R.graph

boxplot

その意味するもの

・データの分布を見る（比較する）
・真ん中の太い線：中央値 Median（50%）、平均ではないので注意
・箱の範囲：25% から 75% の範囲（これを四分範囲と呼ぶ）（つまり、半分のデータはこの範囲に入る）
・ひげの範囲：箱の端から箱の長さの 1.5 倍以内にある「実際の」数値の最大のもの最小のもの
・外れている○印：外れ値（ひげの範囲外のもの）

金先生による図解

■なぜ、平均と標準偏差ではないか

・データのばらつき具合（分布）は多様。（さまざまな分布がある）
・平均（相加平均）（Mean）と標準偏差は、「正規分布」（normal distribution）を前提としている。
・正規分布を前提としていないデータも、これ一つでばらつき具合を見て取れるところが便利。

例１（データの型　その１）

・二種類のデータを比較できるように箱ひげ図を描いてみる。

■ 1) データは、以下の形式（タブ区切り）でテキストファイル保存

group	ms
High	438
High	374
High	313
High	337
High	393
High	432
High	380
High	390
High	354
High	322
High	328
High	305
High	386

High	348
High	271
High	398
High	401
High	380
High	324
High	347
High	350
High	234
High	327
High	375
High	325
High	338
High	366
High	348
High	398
High	290
High	384
High	443
High	303
High	343
High	358
High	393
High	363
High	411
High	389
High	379
High	246
High	408
High	393
High	326
High	405
High	321
High	353

High	361
Low	343
Low	470
Low	337
Low	353
Low	327
Low	326
Low	332
Low	393
Low	395
Low	435
Low	375
Low	311
Low	331
Low	303
Low	331
Low	369
Low	313
Low	351
Low	374
Low	390
Low	362
Low	285
Low	382
Low	298
Low	347
Low	375
Low	380
Low	364
Low	429
Low	375
Low	401
Low	307
Low	401

Low	394
Low	350
Low	351
Low	380
Low	340
Low	398
Low	351
Low	458
Low	427
Low	311
Low	341
Low	389
Low	363
Low	334
Low	386

■ 2) R で読み込む

```
items <- read.table(choose.files(), header=T, sep="\t")
```

■ 3) 各グループの値だけを取り出す

```
hi <- c(items$ms[items$group=="High"])

li <- c(items$ms[items$group=="Low"])
```

例２（データの型　その２）

■ 1) データは、以下の形式（タブ区切り）でテキストファイル保存

Hi	Lo
438	343
374	470
313	337
337	353
393	327
432	326
380	332
390	393

354	395
322	435
328	375
305	311
386	331
348	303
271	331
398	369
401	313
380	351
324	374
347	390
350	362
234	285
327	382
375	298
325	347
338	375
366	380
348	364
398	429
290	375
384	401
443	307
303	401
343	394
358	350
393	351
363	380
411	340
389	398
379	351
246	458
408	427

393	311
326	341
405	389
321	363
353	334
361	386

■ 2) R で読み込む

```
hilo <- read.table(choose.files(), header=T, sep="\t")
```

■ 3) 各グループの値だけを取り出す

```
hi <- hilo$Hi
```

```
lo <- hilo$Lo
```

■ 4) 箱ひげ図を描く

```
boxplot(hi, lo)
```

図に説明を足す

・main でタイトル
・ylab でｙ軸のラベル
・names=c("　", "　") で各箱の名前
・ylim=c(下限値 , 上限値)
・yaxp=c(下限値 , 上限値 , 目盛の数)

```
boxplot(hi, lo, main="High vs. Low", ylab="ms", names=c("High","Low"))
```

■ 5) メニューのファイルから「別名で保存」で、PDF ファイルで保存できる。

・もしくは、以下のコマンドでファイル保存。

```
png("J1boxplot.png")
boxplot(J1, main="Japanese all data", ylab="freq.")
dev.off()
```

外れ値は具体的に何なのか？　boxplot の中身を見てみる。

・boxplot のコマンドの結果を変数に入れてしまう！

```
j1 <- boxplot(J1, main="Japanese all data", ylab="sentence length")
```

・変数 ji の中身が、boxplot の中身
・中身のうち、箱ひげの図の基本的な数値は、$stat で出る。

```
> j1$stat
     [,1]
[1,]    2
[2,]    4
[3,]    7
[4,]   10
[5,]   19
attr(,"class")
      V1
"integer"
```

・上記の、[5,] が箱ひげの「ひげ」の最大値。それより上が「外れ値」となる。
・外れ値は、$out

```
> j1$out
 [1] 21 21 25 29 20 22 20 20 22 20 23 21 20 24 21 21 24 22 22 20 23 22 25 20 20
[26] 23 21 22 21 23 22 24 20 36 21 21 29 22 24 27 25 20 22 23 29 20 21 24 24 21
[51] 20 22 29 22 23 23 20 23 22 27
```

stripchart() と組み合わせる