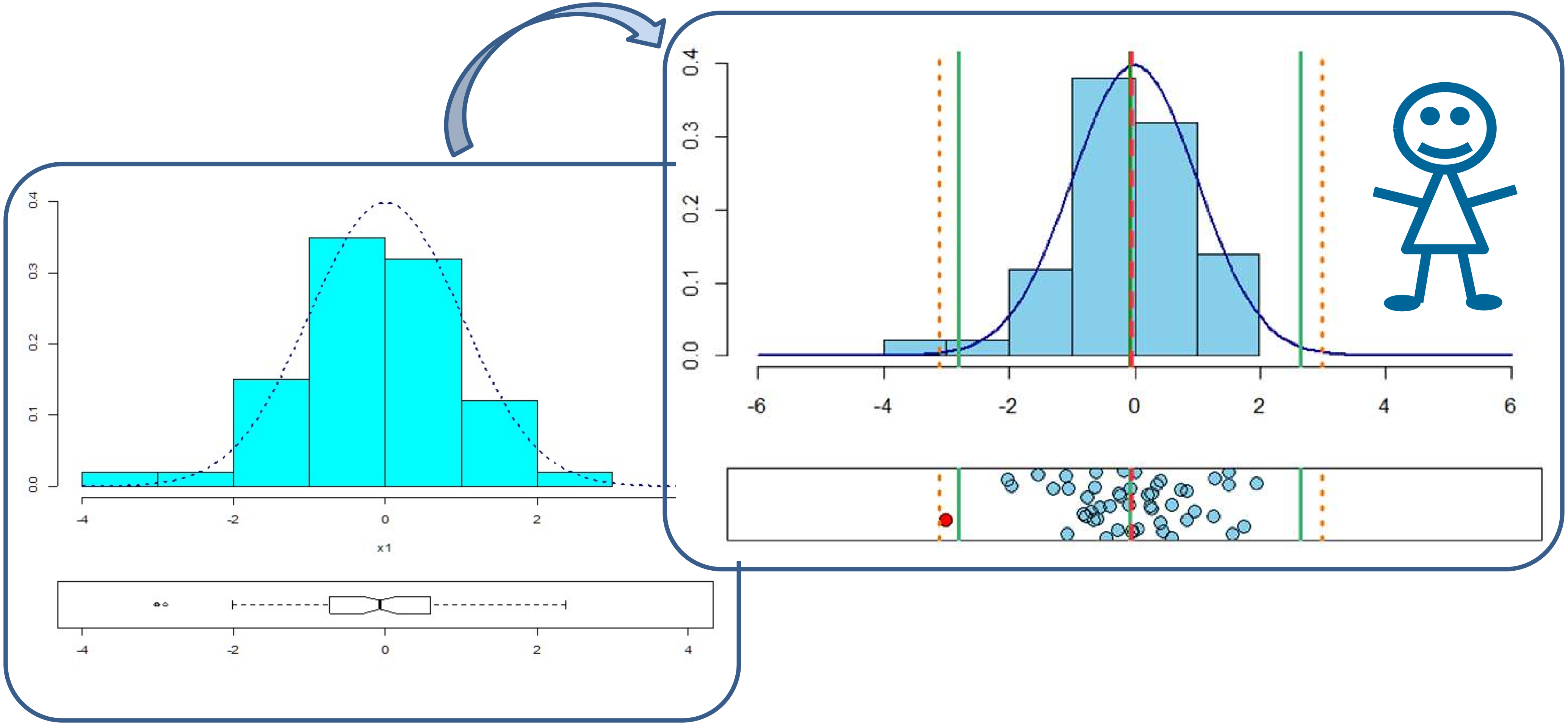
2016.06.04

Univariate data presentation with R

Rによる単変量データのプロット

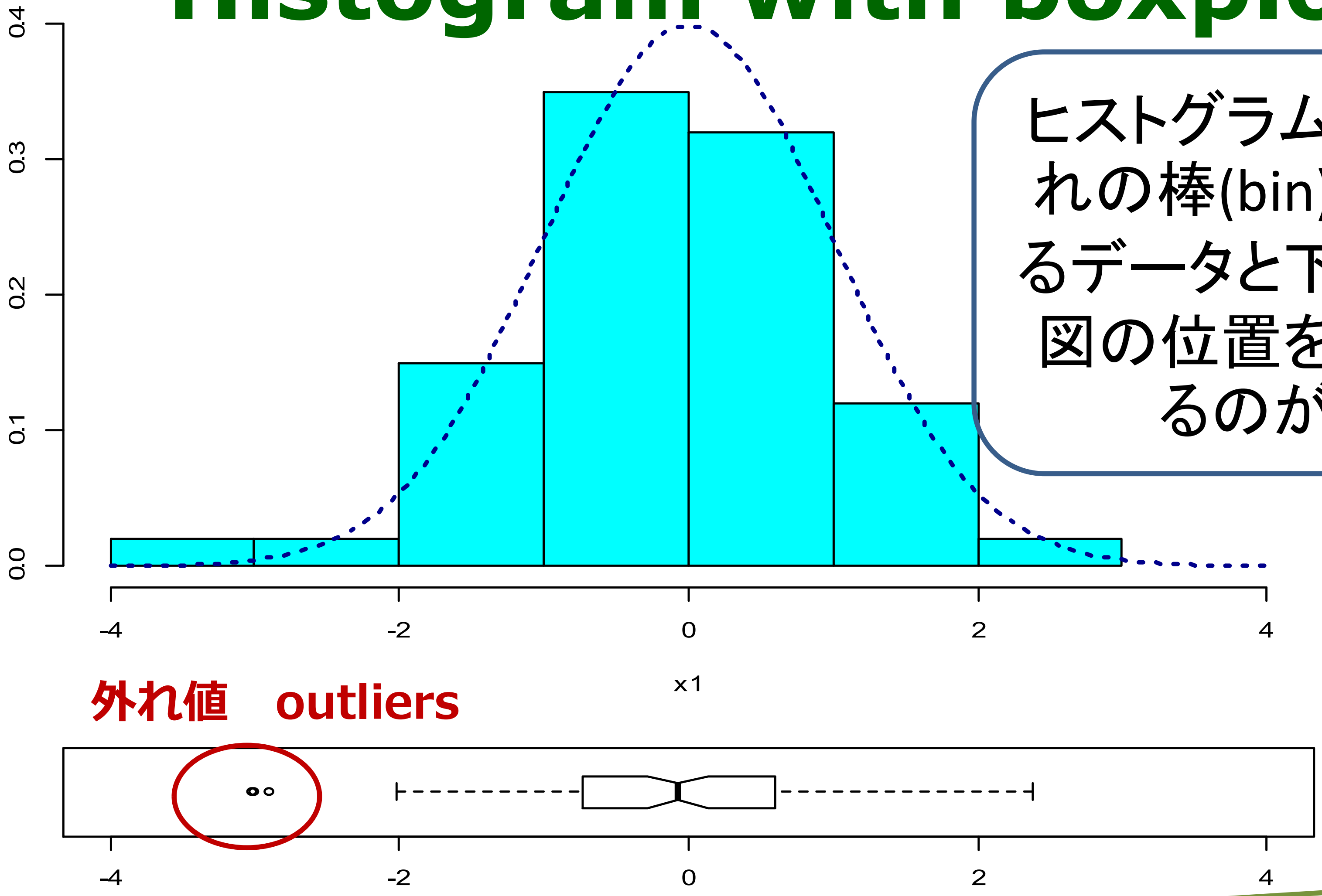
ヒンジによる箱ひげ図から四分位範囲へ

Histogram with box plot and interquartile range

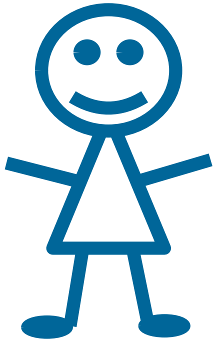


まずは箱ひげ図つきヒストグラム

Histogram with boxplot



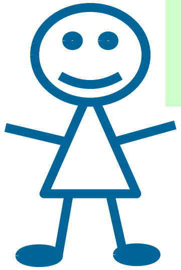
ヒストグラムのそれぞれの棒(bin)に含まれるデータと下の箱ひげ図の位置を揃えているのがミソ!



Data points in the lower window are included to the bins above!

前頁のプロット図のコード
Codes for the plot in the previous slide

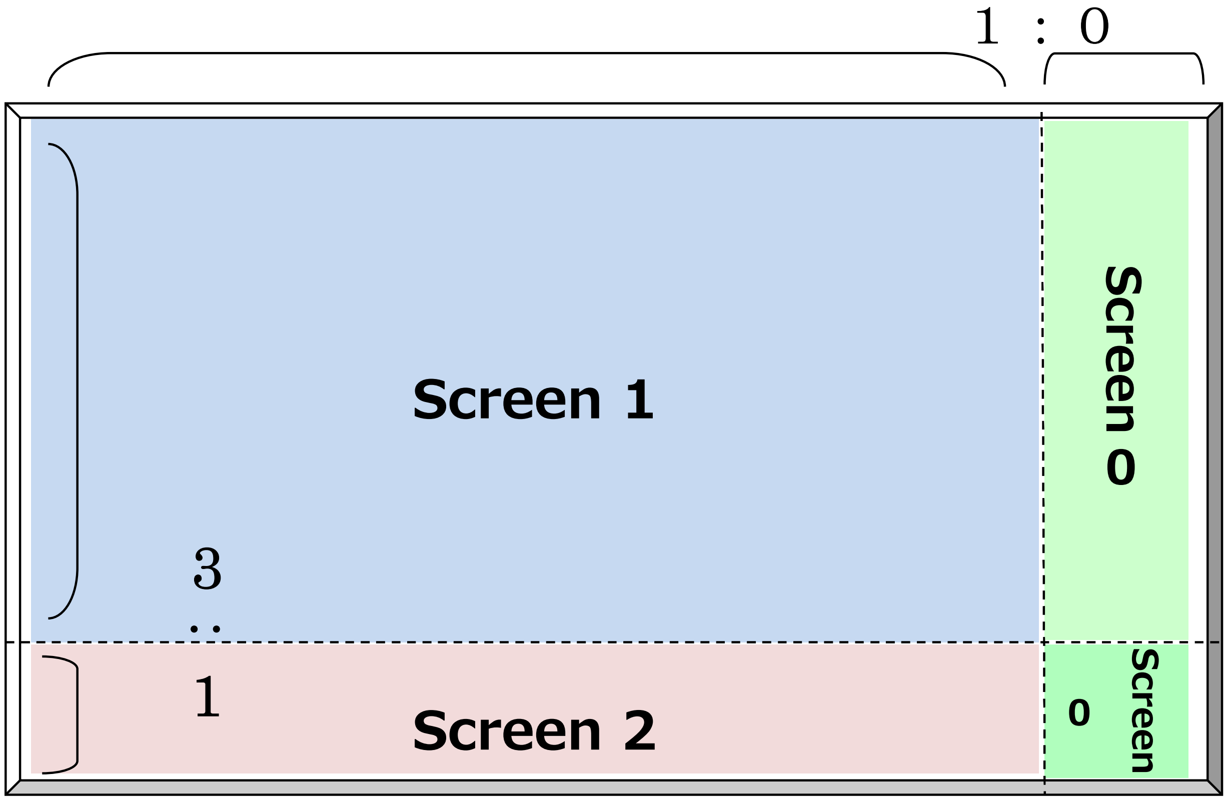
```
require(MASS)
set.seed(8)
x1 <- rnorm(100)                        # Standard normal distribution data
op <- par(no.readonly = TRUE)  # Save current graphic parameters
nf <- layout( matrix( c(1,0,2,0), 2, 2, byrow=T ), c(1,0), c(3,1))
layout.show(nf)                        # Confirm the defined layout
par(mar=c(4,4,4,2))                    # margin(bottom, left, top, right)
  truehist(x1, ylim=c(0, 0.4), xlim=c(-4,4))
  curve(dnorm, col="darkblue", lwd=2, lty=3, add=TRUE)
par(mar=c(4,4,2,2))
  boxplot(x1, notch = TRUE, horizontal = TRUE, ylim=c(-4, 4))
par(op)                                # Restore graphic pamameters
```



[解説1] layoutによる画面分割

Exposition1: Screen separation by "layout"

layout関数は, 任意の比率で画面分割をすることができる。
Function "layout" devides screen with a designated ratio.



nf <- layout(matrix(c(1,0,2,0), 2, 2, byrow=T), c(1,0), c(3,1))

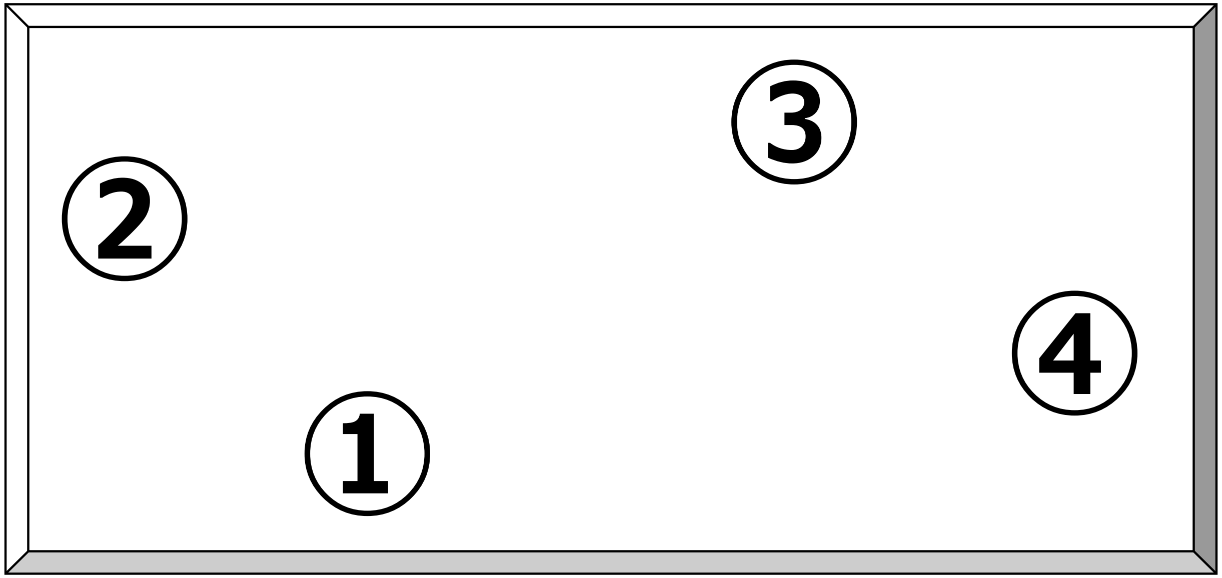
Screen No. **Separation ratios**

[解説2] マージンの指定方法
Exposition2: Margines

下と左は4インチ以上の指定が推奨されている

Down and left may need at least 4 inches.

```
par(mar=c(4,4,2,2))
        # (bottom, left, up, right)
```



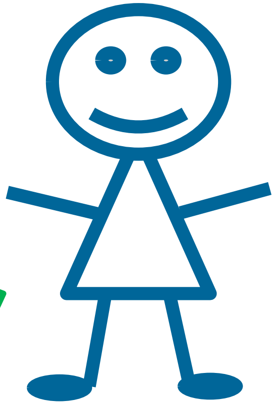
これで十分?

Is this enough?

　単体で使うならこれで十分。ただし、*n*が違うものを複数比較するときにはちょっと難あり。

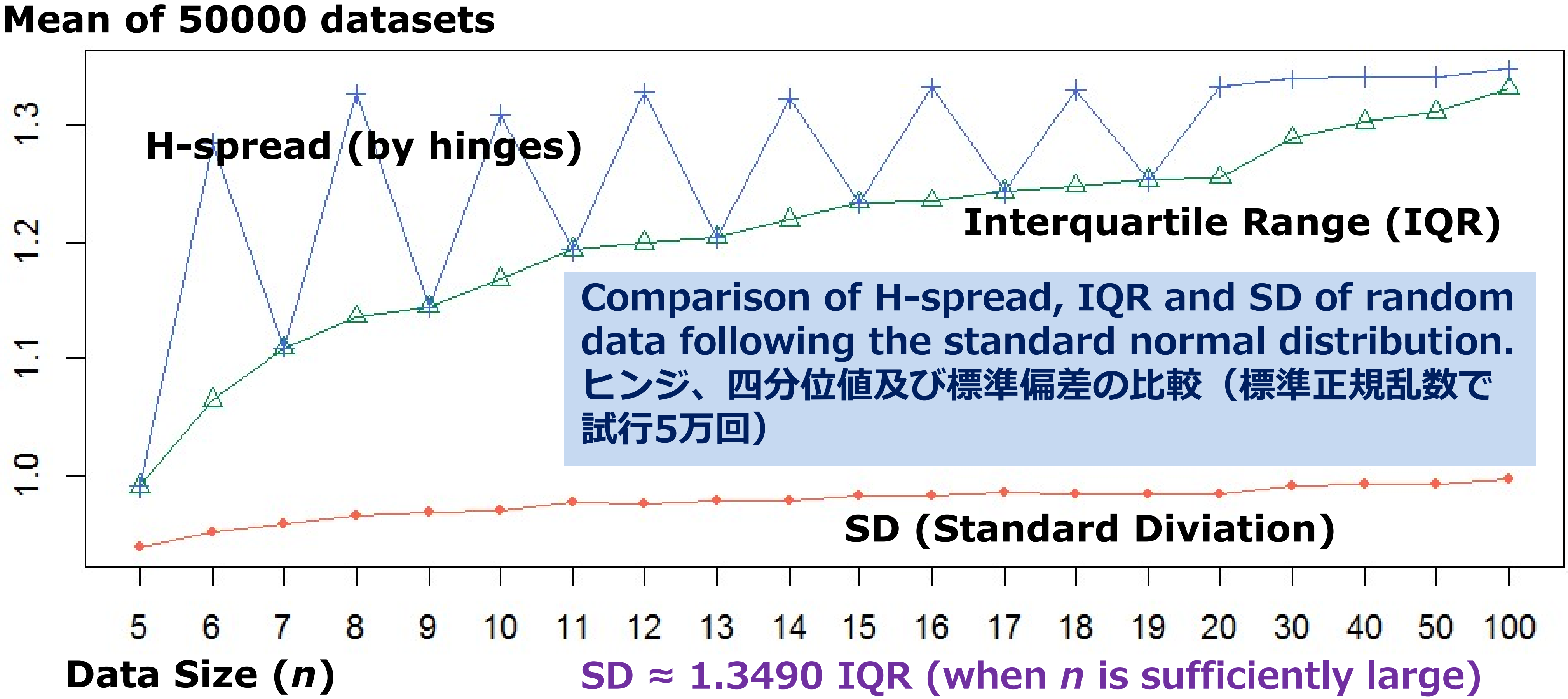
　Rの箱ひげ図は外れ値判定にヒンジを使用し、厳密には*n*が奇数か偶数かで基準が若干違う。

　Well, it's good enough for one dataset, but not for comparing datasets with different data sizes, since R uses hinges for the boundaries of box plot. The boundaries made by hinges slightly differ depends on whether the data size is odd or even, especially when the size is small.



ヒンジの問題点

Problem of H-spread based on hinges



Comparison of H-spread, IQR and SD of random data following the standard normal distribution.
ヒンジ、四分位値及び標準偏差の比較（標準正規乱数で試行5万回）

SD ≈ 1.3490 IQR (when *n* is sufficiently large)

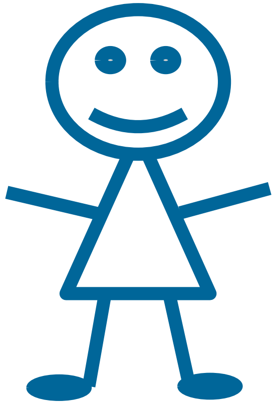
ヒンジによるH-spreadよりも四分位範囲(IQR)の方がnによる振れが少ない。標準偏差はnによる振れが少なく精度も高いが、頑健性がなく、外れ値が存在すれば精度の低下が著しい。

SD is the most stable and efficient among these three, but it’s not suitable for outlier detection since it sores with existence of an outlier.

前頁プロットの作成コード

Source code for the plot in the previous page

```
set.seed(11)              #  for reproductivity
n <- 50000                #  Number of simulation
d <- c(5:20, 30, 40, 50, 100)                        # Data size
m.IQR1 <- m.Hg1 <- m.sd1 <- rep(NA, 6)
for (j in 1:length(d)) {
     IQR1 <- IQR2 <- Hg1 <- sd1 <- rep(NA, n)
     for ( i in 1:n) {
          data <- sort(rnorm(d[j]))                  # Normally distributed datasets
           IQR1[i] <-IQR(data)                       # IQR
          Hg1[i]  <- fivenum(data)[4] - fivenum(data)[2]    # H-spread
          sd1[i]  <- sd(data)                        # SD
     }
     m.IQR1[j] <- mean(IQR1)      # mean of 50000 datasets
     m.Hg1[j]  <- mean(Hg1)
     m.sd1[j]  <- mean(sd1)
}
m.IQR1;    m.Hg1;      m.sd1
ymin <- min(c(m.IQR1, m.Hg1, m.sd1))
ymax <- max(c(m.IQR1, m.Hg1, m.sd1))
dev.new(width=10, height=6)
plot(m.IQR1, type="o", pch=2, col= " springgreen4", ylim=c(ymin, ymax),
      xlab="Data Size", ylab="Width", xaxt="n")
    points(m.Hg1, type="o", pch=3, col="royalblue")
    points(m.sd1, type="o", pch=20, col="tomato")
    axis(1, at=1:length(d), labels=d)
```

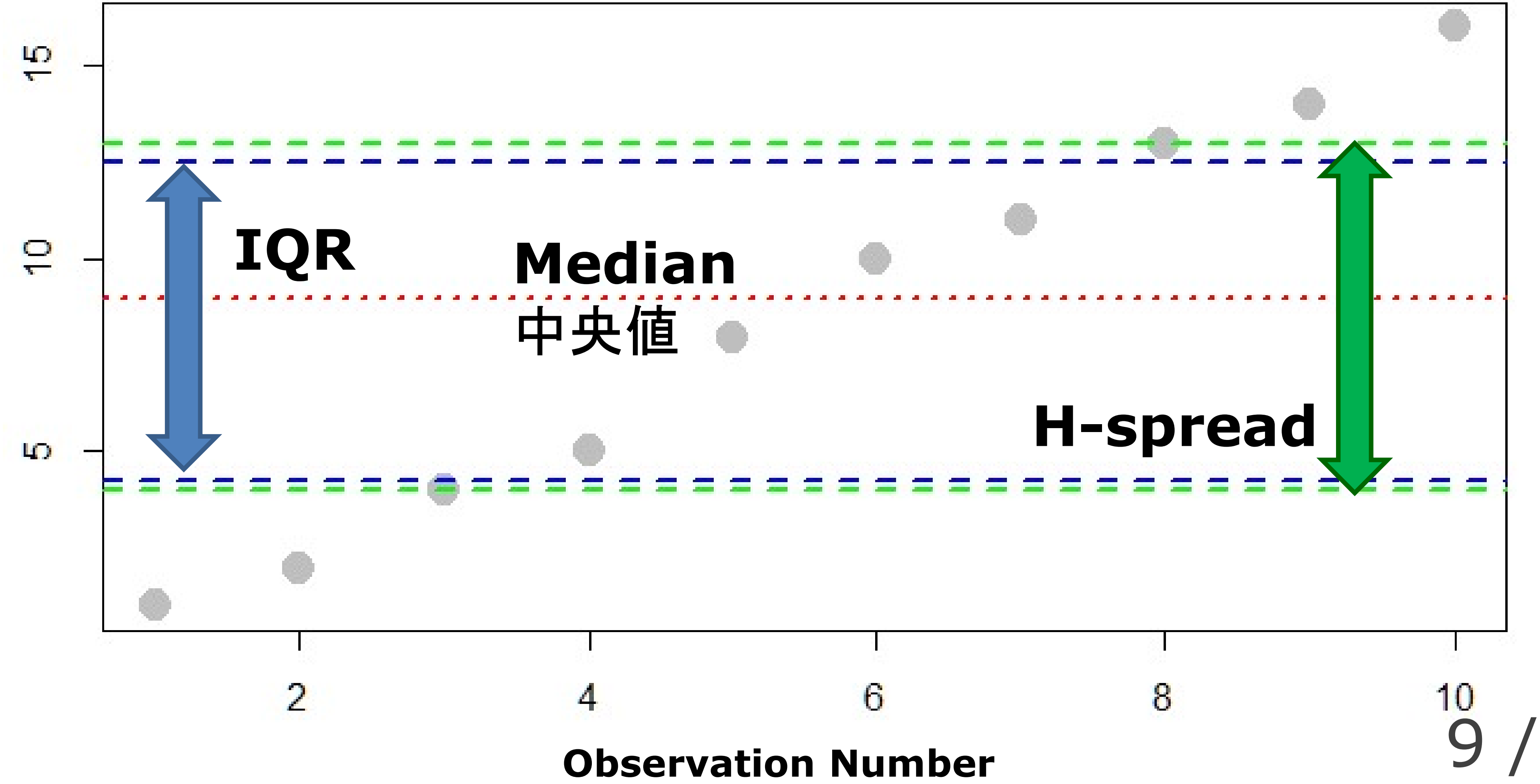


ヒンジと四分位範囲の違い

Difference between H-spread and IQR in R

サイズ10のデータ　size 10

1,2,4,5,8,10,11,13,14,16



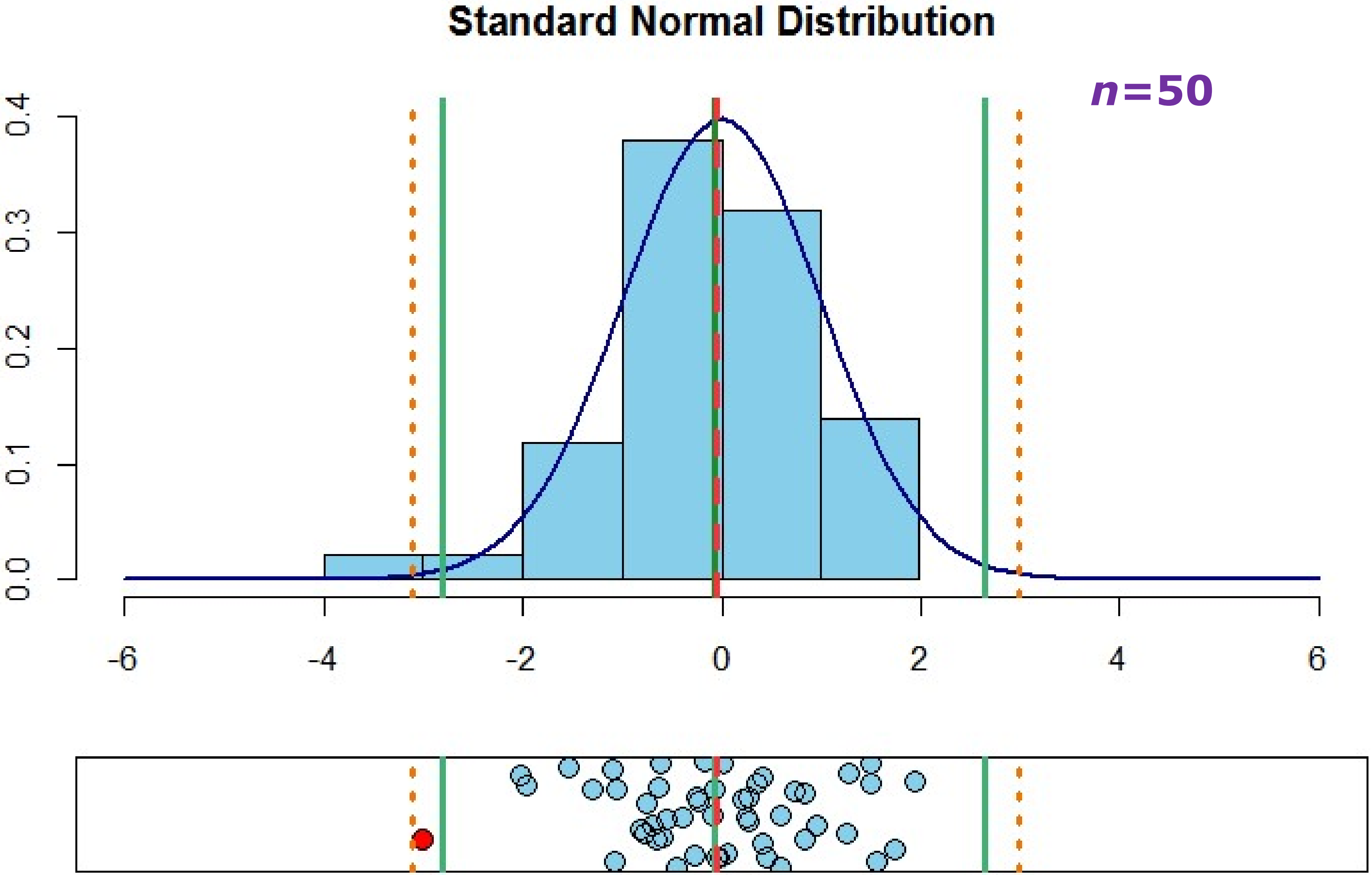
前頁プロットの作成コード

Source code for the plot in the previous page

```
d1 <- c(1,2,4,5,8,10,11,13,14,16)
plot(d1, pch=19, col="gray", cex=2, xlab=“observation
    number", ylab="value")
abline(h=c(quantile(d1)[2], quantile(d1)[4]), col="blue",
    lwd=2, lty=2)
abline(h=c(fivenum(d1)[2], fivenum(d1)[4]), col="green",
    lwd=2, lty=2)
abline(h=quantile(d1)[3], col="red", lty=3, lwd=2)
fivenum(d1)
#[1]  1  4  9 13 16    五数要約/five number summary
quantile(d1)
#    0%   25%   50%   75%  100%
# 1.00  4.25  9.00 12.50 16.00
```

IQRを使った改良プロット

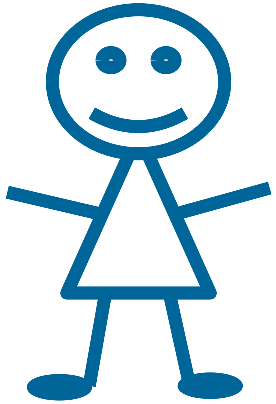
Improvement using IQR



改良プロットの作成コード(1)

Source code for the improved plot (1)

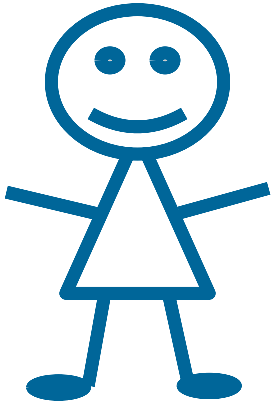
```
set.seed(8)
n1 <- 50                # n=50  : Data size
x1 <- rnorm(n1)    # data following standard normal distribution
dev.new(width=7, height=5)
op <- par()
nf <- layout( matrix( c(1,0,2,0), 2, 2, byrow=T ), c(1,0), c(3,1),)
layout.show(nf)
rg1 <- 2.224        # adjust the IQR range with standard deviation
mn1  <- mean(x1);     sd1  <- sd(x1)
med1 <- median(x1);          IQR1 <- IQR(x1)
fg1 <- rep(1, n1)
jt1 <- jitter(rep(0, n1))         # perturbation
par(mar=c(4,4,4,2))
  require(MASS)                        # for truehist()
  truehist(x1, h=1, ylim=c(0, 0.4), xlim=c(-6,6), col="skyblue",
     xlab="", main="Standard Normal Distribution")
      curve(dnorm, col="darkblue", lwd=2, add=TRUE)
      abline(v=mn1+3*sd1, col="darkorange2", lwd=3, lty=3)
      abline(v=mn1-3*sd1, col="darkorange2", lwd=3, lty=3)
      abline(v=med1, col="forestgreen", lwd=3)
      abline(v=med1 + IQR1 * rg1, col="mediumseagreen", lwd=3)
      abline(v=med1 - IQR1 * rg1, col="mediumseagreen", lwd=3)
      abline(v=mn1, col="firebrick1", lwd=3, lty=2)
```



改良プロットの作成コード(2)

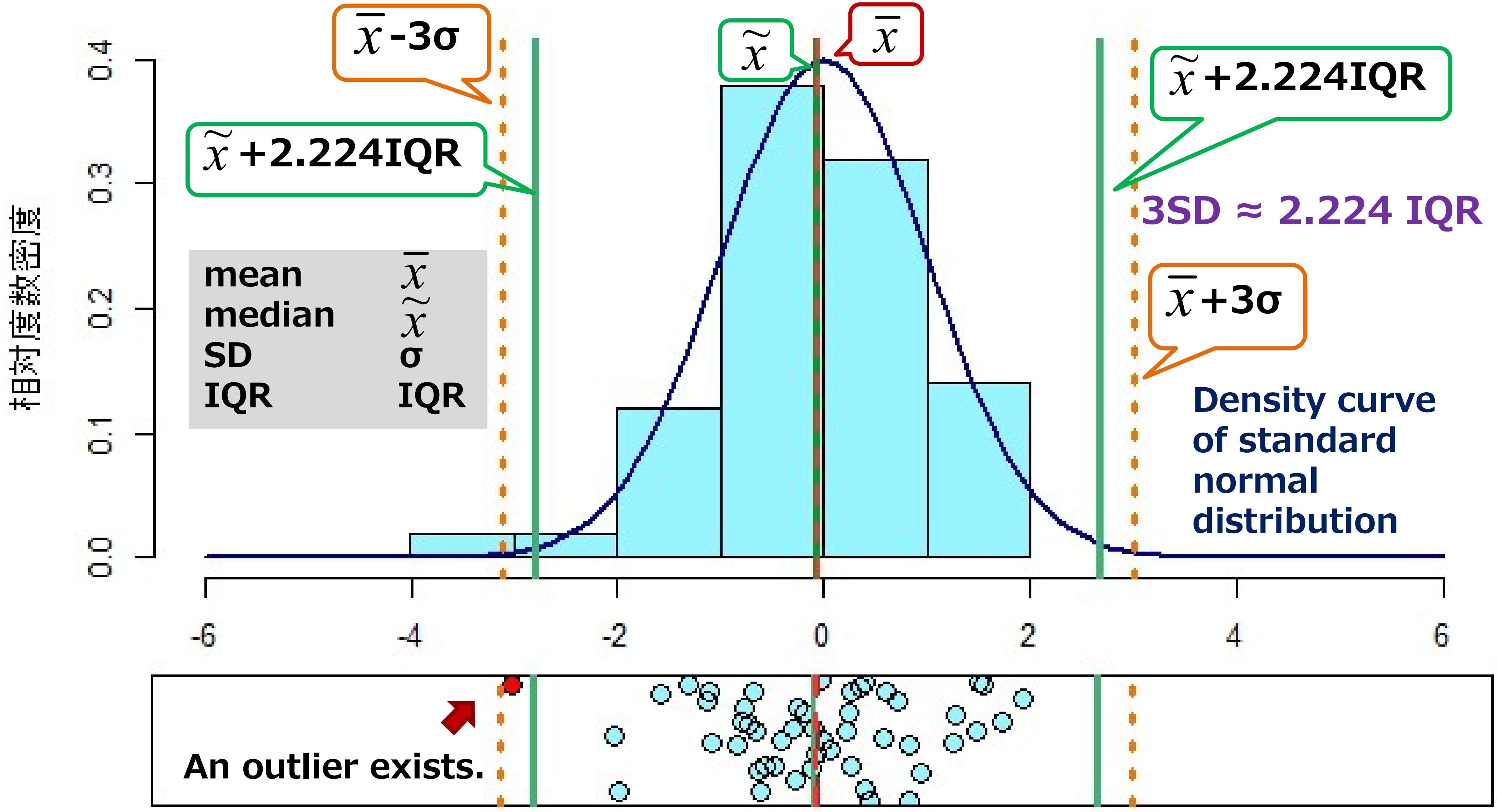
Source code for the improved plot (2)

```
par(mar=c(4,4,0,2))
# find outliers and make them red (color 2)
    fg1[which(x1 < med1 - rg1*IQR1)] <- 2
    fg1[which(x1 > med1 + rg1*IQR1)] <- 2
plot(x1, jt1, xlim=c(-6,6), cex=1.5, pch=19, col=c("skyblue",
      "red")[fg1], axes=F, ylab="", xlab="")
box()       # draw an outer box
points(x1, jt1, cex=1.5, pch=21)
  abline(v=mn1+3*sd1, col="darkorange2", lwd=3, lty=3)
  abline(v=mn1-3*sd1, col="darkorange2", lwd=3, lty=3)
  abline(v=med1, col="mediumseagreen", lwd=3)
  abline(v=med1 + IQR1 * rg1, col="mediumseagreen", lwd=3)
  abline(v=med1 - IQR1 * rg1, col="mediumseagreen", lwd=3)
  abline(v=mn1, col="firebrick1", lwd=3, lty=2)
par(op)
```



50個の正規分布データ

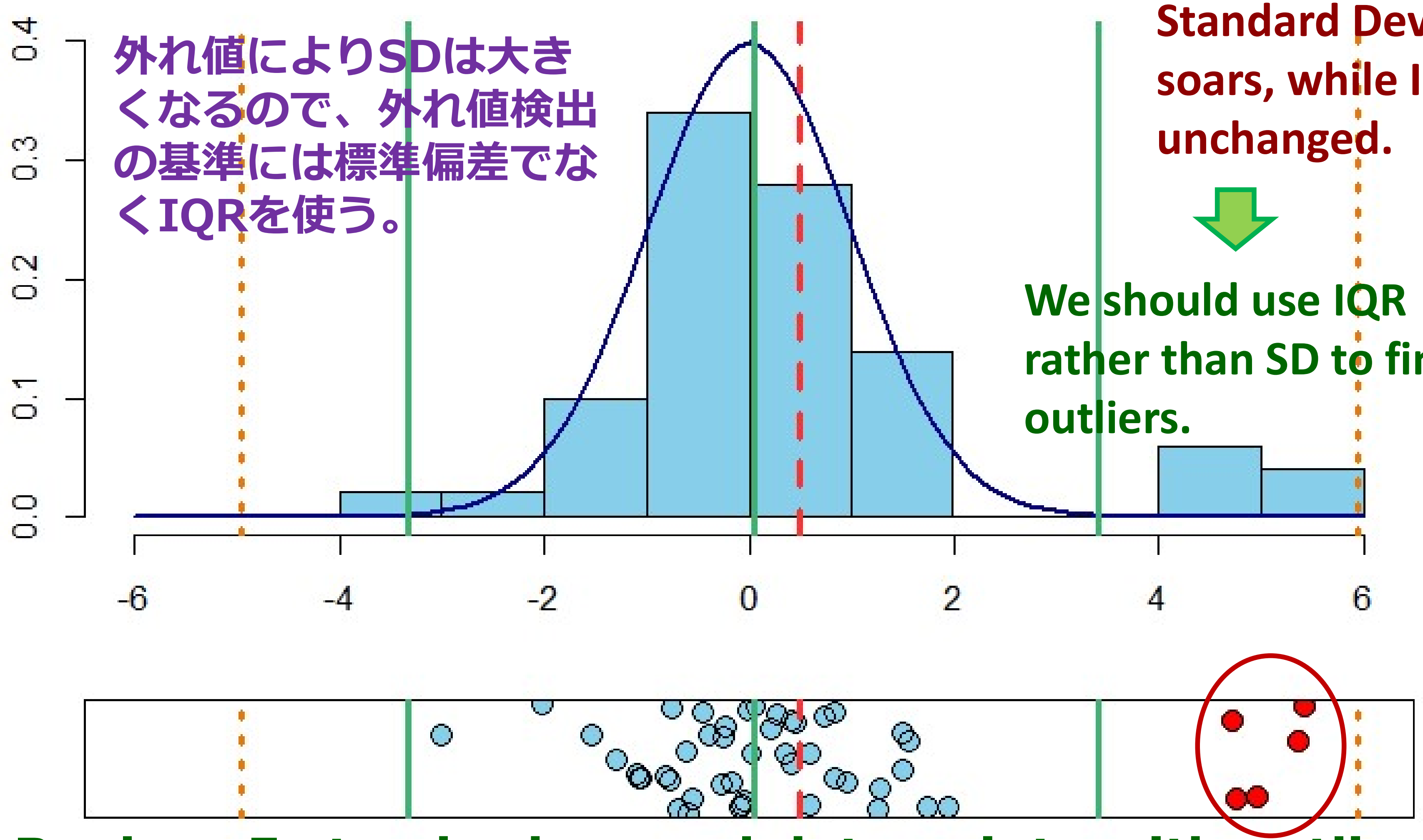
Size 50 of standard normal distribution data



mean	$\bar{x}$
median	$\tilde{x}$
SD	σ
IQR	IQR

人為的に外れ値を入れた場合

Contaminated Case [10% outliers]



外れ値によりSDは大きくなるので、外れ値検出の基準には標準偏差でなくIQRを使う。

Standard Deviation soars, while IQR almost unchanged.

We should use IQR rather than SD to find outliers.

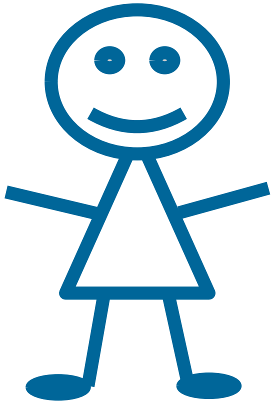
Replace 5 standard normal data points with outliers

標準正規乱数データのうち５個を外れ値に置換

改良プロットの作成コード(1)

Source code with contaminated data (1)

```
set.seed(8)
n1 <- n2 <- 50                    # n=50  : Data size
x1 <- rnorm(n1)    # data following standard normal distribution
x2 <- sample(c(x1[1:(n1-5)], rnorm(5, mean=5, sd=0.5)), n2)
  #  add 5 outliers at the end and shuffle the order
dev.new(width=7, height=5)
op <- par()
nf <- layout( matrix( c(1,0,2,0), 2, 2, byrow=T ), c(1,0), c(3,1),)
layout.show(nf)
rg2 <- 2.224       # adjust the IQR range with standard deviation
mn2  <- mean(x2);             sd2  <- sd(x2)
med2 <- median(x2);           IQR2 <- IQR(x2)
fg2 <- rep(1, n2)
jt2 <- jitter(rep(0, n2))       # for perturbation
par(mar=c(4,4,4,2))
 require(MASS)                  # for truehist()
 truehist(x2, h=1, ylim=c(0, 0.4), xlim=c(-6,7), col="skyblue",
    xlab="", main="Contaminated data (10%)")
     curve(dnorm, col="darkblue", lwd=2, add=TRUE)
     abline(v=mn2+3*sd2, col="darkorange2", lwd=3, lty=3)
     abline(v=mn2-3*sd2, col="darkorange2", lwd=3, lty=3)
     abline(v=med2, col="forestgreen", lwd=3)
```



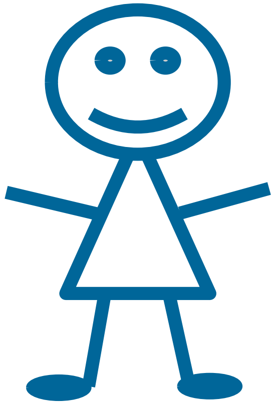
改良プロットの作成コード(2)

Source code with contaminated data (2)

```
    abline(v=med2 + IQR2 * rg2, col="mediumseagreen", lwd=3)
    abline(v=med2 - IQR2 * rg2, col="mediumseagreen", lwd=3)
    abline(v=mn2, col="firebrick1", lwd=3, lty=2)

par(mar=c(4,4,0,2))
# find outliers and make them red (color 2)
    fg2[which(x2 < med2 - rg2*IQR2)] <- 2
    fg2[which(x2 > med2 + rg2*IQR2)] <- 2

plot(x2, jt2, xlim=c(-6,7), cex=1.5, pch=19, col=c("skyblue",
      "red")[fg2], axes=F, ylab="", xlab="")
box()      # draw an outer box
points(x2, jt2, cex=1.5, pch=21)
  abline(v=mn2+3*sd2, col="darkorange2", lwd=3, lty=3)
  abline(v=mn2-3*sd2, col="darkorange2", lwd=3, lty=3)
  abline(v=med2, col="mediumseagreen", lwd=3)
  abline(v=med2 + IQR2 * rg2, col="mediumseagreen", lwd=3)
  abline(v=med2 - IQR2 * rg2, col="mediumseagreen", lwd=3)
  abline(v=mn2, col="firebrick1", lwd=3, lty=2)
par(op)
```



Conclusion / まとめ

- 正規分布データならば、データの位置と散らばりの指標として、平均値と標準偏差(SD)を使うのが一般的で、それは中央値や四分位範囲よりも推定効率が高く計算が楽だから

  As for normally distributed data, mean and standard deviation (SD) are better measures of location and variability compared to median and interquartile range (IQR) since they are efficient and easy to compute, however…

- ただし、中央値やIQRは外れ値に対してロバスト（頑健）だが、平均値や標準偏差はそうではない

  Median and IQR are robust regarding contamination, while mean and SD are not.

- つまり、外れ値検出の目安として使用する散らばりの指標には、平均やSDを使ってはいけない。

  Therefore, we should not use mean and SD for the purpose of outlier detection.

より詳しくは・・・

「統計実務におけるレンジチェックのための外れ値検出方法」

統計研究彙報 第72号 No.3, 2015年3月.

http://www.stat.go.jp/training/2kenkyu/2-2-723.htm

おわり

@ 後楽園，岡山