クレジット：

Mathematics and Informatics Center 文科系のための線形代数・解析II
2020 藤堂 眞治・松尾 泰・藤原 毅夫

ライセンス：

利用者は、本講義資料を、教育的な目的に限ってページ単位で利用することができます。特に記載のない限り、本講義資料はページ単位でクリエイティブ・コモンズ 表示-非営利-改変禁止 ライセンスの下に提供されています。

http://creativecommons.org/licenses/by-nc-nd/4.0/

本講義資料内には、東京大学が第三者より許諾を得て利用している画像等や、各種ライセンスによって提供されている画像等が含まれています。個々の画像等を本講義資料から切り離して利用することはできません。個々の画像等の利用については、それぞれの権利者の定めるところに従ってください。

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

最小二乗法 -1（線形近似）

MATLABのStatistics and Machine Learning Toolboxを使用して，

統計データの扱いを説明する．

データの入力とヒストグラム

データがExcel上に用意されているとする．

エクセルデータ Jap-Math-Sci.xslx

使用するのはこの内の，

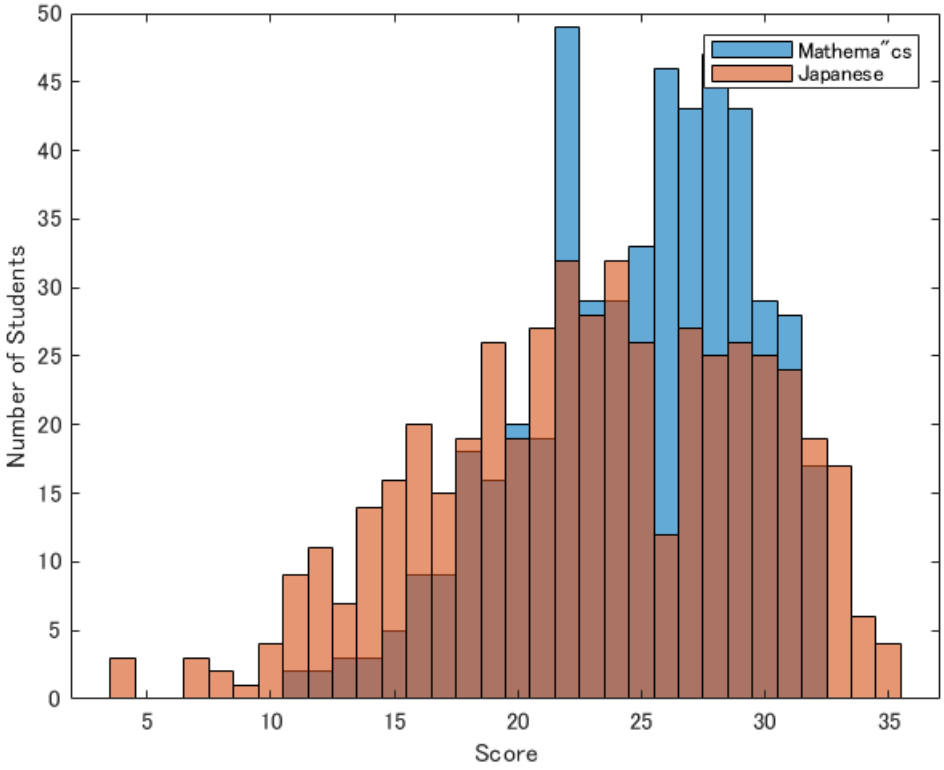
number（個体番号），sex（男女の別），JapA（国語得点），MathA（数学得点）である．

まず，Jap-Math-Sci.xslx からデータを読み込み、必要な列のデータ（5列目および7列目）のヒストグラムを描こう．

xlscファイルを同じディレクトリの中に置いても見つからないというエラーがでることがある．その場合には「ファイルの検索」から，ファイル名を検索し，そのpathを指定する．

```
ds=xlsread('Jap-Math-Sci.xlsx');
histogram(ds(:,5));%Score of Jap.
hold on
histogram(ds(:,7));%Score of Math.
hold off
xlabel('Score')
ylabel('Number of Students')
legend('Mathema"cs','Japanese')
```


Mathematics and Informatics Center **文科系のための線形代数・解析** II 2020 **藤堂 眞治・松尾 泰・藤原 毅夫** CC BY-NC-ND



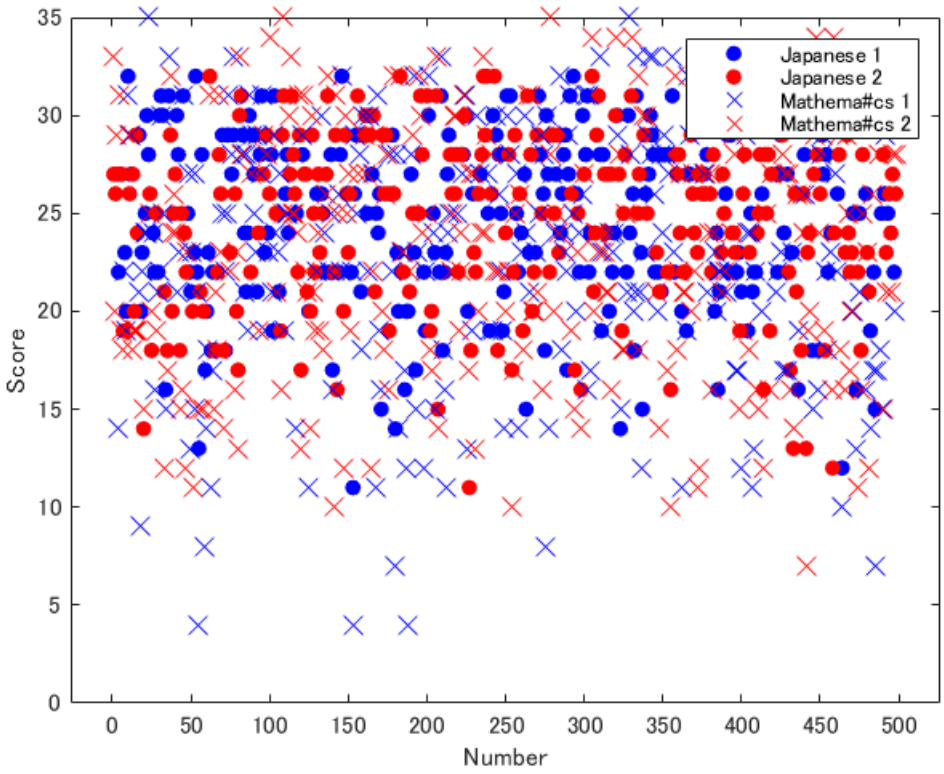
ここでは，ヒストグラムを描くコマンドhistogram を2回呼んで，データMathAとJapAの分布を重ねることにより，分布の違いなどが分かる．

データの散布図

サンプルデータから，第**1**パラメタを**x**軸とし，第**2**パラメタを**y**軸とし散布図を描くコマンド scatter(x,y) が用意されている．

```
ds=xlsread('Jap-Math-Sci.xlsx');
gscatter(ds(:,3),ds(:,5),ds(:,4),...
'br','.',20,'off','Number','Score')
hold on
gscatter(ds(:,3),ds(:,7),ds(:,4),'br','x',10,'off')
hold off
legend('Loca#on','northeastoutside')
legend('Japanese 1','Japanese 2',...
'Mathema#cs 1','Mathema#cs 2')
```


Mathematics and Informatics Center **文科系のための線形代数・解析** II 2020 **藤堂 眞治・松尾 泰・藤原 毅夫**　CC BY-NC-ND



ここでは男女の別を違う色で示したいので，gscatterを使った．

gscatter(x,y,group) はgroup（性別）内のデータに従ってグループ分けされた（x,y）データ（x=Number，y=Score)の散布図を描く．ここではds.sexの中の数1，2（青、赤）に対応する．

平均，分散，相関

与えられたデータの平均値，標準偏差（分散の平方根）および相関係

数の計算は簡単である．mean=平均，std＝標準偏差．

```
ds=dataset('xlsfile', 'Jap-Math-Sci.xlsx');
mMathA=mean(ds.MathA)
```

```
mMathA = 23.0100
```

```
sigMathA=std(ds.MathA)
```

```
sigMathA = 6.4812
```

```
mJapA=mean(ds.JapA)
```

```
mJapA = 24.8377
```

```
sigJapA=std(ds.JapA)
```

```
sigJapA = 4.4945
```


Mathematics and Informatics Center 文科系のための線形代数・解析 II 2020 藤堂 眞治・松尾 泰・藤原 毅夫　CC BY-NC-ND

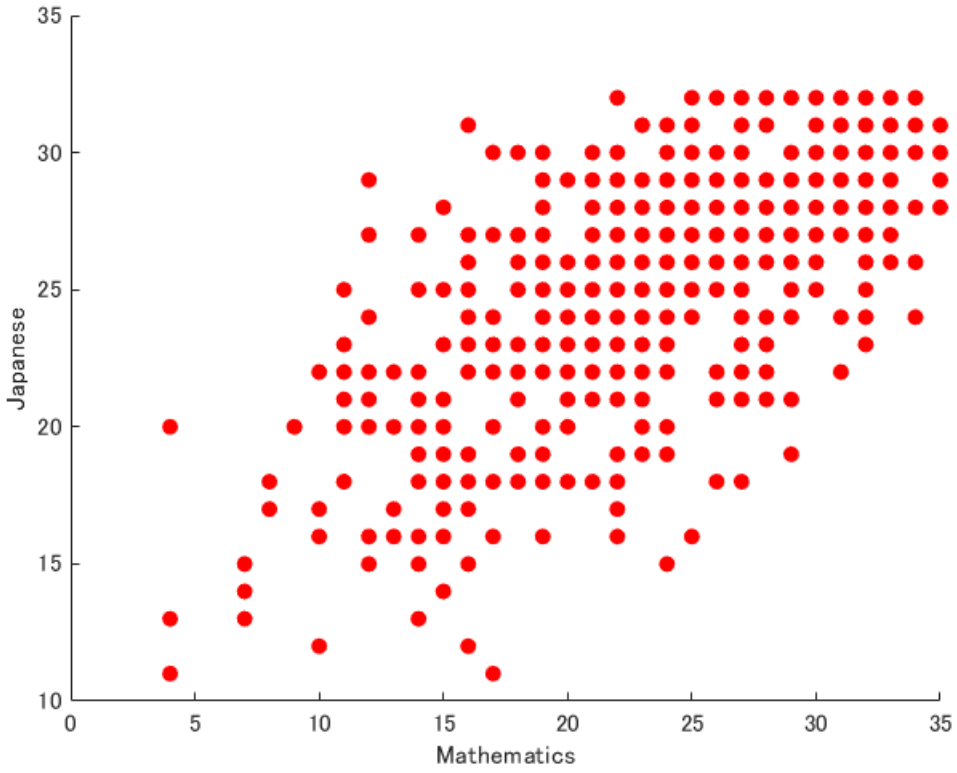
```
cor=corrcoef(ds.MathA,ds.JapA)
```

```
cor = 2×2
    1.0000    0.6607
    0.6607    1.0000
```

相関係数を見ると，例えばここで扱っている例では，MathAとJapAの間には0.66の相関があることを示している．

この相関は小さいのか，あるいは大きいのか．そのようなことを知りたいときには，MathAとJapAの値をそれぞれx軸，y軸にとってデータの分布を見ればよい．

```
scatter(ds.MathA,ds.JapA,'r','o','filled')
xlabel('Mathematics')
ylabel('Japanese')
```



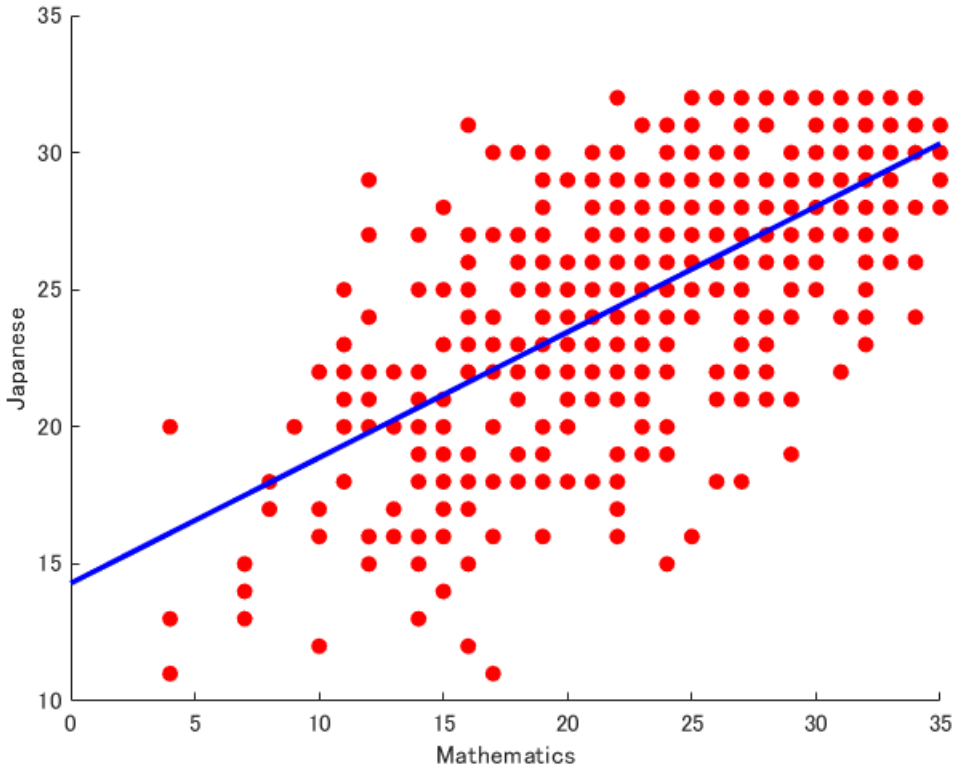
回帰直線

更に、図で分布の様子を見るだけでなく，polyfit を用いてデータを1次式で（最小二乗フィット）表すことを試みる．

```
ds=dataset('xlsfile', 'Jap-Math-Sci.xlsx');
scatter(ds.MathA,ds.JapA,'r','o','filled')
h=lsline;
xlabel('Mathematics')
ylabel('Japanese')
```


Mathematics and Informatics Center **文科系のための線形代数・解析** II 2020 **藤堂 眞治・松尾 泰・藤原 毅夫** CC BY-NC-ND

```
set(h,'linewidth',2,'color','bl');
```




```
polyfit(ds.MathA,ds.JapA,1)
```

```
ans = 1×2
    0.4581   14.2959
```

plyfitは任意の多項式のあてはめができるが,

ここでは n=1 として直線の当てはめをして，国語と数学の成績の相関として

$$\mathrm{JapA} = 0.4581 * \mathrm{MathA} + 14.2959$$

を得た.

課題 1

文科省，厚生労働省，財務省など公的機関のホームページを見て、公的に利用が許されているデータの所在を調べなさい.

課題 2

調べたデータを用いて、ここで行ったようなデータ処理を行え.

Mathematics and Informatics Center **文科系のための線形代数・解析** II 2020 **藤堂 眞治・松尾 泰・藤原 毅夫** CC BY-NC-ND