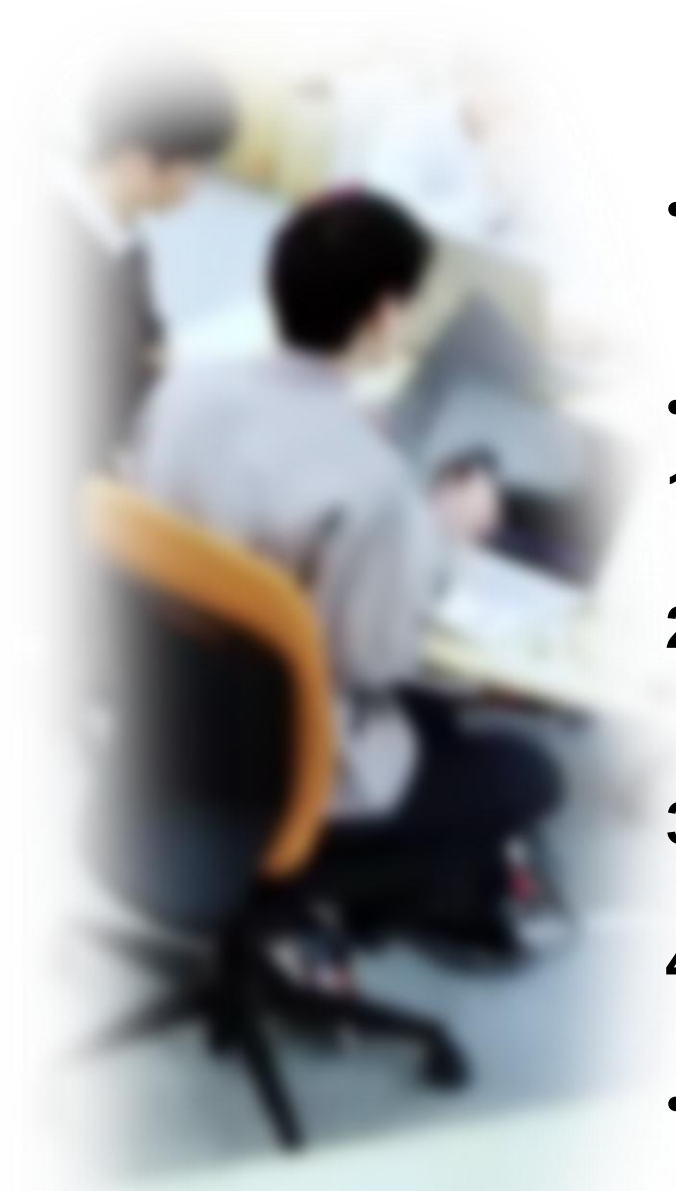
pd-4. 標本の平均、母平均

（Python によるデータサイエンス演習）

URL: https://www.kkaneko.jp/ai/pd/index.html

金子邦彦



- **母集団と標本の関係、標本の平均値・分散値から母平均・母分散を推定する統計的手法の理解とPythonによる実装**
- 【学習内容の構成】

1. **平均**：データの合計をデータの個数で割った値、データ集合の代表値
2. **母集団と標本**：母集団全体を調べることが困難な場合のサンプリングによる一部抽出
3. **標本の平均値**：正規分布を仮定した母平均の推定、標本数nが大きいほど精度向上
4. **標本の分散値**：不偏分散を用いた母分散の推定、t分布による分析

- 前提：Pythonの基本文法、NumPyライブラリの使用方法
- 意義：統計的推定の原理理解、データ分析における標本サイズと推定精度の関係把握

1. 평균

平均

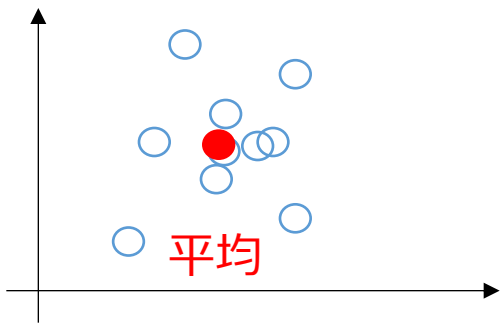
- **平均**は，データの**合計**を，**データの個数**で**割ったもの**

  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:

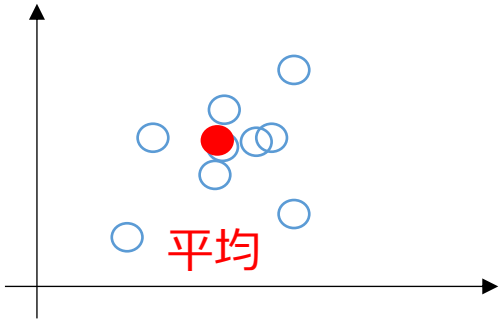
  合計は 120 と 40．4で割って (30, 10)

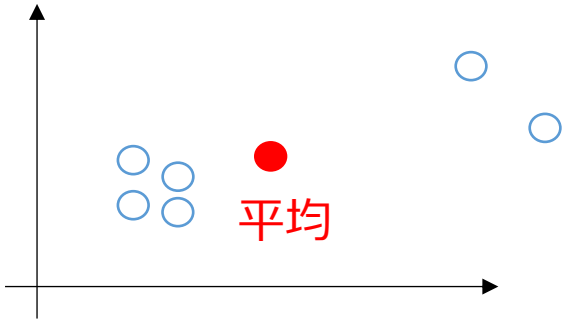


平均は，**データ集合**の**代表**とみることができる場合がある

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

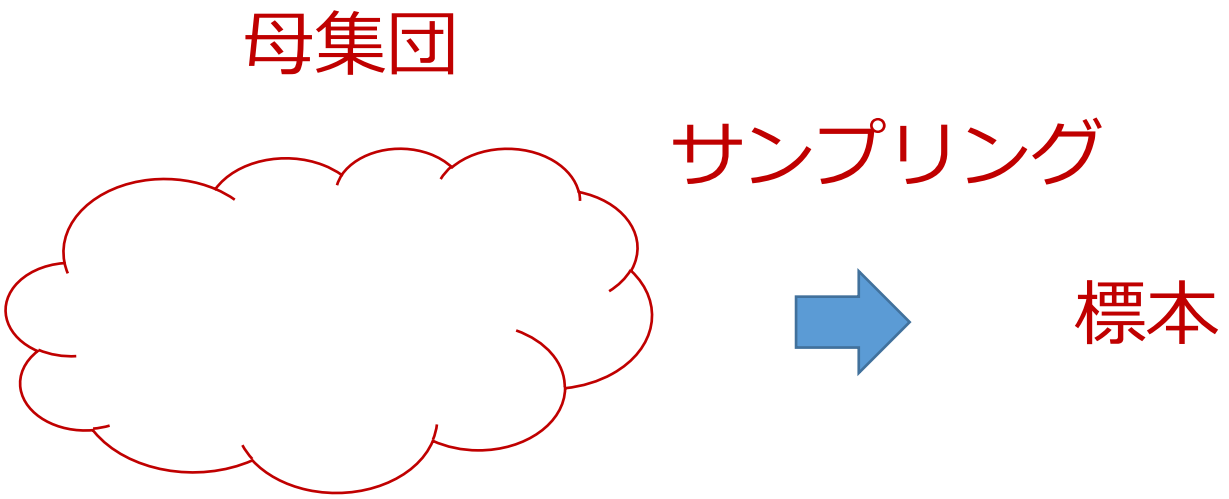
（例）人類全体、２０歳以上の人類全体

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

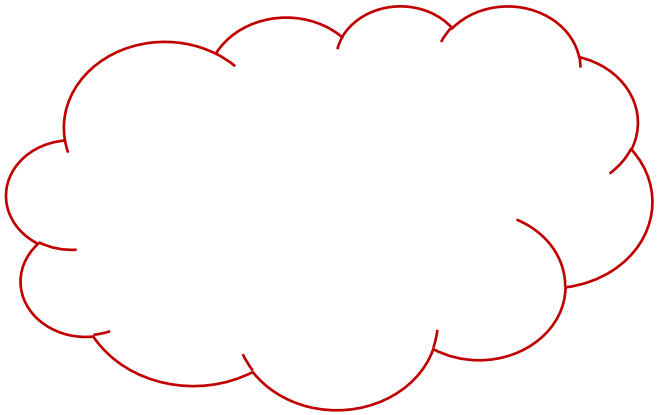
  （例）１０００名をランダムに選ぶ

- **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**。

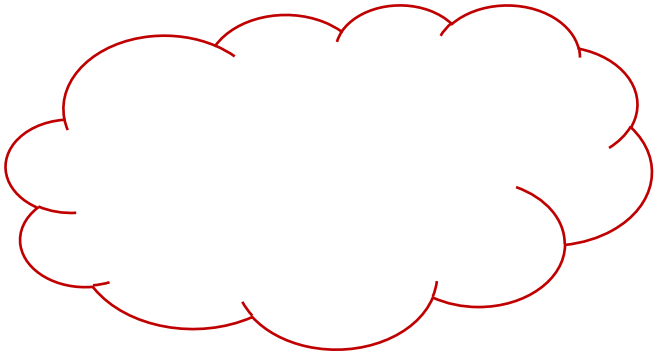
標本から得られたデータを分析し、**母集団全体の性質や傾向を推測**可能。

【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

↓

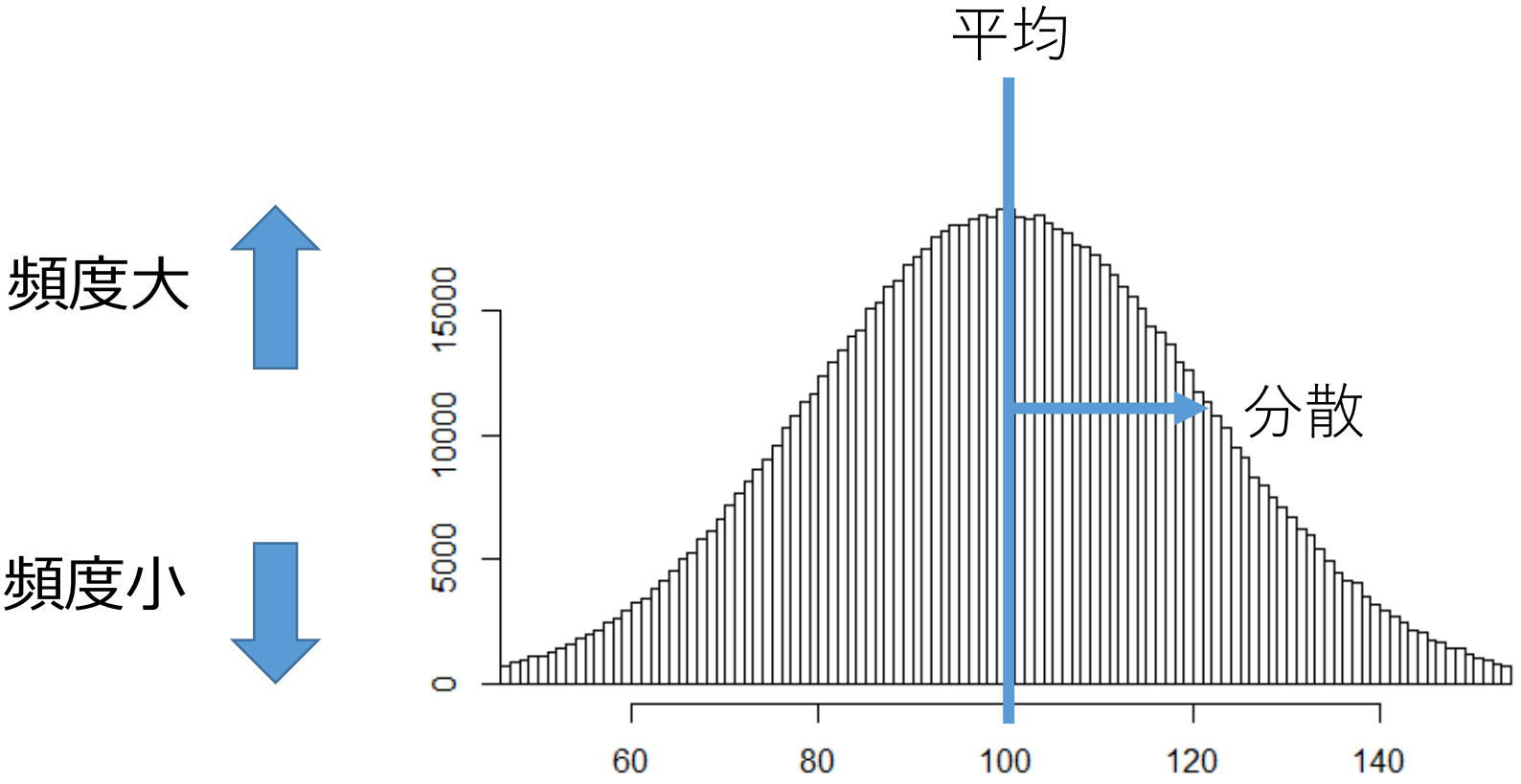
母平均の推定

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。**

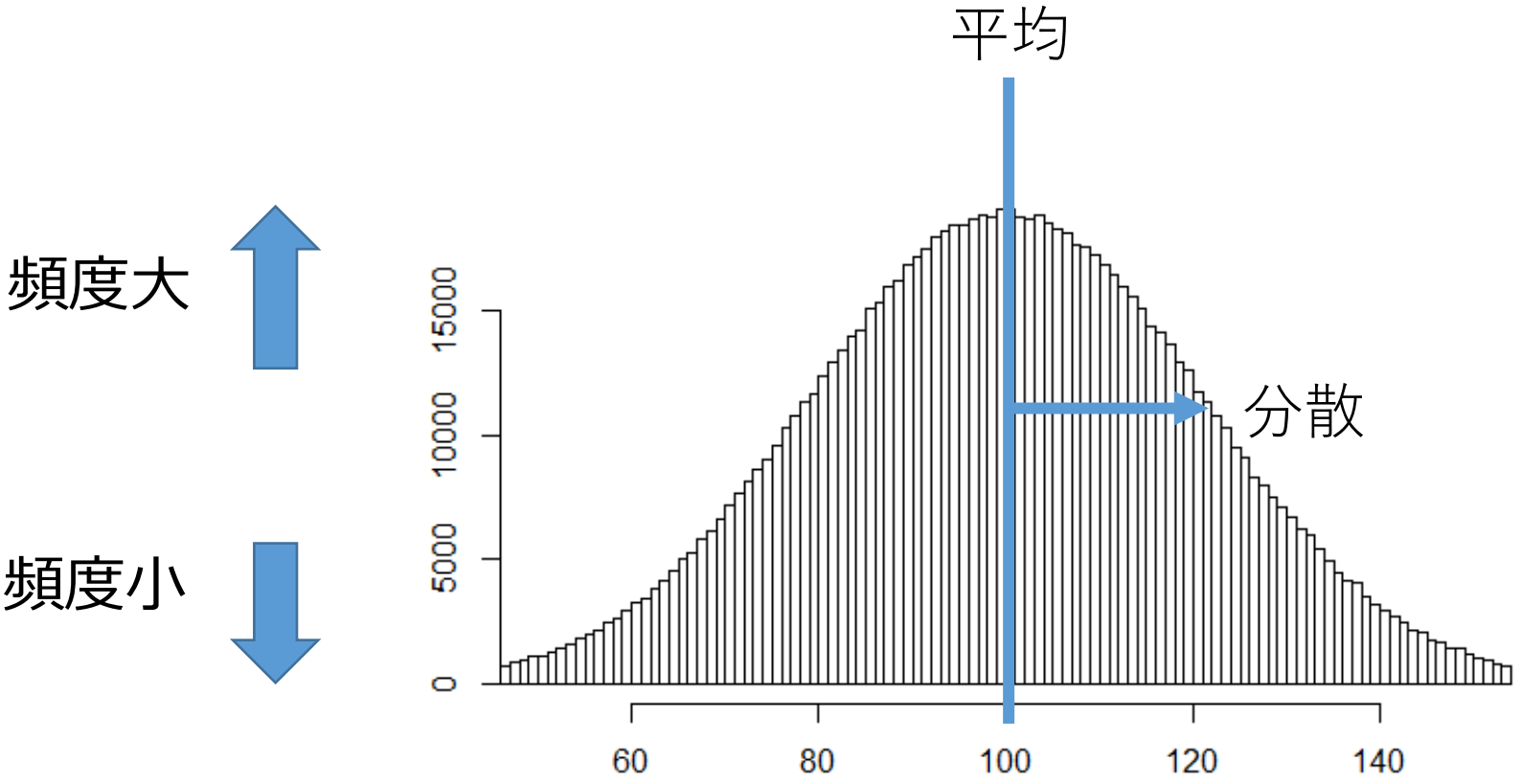
分散は，データの**散らばり度合**を表す



正規分布

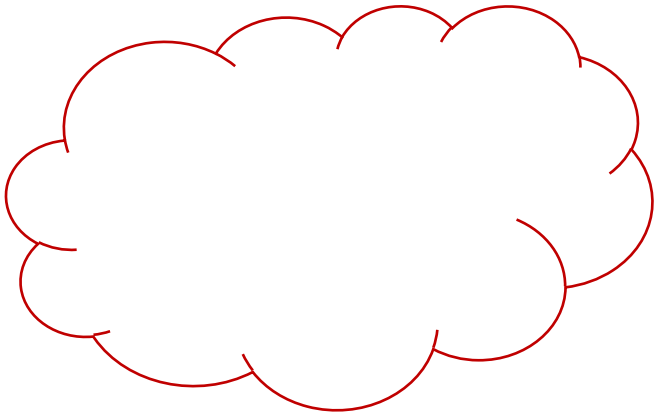
正規分布は，**平均と分散だけで頻度分布を考える。**

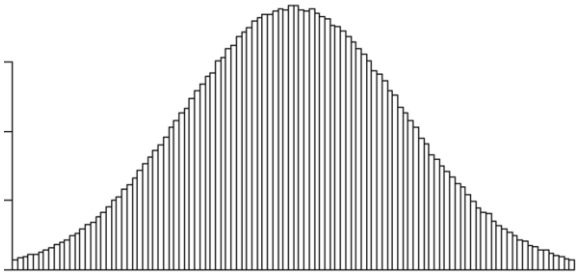
分散は，データの**散らばり度合**を表す



母集団は正規分布であるとし、標本の平均値を算出

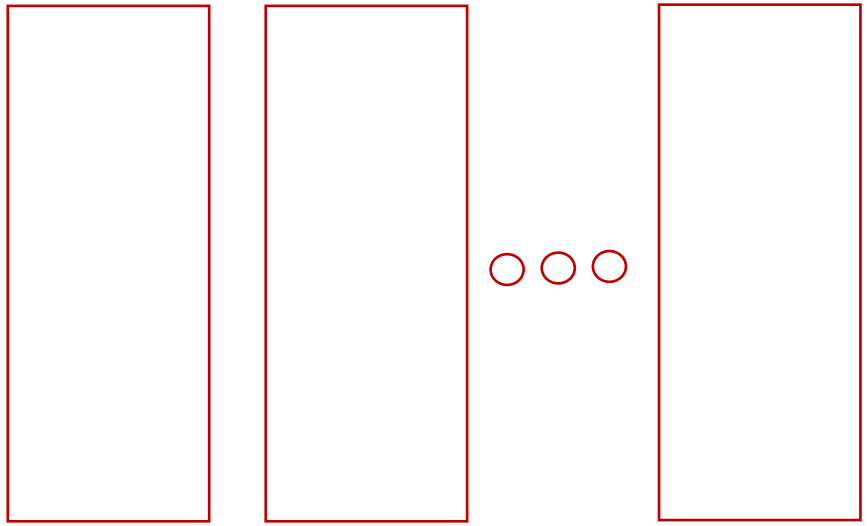
母集団





正規分布

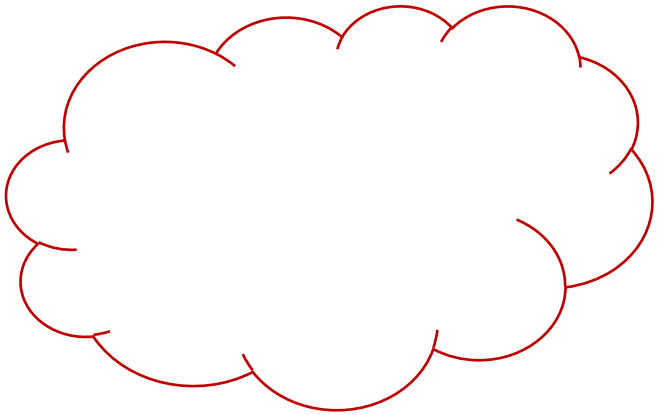
標本（標本数は n）



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

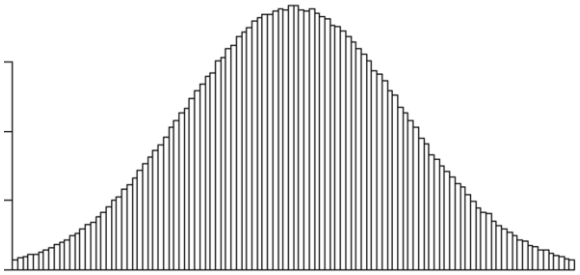
母集団



標本（標本数はｎ，ｎ = 5）

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

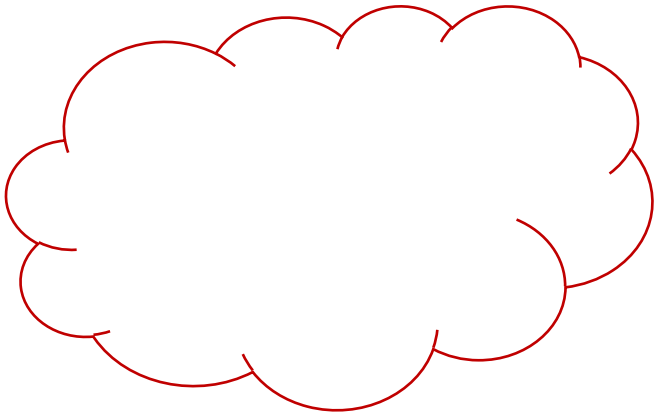
○○○

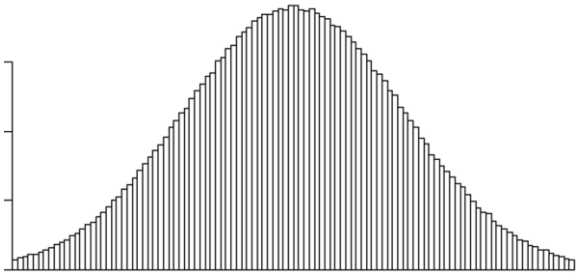


正規分布

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

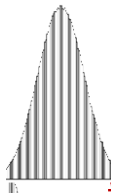
標本（標本数はｎ，ｎ = 5）

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

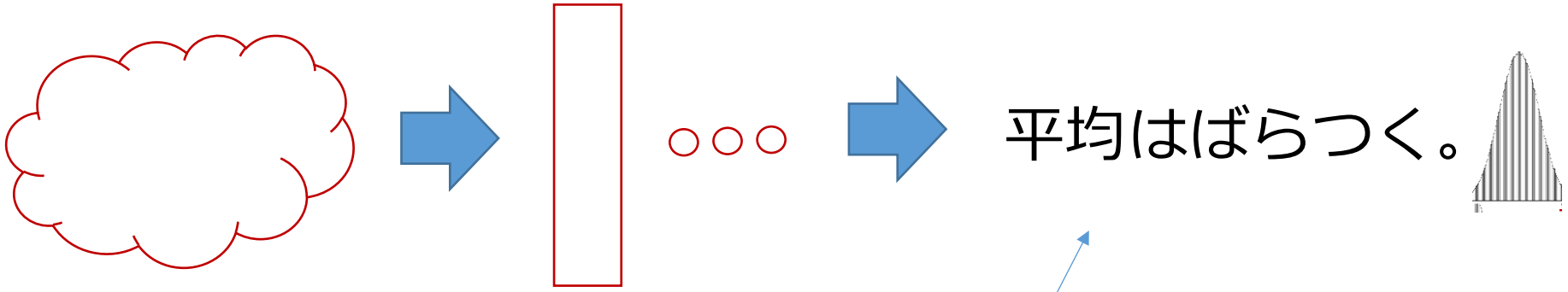
平均はばらつく。

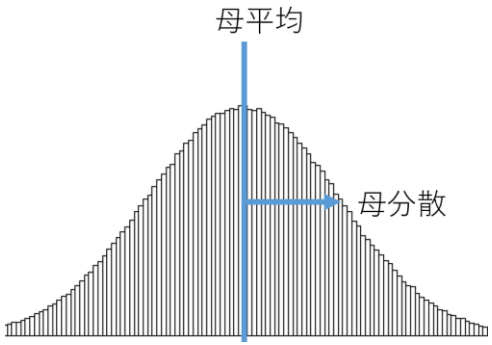


母集団は正規分布であるとし、標本の平均値を算出

母集団

標本（標本数は n ）





正規分布

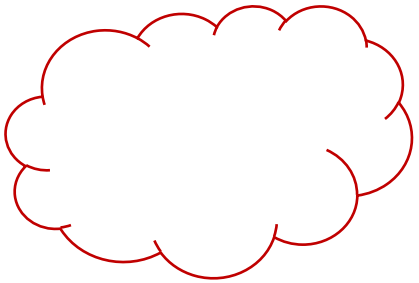
母集団が**正規分布**であるとき、この分布も**正規分布**

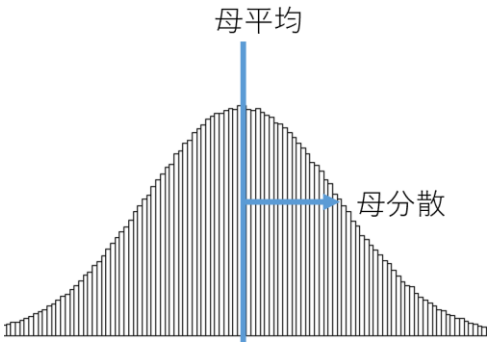
- この**正規分布の平均**
  <母平均>に等しい
- この**正規分布の分散**
  <母分散>／n

- 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

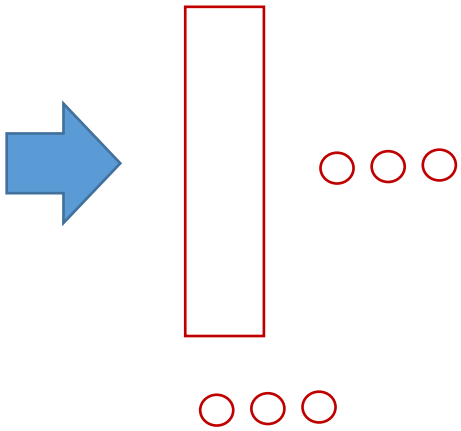
母集団





正規分布

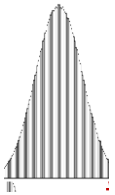
標本（標本数は n ）



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



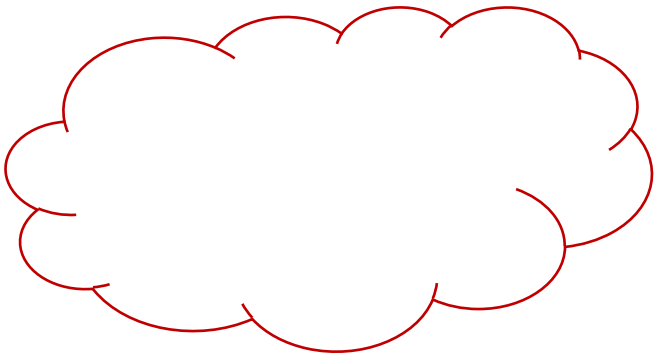
正規分布

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

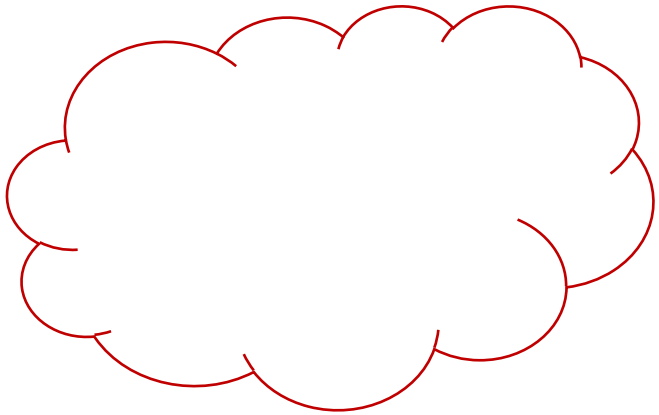
母不偏分散の**推定の精度を分析**する
ために、母集団は t 分布であると仮定
（t 分布は正規分布と少し異なる形）

- **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

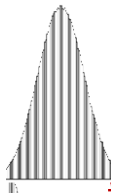
標本（標本数は n ， n = 5）

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

Python による平均，分散，不偏分散

- 平均　　numpy の mean
- 分散，不偏分散　　numpy の var

Python による平均の算出

① ライブラリのインポート

import numpy as np

numpy という数値計算ライブラリをインポート．np という名前で使用できるようにしている

② データの準備

y1 = [10, 40, 30, 40]

リストy1を作成、その中に4つの数値 10, 40, 30, 40 を設定

③ **平均**の算出と表示

print(np.mean(y1))

numpyのmeanを使ってリストy1の平均を算出

平均（Google Colaboratory）

```
import numpy as np

y1 = [10, 40, 30, 40]
y2 = [5, 10, 5, 20]
print(np.mean(y1))
print(np.mean(y2))
```

ライブラリのインポート

データ

平均の算出と表示

コード

```
import numpy as np

y1 = [10, 40, 30, 40]
y2 = [5, 10, 5, 20]
print(np.mean(y1))
print(np.mean(y2))
```

```
30.0
10.0
```

Python による不偏分散の算出

① ライブラリのインポート

import numpy as np

numpy という数値計算ライブラリをインポート．np という名前で使用できるようにしている

② データの準備

y1 = [10, 40, 30, 40]

リストy1を作成、その中に4つの数値 10, 40, 30, 40 を設定

③ **不偏分散**の算出と表示

print(np.var(y1, ddof=1))

numpyのvarを使ってリストy1の平均を算出

不偏分散（Google Colaboratory）

```
import numpy as np

y1 = [10, 40, 30, 40]
y2 = [5, 10, 5, 20]
print(np.var(y1, ddof=1))
print(np.var(y2, ddof=1))
```

ライブラリのインポート

データ

不偏分散の算出と表示

コード

```
import numpy as np

y1 = [10, 40, 30, 40]
y2 = [5, 10, 5, 20]
print(np.var(y1, ddof=1))
print(np.var(y2, ddof=1))
```

```
200.0
50.0
```

正規分布に従うランダムデータ

① ライブラリのインポート

import numpy as np

numpy という数値計算ライブラリをインポート．np という名前で使用できるようにしている

② データの準備

y1 = np.random.normal(100, np.sqrt(20), 1000000)

リストを作成．正規分布に従う．
平均は 100，分散は 20，データの個数は 1000000個

③ **平均**の算出と表示

print(np.mean(y1))

numpyのvarを使ってリストy1の平均を算出

④ **不偏分散**の算出と表示

print(np.var(y1, ddof=1))

numpyのvarを使ってリストy1の平均を算出

正規分布に従うデータの平均と不偏分散（Google Colaboratory）


```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

print(np.mean(y1))
print(np.var(y1, ddof=1))
```

コード

ライブラリのインポート

正規分布に従うランダムデータ

　平均は 100，分散は 20，

　データの個数は 1000000個

平均の算出と表示

不偏分散の算出と表示

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

print(np.mean(y1))
print(np.var(y1, ddof=1))
```

```
99.99953623437364
19.98958992597814
```

標本数が多い場合

不偏分散の値は、

母分散の値に近づく

（大数の法則による）

正規分布に従うデータのヒストグラム（Google Colaboratory）

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

import matplotlib.pyplot as plt
plt.hist(y1)
plt.show()
```

ライブラリのインポート

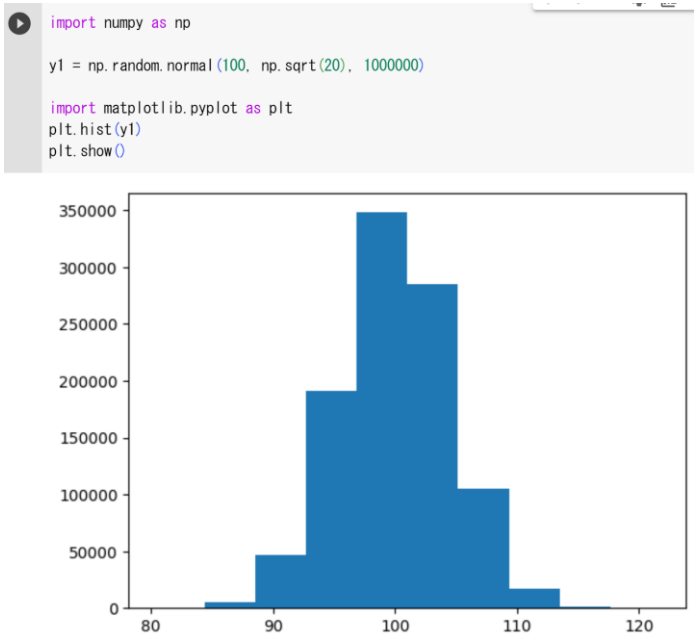
正規分布に従うランダムデータ

　平均は 100,　分散は 20,

　データの個数は 1000000個

ヒストグラム

コード



正規分布に従うランダムデータのサンプリング

① ライブラリのインポート

import numpy as np

numpy という数値計算ライブラリをインポート．np という名前で使用できるようにしている

② データの準備

y1 = np.random.normal(100, np.sqrt(20), 1000000)

リストを作成．正規分布に従う．
平均は 100，分散は 20，データの個数は 1000000個

③ サンプリングを行い 5 個の標本を得る

n = 5
s = np.random.choice(y1, n)
print(s)

numpy の choice を利用して、ランダムに選ぶ

サンプリングを行い標本を得る（Google Colaboratory）

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 5
s = np.random.choice(y1, n)
print(s)
```

ライブラリのインポート

正規分布に従うランダムデータ

　平均は 100，分散は 20，

　データの個数は 1000000個

サンプリングを行い 5 個の標本を得る

コード

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 5
s = np.random.choice(y1, n)
print(s)
```

```
[101.06350089 101.36700607  98.7045995  103.86721812 100.21838381]
```

サンプリングを 10 回繰り返す（Google Colaboratory）

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 5
for i in range(10):
    s = np.random.choice(y1, n)
    print(s)
```

ライブラリのインポート

正規分布に従うランダムデータ

　平均は 100，分散は 20，

　データの個数は 1000000個

サンプリングを行い 5 個の標本を得ることを **10 回繰り返す**

コード

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 5
for i in range(10):
    s = np.random.choice(y1, n)
    print(s)
```

```
[ 99.57333765 100.34407078  96.15176441 100.75205605  98.67265777]
[ 99.24053842 104.28248924  98.42972841 101.83177348  96.18225991]
[104.06978965 109.04532549  95.11935808  98.61405463 108.82277503]
[100.47836989 102.61329203  98.96938859 108.95667046 102.5529762 ]
[103.26935984 103.4297482  105.97976084 103.73104471 105.12278499]
[110.40190064 105.19766063  97.75334877 100.36800352  96.52194538]
[ 95.33534054  95.94443966 105.55636137  98.68895072 101.23234108]
[107.83170478 102.86325636 106.45443893 100.30073682  92.80825688]
[ 99.04004677 109.37320386  93.55374136 103.23749207  93.83266271]
[ 98.25750234 100.82796909  96.4476809   97.33202599  94.74852143]
```

サンプリングを 10 回繰り返し，標本の平均と不偏分散を求める（Google Colaboratory）

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 5
for i in range(10):
    s = np.random.choice(y1, n)
    print(np.mean(s), np.var(s, ddof=1)
```

ライブラリのインポート

正規分布に従うランダムデータ

　平均は 100，分散は 20，

　データの個数は 1000000個

サンプリングを行い 5 個の標本を得ることを **10 回繰り返す．平均と不偏分散を算出**

コード

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 5
for i in range(10):
    s = np.random.choice(y1, n)
    print(np.mean(s), np.var(s, ddof=1))
```

```
99.58453413805664 3.493161376110618
100.46376490954638 15.012769837691028
101.84399180733564 16.830160618354018
99.18766354756495 7.369968918039911
99.06675913470801 28.539880638789406
99.62032310008176 29.422280743584864
99.31891246044485 20.28832363139137
101.15863973704487 15.33457775743577
100.49128029556762 35.09236068784025
98.8665424315428 8.329580662341995
```

サンプリングを 10 回繰り返し，標本の平均と不偏分散を求める
今度は，標本数は 50（Google Colaboratory）


```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 50
for i in range(10):
    s = np.random.choice(y1, n)
    print(np.mean(s), np.var(s, ddof=1))
```

ライブラリのインポート

正規分布に従うランダムデータ

　平均は 100，分散は 20，

　データの個数は 1000000個

サンプリングを行い 50 個の標本を得ることを **10 回繰り返す．平均と不偏分散を算出**

コード

```
import numpy as np

y1 = np.random.normal(100, np.sqrt(20), 1000000)

n = 50
for i in range(10):
    s = np.random.choice(y1, n)
    print(np.mean(s), np.var(s, ddof=1))
```

```
100.58102125384593 23.3539041870206
100.69685306768895 22.723841132919972
99.3468982477124 23.81117222993278
100.45624067433646 16.4795674390004
98.84920984194197 22.410347758483844
99.95754961221583 18.645566912901323
100.51436304868307 19.963596233551662
99.58005570485291 22.03799580718863
100.10091241013656 18.849077329723183
99.92221249329235 24.141975486167212
```

標本数 5 と 50 で結果を比べる

99.58453413805664 3.493161376110618
100.46376490954638 15.012769837691028
101.84399180733564 16.830160618354018
99.18766354756495 7.369968918039911
99.06675913470801 28.539880638789406
99.62032310008176 29.422280743584864
99.31891246044485 20.28832363139137
101.15863973704487 15.33457775743577
100.49128029556762 35.09236068784025
98.8665424315428 8.329580662341995

標本数 5 での平均と不偏分散

100.58102125384593 23.3539041870206
100.69685306768895 22.723841132919972
99.3468982477124 23.81117222993278
100.45624067433646 16.4795674390004
98.84920984194197 22.410347758483844
99.95754961221583 18.645566912901323
100.51436304868307 19.963596233551662
99.58005570485291 22.03799580718863
100.10091241013656 18.849077329723183
99.92221249329235 24.141975486167212

標本数 50 での平均と不偏分散

母集団は、正規分布に従うランダムデータ．平均は 100，分散は 20,

標本数 50 の結果の１つ： 100.58102125384593， 23.3539041870208
精度はどうか
　平均 100.58102125384593 の「100」の部分で誤差がありそう
　分散 23.3539041870208 の「23」の部分で誤差がありそう

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