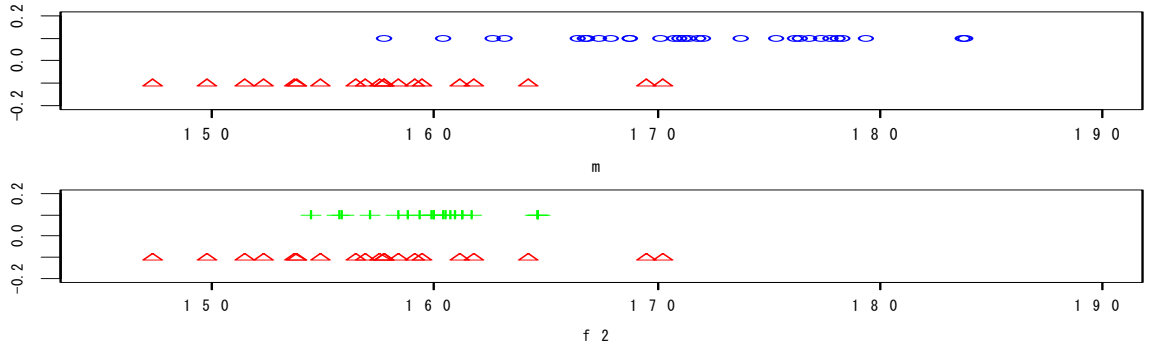
1変数の集計

統計学入門

2008.04

データの分布

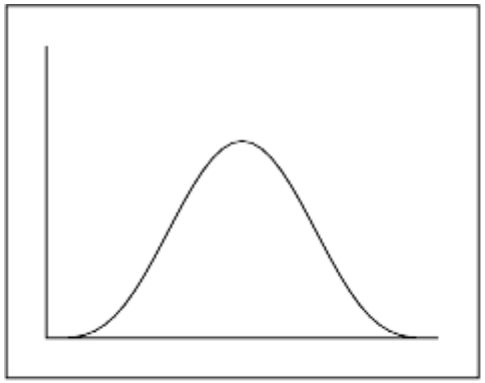
- だいたいどんな値 （代表値）
- 値のバラエティさ　（ばらつき）
- その中でどんな値が多いか （ヒストグラム）



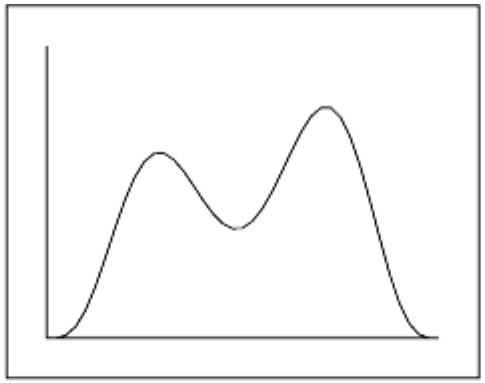
- 分布形状
  - 単峰性、双峰性、多峰性
  - 左に偏ったＪ型分布、左右対称分布
    右に偏ったＬ型分布

分布の形状

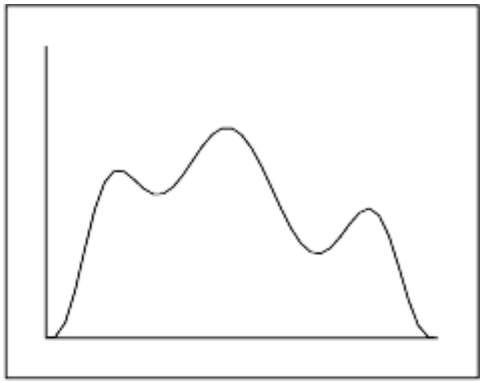
単峰性



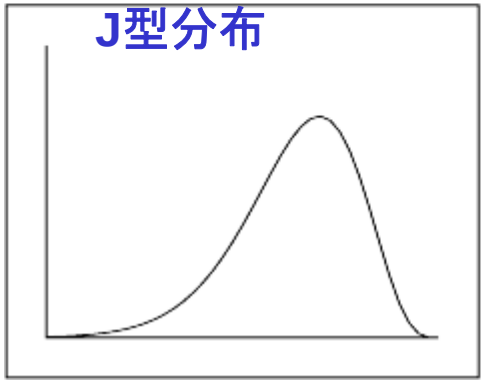
双峰性



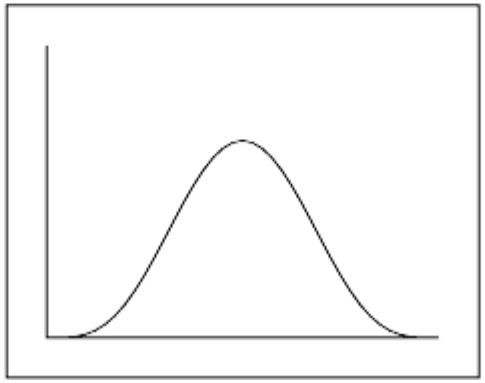
多峰性



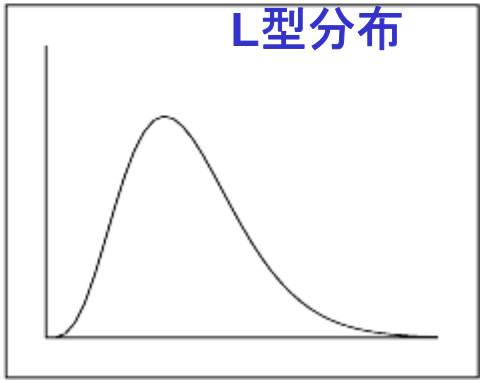
左に裾を引いている
J型分布



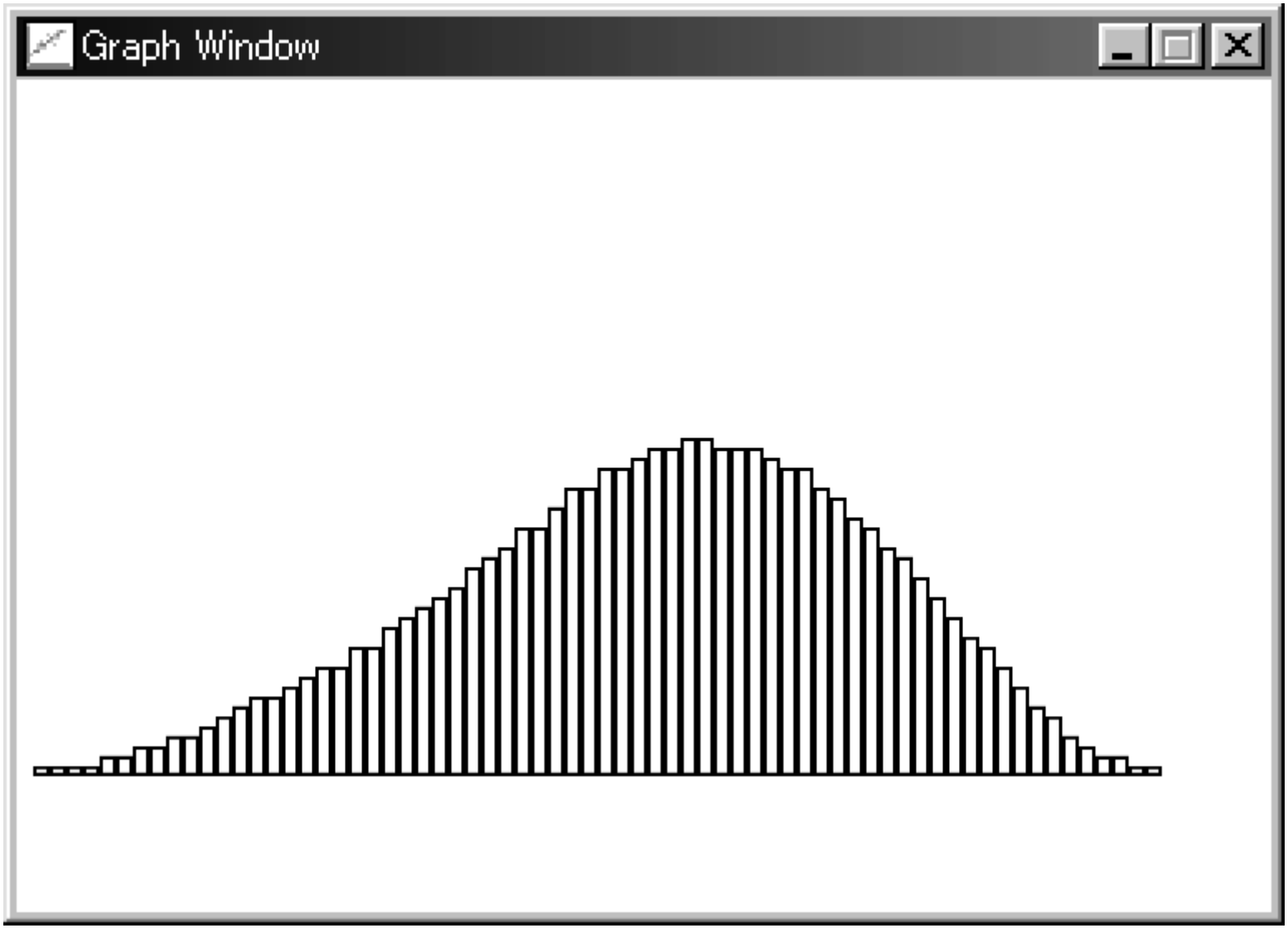
対称性



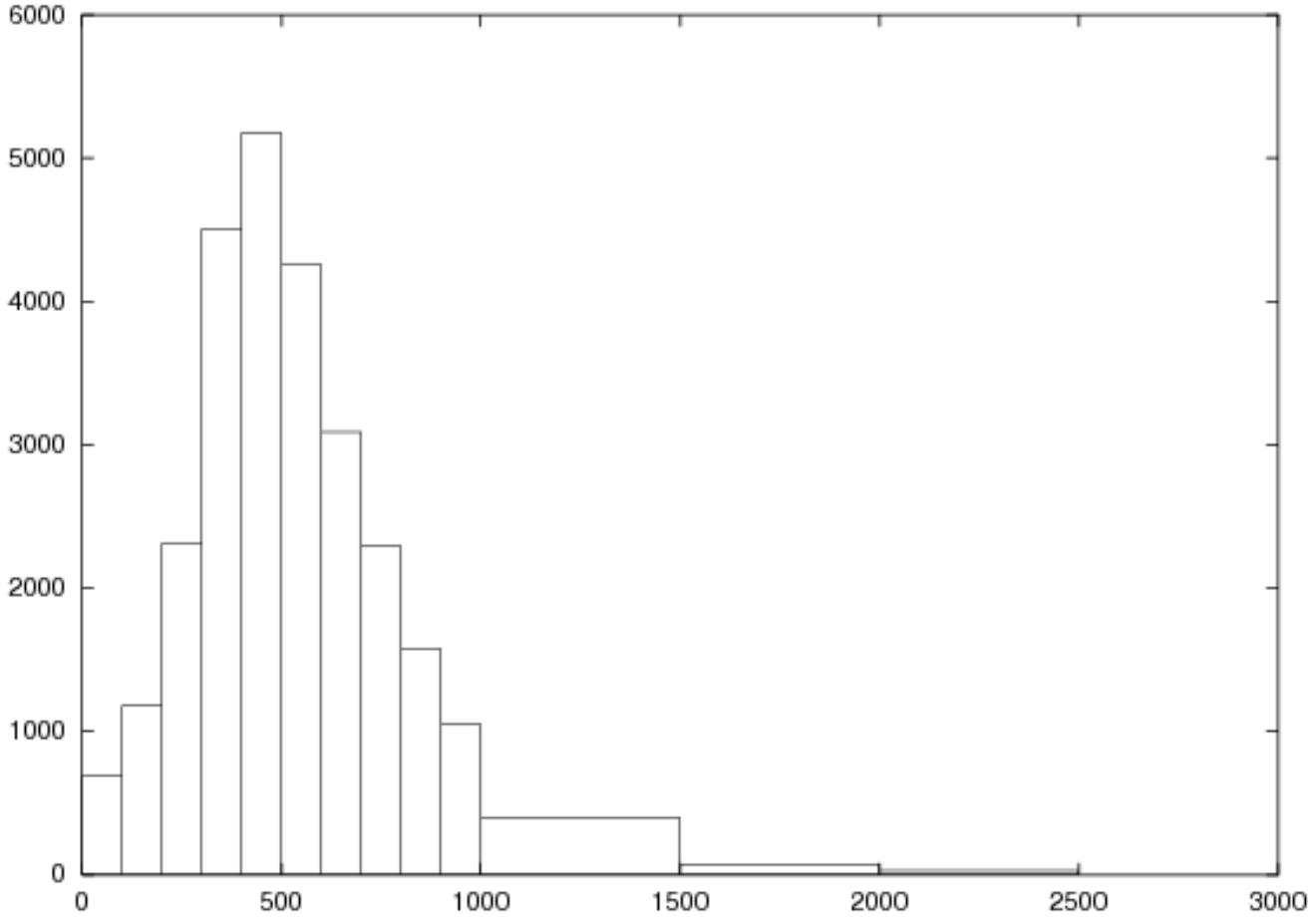
右に裾を引いている
L型分布



共通一次試験(1980年)



民間給与実態統計調査（平成9年）



データの縮約

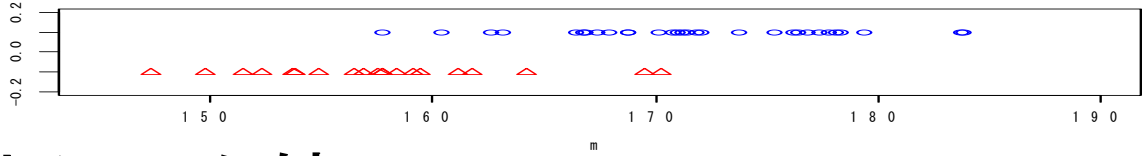
- n個のデータ
  $x_1, x_2, \cdots x_n$
  を数個の値にまとめる！

- 代表値（位置、location）
- ばらつき（広がり、dispersion）

- ５数要約
- グラフ表現

代表値

- データが概ねどんな値か
- （数直線上の）どのあたりの位置にあるか



- （算術）平均値　arithmetic mean
- 中央値　median
- 最小値　minimum
- 最大値　maximum

「良い」代表値とは

- ある定めた基準の元で「良い」
- 代表値とはn個の様々な値を、1個の値に置き換えること
- 置き換えたときの誤差
- 誤差は無いほどよい
- できるだけ小さく
- 最小化しよう

データ	代表値	誤差	絶対誤差	2乗誤差
x_1	a	x_1-a	$\lvert x_1-a\rvert$	$(x_1-a)^2$
x_2	a	x_2-a	$\lvert x_2-a\rvert$	$(x_2-a)^2$
⋮	⋮	⋮	⋮	⋮
x_n	a	x_n-a	$\lvert x_n-a\rvert$	$(x_n-a)^2$
和			$\Sigma \lvert x_n-a\rvert$	$\Sigma (x_n-a)^2$

平均値(mean)

- mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

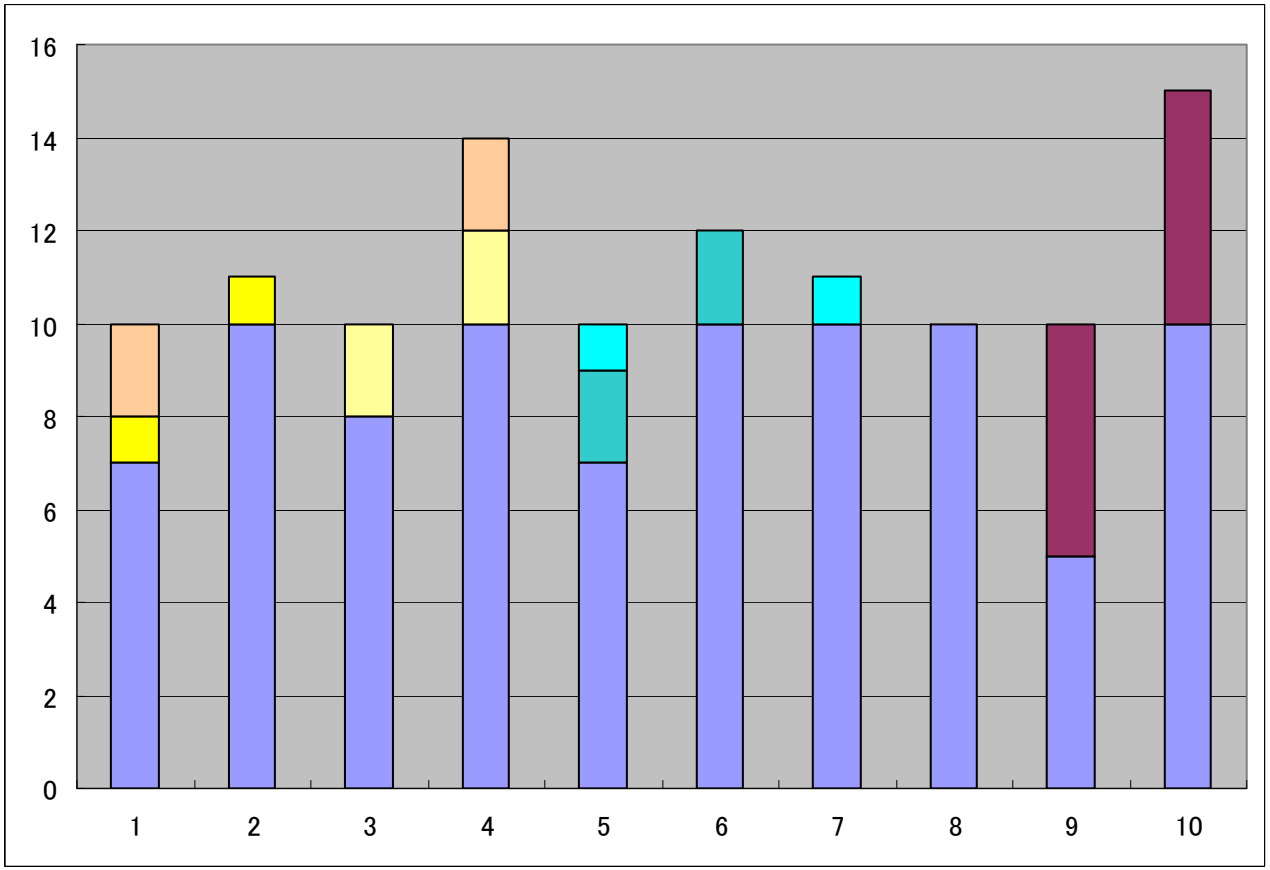
- 代表値として一番良く使われる
- 最小2乗法(Least Square Methods)の意味で最良

誤差の2乗

$Q(a)=(x_1-a)^2+(x_2-a)^2+\cdots+(x_n-a)^2$

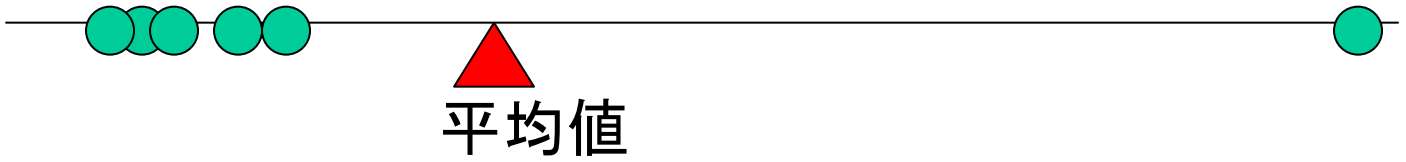
を最小にする a

平均値



中央値(median)

- 平均値の問題点
  はずれ値(outlier)の影響を受けやすい



- 多数のデータからは離れた値になっている

- 集団の代表値としては中央値の方が妥当
  $Me = x_{((n+1)/2)}$ ないしは $(x_{(n/2)} + x_{(n/2+1)})/2$

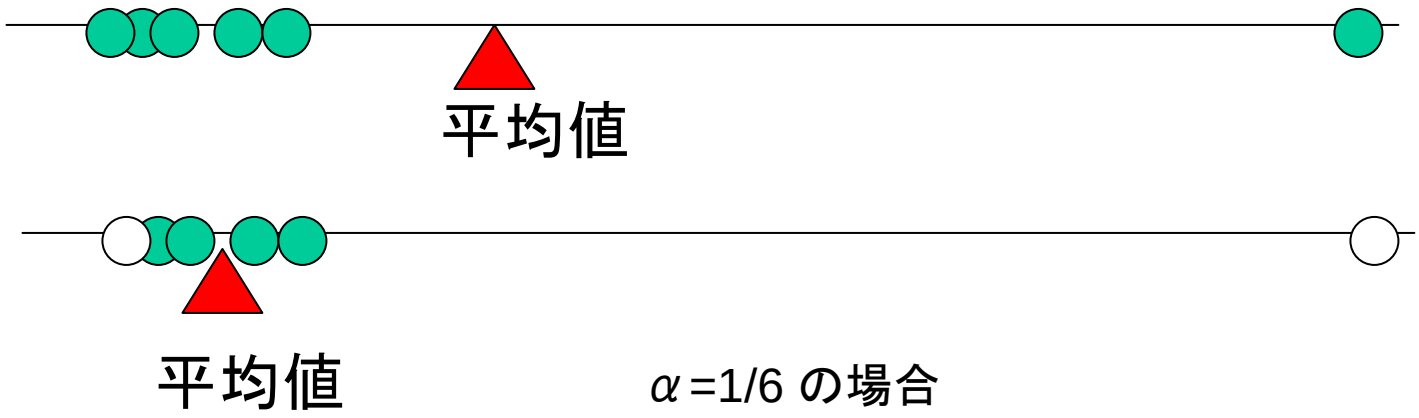
平均値と中央値との関係

- 平均値は外れ値の影響を受けやすい
- 中央値は外れ値に頑健(robust)

http://bstat.f7.ems.okayama-u.ac.jp/~yan/dataplot/

平均値の改良

- truncated mean, trimmed mean（切捨て平均）
  - 大きい方 α、小さい方 α のデータを捨てて残りの1-2 α のデータで平均値を計算する
  - α としては0.25が良く使われる



α =1/6 の場合

最小値、最大値

- 洪水対策
  　過去のデータでの最大降水量に対して対策を立てる

- 許容量
  　被害が起きた最小の値に対して対策を立てる

5数要約（fivnum）

five-number summary

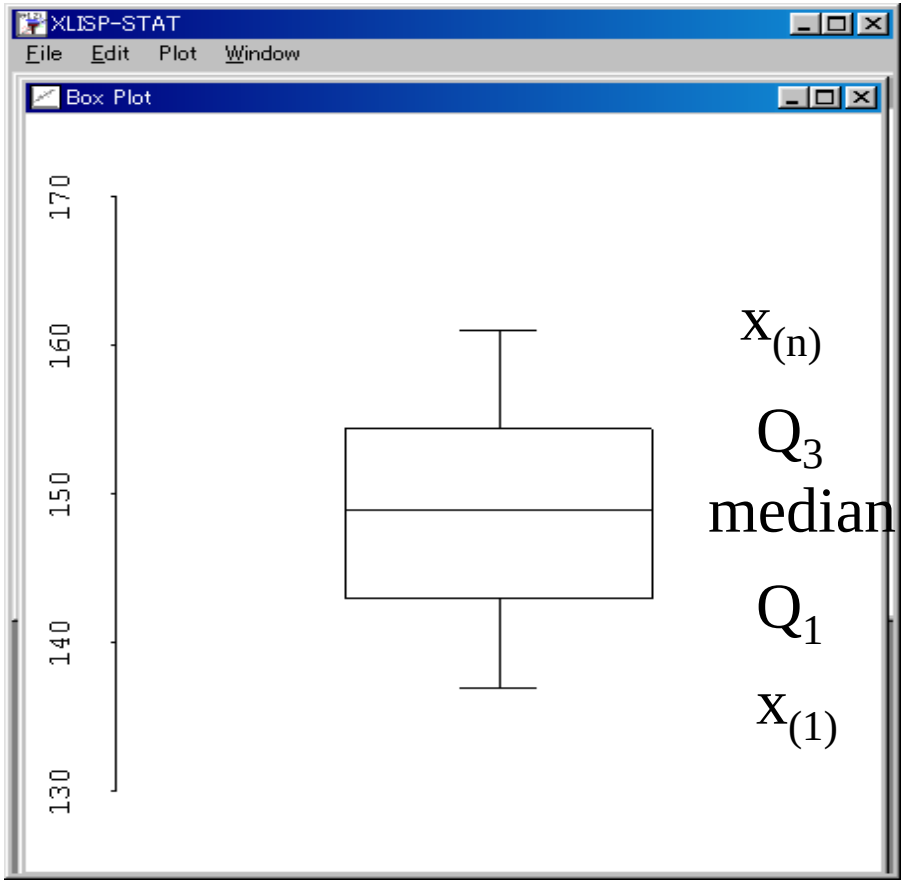
- 大きさの順序に並び換え

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

- $x_{(1)}$　最小値
- $x_{(n/4)}$　第１四分位値(Q_1)
- $x_{(2n/4)}$　第２四分位値(Q_2)＝中央値(Me)
- $x_{(3n/4)}$　第３四分位値(Q_3)
- $x_{(n)}$　最大値

箱髭図(boxplot)

- 5数要約のグラフ表現

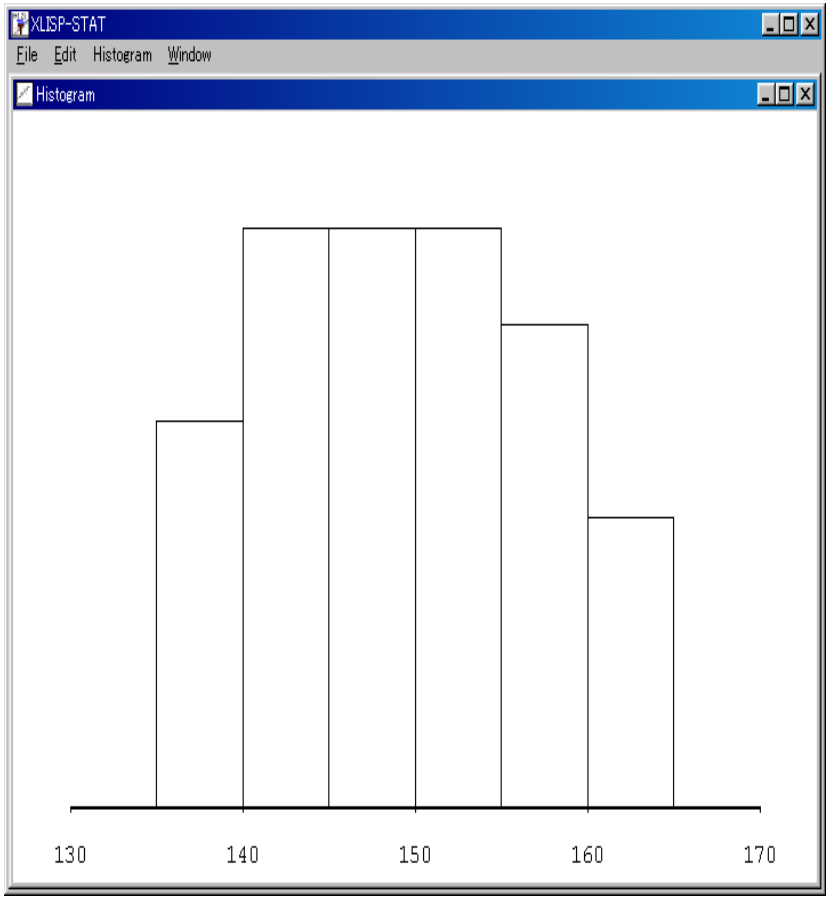


(boxplot height)

度数分布表とヒストグラム
frequency table and histogram

- 階級数 k
  - $\sqrt{n}$
  - $1 + \log n/\log 2$
  - 10〜20
- 階級の幅 w
  - $w=(x_{(n)}-x_{(1)})/k$
- 端点 a_0

  $a_0 < x_{(1)} < a_0 + w/2$



度数分布表の追加情報

階級	階級値	度数	相対度数	累積度数	累積相対度数
$a_0 \sim a_1$	m_1	f_1	f_1/n	f_1	$(f_1)/n$
$a_1 \sim a_2$	m_2	f_2	f_2/n	f_1+f_2	$(f_1+f_2)/n$
$a_{k-1} \sim a_k$	m_k	f_k	f_k/n	$f_1+\cdots+f_k$	$(f_1+\cdots+f_k)/n$
		n	1		

演習

- 統計学168人の試験の成績は
  最小値　24点
  最大値　87点
  であった。度数分布表を作るための階級を定めよ。
- 169人の身長を調査したところ
  最小値　143.2cm
  最大値　175.7cm
  であった。度数分布表を作るための階級を定めよ。

ばらつき(dispersion)の尺度

- 代表値は同じでも、分布が異なる

> (mean height)

151.57142857142856

> (mean height2)

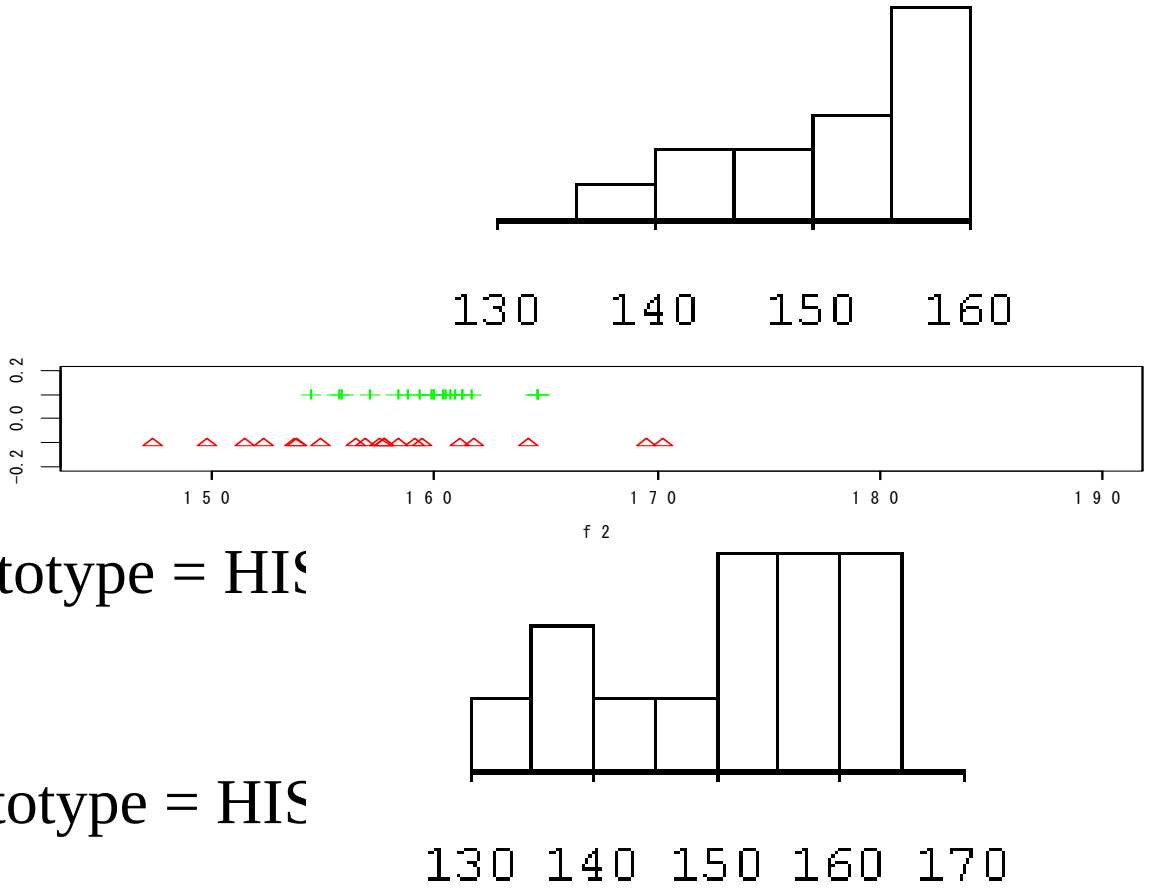
151.3571428571429

> (histogram height)

#<Object: 779b00, prototype = HIS

> (histogram height2)

#<Object: 7810ac, prototype = HIS



height2 ~tarumi/lispstat/height2.lsp

バラツキの尺度

- 範囲(range)
- 四分位範囲(interquartile range)
- 平均偏差(mean deviation)

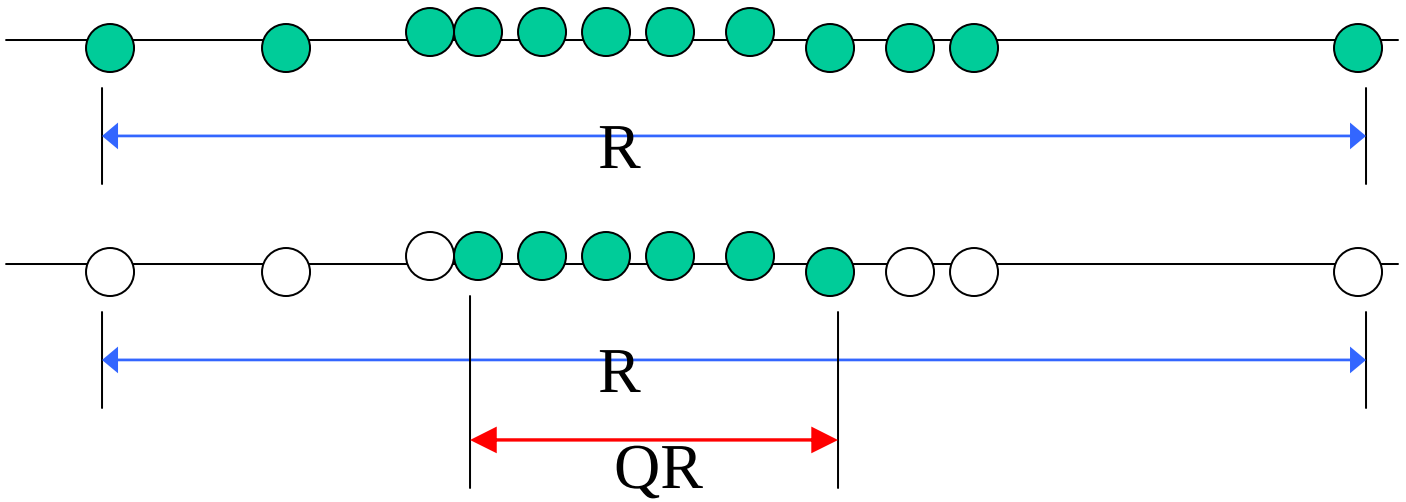
- 分散(variance)
- 標準偏差(standard deviation)

範囲(range)
四分位範囲(quartile range)

- 最小値から最大値までの幅

  $R = x_{(n)} - x_{(1)}$

- Outlier の影響を受けやすい
- 両端 25%のデータを捨てた真中の50%のデータでの範囲＝四分位範囲 $= Q_3 - Q_1$



平均偏差(mean deviation)

- 偏差

$$d_i = x_i - \bar{x}$$

- 平均偏差

$$d = \bar{d} = \frac{1}{n}\sum_{i=1}^{n} | d_i | = \frac{1}{n}\sum_{i=1}^{n} |x_i - \vec{x}|$$

分散(variance)
標準偏差(standard deviation)

- 分散　偏差2乗の平均

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{n}(\sum_{i=1}^{n} x_i{}^2) - (\bar{x})^2$$

- 標準偏差　　分散の平方根

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

不偏分散(unbiased variance)

- 標本分散としては不偏分散 u^2 を使うことも多い

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

演習

- height
  - 148, 160, 159, 153, 151, 140, 156, 137, 149, 160, 151, 157, 157, 144
  - 和 2122 2乗和 322338
- height2
  - 138, 162, 158, 151, 145, 134, 160, 137, 151, 163, 152, 163, 158, 147
  - 和 2119 2乗和 322019
- weight
  - 41, 49, 45, 43, 42, 29, 49, 31, 47, 47, 42, 39, 48, 36
  - 和 588　2乗和 25226