;
  define statgraph matrix;
    begingraph;
      layout lattice / rows=2 columns=3
                order=rowmajor/columnmajor;
        プロットの種類①;
        ・・・
        プロットの種類⑥;
      endlayout;
    endgraph;
  end;
run;
```

rows: 行数, columns: 列数

今回の例では**2行3列**のプロットになる



order: プロットを配置する順番

以下のいずれかを選択可能

rowmajor

columnmajor

GTLの基礎 (layout lattice)

<構文>

```
proc template;
  define statgraph matrix;
    begingraph;
      layout lattice / rows=2 columns=3
                order=rowmajor/columnmajor;
        プロットの種類①;
        ・・・
        プロットの種類⑥;
      endlayout;
    endgraph;
  end;
run;
```

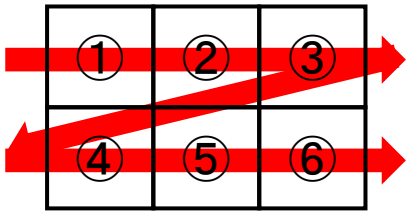
rows: 行数, columns: 列数

今回の例では2行3列のプロットになる

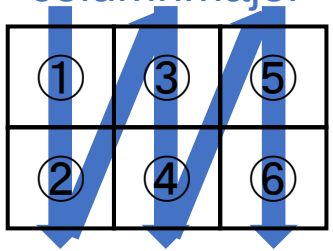
order: プロットを配置する順番

以下のいずれかを選択可能

rowmajor



columnmajor

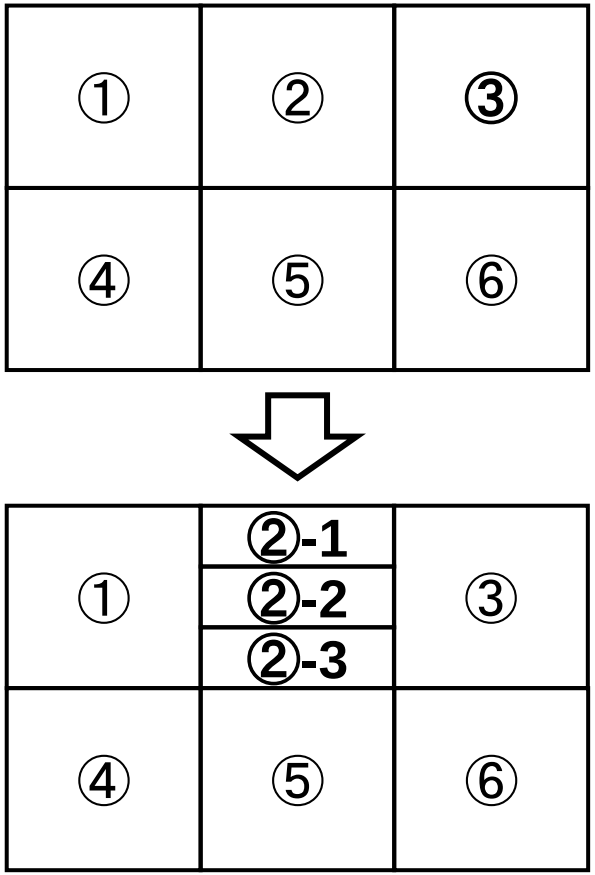


GTLの基礎 (layout latticeのネスト)

<構文>

```
layout lattice / rows=2 columns=3
                 order=rowmajor;
  プロットの種類①;
  layout lattice / rows=3 columns=1;
    プロットの種類②-1;
    プロットの種類②-2;
    プロットの種類②-3;
  endlayout;
  ・・・
  プロットの種類⑥;
endlayout;
```

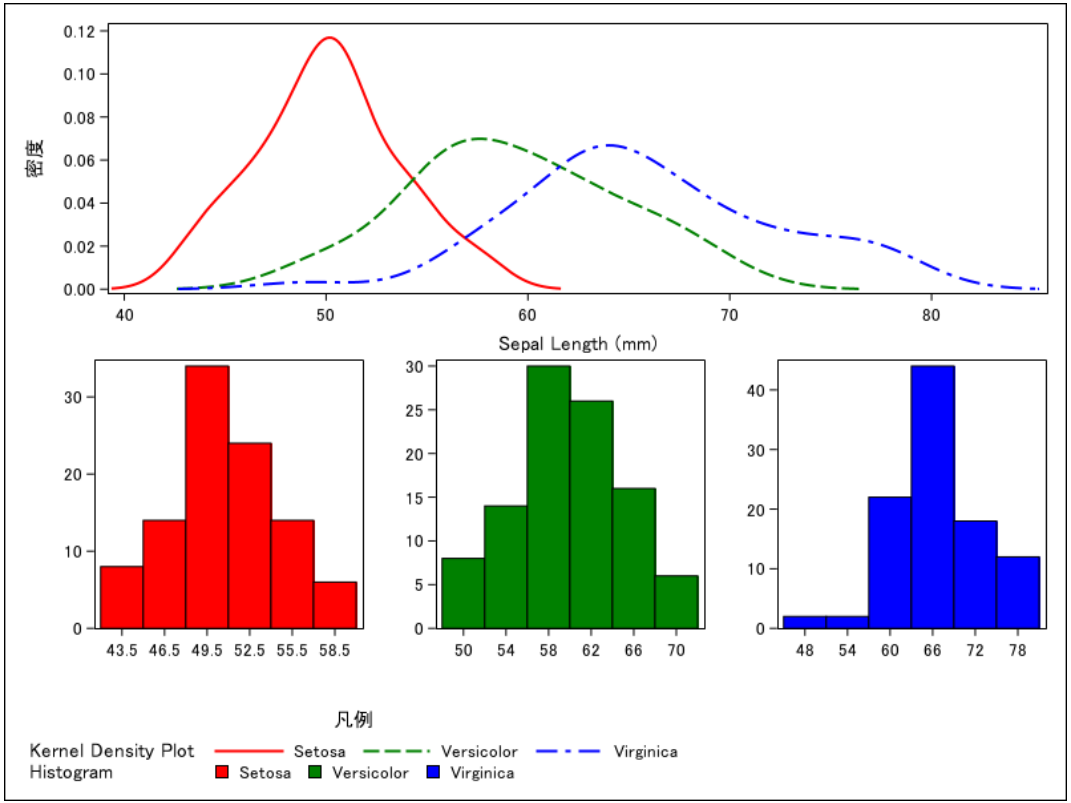
②



GTLの基礎 (実装例:プログラム+出力結果)

<カーネル密度+種類別ヒストグラム>

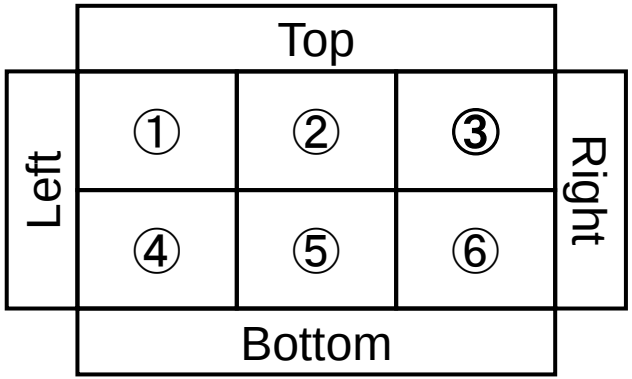
```
proc template ;
  define statgraph KERNEL_HIST;
   begingraph /
   datacontrastcolors=(red green blue)
   datacolors=(red green blue);
    /*2行1列のレイアウト設定*/
    layout lattice / rows=2 columns=1 order=rowmajor;
     *プロットの種類①: カーネル密度推定;
     layout overlay;
       densityplot SepalLength / kernel() group=Species  name="a1";
     endlayout;
     /*プロットの種類②に1行3列のレイアウト設定*/
     layout lattice / rows=1 columns=3 order=columnmajor ;
       layout overlay / xaxisopts=(label=" ") yaxisopts=(label=" ");
         *プロットの種類②-1: SetosaのSepalLengthのヒストグラム;
         histogram eval(ifn(Species="Setosa",SepalLength,.)) / fillattrs=(color=red)  name="Setosa";
       endlayout;
       layout overlay / xaxisopts=(label=" ") yaxisopts=(label=" ");
         *プロットの種類②-2: VersicolorのSepalLengthのヒストグラム;
         histogram eval(ifn(Species="Versicolor",SepalLength,.)) / fillattrs=(color=green)  name="Versicolor";
       endlayout;
       layout overlay / xaxisopts=(label=" ") yaxisopts=(label=" ");
         *プロットの種類②-3: VirginicaのSepalLengthのヒストグラム;
         histogram eval(ifn(Species="Virginica",SepalLength,.)) / fillattrs=(color=blue) name="Virginica";
       endlayout;
     endlayout;
    endlayout;
    /*凡例設定*/
    layout globallegend / type=column title="凡例" halign=left border=false;
      discretelegend "a1" / title="Kernel Density Plot";
      discretelegend "Setosa" "Versicolor" "Virginica" / title="Histogram";
    endlayout;
   endgraph;
  end;
run;
```



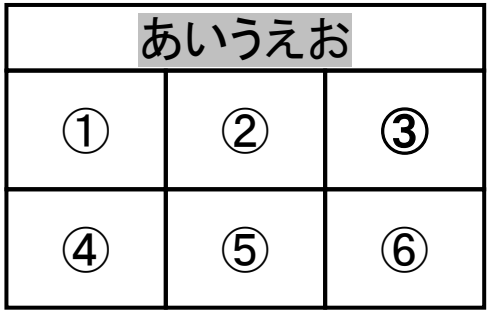
GTLの基礎 (sidebarによる軸ラベルの設定)

<構文>

```
layout lattice / rows=2 columns=3
                 order=rowmajor;
  sidebar / align=Top (or Left, Right, Bottom);
    halign=Center (or Left, Right)
    entry “あいうえお”/
    textattrs=(size=18)
    opaque=True (or False) (背景透過の有無)
    backgroundcolor=lightgray; (背景色の設定)
  endsidebar;
```



出力イメージ



GTLの基礎 (sidebarによる軸ラベルの設定)

<構文>

```
layout lattice / rows=2 columns=3
                 order=rowmajor;
  sidebar / align=Top (or Left, Right, Bottom);
    layout lattice / rows=1 columns=3;
      entry halign=Center “あい”;
      entry halign=Center “うえ”;
      entry halign=Center “おか”;
    endlayout;
  endsidebar;
```

出力イメージ

あい	うえ	おか
①	②	③
④	⑤	⑥

目次

GTLによる散布図行列の作成方針

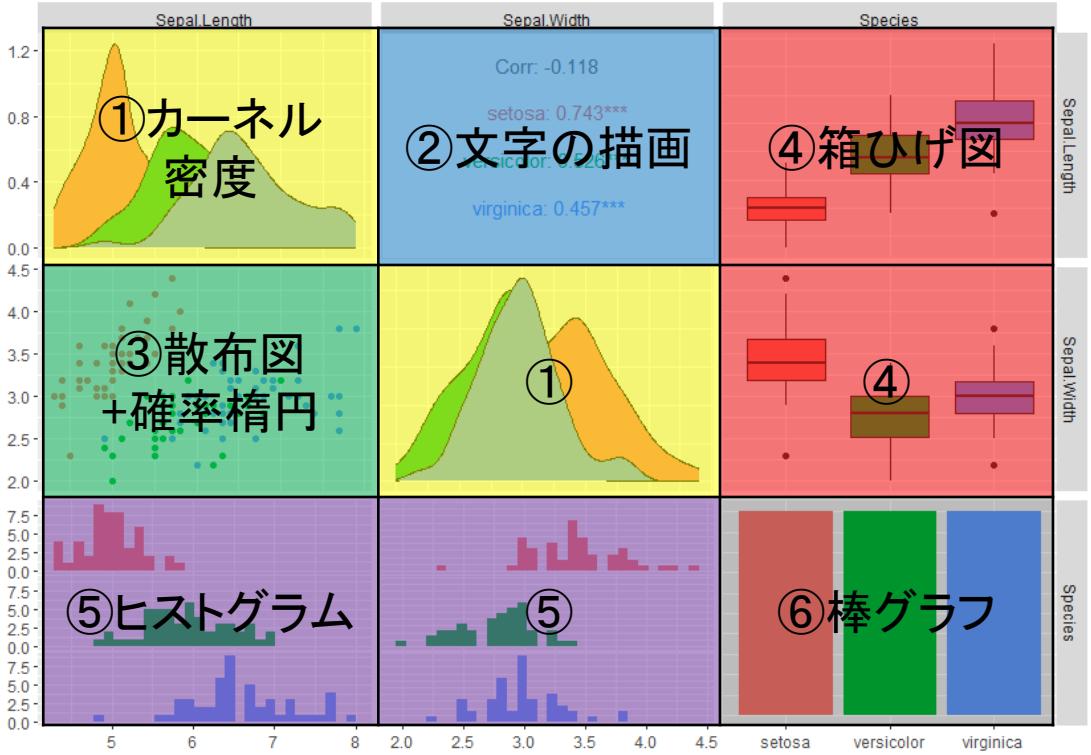
layout ステートメントは
以下を使用

- layout lattice (3×3)
- layout overlay

黒枠の各行列のラベルは、
sidebarを使用

赤枠の箇所について、
layout lattice(3×1)をネスト
することで描画

GTLによる散布図行列の作成方針



描画には以下を使用

① densityplot

② entry (相関係数を表示)

③ scatterplot+ellipse

④ boxplot

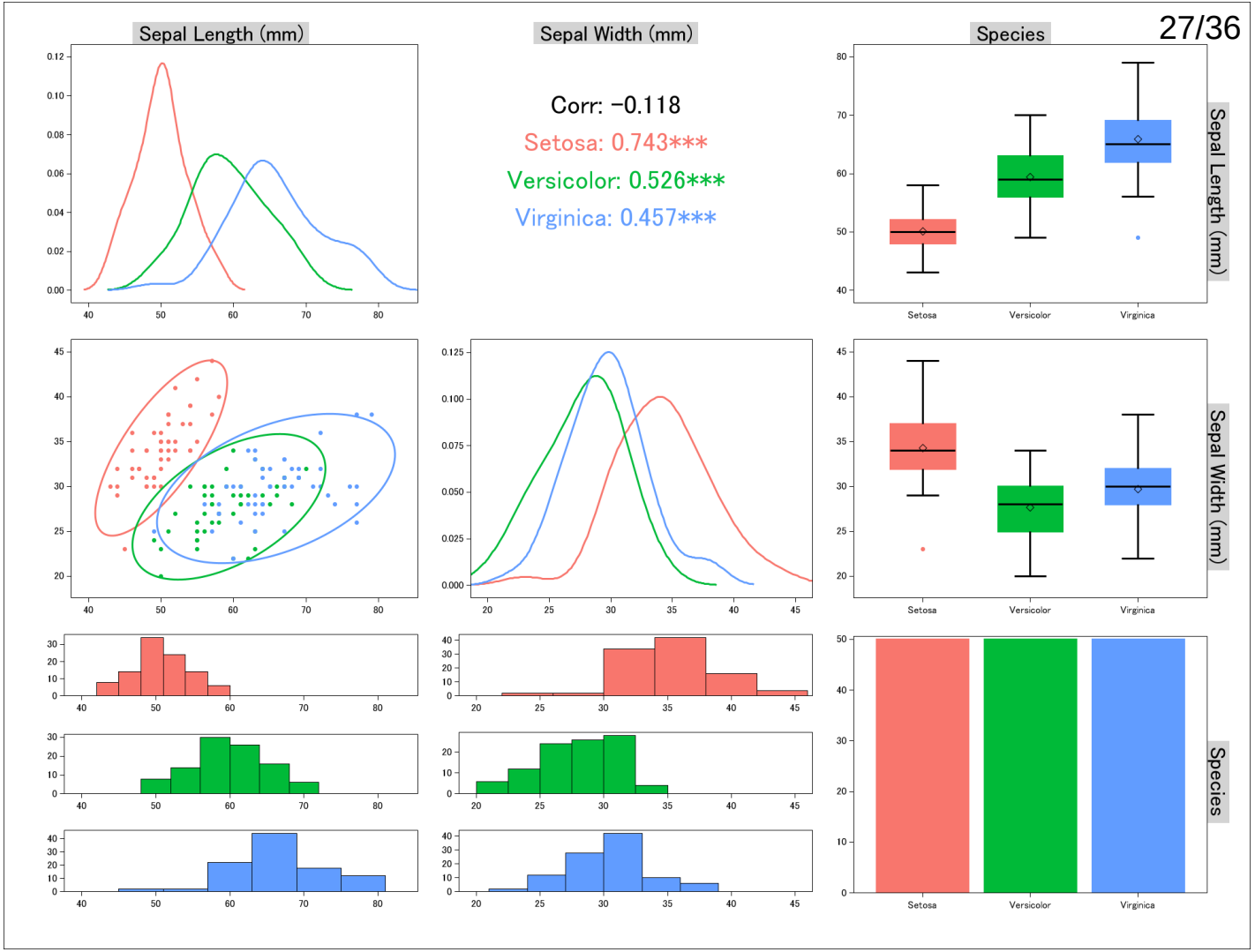
⑤ histogram

⑥ barchart

②はtextステートメントの1種で、プロット領域に文字列を記載できる

目次

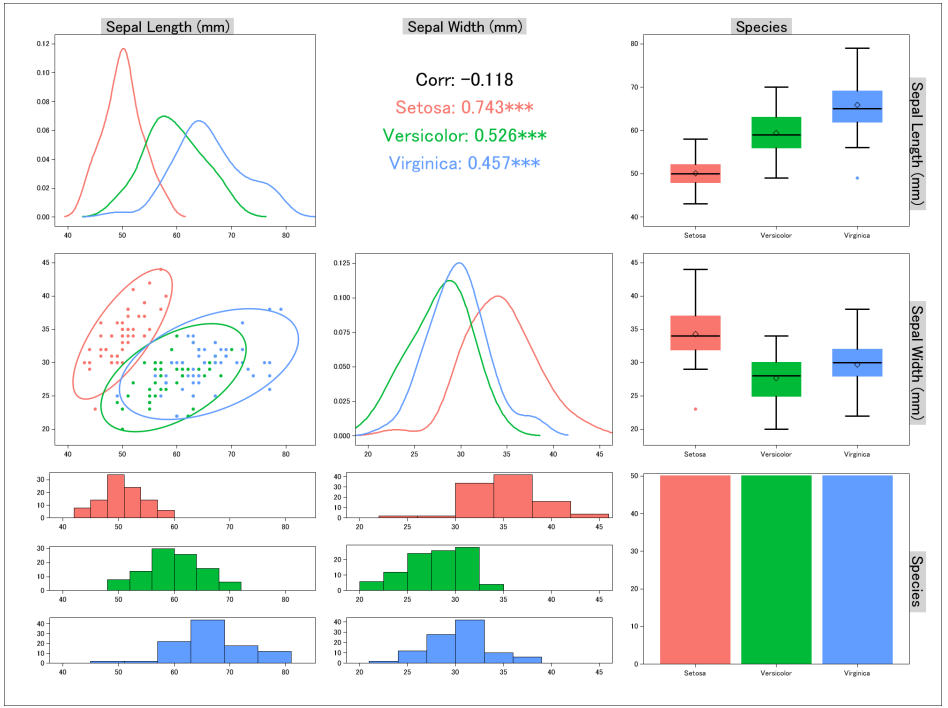
出力結果


Sepal Length (mm)
Sepal Width (mm)
Species
Corr: −0.118
Setosa: 0.743***
Versicolor: 0.526***
Virginica: 0.457***
Sepal Length (mm)
Sepal Width (mm)
Species
Setosa
Versicolor
Virginica

出力結果 (比較)

Rによる出力

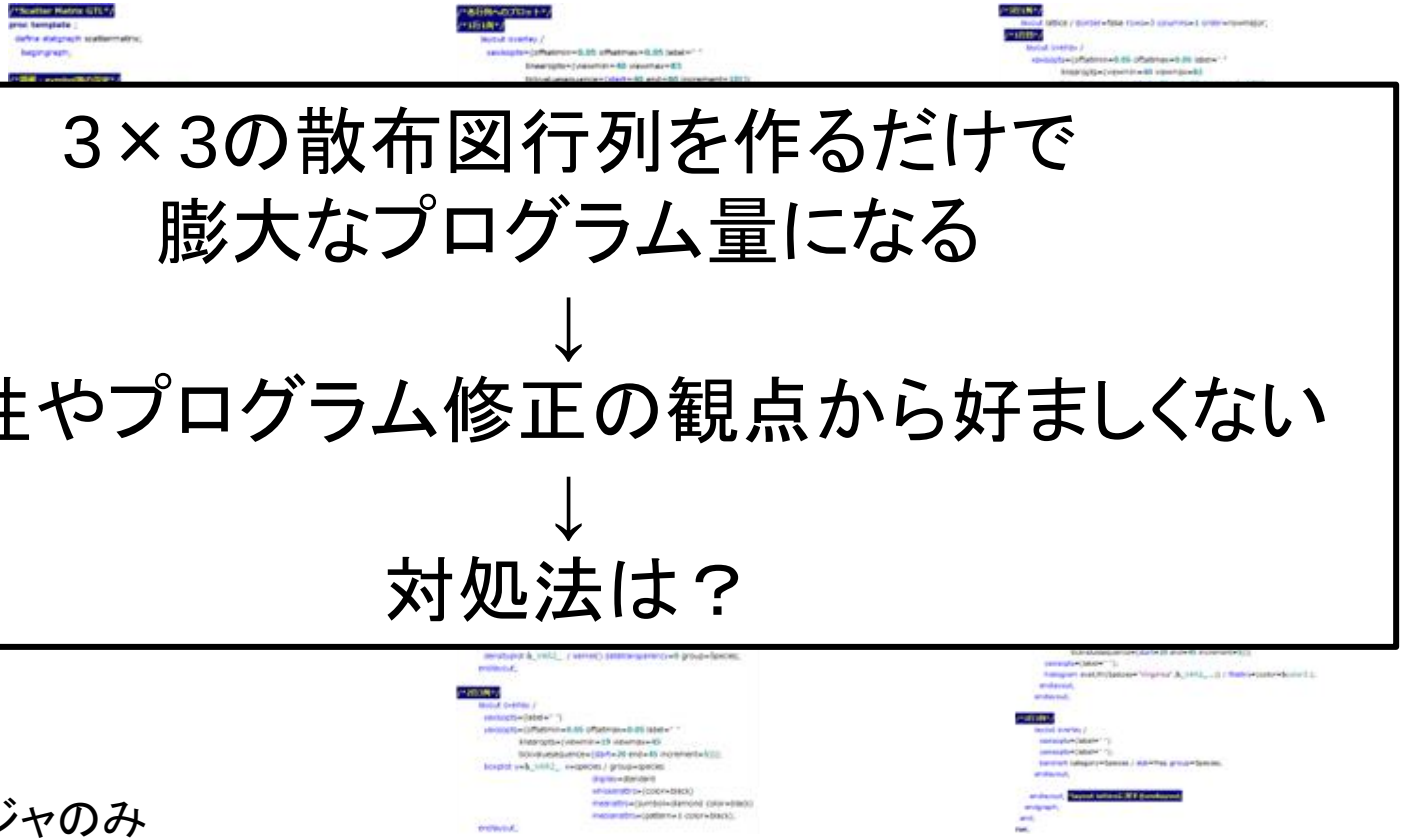
SASによる出力



出力に使用したプログラムと課題

※templateプロシジャのみ

出力に使用したプログラムと課題



３×３の散布図行列を作るだけで
膨大なプログラム量になる

↓

可読性やプログラム修正の観点から好ましくない

↓

対処法は？

※templateプロシジャのみ

目次

プログラムの簡略化について

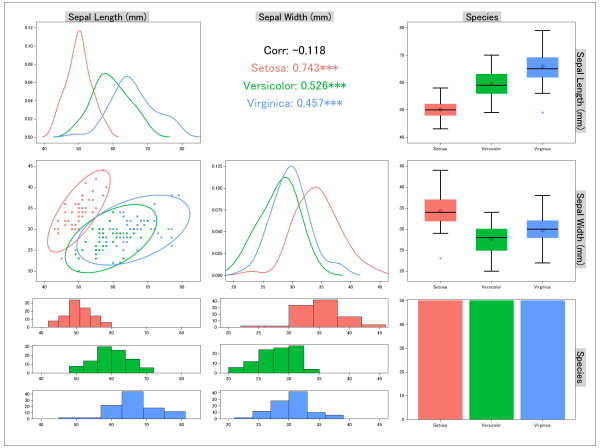
下記のようにマクロ化することで、templateステートメントの記載を簡潔にできる

```
layout overlay /
  xaxisopts=(label=" ") yaxisopts=(label=" ");
  densityplot SepalLength / kernel() datatransparency=0 group=Species;
endlayout;
```

⇩

```
%macro density (_VAR_=);
  layout overlay /
    xaxisopts=(label=" ") yaxisopts=(label=" ");
    densityplot &_VAR_. / kernel() datatransparency=0 group=Species;
  endlayout;
%mend density;
```

SASによる出力(再掲)



プログラムの簡略化について

下記のようにマクロ化することで、templateステートメントの記載を簡潔にできる

```
proc template ;
 define statgraph scattermatrix;
  begingraph;
   layout lattice / border=false rows=3 columns=3 order=rowmajor;
    %density(VAR=SepalLength);
    %corr;
    %boxplot(VAR=SepalLength);
    %scatter(VAR_X=SepalLength, VAR_Y=SepalWidth);
    %density(VAR=SepalWidth);
    %boxplot(VAR=SepalWidth);
    %histogram(VAR=SepalLength);
    %histogram(VAR=SepalWidth);
    %bar(VAR=Species);
   endlayout;
  endgraph;
 end;
run;
```

描画するプロット数分で書ける！

(細かい設定を加えるならもう少し複雑に…)

目次

まとめ

- 本発表では、柔軟に且つ多様な情報を含めた散布図行列について、SASのGTLを用いた実装方法を紹介した。

- layout lattice等を活用することで、イメージ通りの散布図行列を作成できた。一方で、細かいオプション指定によりプログラムの量が増えると、可読性やメンテナンスの問題が生じるので、マクロ化する等の工夫が必要である。

- 5×5以上の散布図行列では1つ1つの描画が小さくなり視認性が下がるので、figureのサイズや表示する変数の数に注意すること。

参考文献等

- SAS Help (SGSCATTERプロシジャ・GTL)
  https://documentation.sas.com/doc/ja/pgmsascdc/v_042/grstatproc/n1je1qlb5bvtypn1rhbeyqw5h9rc.htm
  https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatgraph/titlepage.htm

- 舟尾 暢男 SAS Graph Template Language（GTL） 超入門
  http://nfunao.web.fc2.com/files/Intoroduction_to_SAS_GTL.pdf

- 高浪洋平. SGプロシジャとGTLによるグラフの作成とODS PDFによる統合 解析帳票の作成 ～TQT試験における活用事例～. (SASユーザー総会2011)
  https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups11-b-19.pdf

- 日高優, 魚住龍史, 堀隆正 Graph Template Language（GTL）を用いた臨床試験におけるグラフデザインの応用 (SASユーザー総会2021)
  https://www.sas.com/content/dam/SAS/documents/event-collateral/2021/ja/sas-users-group-2021/slides/o-02-hidaka.pdf

- CRAN ggplot2 package
  https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

ご清聴ありがとうございました！