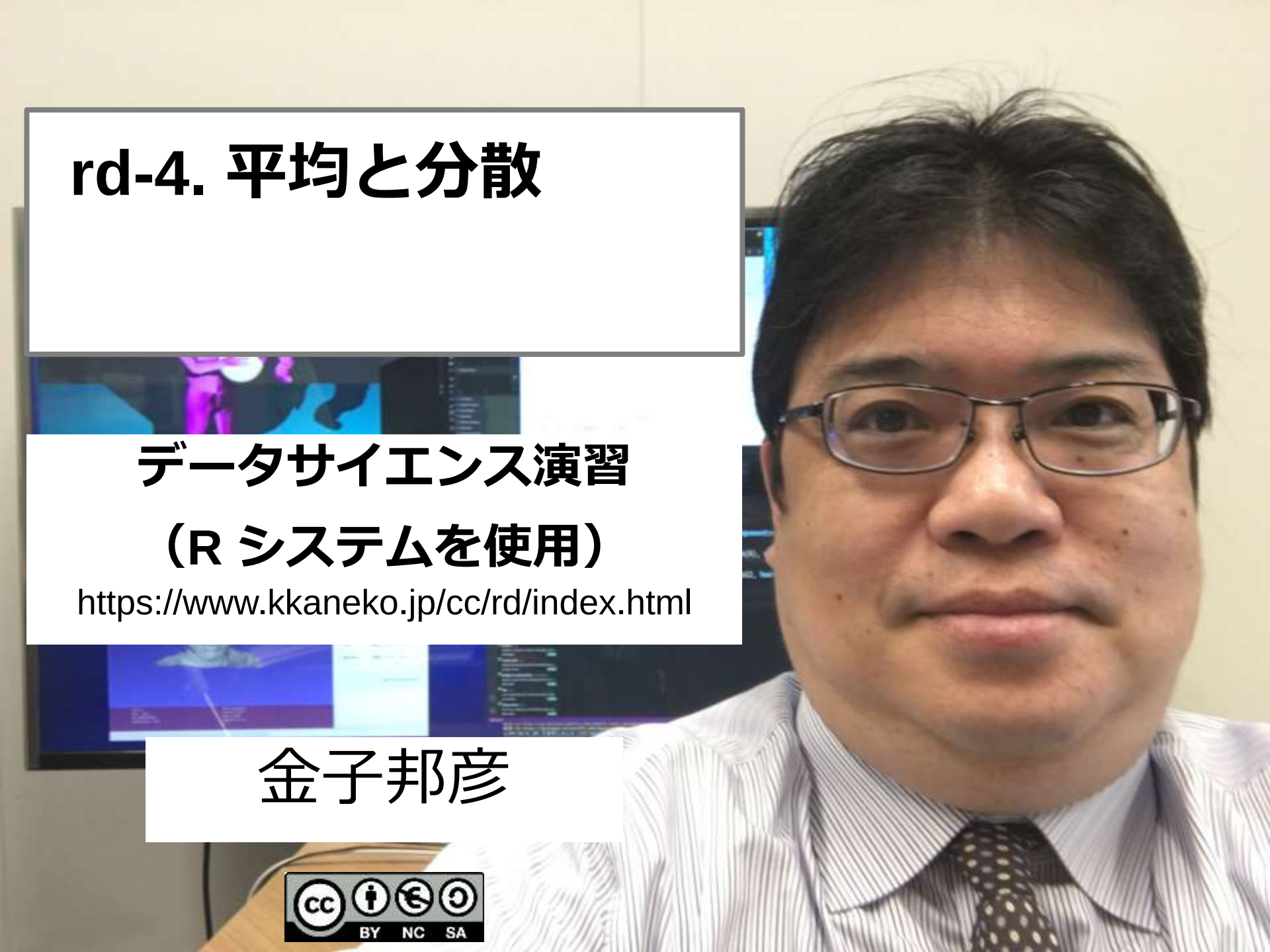
rd-4. 平均と分散



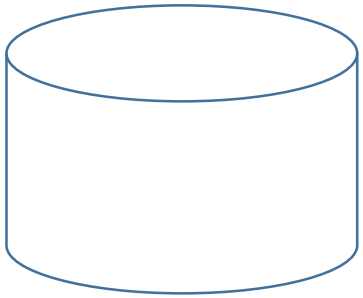
データサイエンス演習

（R システムを使用）

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

金子邦彦

「1,000,000個の中から
ランダムに標本を選ぶ」



サイズ：
1,000,000

標本

128
104
124
85
120

平均　112.2
不偏分散 314.2

80
80
126
122
79

平均　97.4
不偏分散 591.8

標本

118
110
96
85
109

平均　103.6
不偏分散 170.3

標本

アウトライン

4-1 変数

4-2 平均と不偏分散

4-3 母平均と母分散

4-1 変数

変数の例

変数が３つ

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

元データ

プログラムでの変数

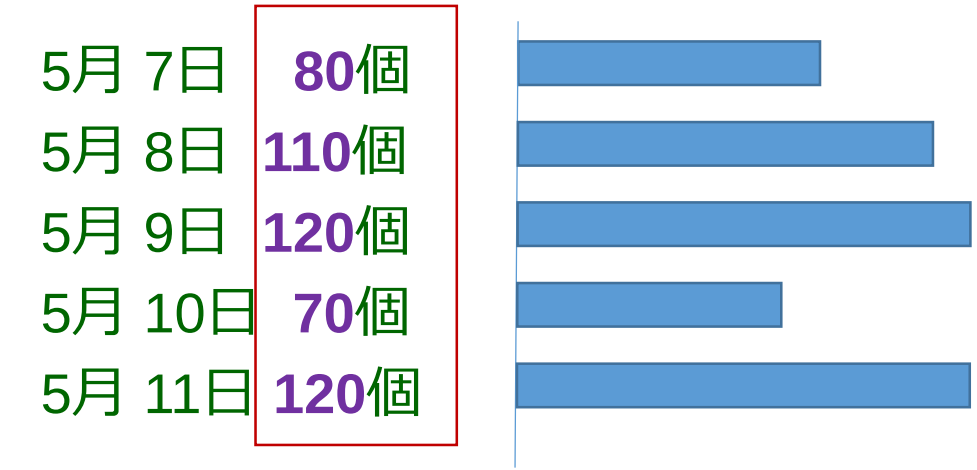
◆　値を１つ保持するためのもの

ここで説明する変数

◆　変化する値のこと

変数の例

値が変化する何か



＜変数＞

（例）
日本人全員
（超巨大だったり）
1900年から2100年までの人口変化
（未知だったり）

＜標本＞

（例）
ランダム抽出された３０人
2016年5月8日から11日までのデータ

変数と標本の例

各標本のデータ数（標本の大きさ）
を決め，標本を得る

値が変化する何か

<変数>

標本１	標本２	標本３	標本４
128	118	80	127
104	110	80	72
124	96	126	111
85	85	122	82
120	109	79	81

◆ 標本をとるたびに**違う値**

4-2 平均と不偏分散

平均と不偏分散

128
104
124
85
120

標本１

数値データの集まり

■ 平均とは
すべての数値を足して，データの個数で割った値

■ 不偏分散とは
数値データの散らばり具合を表す数値の１つ

平均と不偏分散

値が変化する何か

＜変数＞

標本１	標本２	標本３	標本４
128	118	80	127
104	110	80	72
124	96	126	111
85	85	122	82
120	109	79	81

それぞれの平均と不偏分散を求めると

	標本１	標本２	標本３	標本４
平均	112.2	103.6	97.4	94.6
不偏分散	314.2	170.3	591.8	543.3

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
> 
```

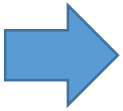
R での平均と不偏分散

- 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)
```

4-3 母平均と母分散

母平均と母分散

値が変化する何か

<変数>

母平均：変数値の平均
母分散：変数値の分散

標本１	標本２	標本３	標本４
128	118	80	127
104	110	80	72
124	96	126	111
85	85	122	82
120	109	79	81
平均			
112.2	103.6	97.4	94.6
不偏分散			
314.2	170.3	591.8	543.3

変数に関するもの	標本に関するもの
母平均	平均
母分散	不偏分散

標本や不偏分散を使って，
母平均や母分散を**推定**する

標本データから，
平均，不偏分散が求まる

今から行うことのイメージ

値が変化する何か
＜変数＞

たくさんの標本

↓

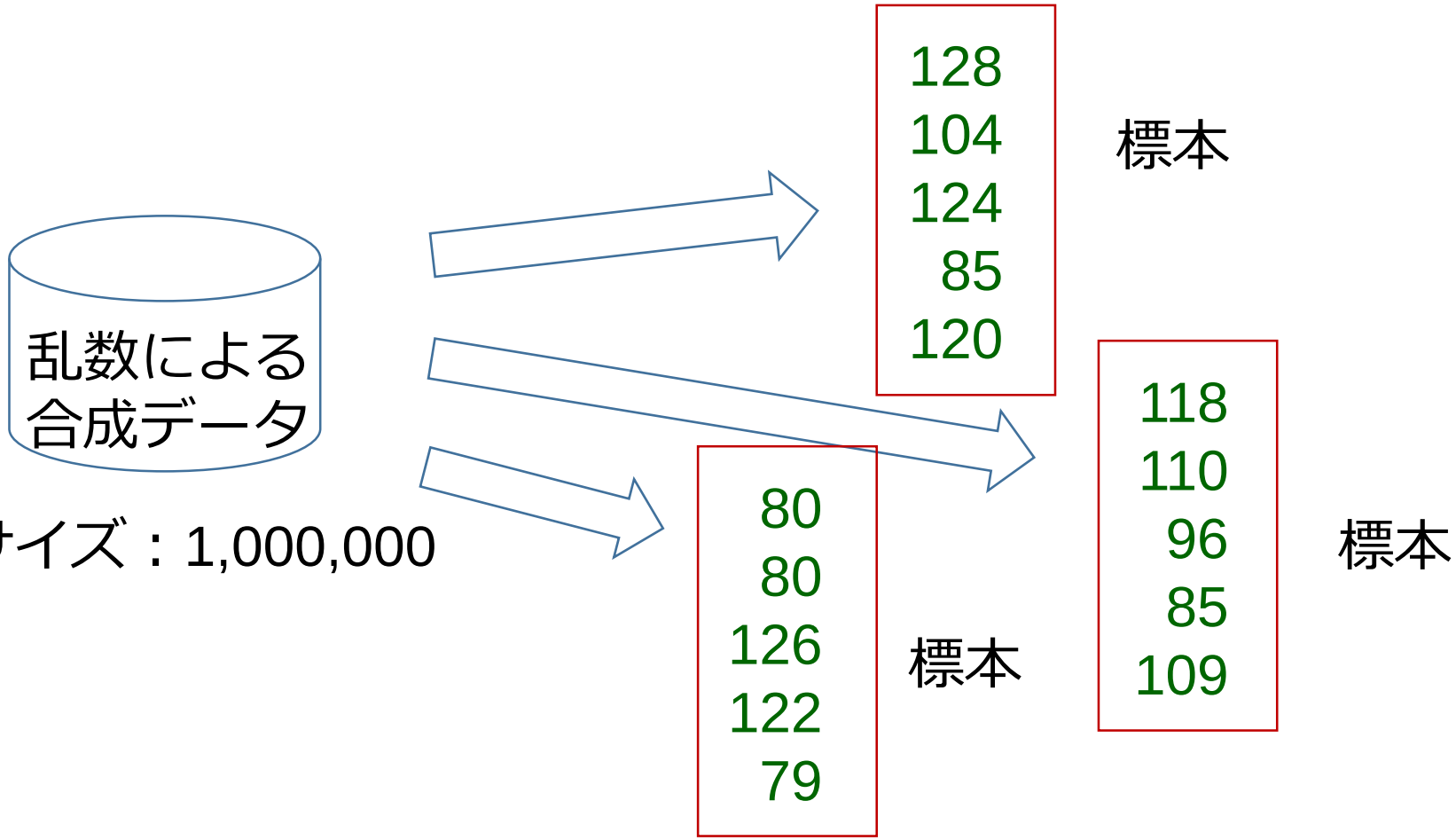
平均，不偏分散の算出

↓

母平均，母分散の推定

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」

128

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))]

乱数による
合成データ

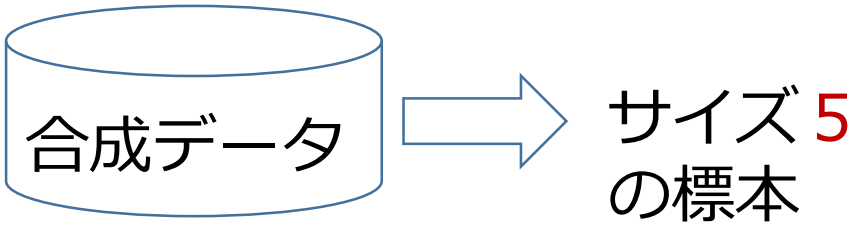
サイズ：1,000,0

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

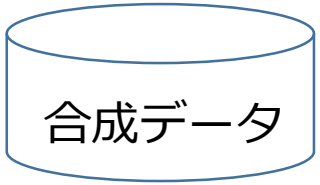
```
> x[floor( runif(5, 1, 1000000+1) )]
[1] 102  79 101  91 103
> x[floor( runif(5, 1, 1000000+1) )]
[1] 110 110 106 115  90
> x[floor( runif(5, 1, 1000000+1) )]
[1] 114 114 112  98 103
>
```

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)
```

平均と不偏分散

合成データからランダムに
5個選び標本を作る

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

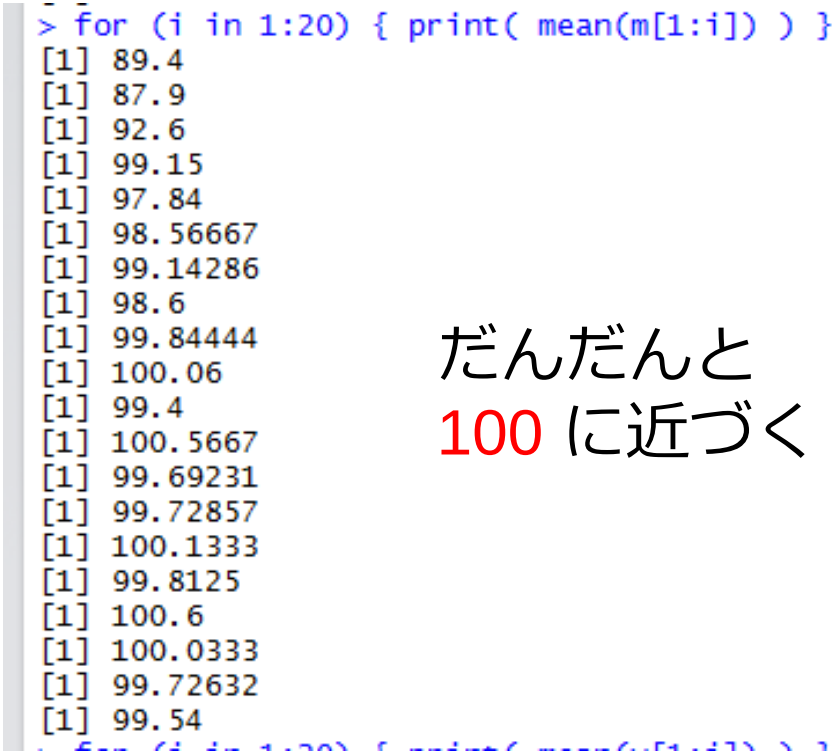
その平均：358.7667

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

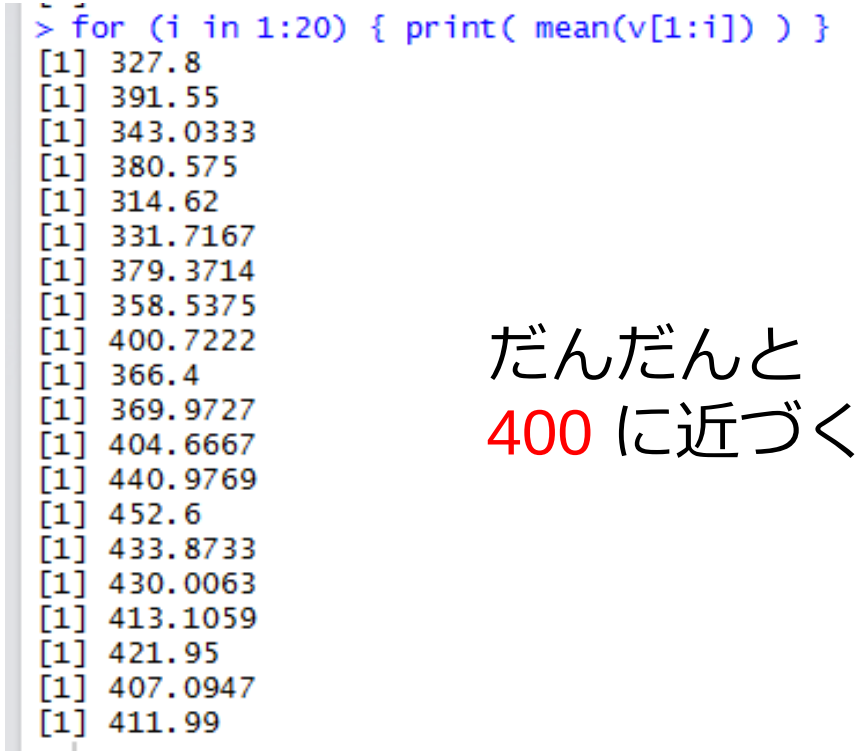
各標本の平均値や不偏分散値を集めて，平均をとる


```
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]) ) }
```



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



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
```

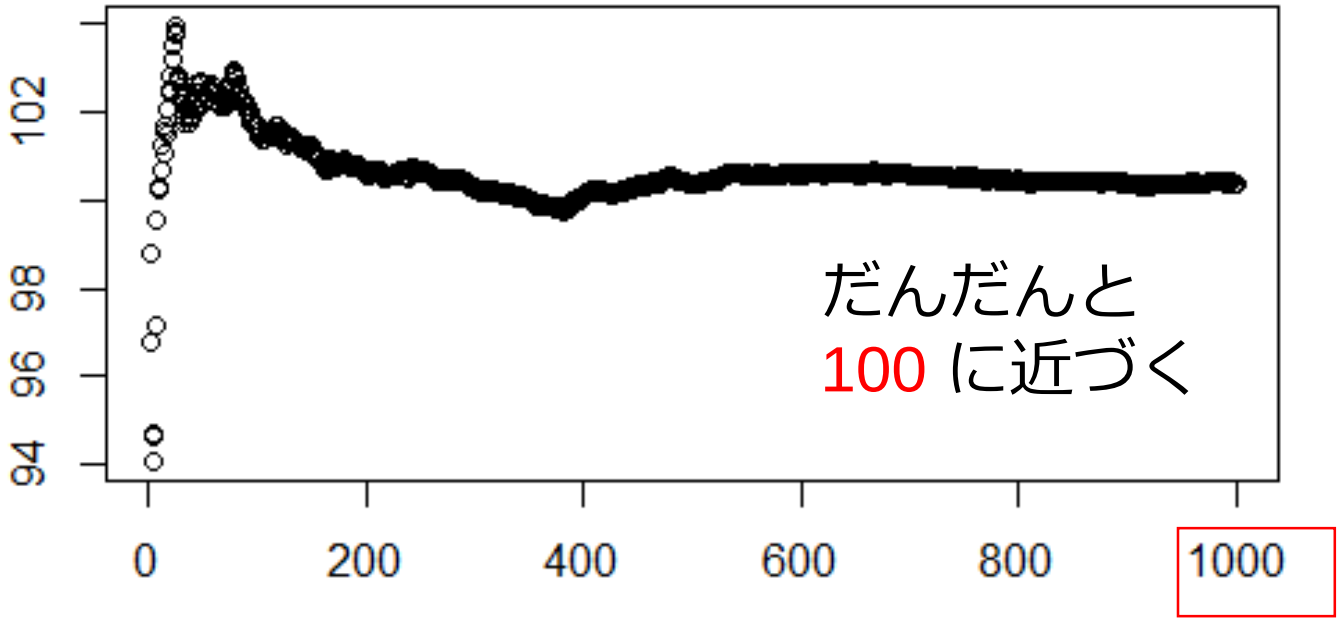
だんだんと
400 に近づく

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

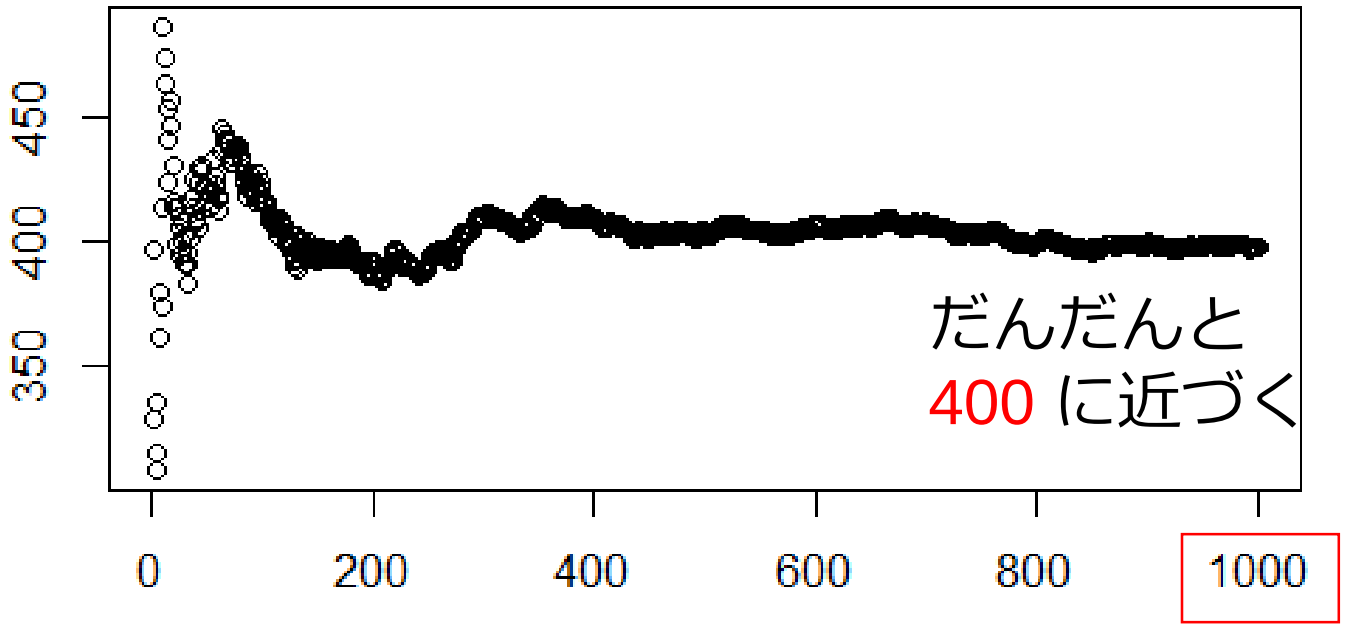
ランダムなので，毎回違う結果が出る

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

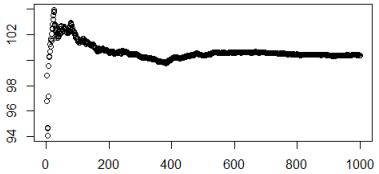


各標本の平均値を集めて**総平均**を求める

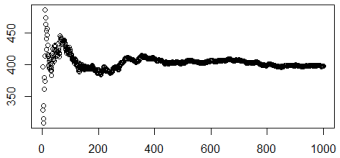
標本の個数を 20 から 1000 の間で変えて，総平均を求めてみる



各標本の不偏分散値を集めて**総平均**を求める



だんだんと
100 に近づく



だんだんと
400 に近づく

変数	標本
母平均 その値は 100 であると**推定**	平均 ※ R では mean
母分散 その値は 400 であると**推定**	不偏分散 ※ R では var

標本や不偏分散を使って，
母平均や母分散を**推定**する

標本はデータなので，
平均は不偏分散は求まる