

記述統計

Code ▾

descriptive statistics

度数分布表

- 度数分布表とは?
 - data を 階級 に分けて階級ごとの 度数 を数えた表

histogram

- 度数分布表から縦軸に 度数 横軸に 階級 をとり グラフ化 したもの
 - 集団の特徴を定義付けていく

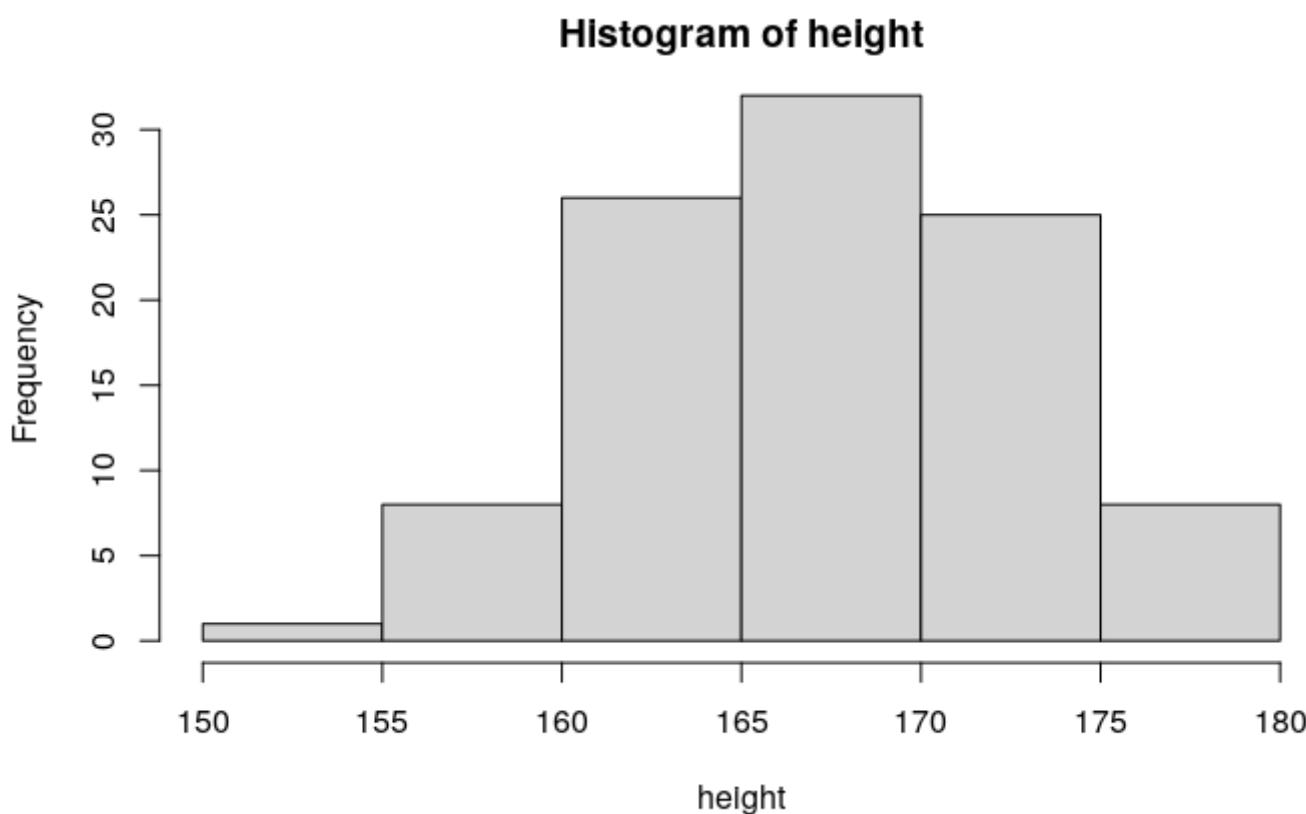
Hide

```
height =
```

Warning message:
In grSoftVersion() :
 unable to load shared object '/usr/local/lib/R/modules//R_X11.so':
 libXt.so.6: cannot open shared object file: No such file or directory

Hide

```
rnorm(100, 167, 5)  
y = hist(height)
```



Hide

```
y
```

```
$breaks  
[1] 150 155 160 165 170 175 180  
  
$counts  
[1] 1 8 26 32 25 8  
  
$density  
[1] 0.002 0.016 0.052 0.064 0.050 0.016  
  
$mids  
[1] 152.5 157.5 162.5 167.5 172.5 177.5  
  
$xname  
[1] "height"  
  
$equidist  
[1] TRUE  
  
attr("class")  
[1] "histogram"
```

基本統計量

分布の基本的な特性を数値で表した指標

- **代表値** (分布の中心を表す指標)
 - 平均 : (算術平均, 幾何平均, 調和平均, 加重平均)
 - 中央値(メジアン)
 - 最頻値(モード)
- **散布度** (分布のばらつきを表す指標)
 - 範囲(レンジ)
 - 四分位範囲
 - 平均偏差
 - 分散
 - 標準偏差
 - 変動係数
 - 標準化得点

算術平均

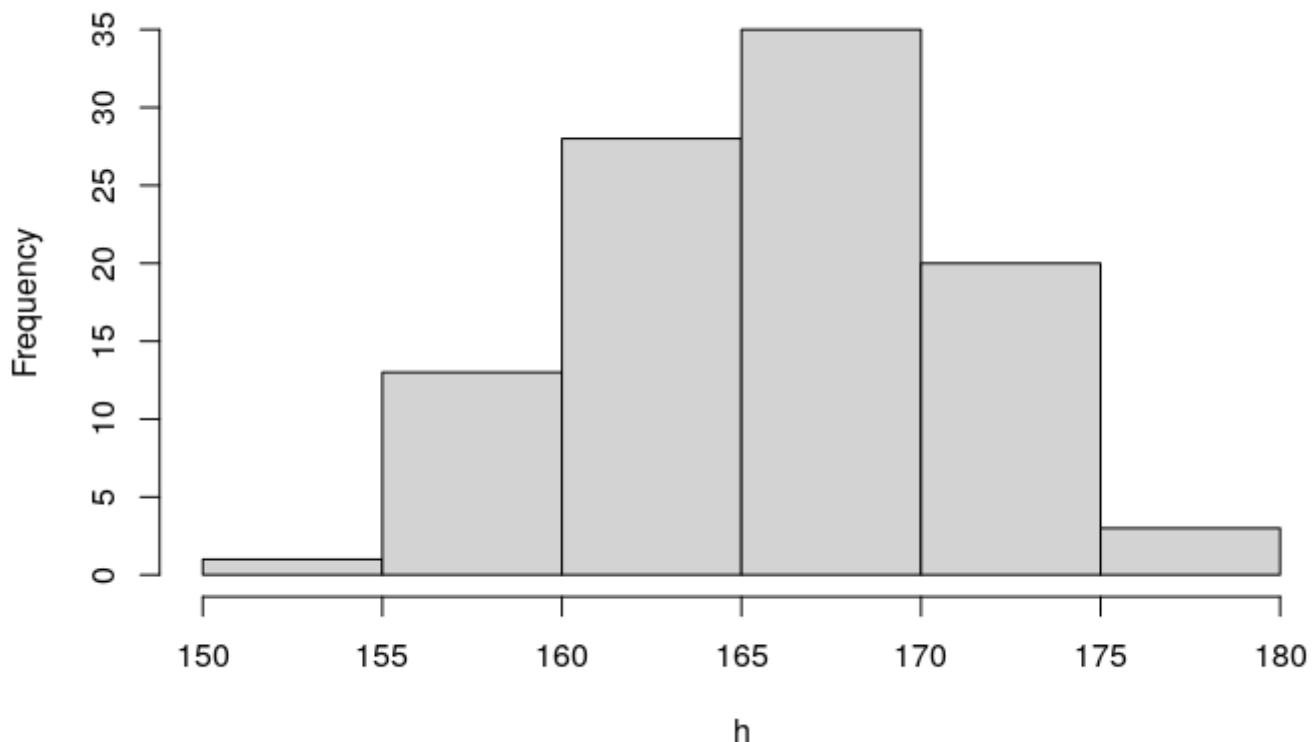
中心を表す指標

- 全ての data を足して個数で割った値 = **平均** = μ
 - **長所**
 - 一般的, 計算が簡単, 数値の意味の理解のし易さ
 - **短所**
 - 外れ値の影響を受けやすい, 分布の形状に依存する

Hide

```
h = as.integer(height)  
hist(h)
```

Histogram of h



[Hide](#)

```
mean(h)
```

```
[1] 166.62
```

[Hide](#)

```
median(h)
```

```
[1] 167
```

- 平均 : 166.62 cm | 中央値 : 167 cm
 - 解析する histograme の形を考慮して mean() , median() を使い分けていく
- summary() : 一度に様々な数値を取得できる
 - Min : 最小 | 1st : 25% | Median : 中央値 | Mean : 平均 | 3rd : 75% | Max : 最大値

[Hide](#)

```
summary(h)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.  
152.0 163.0 167.0 166.6 170.0 178.0
```

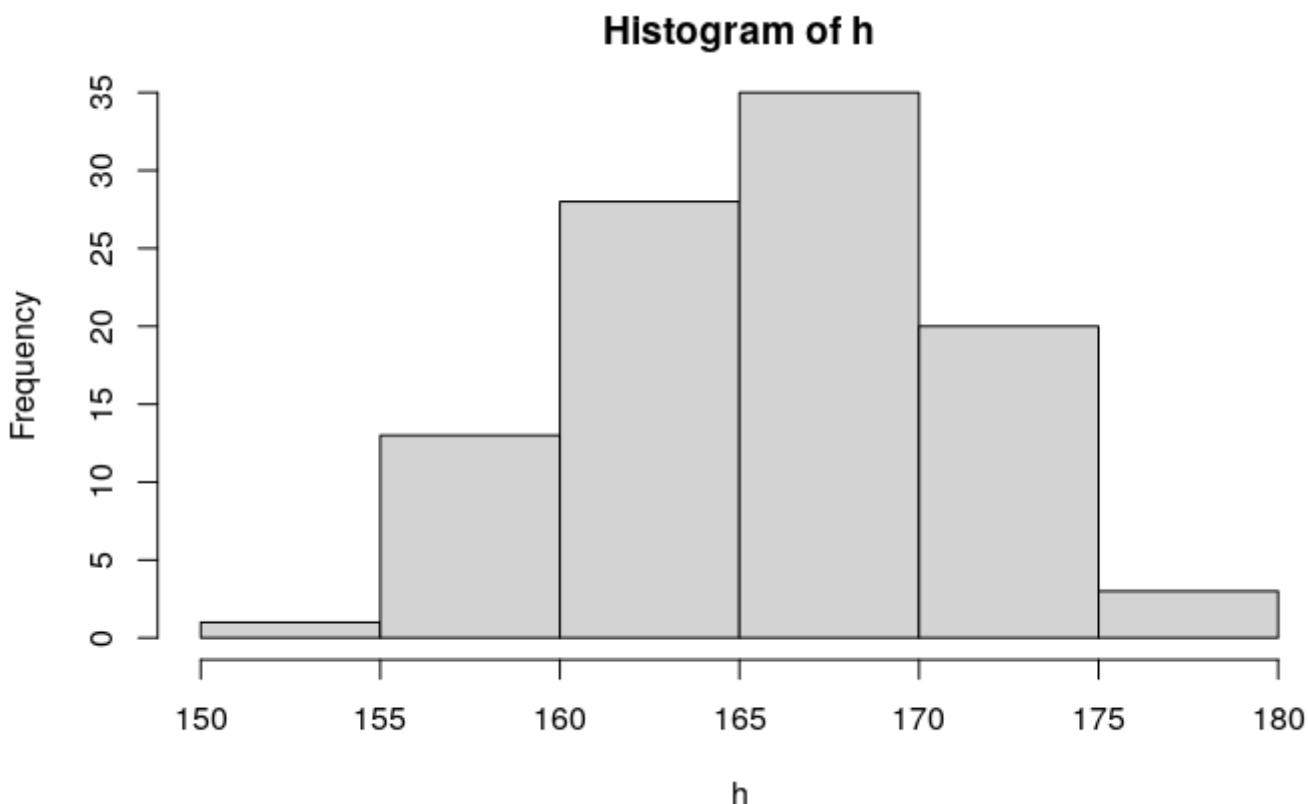
分散・標準偏差

ばらつきを表す指標

- 平均との差の二乗平均を求めた値 分散(σ^2) その平方根 標準偏差(σ)
 - 長所
 - バラつきの数値で最も一般的, 理論的に扱いやすい
 - 短所
 - 分散は単位が分かりづらい

[Hide](#)

```
hist(h)
```



Hide

```
var(h)
```

```
[1] 29.06626
```

Hide

```
sd(h)
```

```
[1] 5.391314
```

- 分散 : 29.0662626 | 標準偏差 : 5.3913136

- ※ R言語の分散は 不偏分散 なので注意!

標準化

data の 平均値を 0 , 分散を 1 に変換する操作

$$\frac{X - \mu}{\sigma}$$

(それぞれの値 - 平均) \div 標準偏差 = 標準化
↓

標準化により異なる集団も全て 標準得点 で比較できる

||

scale を合わせて比較することが出来る

Hide

```
x
```

```
[1] 160 165 165 164 158 166 161 173 165 167 178 167 170 170 171 169 170  
[18] 163 166 166 167 169 170 172 165 166 162 170 169 158 169 169 175 152  
[35] 162 176 160 170 165 173 173 167 158 165 162 171 171 161 167 169 175  
[52] 175 174 167 169 168 164 168 168 169 178 162 156 161 170 169 164 157  
[69] 163 170 164 160 169 163 173 164 165 164 157 173 163 159 170 160 162  
[86] 166 161 172 171 160 171 171 174 175 166 156 163 161 175 170
```

Hide

```
head(scale(h))
```

```
[,1]  
[1,] -1.2279011  
[2,] -0.3004834  
[3,] -0.3004834  
[4,] -0.4859669  
[5,] -1.5988682  
[6,] -0.1149998
```

Hide

```
var(scale(h))
```

```
[,1]  
[1,] 1
```

Hide

```
mean(scale(h))
```

```
[1] -8.365444e-16
```

- 標準化 : scale()
 - 標準化後の 分散 : 1
 - 標準化後の 平均 : -8.3654438^{-16}

正規分布

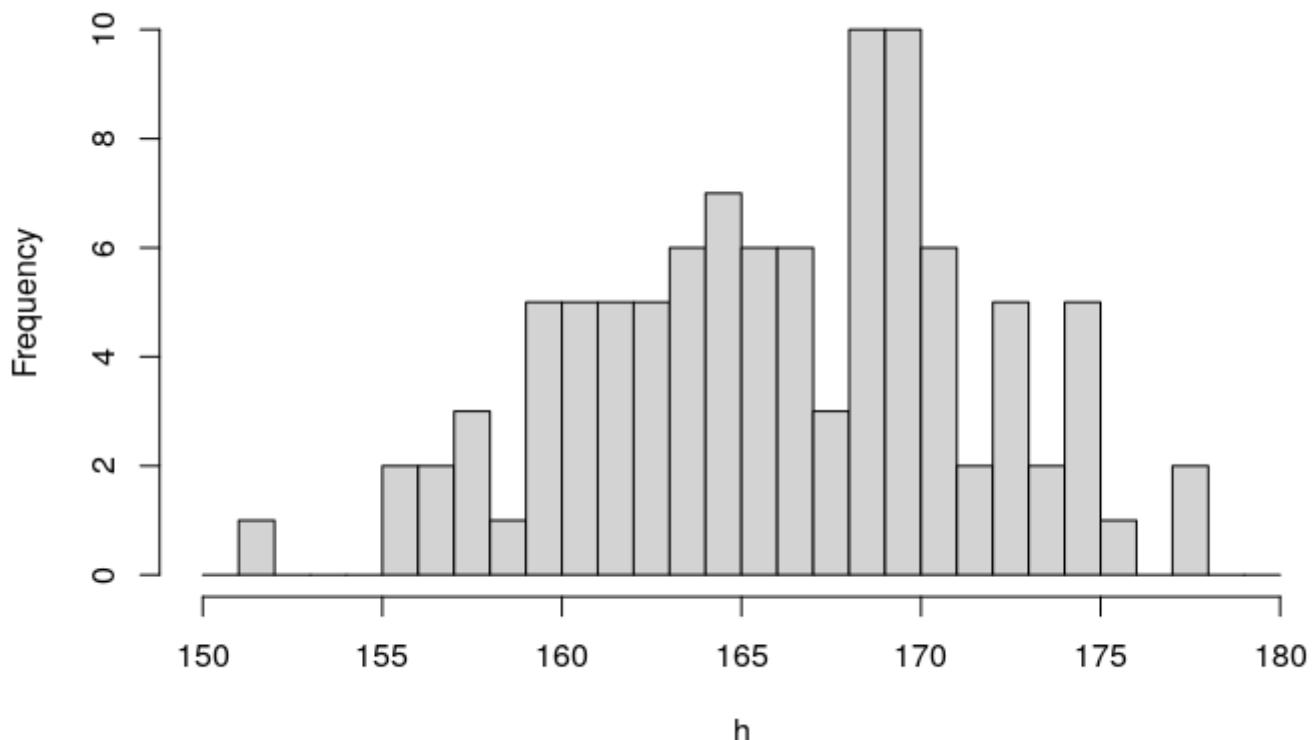
左右対称つり鐘型をした 確率分布 多くの現象がこの分布に従う
→ 身長, 植物の大きさ, testの点数, 工業製品の誤差 etc...

- 確率分布
 - グラフの 面積が確率 なるようなグラフ
 - 全部足すと 1 になる = 確率 100%
 - 平均値 と 標準偏差 で形が決まる
 - 数学的に一意に決まるグラフ

Hide

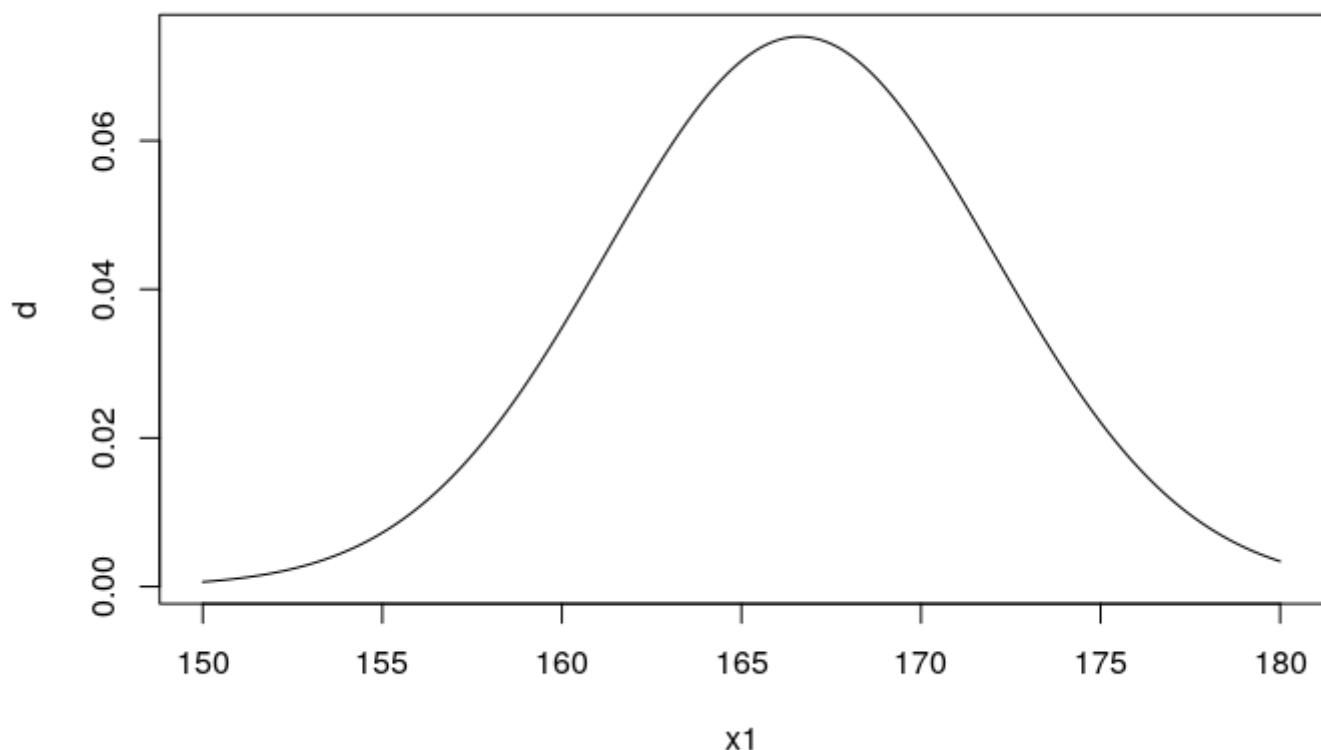
```
hist(h, breaks = seq(150, 180, 1))
```

Histogram of h



[Hide](#)

```
m <- mean(h)
s <- sd(h)
x1 <- seq(150, 180, 0.01)
d = dnorm(x1, mean = m, sd = s)
plot(x1, d, type = "l")
```



- 平均 : 166.62 cm | 標準偏差 : 5.3913136

正規分布から分かること

分布のある範囲に どれだけのdataが含まれているか が分かる

面積が分かるという事は

||

その範囲に入る dataの確率 が分かる

[Hide](#)

```
sc <- 850
m1 <- 582.6
s1 <- 172.7
n <- as.integer(103955)
Z = (X1 - m1)/s1
```

自分の得点 : $X = 850$ | 平均点 $\mu : 582.6$ | 標準偏差 $\sigma : 172.7$ | 標準化 : 1.5483497 | 総数(人) : 103955

標準化

$$Z = \frac{\text{得点} - \text{平均値}}{\text{標準偏差}}$$

公式

$$Z = \frac{X - \mu}{\sigma} = \frac{850 - 582.6}{172.7} \simeq 1.55$$

正規分布の面積を求める

- Rの場合 pnorm()では、無限大の - (マイナス)方向から 1.55 までを求める
 - なので - 0.5 で 0 から - 方向の面積を引く

[Hide](#)

```
d = pnorm(Z, mean = 0, sd = 1)
d
```

[1] 0.9392309

[Hide](#)

```
d2 = pnorm(Z, mean = 0, sd = 1) - 0.5
d2
```

[1] 0.4392309

[Hide](#)

```
N = 1 - d
N
```

[1] 0.06076906

[Hide](#)

```
N1 = as.integer(n*(1 - d)*100)
N1
```

[1] 631724

[Hide](#)

- 無限大の - (マイナス)方向から 1.55 までの面積 : 0.9392309
- 0.5 で 0 から - 方向の面積を引く : 0.4392309
- 上位の面積 : 0.0607691 | 上位からの順位 : 631724