データの型・タイプ・欠測値・データの取得方法

データの型・タイプ

『遺伝統計学の基礎』の該当ページ配布

データの次元

データの欠測

統計学的説明と対処法

欠測のメカニズム別分類

- Missing completely at random (MCAR)
- この場合は、解析結果に影響は出ない
- しかしながら、このような場合はほとんどない
- Missing at random (MAR)
- MCARではないけれど、説明変数の情報によってデータ欠測率が予測できて、その率の下でランダムな場合
- その仮定の下では(仮定が正しく、仮定を用いて適切に補正できれば)、解析結果に影響を出さずにできる
- しかしながら、予測できるという仮定が正しいという保証がないので、あくまでも、仮定に立脚したものであり、その仮定が間違っていれば、結果を誤った方向に導く
- 仮定がまずまず妥当そうなら、MCARでないところをMCARと思って解釈するよりはまし、と考える
- Missing not at random (MNAR)
- 現実的にはこの場合ばかり
- 最も解析結果に影響を及ぼす場合は、Y ~ X で考えているときに、Yが大きいほど、Xの値に応じて欠測率が変わる場合など

対処法

- データ解析手法は、値がすべて詰まっていることを前提にしたものがほとんど
- なんとかして「統計的に妥当な」穴埋めがしたい
- Imuputation法(推定法)
- 削除法(欠測値がある標本を使わない)。。。それ自体が結果に影響を及ぼしうるという問題もある。また、変数が増えると、ほとんどすべての標本が使えなくなるリスクもある
- 欠測値がないことを前提にした解析手法の開発

臨床試験におけるルール

研究における欠測値の値と臨床試験におけるそれとでは、少しスタンスが異なる。

- 臨床試験では、保守的(患者さんへの悪影響を避けたい・無駄な医療費増を避けたい、という力)であることが、研究よりも強く求められる一方で、認可してほしいという力も研究に比べると大規模である
- 統一ルールが必要になる

講義内課題

配布資料 臨床試験の欠測データの取り扱いに関する最近の展開と今後の課題について- NAS レポート，EMA ガイドライン，estimandと解析方法の概説-(http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/pdf/data_140704_1.pdf)を速読し、A4 1 枚程度にまとめなさい。

まとめは、PandAにアップロードしなさい。

予定：１０分

データを取得したらするべきこと

- ステップ 1
  - サンプル数 n x 項目数 mにする　+ 項目はカテゴリか量的かに 2 分する
- ステップ 2　個々の項目の分布を知る
  - 量的項目はヒストグラム(をひたすら)描く。全項目で描く。“hist(X)”
  - 多すぎてさすがに・・・というときは、ソートプロットして個々の分布を曲線にした上で、多数の項目からの多数の曲線を一まとめに描く。“matplot(apply(X,2,sort),type=”l“)
  - カテゴリカル項目は項別集計する。全部する。“table(X)”
- ステップ 3　2 項目のペアでの分布・集計をする
  - カテゴリ x 量
  - カテゴリ x カテゴリ： 2 元テーブル “table(X[,1],X[,2])”
  - 量 x 量：コプロット“pairs(X)”
- ステップ 4　項目ペアで相関を計算して、相関係数行列の全体像を描く
  - カテゴリカルか量的かもいったん忘れてひたすら相関係数を作って行列にして絵を描く
  - 全体の様子を知るためには、全体を 1 つの絵にするのが一番 “cor.mat <- cor(X);image(cor.mat)”
- ステップ 5　サンプル間の違い評価する
  - 距離行列を作る “dist.out <- dist(t(X))”
  - 距離行列はクラスタリングする “plot(hclust(dist.out)))”
  - 内積を取っておく “ip.mat <- X %*% t(X)"
  - 内積を取ったら固有値分解して固有値をプロットする “eigen.out <- eigen(ip.mat); plot(eigen.out[[1]])”

実践編

サンプル数

```
n.sample <- 1000
n.sample
```

```
## [1] 1000
```

第 1 変数は 2 値型

```
X1 <- sample(0:1, n.sample, replace = TRUE, prob = c(0.3,0.7) )
```

データがとれたら、分布を確認する

```
table(X1)
```

```
## X1
##   0   1
## 312 688
```

第２説明変数も、２値型

```
# X2をX1とは、独立に作る
X2 <- sample(0:1, n.sample, replace = TRUE, prob = c(0.6,0.4) )
```

```
# X1, X2間に関連を導入する
X2[which(X1==1)] <- sample(0:1,n.sample,replace = TRUE, prob=c(0.55,0.45))
```

```
## Warning in X2[which(X1 == 1)] <- sample(0:1, n.sample, replace = TRUE, prob
## = c(0.55, : 置き換えるべき項目数が、置き換える数の倍数ではありませんでした
```

```
table(X2)
```

```
## X2
##   0   1
## 556 444
```

２値型２変数の関係は、すぐに２ｘ２表形式で確認する

```
X1X2 <- data.frame(X1,X2)
table(X1X2)
```

```
##    X2
## X1    0   1
##   0 185 127
##   1 371 317
```

これだけでは、関係があるのかないのか、関係が強いのか弱いのかわからないので、最も単純な処理をする

```
chisq.test(table(X1X2))
```

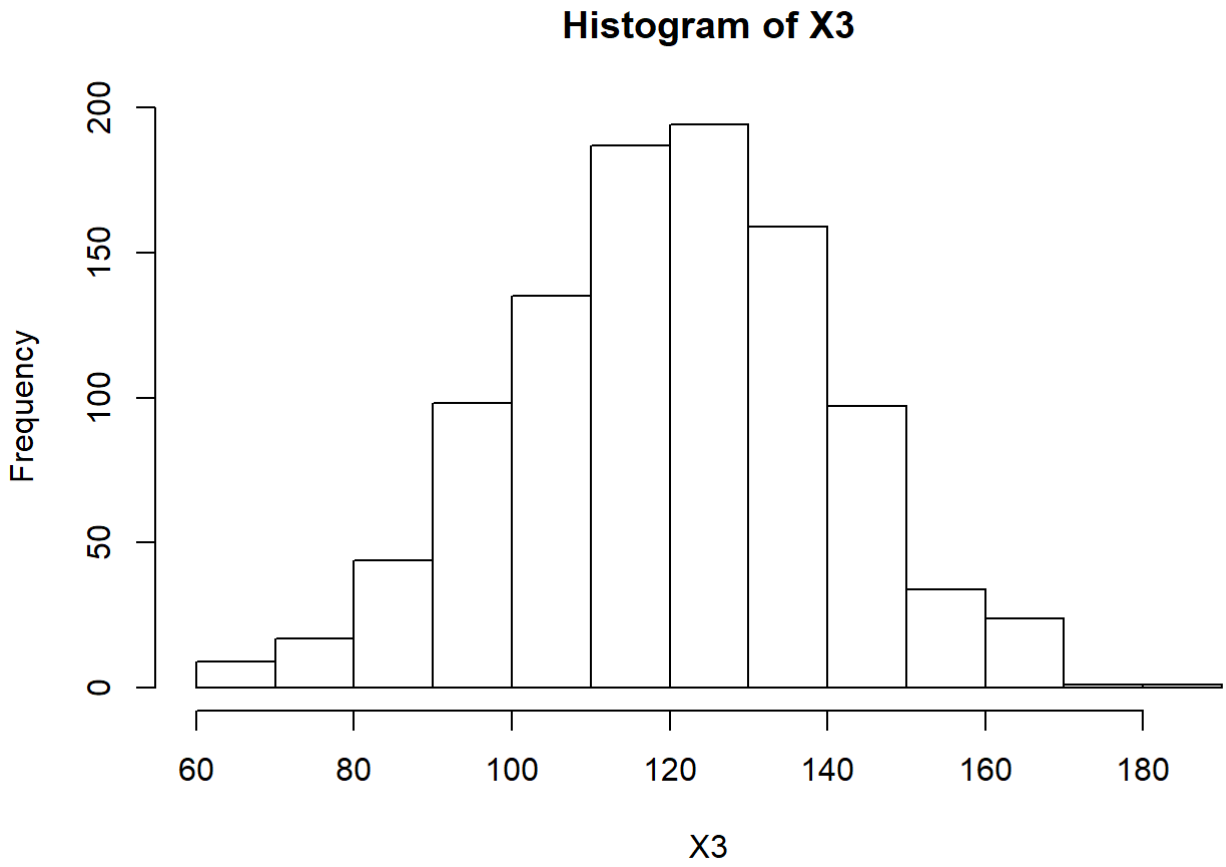
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(X1X2)
## X-squared = 2.2951, df = 1, p-value = 0.1298
```

X3,...,X6は相互に関係のある量的変数

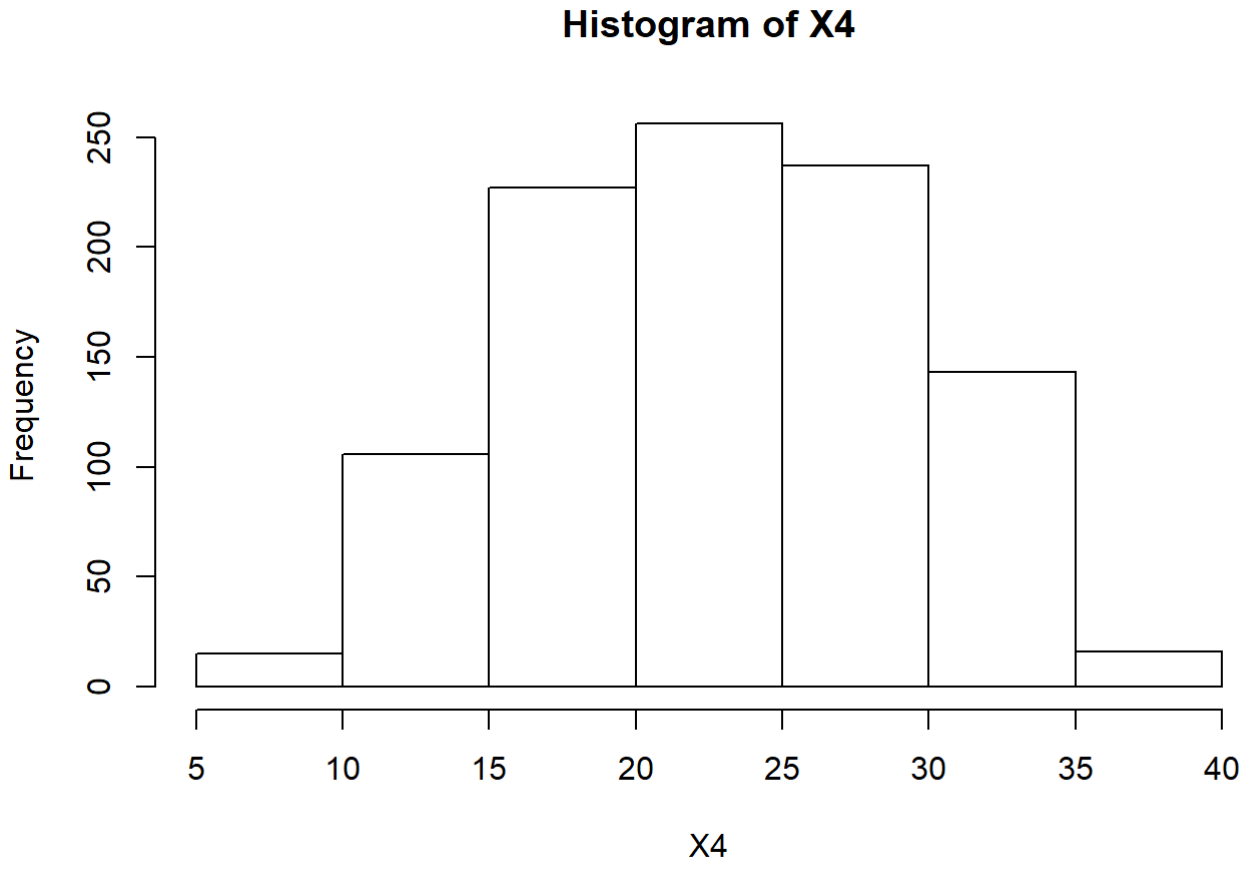
```
X3 <- rnorm(n.sample,120,20)
X4 <- c(rnorm(n.sample*0.7,20,5),rnorm(n.sample*0.3,30,3))
X5 <- rbeta(n.sample,2,4)
X6 <- X5 * 5 + X4*0.1 + rnorm(n.sample,0,0.01)
```

それぞれの分布をみる

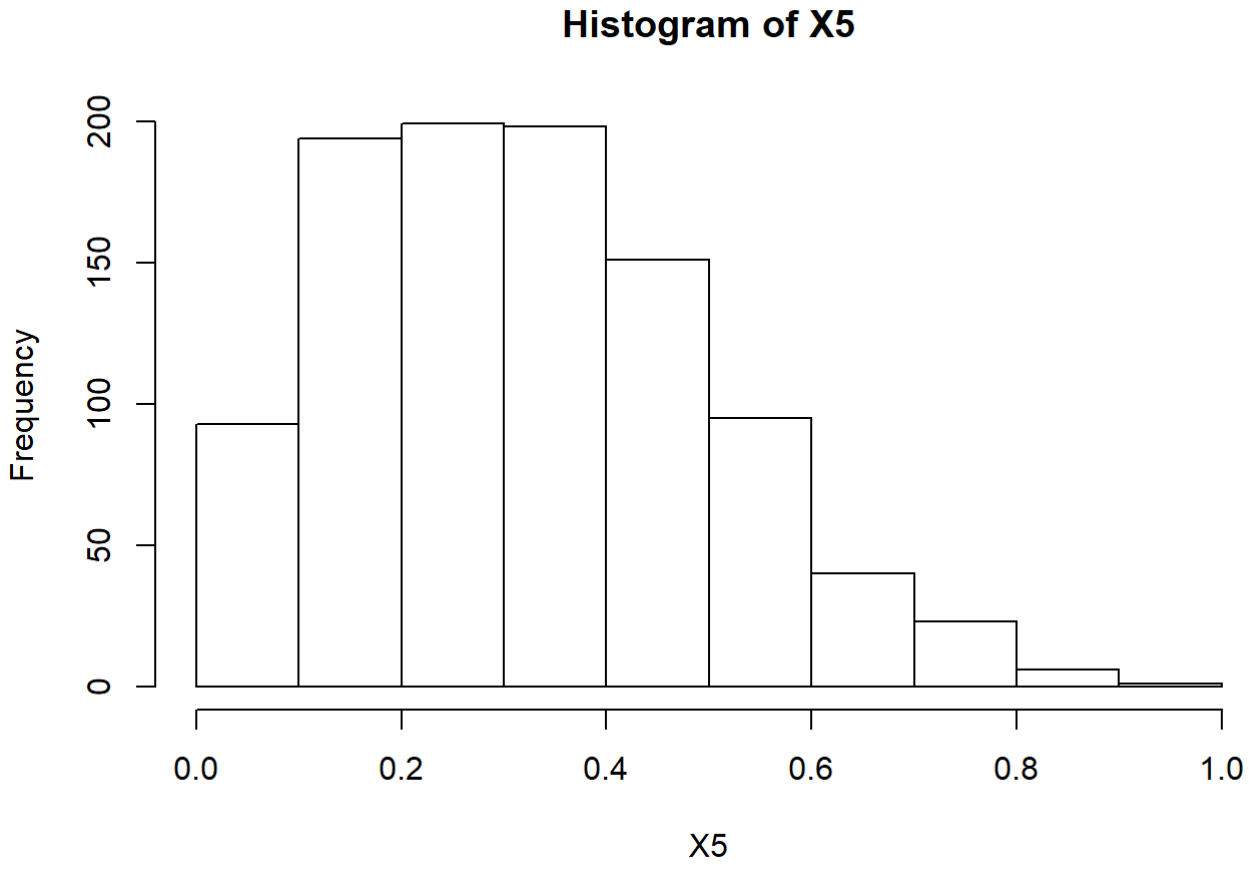
```
hist(X3)
```


Histogram of X3
Frequency
0
50
100
150
200
60
80
100
120
140
160
180
X3

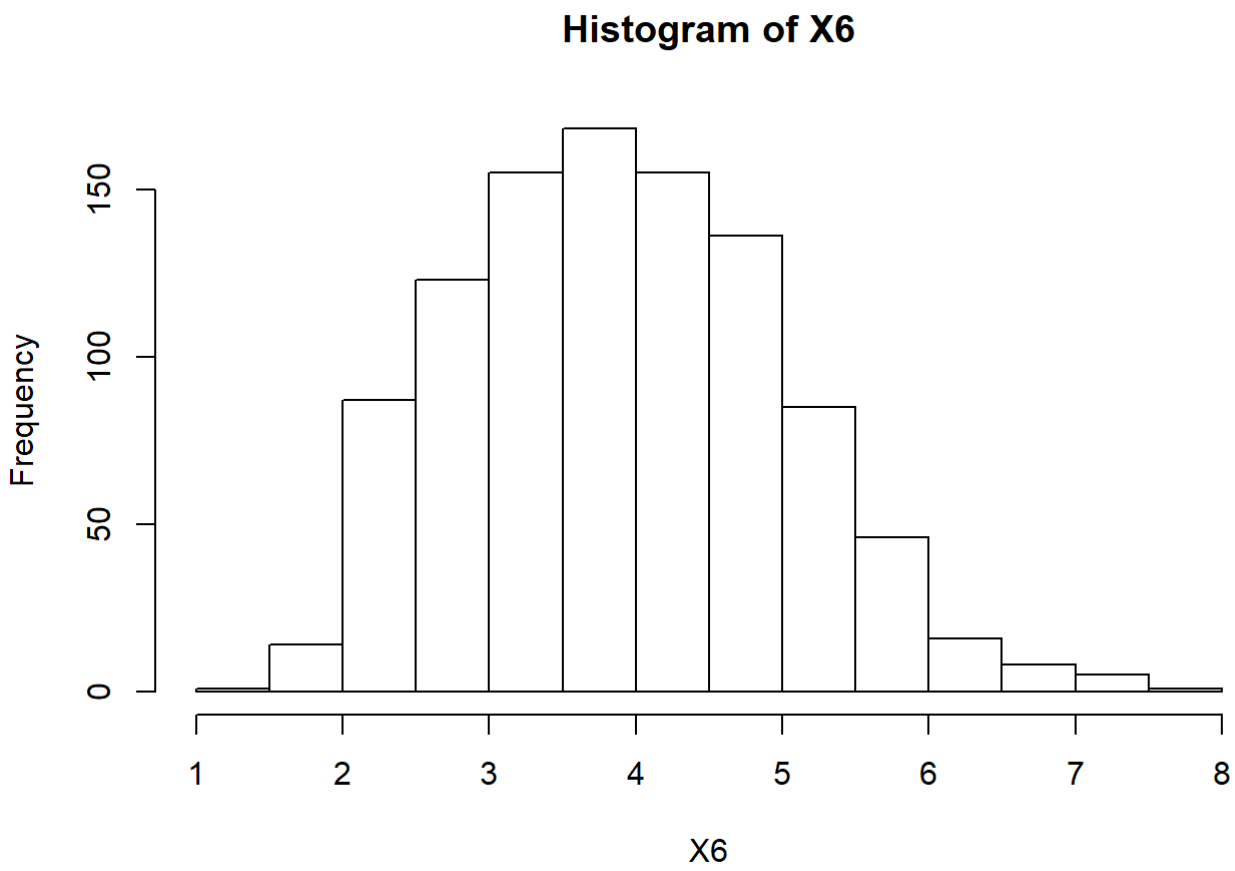
```
hist(X4)
```


Histogram of X4
Frequency
0
50
100
150
200
250
5
10
15
20
25
30
35
40
X4

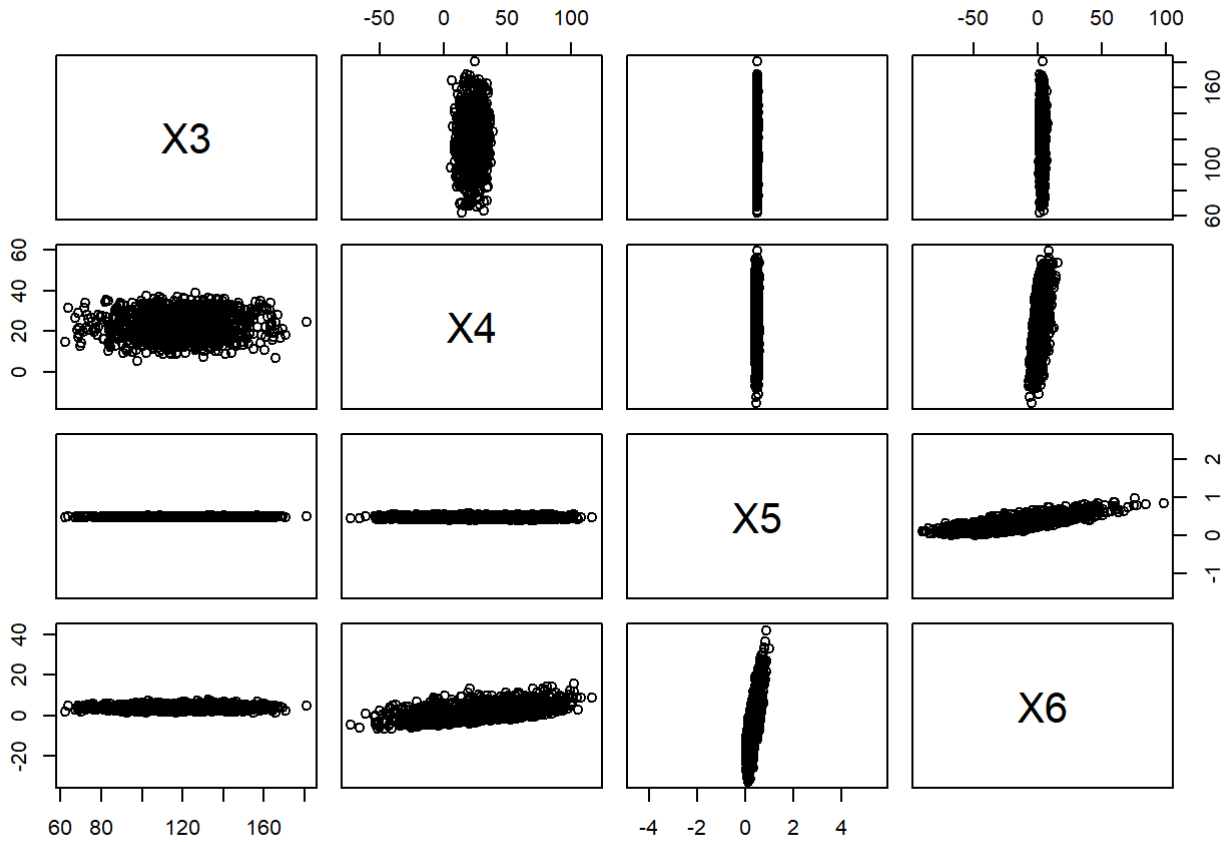
```
hist(X5)
```




```
hist(X6)
```

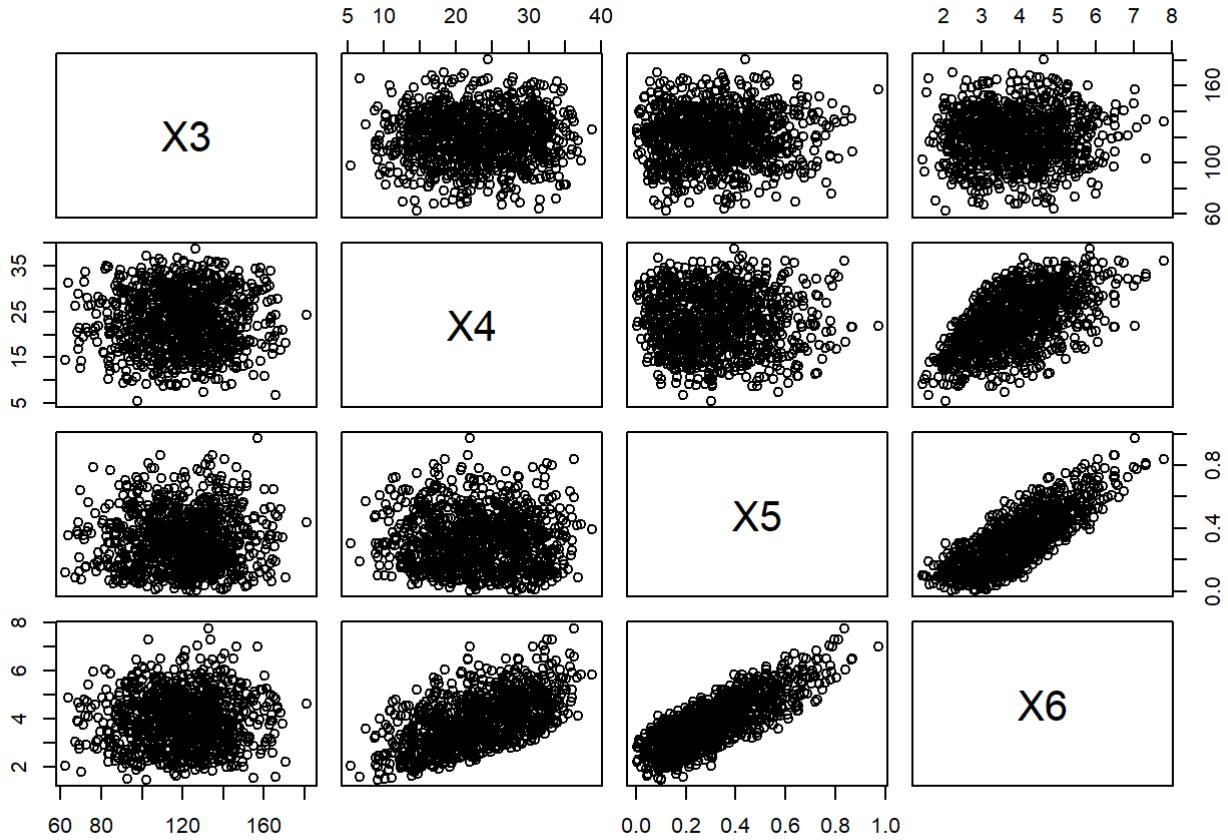



```
pairs(cbind(X3,X4,X5,X6),asp=TRUE)
```



変数ごとに分散が異なると、わかりにくい場合もあるので、低分散変数の表示幅を狭くして表示するのもよい

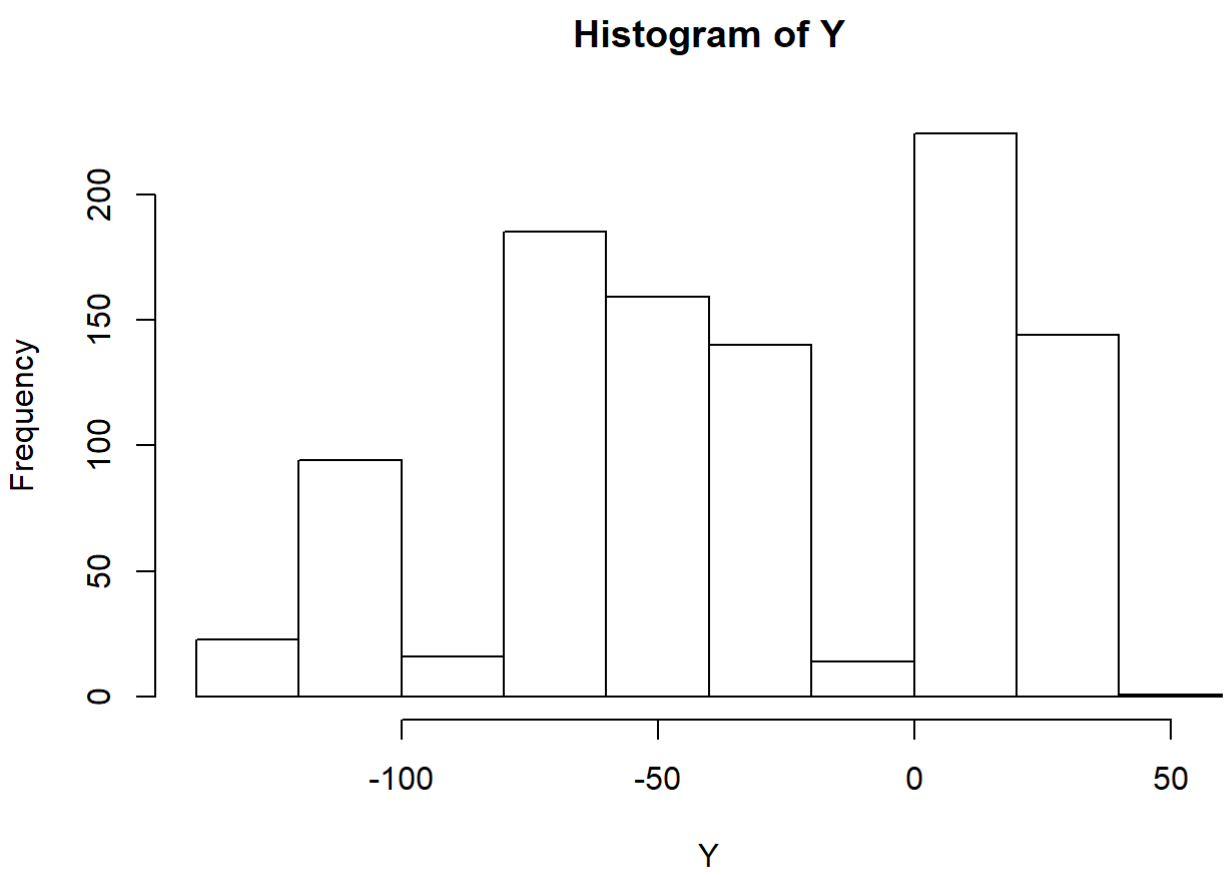
```
pairs(cbind(X3,X4,X5,X6))
```



被説明変数Yは、X1,...,X6に依存して決まる値とする

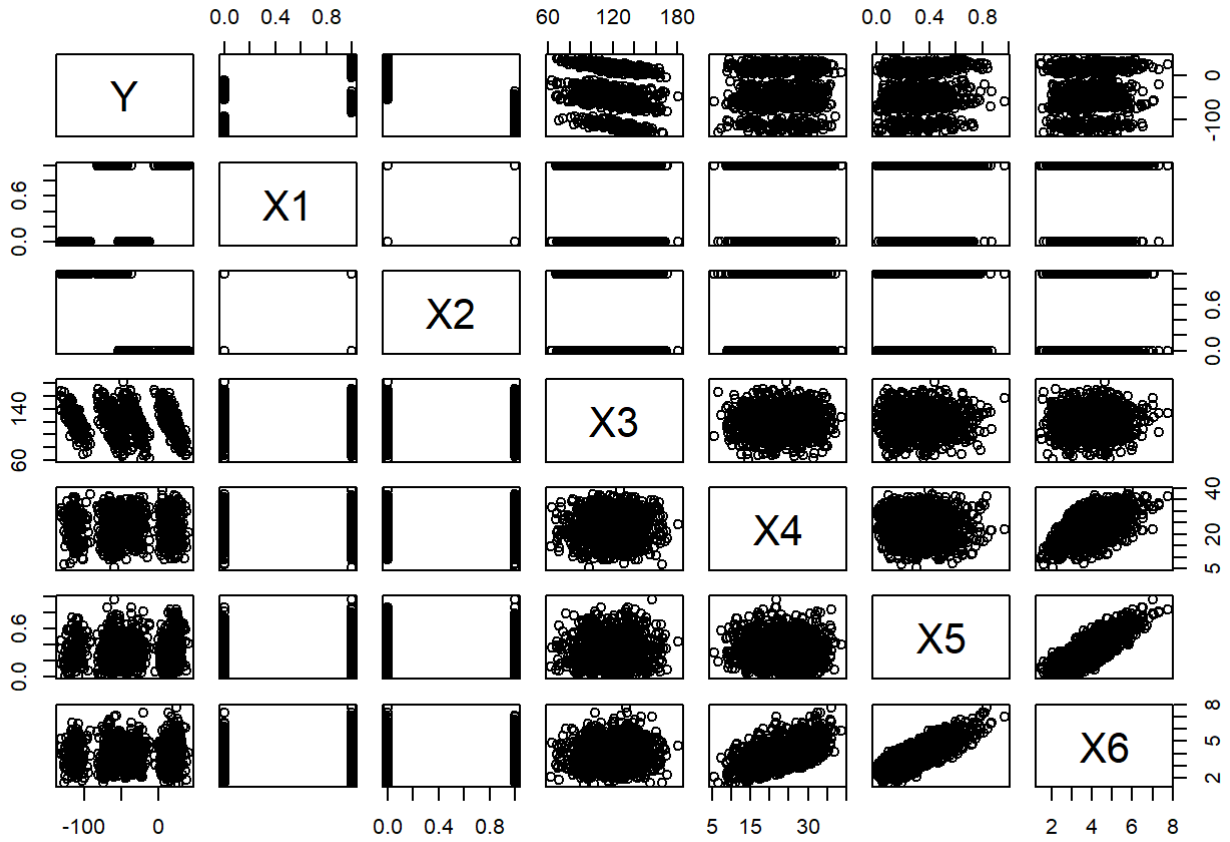
```
Y <- X1 * 50 + X2 * (-80) + X3 *(-0.3) + X4 *0 + X5 * 10 + X6 * 0.2 + rnorm(n.sample,0,5)
```

```
hist(Y)
```

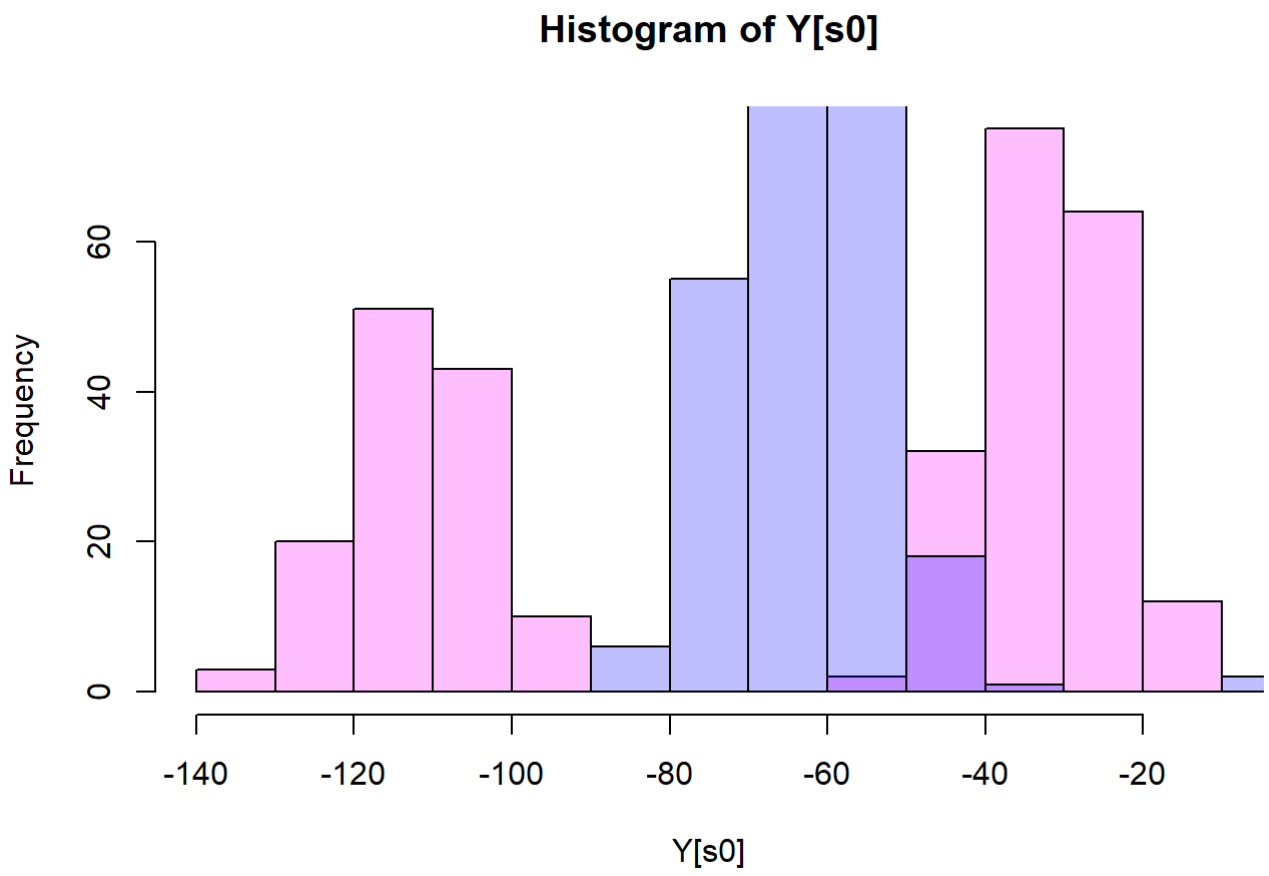


とにかくまとめて視覚化

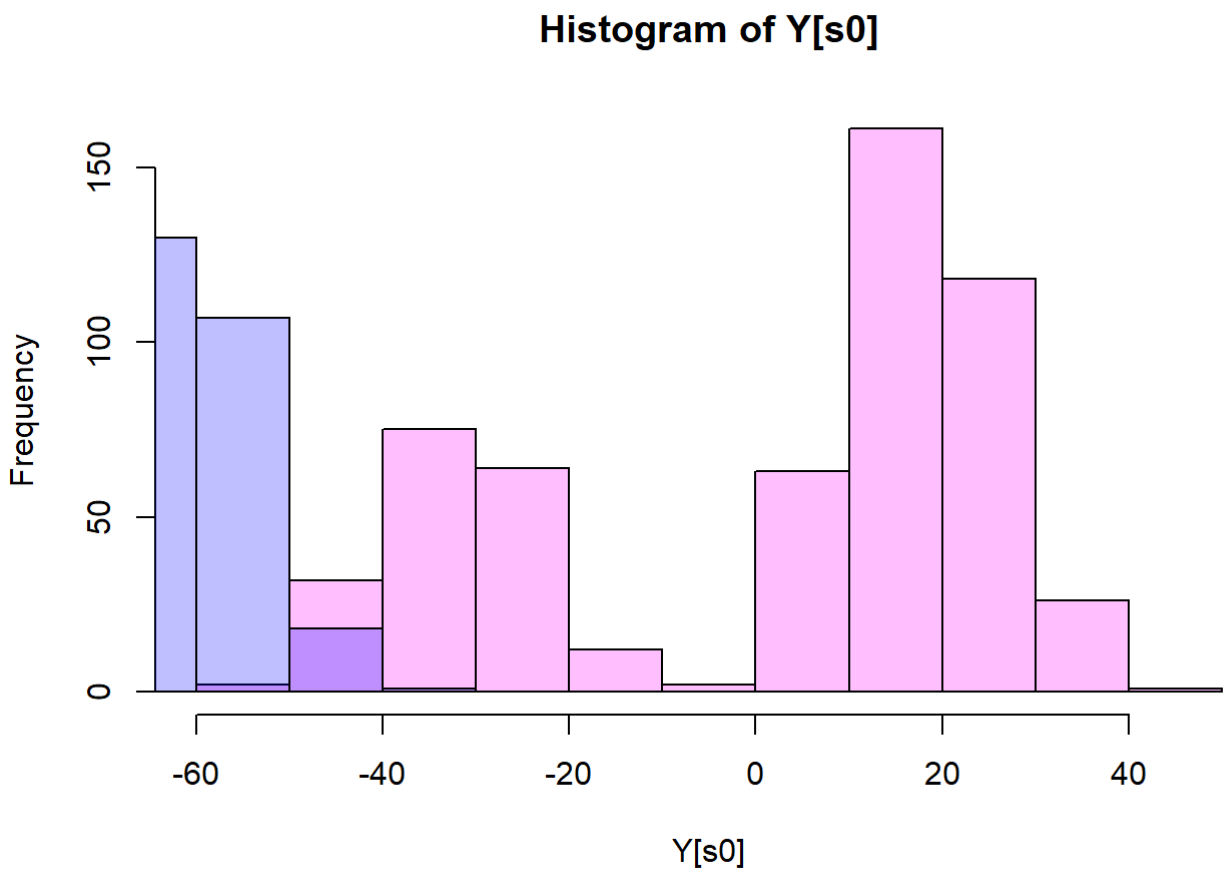
```
Whole <- cbind(Y,X1,X2,X3,X4,X5,X6)
pairs(Whole)
```




```
s0 <- which(X1==0)
s1 <- which(X1==1)
hist(Y[s0],col = "#ff00ff40")
hist(Y[s1],col = "#0000ff40",add=TRUE)
```




```
s0 <- which(X2==0)
s1 <- which(X2==1)
hist(Y[s0],col = "#ff00ff40")
hist(Y[s1],col = "#0000ff40",add=TRUE)
```


Histogram of Y[s0]
Frequency
0
50
100
150
-60
-40
-20
0
20
40
Y[s0]