Code

推測統計(点,区間)

inferential statistics

推測統計

少ない data から大きな集団の特徴を掴む

母集団と**標本**と**サンプリング**

↓

推定・検定

母集団と標本

- **母集団** → 情報を得たい対象全体
- **標本** → 母集団の一部

標本抽出のサンプリング

母集団	標本
母平均： μ	標本平均： $\overline{X}$
母分散： σ^2	標本分散： s^2
母標準偏差： σ	標本標準偏差： s

- 母集団の平均, 分散, 標準偏差はわからない
  - 全てのdataが手元にない為 → 求められない値
- 標本の平均, 分散, 標準偏差は分かる
  - 手元にある data

↓

推定・点推定・区間推定を予測
推定母平均： $\hat{\mu}$
推定母分散： $\hat{\sigma}^2$
推定母標準偏差： $\hat{\sigma}$

- 標本から仮説が正しいかを判断
  - **仮説検定**

Hide

```
n <- 10000
```

```
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

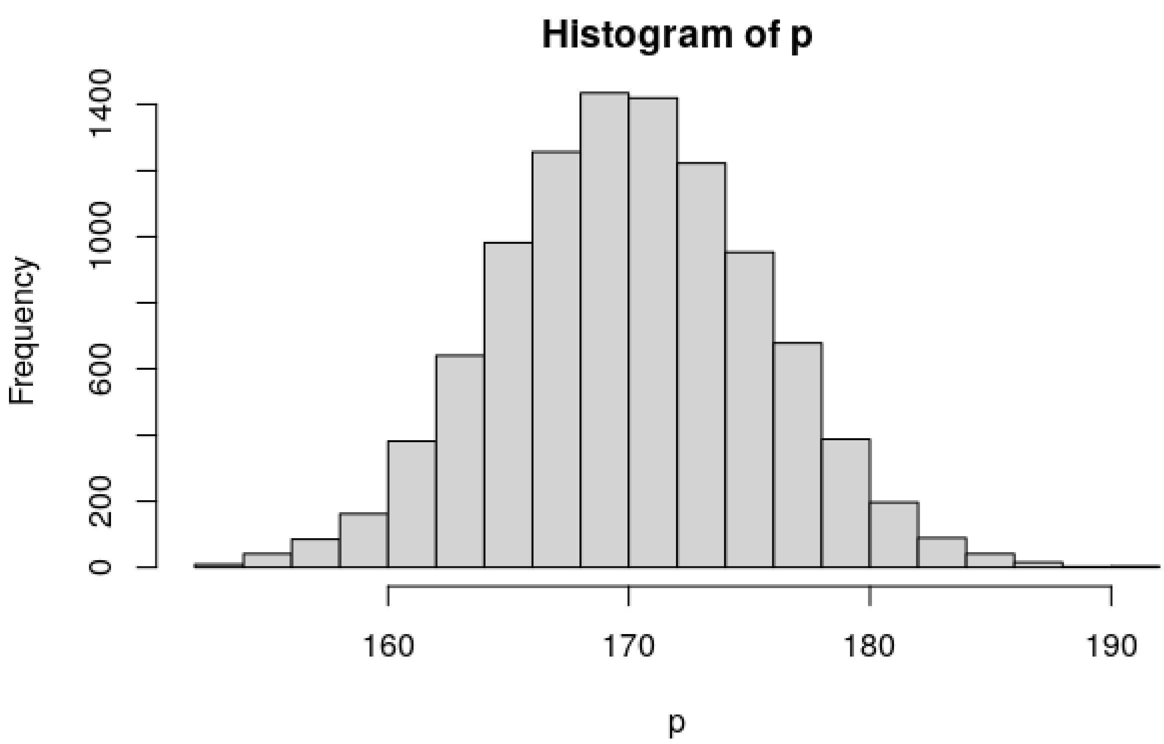
```
m <- 170
s <- 5.5

# 乱数発生させる : 1万人の男性の身長
p <- rnorm(n, mean = m, sd = s)
head(p, 50)
```

```
 [1] 174.8529 174.6029 167.2186 171.3273 167.9473 167.9336
 [7] 174.3294 169.6803 163.8044 163.6549 168.0520 170.2827
[13] 160.0403 171.6271 166.3651 165.0358 170.9178 177.5027
[19] 165.8259 176.3748 163.2458 175.5495 172.3087 172.8013
[25] 173.7150 168.7191 171.9865 169.2947 168.2785 179.4040
[31] 172.6773 169.3132 172.5737 169.9297 174.3042 167.9303
[37] 170.1624 170.2715 167.1509 168.2889 177.3980 167.9464
[43] 181.4072 164.0399 165.5930 163.4960 172.1965 167.3100
[49] 176.6203 166.7780
```

Hide

```
hist(p)
```



母集団から random に data を抽出

- 100個の data を取得

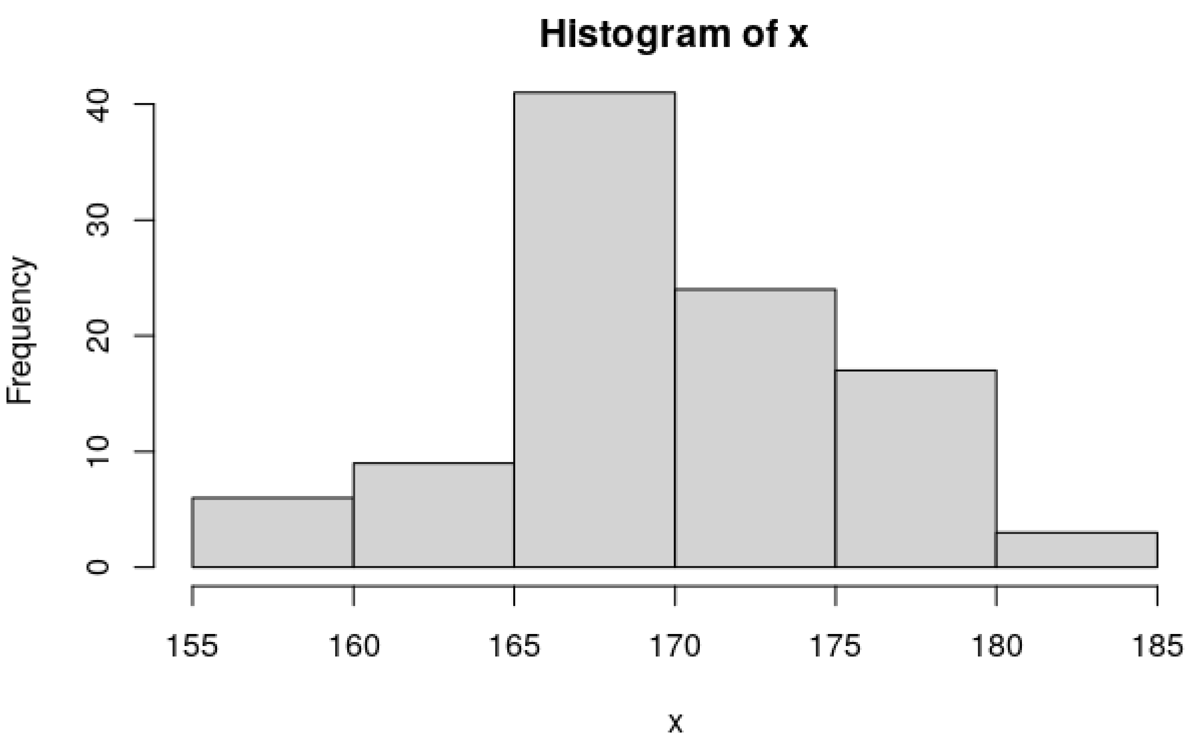
Hide

```
x <- sample(p, 100)
x
```

```
 [1] 166.8531 165.5551 177.4710 169.5845 167.3851 172.0227
 [7] 180.3147 168.8445 169.1520 164.6816 165.2264 169.3671
[13] 166.0296 166.3230 166.5556 164.2880 179.3665 178.4868
[19] 167.2627 167.3896 175.7166 166.4897 169.8904 167.9106
[25] 167.8710 177.8596 170.8507 165.7839 170.0462 169.5845
[31] 160.7666 176.1840 171.8486 165.6940 160.3004 168.3398
[37] 167.5284 168.2589 165.5710 164.1009 162.3450 168.0342
[43] 159.2725 172.8484 180.5820 164.5885 157.3365 176.1893
[49] 165.1594 173.4585 175.8635 169.0933 173.6280 158.8514
[55] 174.1865 174.6385 169.6264 174.4522 172.2184 177.5404
[61] 172.8549 165.2675 172.8979 176.0956 168.0372 174.0632
[67] 174.4497 171.3153 178.5327 169.2013 157.8129 168.1564
[73] 165.9869 175.6276 175.4683 162.8057 181.1153 159.9572
[79] 168.0516 169.1127 159.0311 174.1849 169.3587 172.4592
[85] 170.6753 165.5698 174.2368 167.6649 172.3399 175.0637
[91] 166.7780 179.9438 167.1342 173.9413 176.7806 172.0095
[97] 178.5122 171.5525 164.2081 169.3286
```

Hide

```
hist(x)
```



Hide

```
mean(x)
```

```
[1] 169.9025
```

Hide

```
sd(x)
```

```
[1] 5.514522
```

x の data の値

平均 : 169.9025136
標準偏差 : 5.5145216

点推定

母集団の平均, 分散, 標準偏差を **ピンポイント** で推定

- **母集団** -> 全国の中学生test
- **標本** -> 82, 35, 69 点(3人の data)

全国	標本
母平均： μ	標本平均： $\overline{X}$
母分散： σ^2	標本分散： s^2
母標準偏差： σ	標本標準偏差： s

計算

$$\begin{aligned}\overline{X} &= \frac{82-35-69}{3} \\ &= 62\end{aligned}$$

$$\begin{aligned}s^2 &= \frac{(82-62)^2+(35-62)^2+(69-62)^2}{\underbrace{2}_{sample\text{数}-1(\text{不偏分散})}} \\ &= 589\end{aligned}$$

$$s = \sqrt{589} \simeq 24.3$$

別の標本や sample数が変わると推定した値も変わるのでは…

↓

区間推定

分散と不偏分散の違い

- 平均を求める 分母の値 が異なる -> **sample数 - 1**
  - **不偏分散の特徴**
    - **母分散の期待値** と一致する -> 少数 sample から正確に推定
    - data 数が多い時は変わらない

乱数生成

- **sample数** : 3000000 | **平均値** : 70 | **標準偏差** : 10

Hide

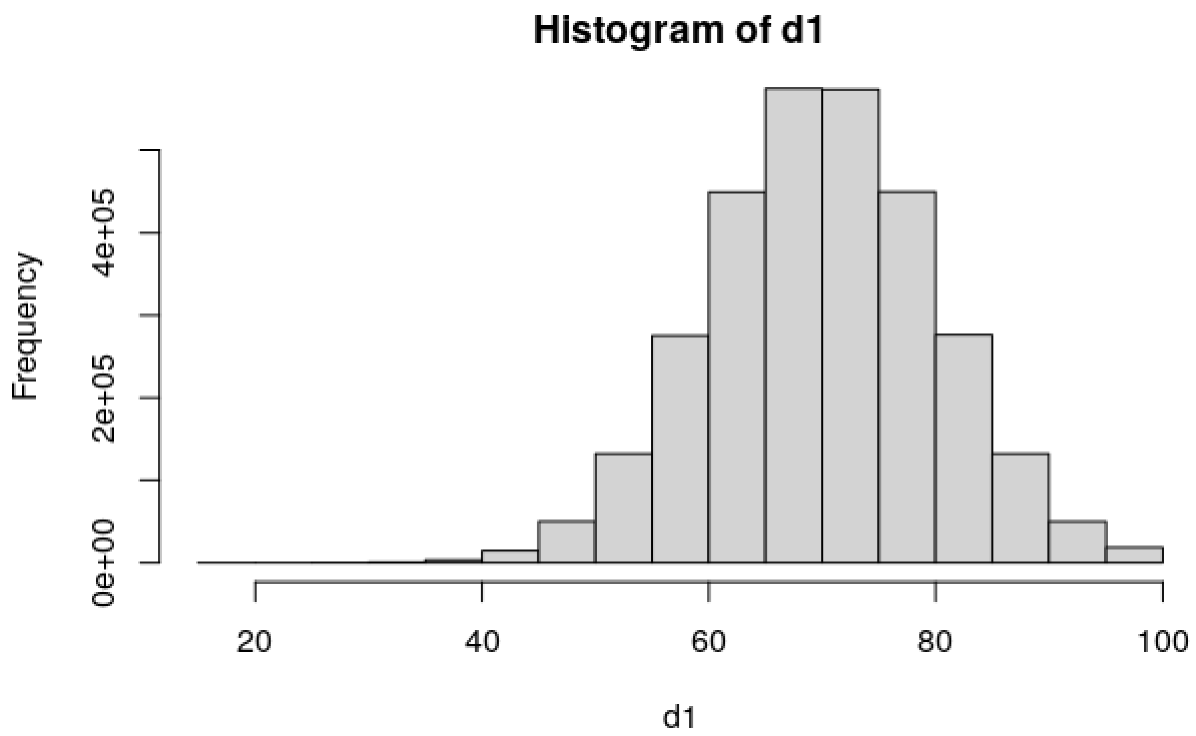
```
n1 <- 3000000
m1 <- 70
s1 <- 10

d1 <- rnorm(n1, mean = m1, sd = s1)
d1 = ifelse(d1 >= 100, 100, d1)
head(d1, 50)
```

```
 [1] 79.63742 74.23461 80.52806 63.81331 54.50828 66.00441
 [7] 67.82964 71.36845 65.08166 66.83498 53.93642 63.74710
[13] 81.34527 78.27128 59.05628 59.70945 63.90869 77.45391
[19] 67.88694 72.84260 43.71903 69.32470 46.03974 76.34489
[25] 75.73658 83.93428 76.88751 65.55255 63.61467 57.74589
[31] 78.00479 54.39060 71.92161 71.53311 69.28176 67.04884
[37] 77.93478 77.22868 60.05815 67.71472 78.03465 55.14857
[43] 68.41044 81.10176 65.02315 68.32329 74.87096 80.05032
[49] 71.35586 58.55932
```

Hide

```
hist(d1)
```



Hide

```
x1 <- sample(d1, 3)
x1
```

```
[1] 94.25051 88.28961 82.06522
```

母集団から **sample**　３つ取得

- 94.2505073
- 88.2896115
- 82.0652169

Hide

```
mean(x1)
```

```
[1] 88.20178
```

- **平均** : 88.2017786

Hide

```
sd(x1)
```

```
[1] 6.09312
```

- **標準偏差** : 6.09312

Hide

```
var(x1)
```

```
[1] 37.12611
```

- **不偏分散** : 37.126111

Hide

```
sum((x1-mean(x1))**2)/3
```

```
[1] 24.75074
```

- **分散** : 24.7507407

区間推定

母集団から sample を 1つとる場合

- どれくらいの　**確率**　でどれくらいの　**範囲**　に現れるか？

公式

$$-1.96 \leq \frac{X-\mu}{\sigma} \leq 1.96$$
$$-1.96\sigma + \mu \leq X \leq 1.96\sigma + \mu$$

- 母集団から　**95%の確率** で
  - $-1.96\sigma + \mu \leq X \leq 1.96\sigma + \mu$　の範囲の data が sample される

Hide

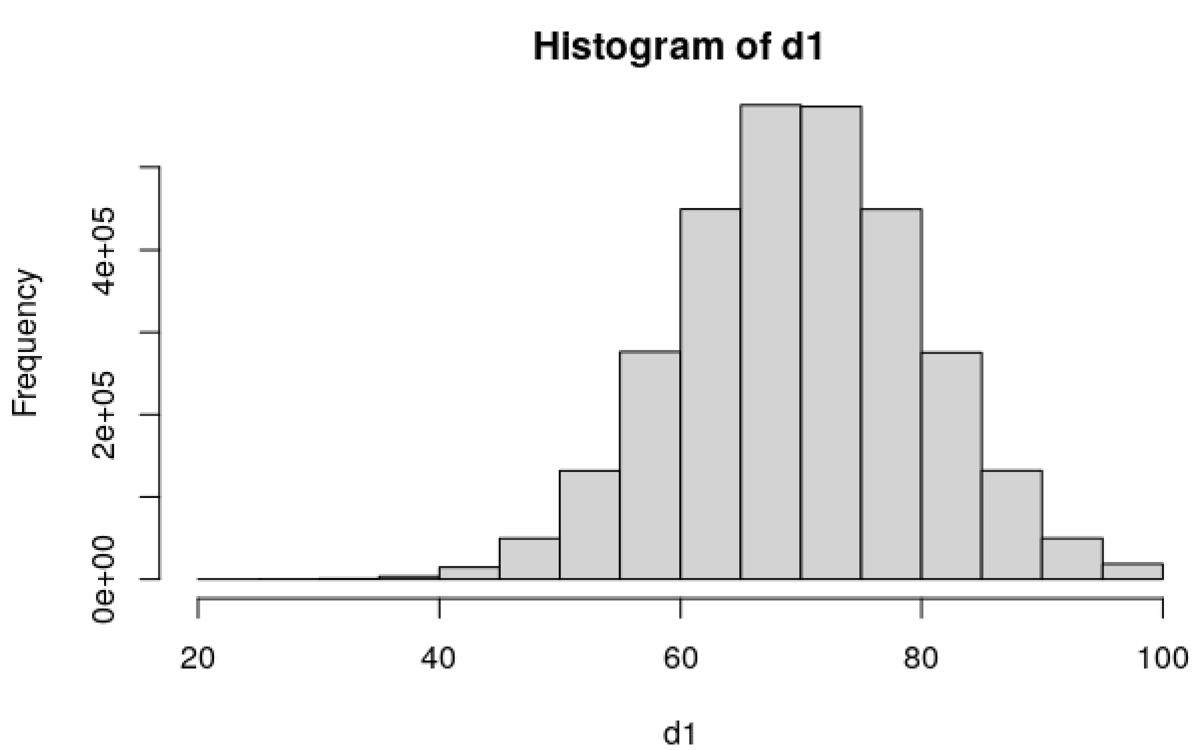
```
n1 <- 3000000
m1 <- 70
s1 <- 10

d1 <- rnorm(n1, mean = m1, sd = s1)
d1 = ifelse(d1 >= 100, 100, d1)
head(d1, 50)
```

```
 [1] 60.75802 84.92520 62.55668 67.07477 81.17898 60.79139
 [7] 62.84295 68.82550 73.27963 62.68304 58.13987 92.11917
[13] 71.18879 53.22491 65.36269 55.91935 85.07858 69.97182
[19] 73.37999 67.48620 70.10253 77.53683 64.97741 89.65119
[25] 68.53228 58.48024 77.85446 58.34877 58.82682 82.53943
[31] 41.76999 75.71214 62.25389 84.82113 60.80362 60.39202
[37] 78.27873 68.11545 45.93541 77.38940 71.64001 87.48785
[43] 78.32443 75.85206 60.06417 70.56833 78.99659 60.68052
[49] 70.10269 69.15931
```

Hide

```
hist(d1)
```



区間推定

- 95%の確率で信頼区間の数値が取得される

$$-1.96\sigma + \mu \leq X \leq 1.96\sigma + \mu$$

- **sample数** : 3000000
- **母平均** :　70
- **標準偏差** : 10

- **母平均** : 70
- **標本(sample)** : 76.937679
- **信頼区間(95%)** : 57.337679 ~ 96.537679

母平均から信頼区間を求める

↓

妥当であるか? or 不適当?

||

どれだけ data が正確であるかを **確率 & 範囲** で判断できる

母平均と標本平均

標本の統計量の分布を　**標本分布**　->　今は **標本平均の標本分布**

- 標本平均の平均 :

$$\mu$$

- 標本平均の標準偏差 :

$$\frac{\sigma}{\sqrt{n}}$$

標準化を行う

$$\text{標準化} \quad = \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

推定区間

- sample 数が 1 -> n数個　へと大きくなると推定区間が変わる

$$-1.96 \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$$

$$-1.96\frac{\sigma}{\sqrt{n}} + \overline{X} \leq \mu \leq 1.96\frac{\sigma}{\sqrt{n}} + \overline{X}$$

sample 数が 1 の時より

↓

標本平均の標準偏差が n が増えるごとに　**小さくなる**

||

信頼区間が狭くなる

区間推定を行う

- **sample数** : 3000000
- **母平均** :　70
- **標準偏差** : 10

Hide

```
nn1 <- 1
nn2 <- 15

sp1 <- sample(d1, nn1)
sp1
```

```
[1] 61.15097
```

Hide

```
sp2 <- sample(d1, nn2)
sp2
```

```
 [1] 76.92950 60.10347 57.05994 77.53548 75.80153 44.60846 80.26383 65.19795 58.04432
[10] 88.54183 65.19209 89.94140 74.17388 55.16521 56.38517
```

- **標本平均の平均**:

$$\mu$$

= 68.3296036

Hide

```
sp2m = mean(sp2)
sp2m
```

```
[1] 68.3296
```

公式

$$-1.96\frac{\sigma}{\sqrt{n}} + \overline{X} \leq \mu \leq 1.96\frac{\sigma}{\sqrt{n}} + \overline{X}$$

sample数 1個

- **信頼区間(95%)**: 41.550967 ~ 80.750967
- **区間の長さ**: 39.2

Hide

```
rl1 <- sp1 - 1.96*(s1/sqrt(nn1))
ru1 <- sp1 + 1.96*(s1/sqrt(nn1))
c(rl1, ru1, ru1 - rl1)
```

```
[1] 41.55097 80.75097 39.20000
```

sample数 15個

- **信頼区間(95%)**: 63.2689054 ~ 73.3903019
- **区間の長さ**: 10.1213965

Hide

```
rl2 <- sp2m - 1.96*(s1/sqrt(nn2))
ru2 <- sp2m + 1.96*(s1/sqrt(nn2))
c(rl2, ru2, ru2 - rl2)
```

```
[1] 63.26891 73.39030 10.12140
```

区間推定とsample数

全国中学生testの結果から標本を抽出。標本から得られる値から母集団を推定する

母集団
母平均： μ 母標準偏差： σ

標本	値
標本平均： $\overline{X}$	70点
標本標準偏差： s	10点
標本数： n	400

母平均の区間推定

- 標本平均： $\overline{X}$： 70点
- 標本標準偏差： s： 10点
- 標本数： n： 400

1. 母集団の何が知りたいのか？ ＝ **母平均**

↓

$$-1.96 \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96$$

2. **標本平均の標本分布 ＝ 正規分布**

↓

$$-1.96\frac{\sigma}{\sqrt{n}} + \overline{X} \le \mu \le 1.96\frac{\sigma}{\sqrt{n}} + \overline{X}$$

実際は母標準偏差を手に入れることはできない…ので

||

sample数が多い場合は **標本標準偏差** を近似値として使用

↓

$$70 - 1.96\frac{10}{\sqrt{400}} \le \mu \le 70 + 1.96\frac{10}{\sqrt{400}}$$

$$69.02 \le \mu \le 70.98$$

3. どれくらいの正確さ？ ＝ **95%**

||

母平均 ＝ **69.02 ~ 70.98**

Hide

```
n3 <- 400
x3 <- rnorm(n3, mean = 70, sd = 10)
```

- **標本平均** : 69.5510083点

Hide

```
x3m <- mean(x3)
x3m
```

```
[1] 69.55101
```

- **標本標準偏差** : 9.4499627

Hide

```
x3sd = sd(x3)
```

- **区間推定**
  - **下限値** : 68.624912
  - **上限値** : 70.4771047

Hide

```
rlx3 <- x3m - 1.96*(x3sd/sqrt(n3))
rux3 <- x3m + 1.96*(x3sd/sqrt(n3))
c(rlx3, rux3)
```

```
[1] 68.62491 70.47710
```