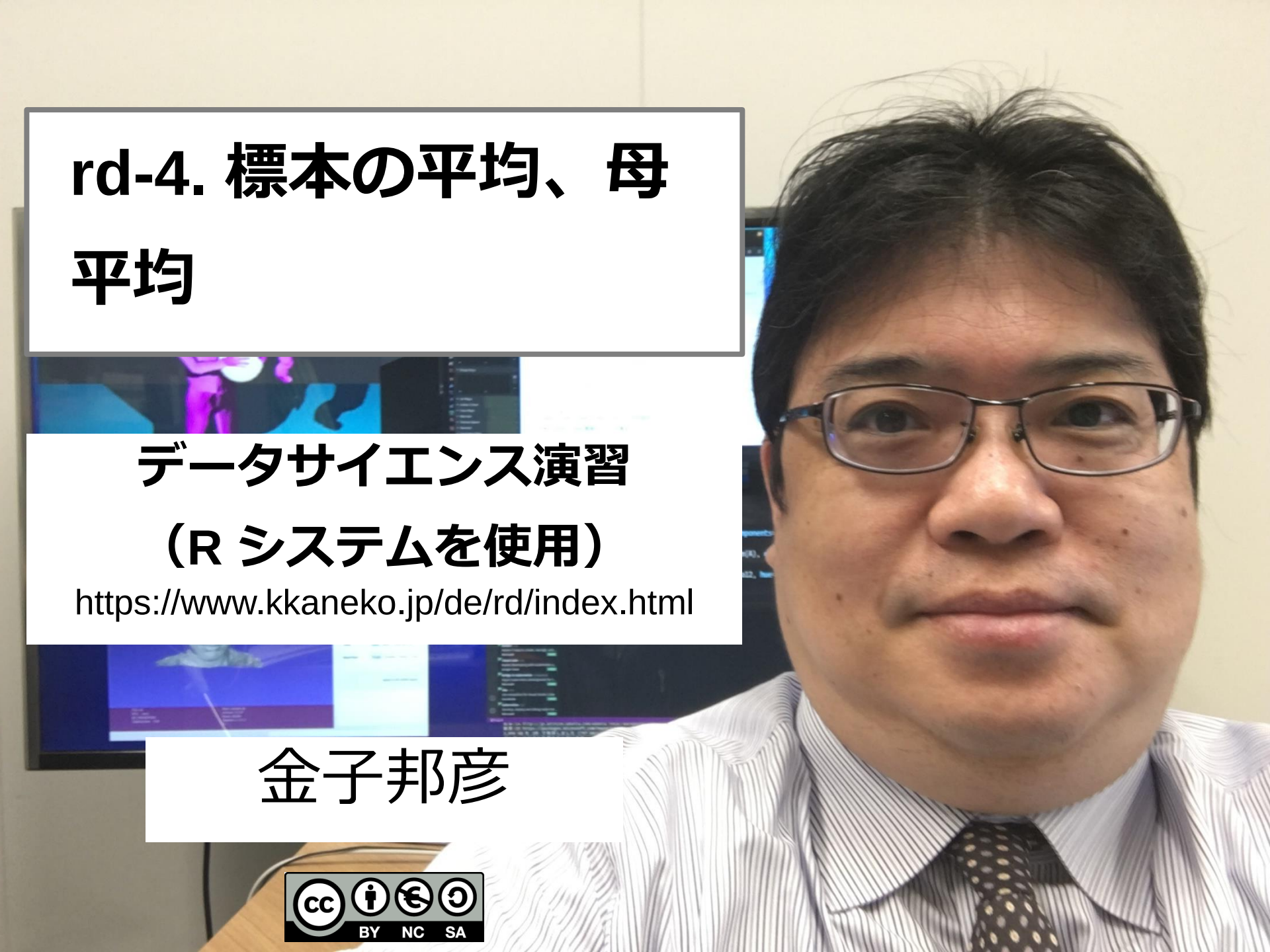
rd-4. 標本の平均、母平均

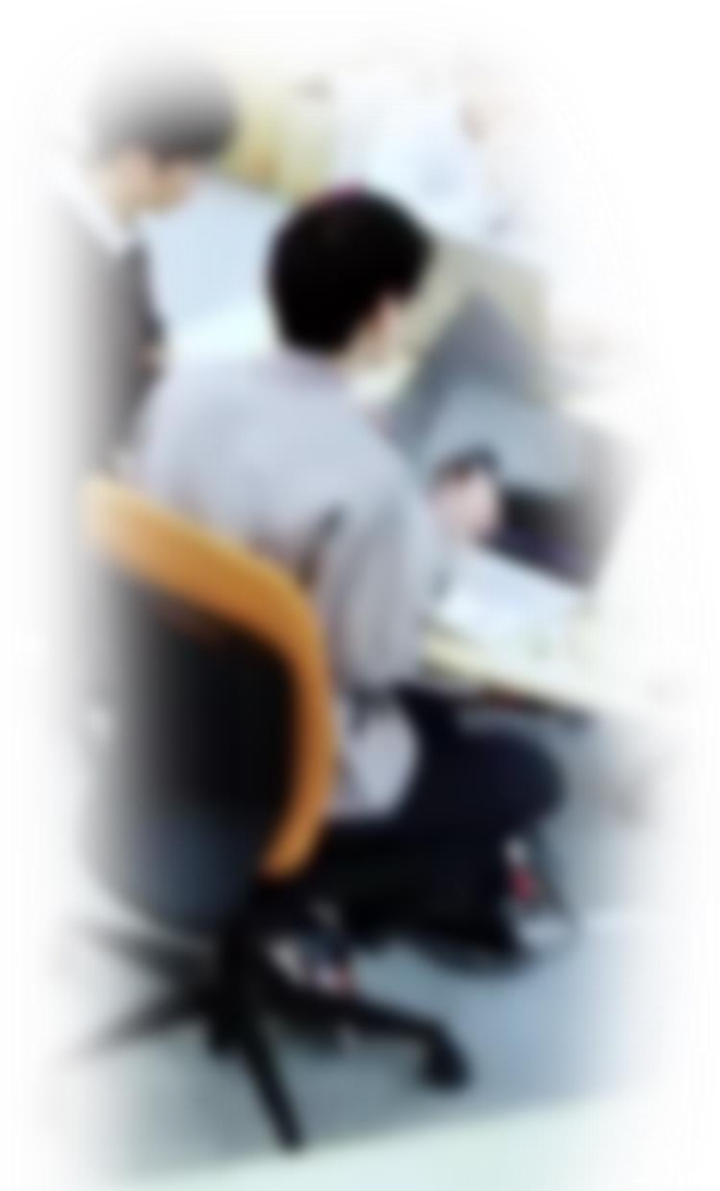


データサイエンス演習

（R システムを使用）

https://www.kkaneko.jp/de/rd/index.html

金子邦彦



アウトライン

1. 平均
2. 母集団と標本
3. 標本の平均値
4. 標本の分散値
5. 演習

1. 평균

平均

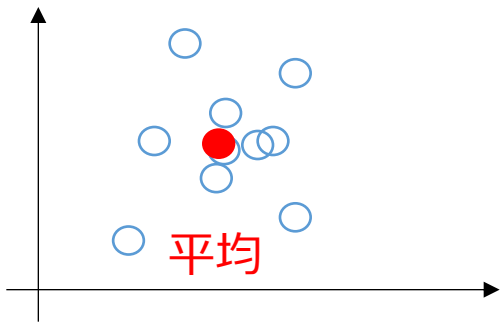
- **平均**は，データの**合計**を，**データの個数**で**割ったもの**

  10, 40, 30, 40 の**平均**: **120** ÷ **4** で **30**

- **複数の値の組**の**平均**を考えることもある

  (10, 5), (40, 10), (30, 5), (40, 20) の平均:

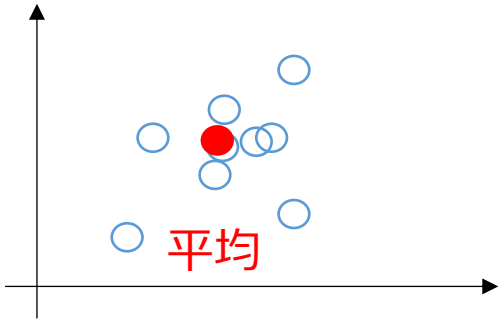
  合計は 120 と 40．4で割って (30, 10)

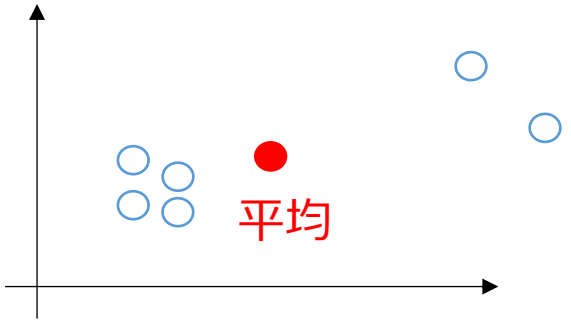


平均は，**データ集合**の**代表**とみることができる場合がある

計測に**誤差**があるとき，
複数の計測を繰り返し，**平均**をとることで，**誤差を軽減**できることも

平均を使うときの注意点





このような平均に，
意味があるでしょうか？

データの分布によっては，**平均**では**役に立たない**こともある．
（平均は万能ではない）

2. 母集団と標本

母集団

母集団は，調査や研究の対象となる**全体の集団**のこと

- **母集団**の**把握と理解**が重要

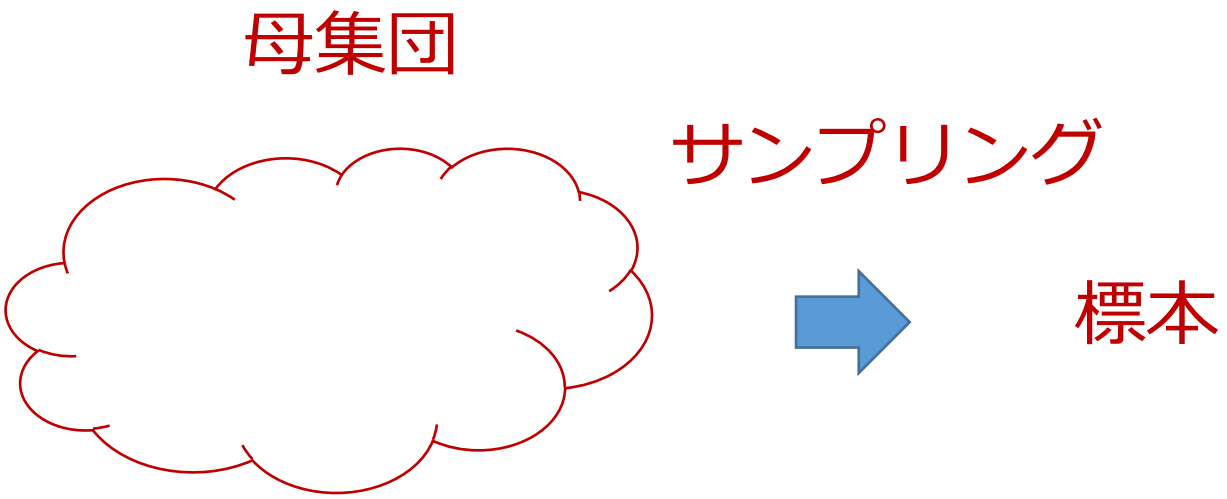
（例）人類全体、２０歳以上の人類全体

サンプリングと標本

- **母集団全体を調べることが困難な場合**、**サンプリング**を適切に行う

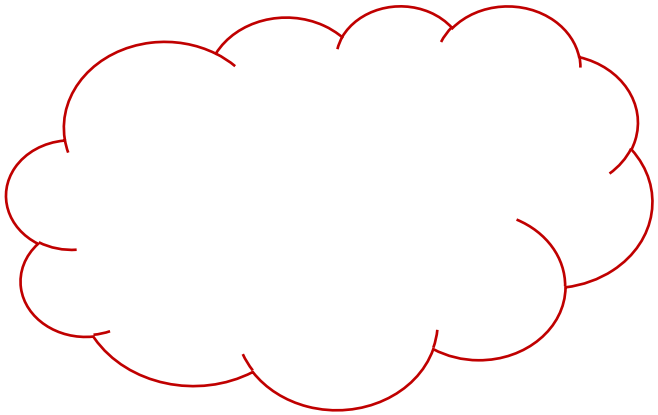
  （例）１０００名をランダムに選ぶ

- **サンプリング**は、**母集団から一部を選ぶ**こと。
- 母集団全体を調べるのでなく、**一部を調べる**ことになる。
- **標本**は、**サンプリングで選ばれたもの**のこと。



サンプリングと標本

母集団



あるときの標本

128
104
124
85
120

平均
112.2

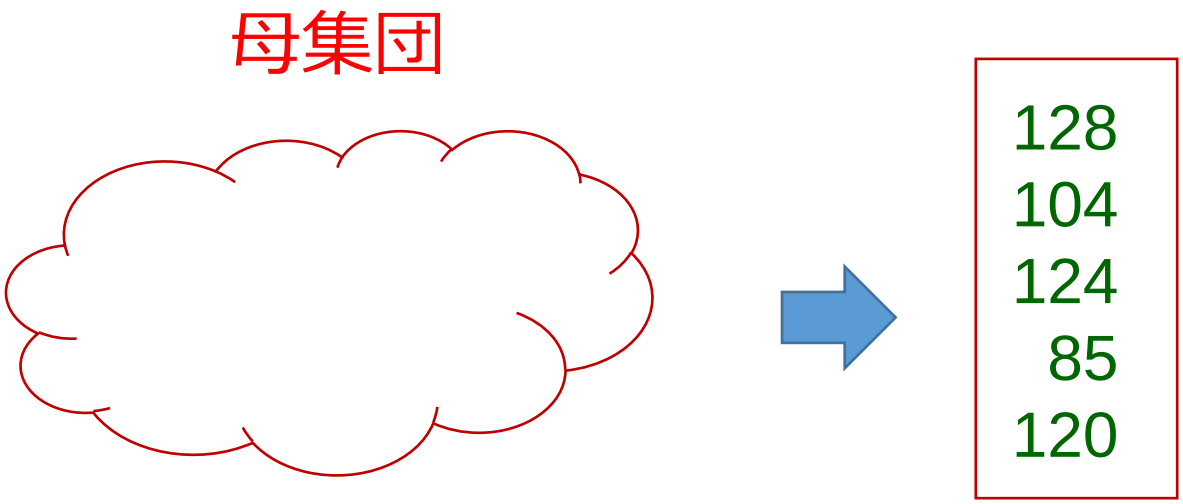
別の標本

118
110
96
85
109

平均
103.6

選ばれた標本によっては、
値が違い、平均なども異なってくる

十分な数の標本が必要



- 標本の大きさが**小さい**と、**結果の信頼性が下がる**
- **十分な数の標本を得る**ことが重要
- 標本の大きさの決定は簡単に決めることができない
- 母集団の特徴、調査や研究の目的に**よって，適切な標本の大きさは変わる**ことに注意しよう

まとめ

- **母集団**：調査や研究の対象となる**全体の集団**
- **サンプリング**：

母集団全体を調べることが困難な場合、**母集団から一部を選ぶサンプリング**を行う。

母集団の特徴や性質を**推測**することが可能となる。

- **標本**：

標本は、母集団からサンプリングで選ばれた**母集団の一部**。

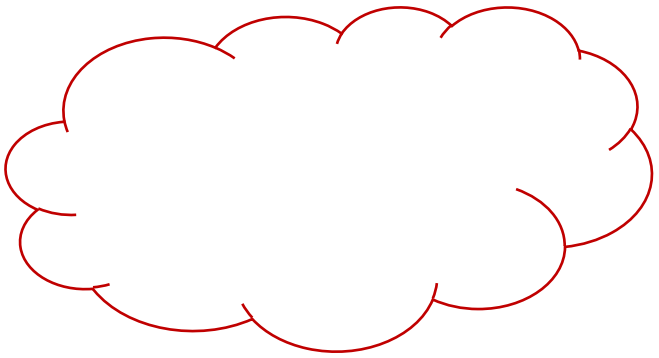
標本から得られたデータを分析し、**母集団全体の性質や傾向を推測**可能。

【注意点】**十分な標本サイズの確保**が必要。ランダムに選択するなどの考慮が重要。

3. 標本の平均値

今から行うことのイメージ

母集団



母集団の平均を
母平均という

たくさんの標本

↓

平均の算出

↓

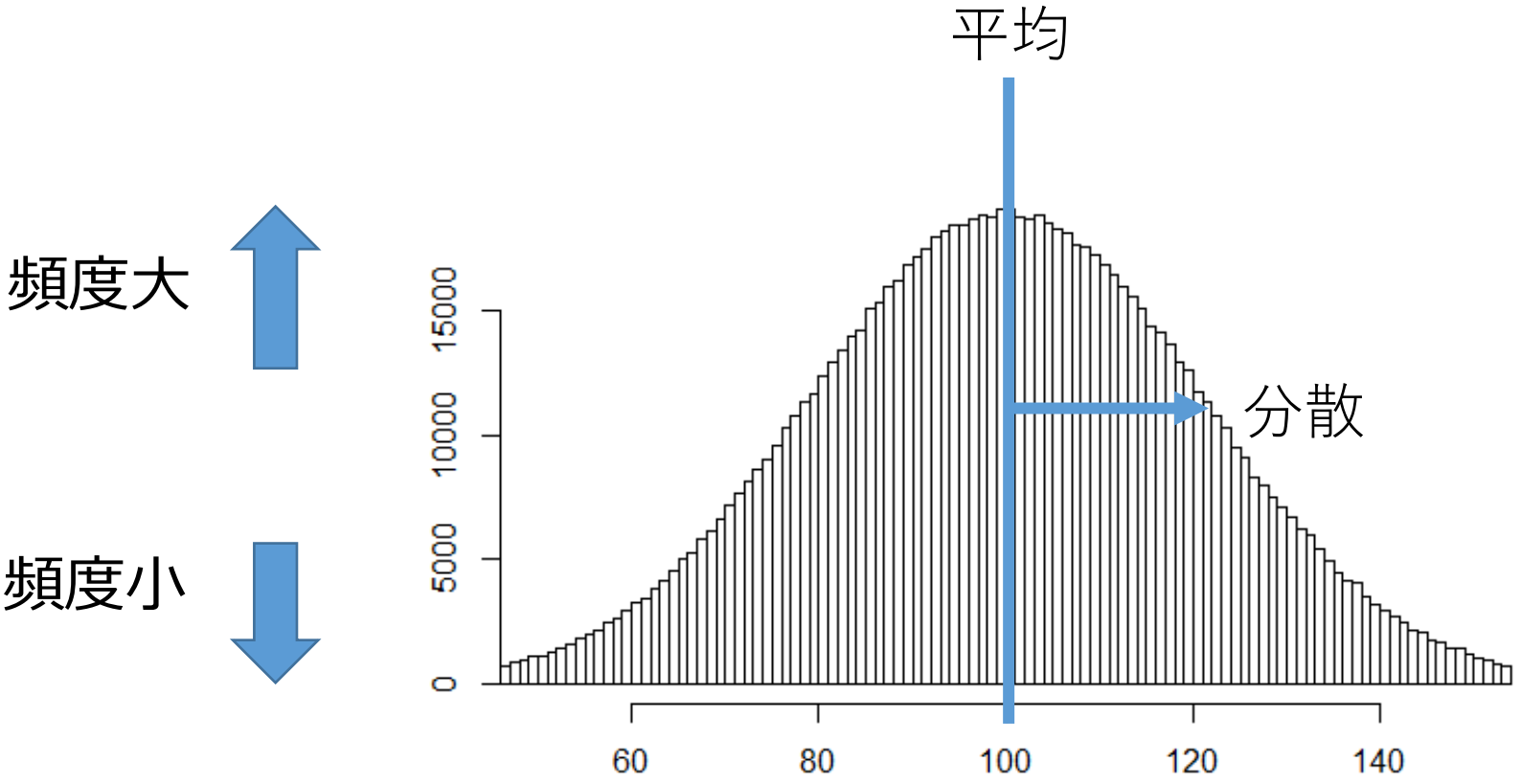
母平均の推定

母平均の**推定の精度を分析**する
ために、母集団は正規分布である
と仮定

正規分布

正規分布は，**平均と分散だけで頻度分布を考える。**

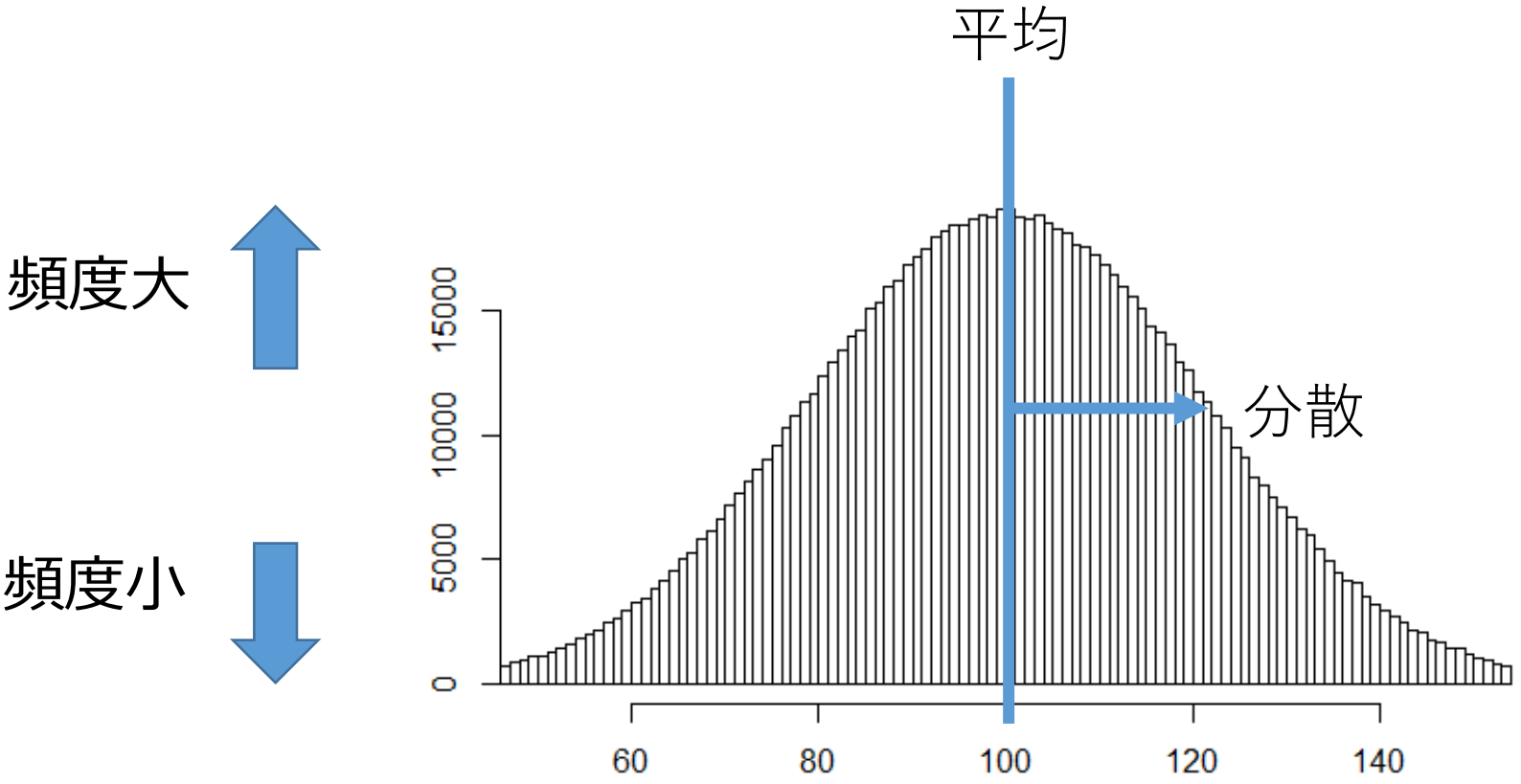
分散は，データの**散らばり度合**を表す



正規分布

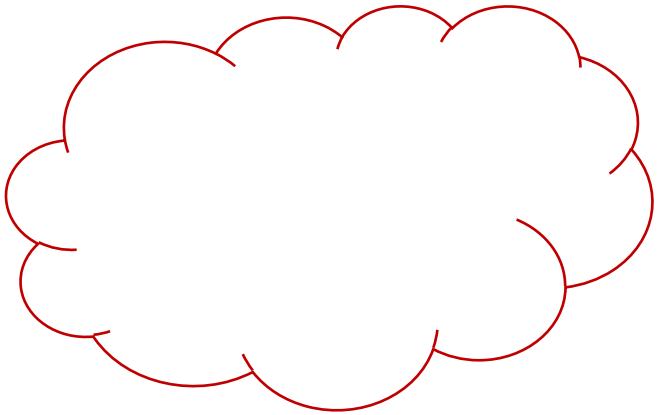
正規分布は，**平均と分散だけで頻度分布を考える。**

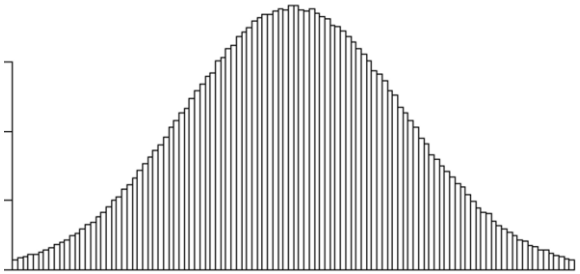
分散は，データの**散らばり度合**を表す



母集団は正規分布であるとし、標本の平均値を算出

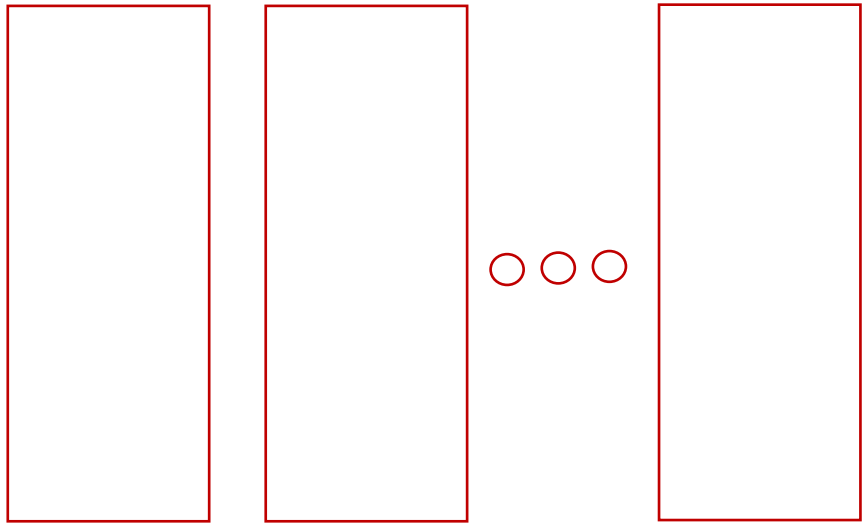
母集団





正規分布

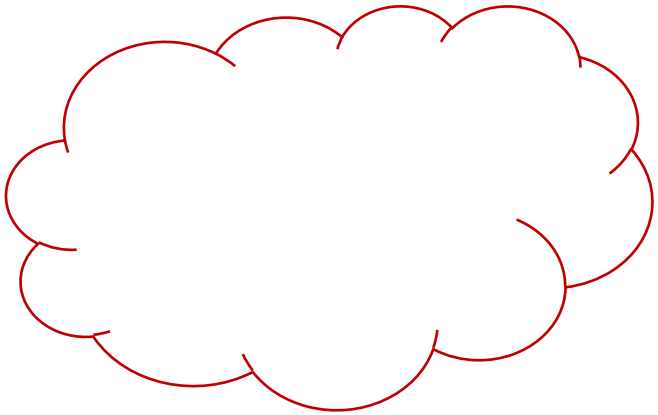
標本（標本数は n）



平均　（n 個の数の平均）

母集団は正規分布であるとし、標本の平均値を算出

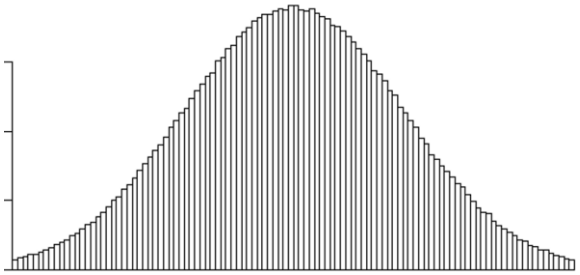
母集団



標本（標本数はｎ，ｎ＝5）

128	80	118
104	80	110
124	126	96
85	122	85
120	79	109

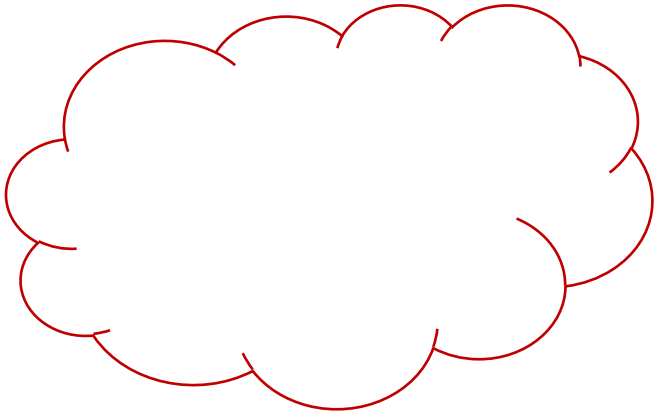
○○○

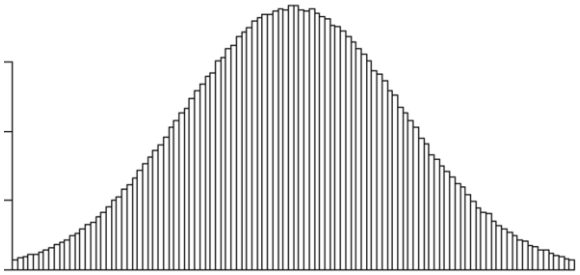


正規分布

母集団は正規分布であるとし、標本の平均値を算出

母集団





正規分布

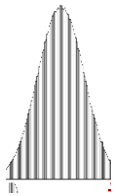
標本（標本数はｎ，ｎ＝5）

128	80	118
104	80	110
124	126	96
85	122	85
120	79	109

○○○

平均 112.2　平均 97.4　平均 103.6

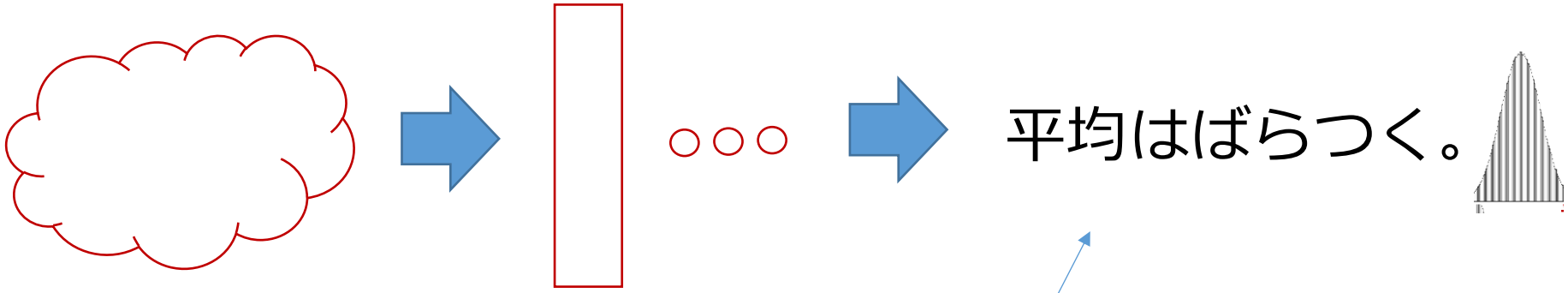
平均はばらつく。

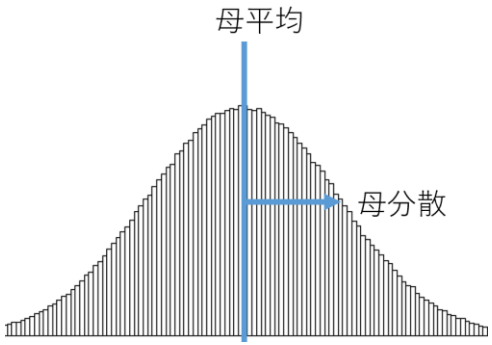


母集団は正規分布であるとし、標本の平均値を算出

母集団

標本（標本数は n ）





正規分布

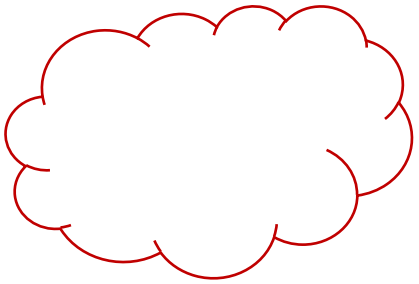
母集団が**正規分布**であるとき、この分布も**正規分布**

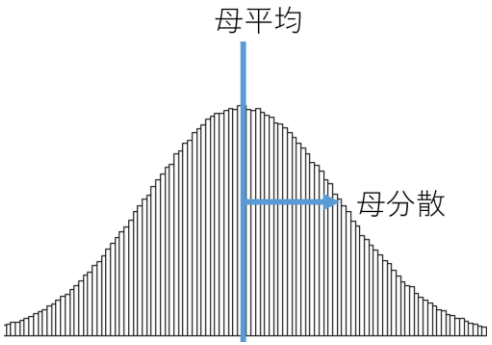
- この**正規分布の平均** **<母平均>** に等しい
- この**正規分布の分散** **<母分散>／n**

- 母集団の平均は、**母平均**という
- 母集団の分散は、**母分散**という

まとめ

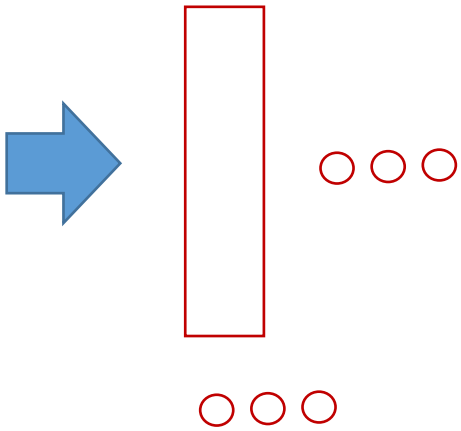
母集団





正規分布

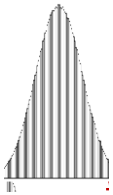
標本（標本数は n ）



平均

この平均から、母平均を推定したい

母分散が小さいほど精度がよい。n が大きいほど精度がよい



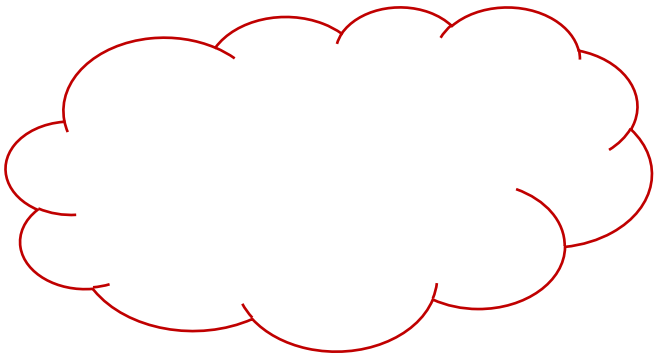
正規分布

この正規分布の＜分散＞は、＜母分散＞／n

4. 標本の分散値

今から行うことのイメージ

母集団



母集団の不偏分散を
知りたい

たくさんの標本

↓

不偏分散の算出

↓

母集団の不偏分散の推定

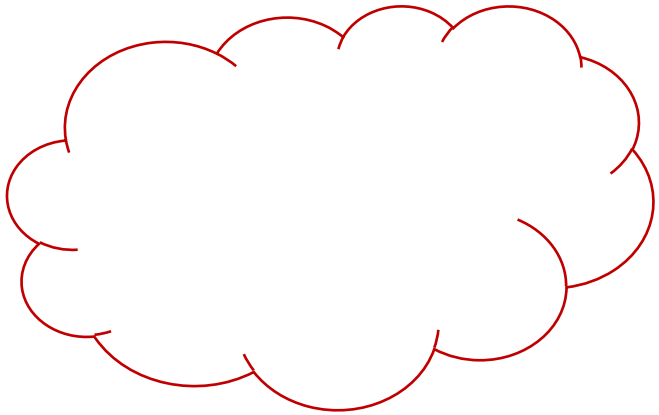
母不偏分散の**推定の精度を分析**する
ために、母集団は t 分布であると仮定
（t 分布は正規分布と少し異なる形）

- **分散**は，データの**散らばり度合**を表す
- 母分散（母集団の分散）は、標本からは推定できないもの

- 母分散の代わりに、不偏分散を用いる

標本の分散値を算出

母集団



t 分布

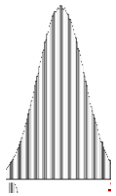
標本（標本数は n ， n = 5）

128	80	118
104	80	110
124	126	96
85	122	85
120	79	109

○○○

不偏分散 314.2　170.3　591.8

求まった値はばらつく。
・その分布の平均は、
元の母集団の不偏分散に等しい
・n が大きいほど精度がよい



5. 演習

R のベクトル

ベクトルとは，**データの並び**のこと．
各要素に番号（添え字）がある．

- コンストラクタ（ベクトルデータの組み立て）

  c や numeric など

```
> p <- c(100, 200, 300, 400)
> print(p)
[1] 100 200 300 400
> 
```

```
> p <- numeric(10)
> print(p)
 [1] 0 0 0 0 0 0 0 0 0 0
> 
```

- 添え字によるアクセス []

```
> print(p)
[1] 100 200 300 400
> p[1]
[1] 100
> p[2]
[1] 200
> p[3]
[1] 300
> p[4]
[1] 400
> 
```

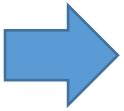
R での平均と不偏分散

- 平均　　mean
- 不偏分散　　var

※ 不偏分散は，標本値のばらつきを表す値

R での平均と不偏分散

128	118	80	127
104	110	80	72
124	96	126	111
85	85	122	82
120	109	79	81

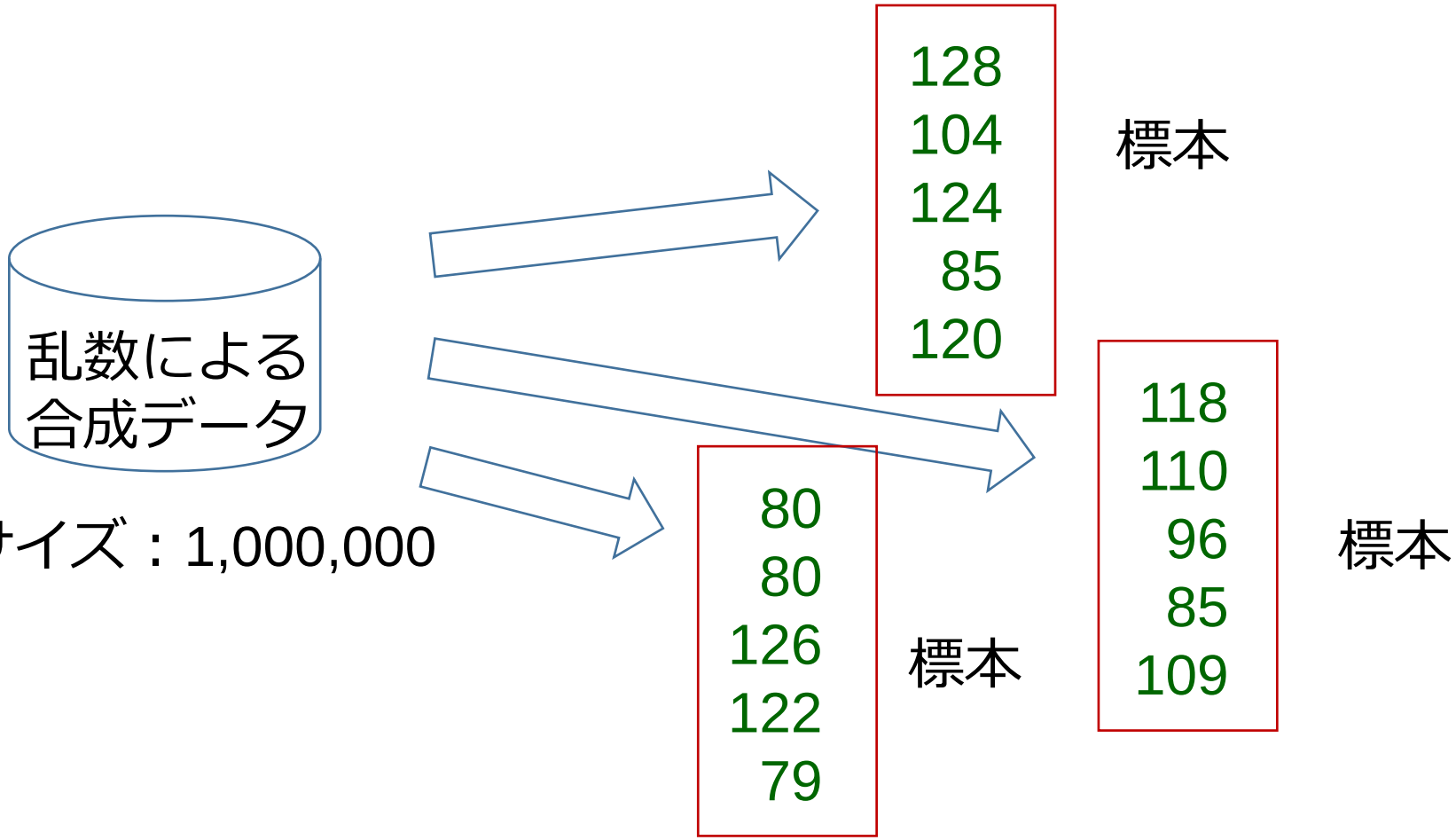


```
> c1 <- c(128, 104, 124, 85, 120)
> c2 <- c(118, 110, 96, 85, 109)
> c3 <- c(80, 80, 126, 122, 79)
> c4 <- c(127, 72, 111, 82, 81)
> mean(c1)
[1] 112.2
> mean(c2)
[1] 103.6
> mean(c3)
[1] 97.4
> mean(c4)
[1] 94.6
> var(c1)
[1] 314.2
> var(c2)
[1] 170.3
> var(c3)
[1] 591.8
> var(c4)
[1] 543.3
>
```

```
c1 <- c(128, 104, 124, 85, 120)
c2 <- c(118, 110, 96, 85, 109)
c3 <- c(80, 80, 126, 122, 79)
c4 <- c(127, 72, 111, 82, 81)
mean(c1)
mean(c2)
mean(c3)
mean(c4)
var(c1)
var(c2)
var(c3)
var(c4)
```

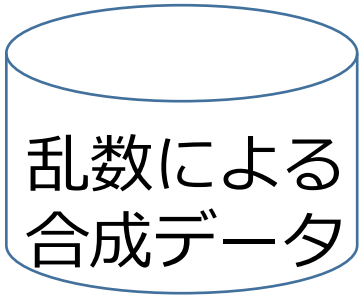
今から行うこと

「1,000,000個の中から
ランダムに標本を選ぶ」



今から行うこと

「1,000,000個の中から
ランダムに標本を選ぶ」



サイズ：1,000,00

R では
ベクトルデータ x の 1,000,000個の中から
ランダムに5個選びたいときは

```
x[floor( runif(5, 1, 1000000+1) )]
```

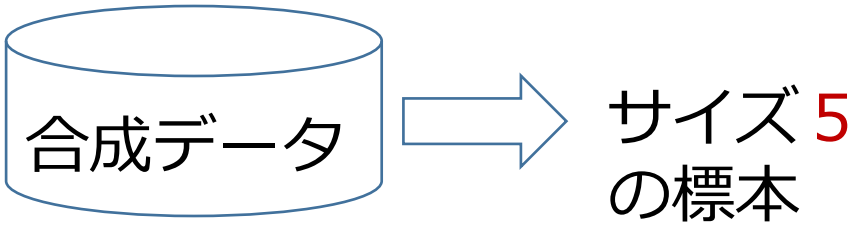
128

80
126
122
79

標本

85
109

合成データからランダムに5個選び標本を作る



タイプ：数値
サイズ：1,000,000

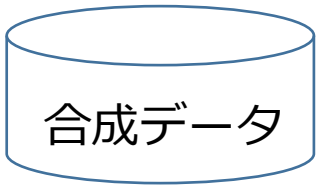
```
> x[floor( runif(5, 1, 1000000+1) )]
[1] 102  79 101  91 103
> x[floor( runif(5, 1, 1000000+1) )]
[1] 110 110 106 115  90
> x[floor( runif(5, 1, 1000000+1) )]
[1] 114 114 112  98 103
>
```

毎回違う結果が出る

```
x <- round( rnorm(1000000, mean=100, sd=20) )
x[floor( runif(5, 1, 1000000+1) )]
x[floor( runif(5, 1, 1000000+1) )]
x[floor( runif(5, 1, 1000000+1) )]
```

乱数による合成データの生成

標本を20個作り，各標本の平均や不偏分散を求める



サイズ 5
の標本を
20個

各標本の
平均や
不偏分散

タイプ：数値
サイズ：1,000,000

```
> print(m)
 [1]  89.4  86.4 102.0 118.8  92.6 102.2
 [7] 102.6  94.8 109.8 102.0  92.8 113.4
[13]  89.2 100.2 105.8  95.0 113.2  90.4
[19]  94.2  96.0
> print(v)
 [1] 327.8 455.3 246.0 493.2  50.8 417.2
 [7] 665.3 212.7 738.2  57.5 405.7 786.3
[13] 876.7 603.7 171.7 372.0 142.7 572.3
[19] 139.7 505.0
>
>
```

毎回違う結果が出る

```
x <- round( rnorm(1000000, mean=100, sd=20) )
m <- numeric(20)
v <- numeric(20)
for (i in 1:20) {
  s <- x[floor( runif(5, 1, 1000000+1) )]
  m[i] <- mean(s)
  v[i] <- var(s)
}
print(m)
print(v)
```

合成データからランダムに
5個選び標本を作る

平均と不偏分散

各標本の平均値を比べる

標本の例

128	118	80	127
104	110	80	72
124	96	126	111
85	85	122	82
120	109	79	81

標本２個の各平均値　112.2　103.6

総平均：107.9

標本３個の各平均値　112.2　103.6　97.4

総平均：104.4

標本４個の各平均値　112.2　103.6　97.4　94.6

総平均：101.95

各標本の不偏分散値を比べる

標本

128	118	80	127
104	110	80	72
124	96	126	111
85	85	122	82
120	109	79	81

標本２個の各不偏分散値　314.2　170.3

その平均：242.25

標本３個の各不偏分散値　314.2　170.3　591.8

その平均：358.7667

標本４個の各不偏分散値　314.2　170.3　591.8　543.3

その平均：404.9

各標本の平均値や不偏分散値を集めて，平均をとる


```
x <- round( rnorm(1000000, mean=100, sd=20) )
m <- numeric(20)
v <- numeric(20)
for (i in 1:20) {
  s <- x[floor( runif(5, 1, 1000000+1) )]
  m[i] <- mean(s)
  v[i] <- var(s)
}
for (i in 1:20) { print( mean(m[1:i]) ) }
for (i in 1:20) { print( mean(v[1:i]) ) }
```

```
> for (i in 1:20) { print( mean(m[1:i]) ) }
[1] 89.4
[1] 87.9
[1] 92.6
[1] 99.15
[1] 97.84
[1] 98.56667
[1] 99.14286
[1] 98.6
[1] 99.84444
[1] 100.06
[1] 99.4
[1] 100.5667
[1] 99.69231
[1] 99.72857
[1] 100.1333
[1] 99.8125
[1] 100.6
[1] 100.0333
[1] 99.72632
[1] 99.54
```

だんだんと
100 に近づく

各標本の平均値を集めて
平均を求める

```
> for (i in 1:20) { print( mean(v[1:i]) ) }
[1] 327.8
[1] 391.55
[1] 343.0333
[1] 380.575
[1] 314.62
[1] 331.7167
[1] 379.3714
[1] 358.5375
[1] 400.7222
[1] 366.4
[1] 369.9727
[1] 404.6667
[1] 440.9769
[1] 452.6
[1] 433.8733
[1] 430.0063
[1] 413.1059
[1] 421.95
[1] 407.0947
[1] 411.99
```

だんだんと
400 に近づく

各標本の不偏分散値を集めて
平均を求める

ランダムなので，毎回違う結果が出る

```
> for (i in 1:20) { print( mean(m[1:i]) ) }
[1] 90
[1] 94.3
[1] 102.8
[1] 101.7
[1] 103.24
[1] 103.7
[1] 102.5714
[1] 104.4
[1] 105.9778
[1] 105
[1] 105.6909
[1] 105.75
[1] 106.1692
[1] 105.5286
[1] 106.0133
[1] 106.175
[1] 105.3529
[1] 105.2
[1] 105.7895
[1] 106.97
```

だんだんと
100 に近づく

何度やっても同じ

各標本の平均値を集めて
平均を求める

```
> for (i in 1:20) { print( mean(v[1:i]) ) }
[1] 649
[1] 571.15
[1] 593
[1] 500.075
[1] 452.72
[1] 410.9333
[1] 449.4
[1] 405.4375
[1] 524.4222
[1] 546.5
[1] 519.6182
[1] 502.7583
[1] 473.9462
[1] 468.8929
[1] 511.3133
[1] 489.8125
[1] 480.7176
[1] 463.9167
[1] 461.0684
[1] 457.33
```

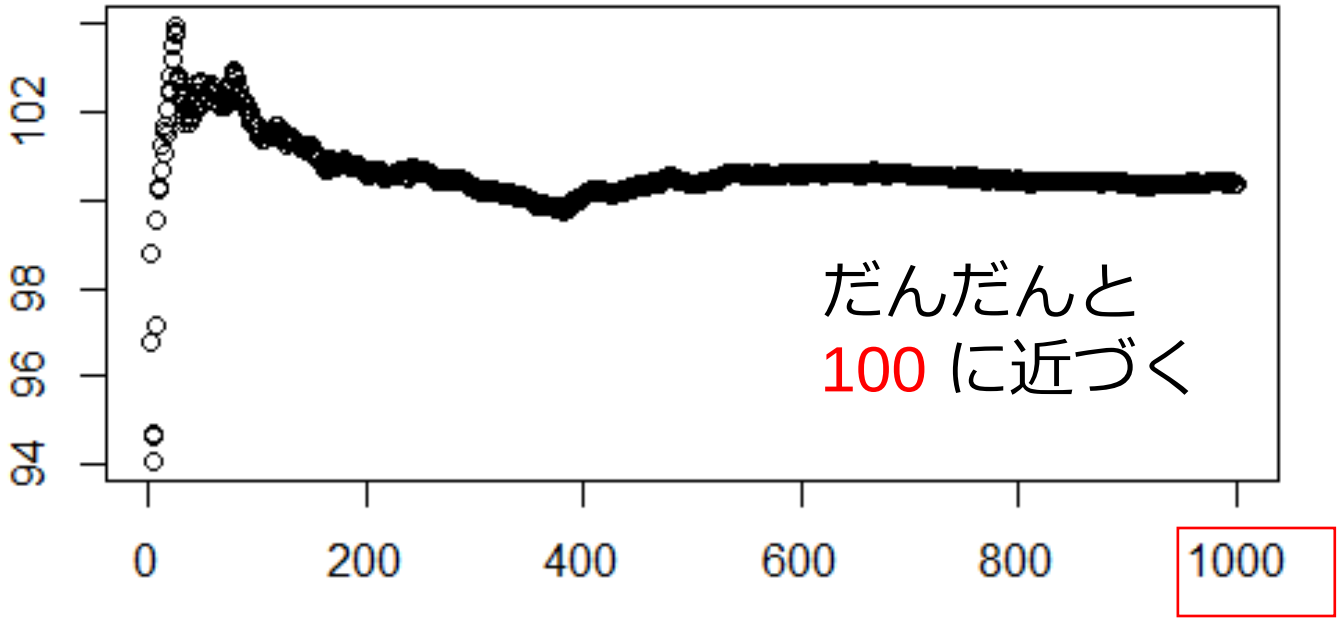
だんだんと
400 に近づく

何度やっても同じ

各標本の不偏分散値を集めて
平均を求める

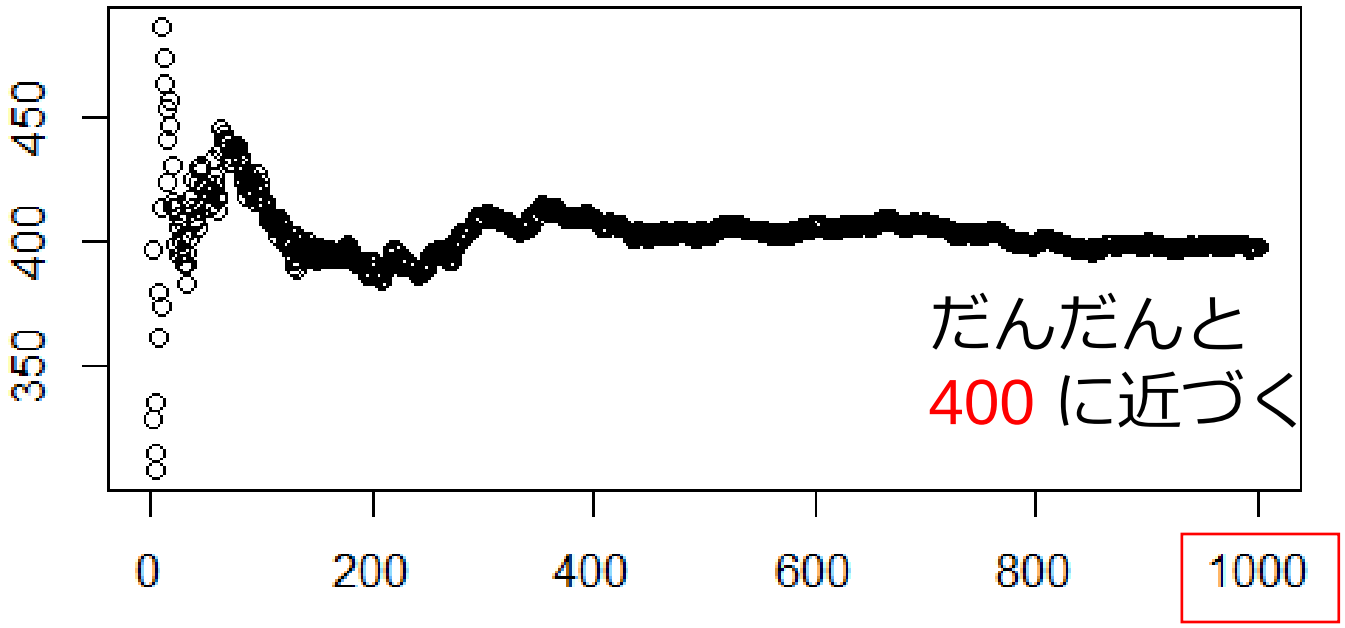
ランダムなので，毎回違う結果が出る

標本の個数を 20 から 1000 の間で変えて，総平均を求めてみる



各標本の平均値を集めて**総平均**を求める

標本の個数を 20 から 1000 の間で変えて，総平均を求めてみる



各標本の不偏分散値を集めて**総平均**を求める

標本の平均から母平均を推定

標本の平均から**母平均**を推定するときに気を付けること

- **標本の大きさ**
  標本の大きさは、母平均の推定精度に大きく影響．標本の大きさが大きいほど精度が向上
- **誤差の認識**
  標本の平均から母集団を推定する際は、必ず誤差が発生する（論文などに細かすぎる値を書かないこと）
- **サンプリングはランダムに**
  母集団を正確に反映する標本を得ることが重要
- **母集団のデータの分布の確認**
  正規分布か確認．統計手法では（t 検定など）、正規分布を前提としている場合がある
- **外れ値の考慮**
  外れ値は，平均値に大きく影響する．外れ値は取り除くか適切に書き換える