2019年7月8日更新

9. 統計学I

東北大学 大学院工学研究科
嶋田 慶太 shimada@tohoku.ac.jp

TOHOKU UNIVERSITY

目次

- 平均･分散･期待値
- 二項分布
- ポアソン分布

統計学の役割

･サンプリングした集団の性質について調べる ▸ 記述統計学

･サンプリングをもとに母集団の性質を推定する ▸ 推測統計学

何をしたいのか意識しないと辛い学問かも（個人の感想）

TOHOKU
UNIVERSITY

平均・分散・期待値

統計量と統計学

統計的データがあった場合，

第1データとしては，

平均	中央値	最頻値
Mean	Median	Mode

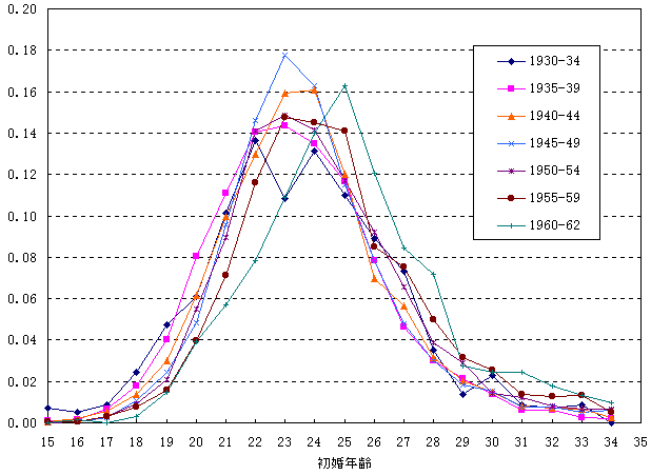
がよく用いられる．
（最大値・最小値も）

次のデータとして，

分散
Dispersion

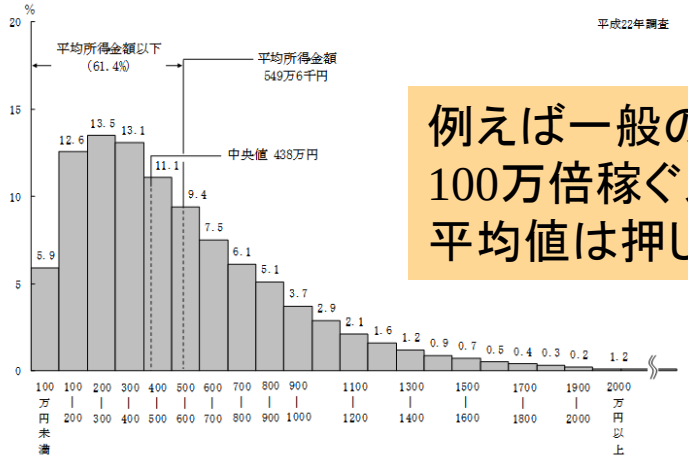
が重視されることが多い．

初婚年齢



http://www.mhlw.go.jp/shingi/0112/s1211-3a.html

平均所得



例えば一般の人の
100万倍稼ぐ人がいると，
平均値は押し上げられる．

http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa10/2-2.html

Octaveの統計関数(１)

平均: mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\{1\}^T\{x_i\}}{n}$$

◂最小二乗法で書いたベクトル表現

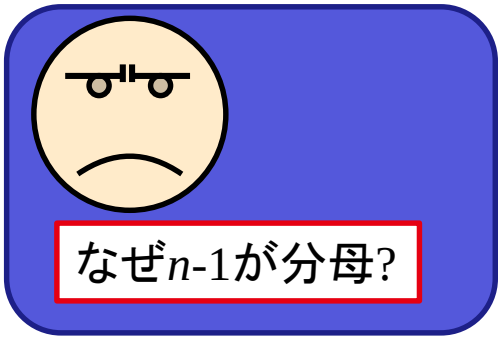
分散: var　不偏分散と呼ばれる

$$V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{\{x_i-\mu\}^T\{x_i-\mu\}}{n-1}$$

標準偏差: std

$$\sigma = \sqrt{V}$$

中央値:median



正規分布と一様分布

```
>> X = randn(10000,1);
>> mean(X)
ans =  0.0034172
>> var(X)
ans =  1.0268
>> X = rand(10000,1);
>> mean(X)
ans =  0.50384
>> var(X)
ans =  0.083720
```

std関数と定義からの検証

```
>> X = randn(10000,1);
>> std(X)
ans =  0.99576
>> sqrt(var(X))
ans =  0.99576
>> median(X)
ans =  -0.0051996
```

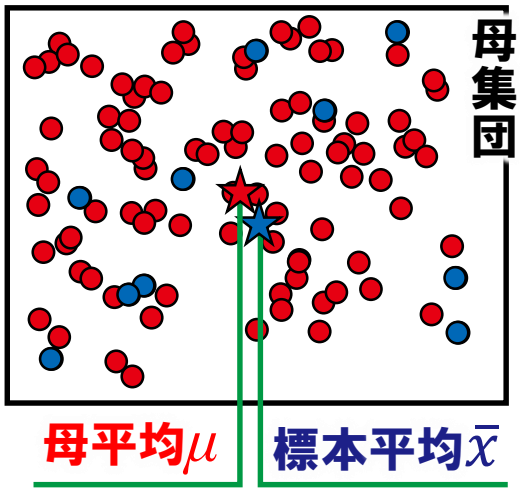
二つの分散

母分散 (Population variance)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

対象とする集合すべての要素の平均から求めれる

▸ 全数調査が容易ならこれで対応するが,
現実には無作為抽出した標本から母集団を推定.



N 個の母集団の要素から n 個の要素を無作為抽出

標本分散 (Sample variance)

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

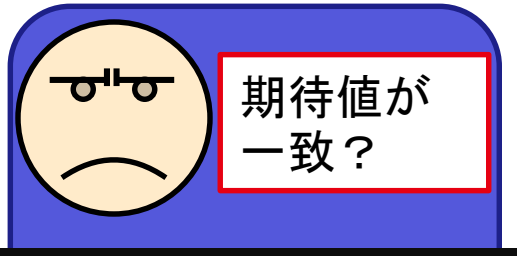
$\bar{x}$: 標本平均

一般に $\bar{x} \neq \mu$ であり,
s^2 は小さく見積もられる
▸ $\sigma^2 > s^2$ となりやすい.

不偏分散 (Unbiased variance)

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

分母をn-1とすることで補正.
▸ 期待値は母分散に一致する



気になる場合は
このスライドの最後に

期待値

期待値 (Expected value)

$$E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i)$$

確率による重み付き平均

例: サイコロを振って（出目 ×1000円）がもらえるゲームをした場合，得する参加金額

$$E[X] = 1000 \times \frac{1}{6} + 2000 \times \frac{1}{6} + 3000 \times \frac{1}{6} + 4000 \times \frac{1}{6} + 5000 \times \frac{1}{6} + 6000 \times \frac{1}{6} = 3500$$

3500円以下の参加金額ならそのうち得する．（賭博罪になるので実際はダメだが．）

モンテカルロ法によるシミュレーション

乱数を用いたシミュレーション法

```
>> X=rand(10000,1);
>> Y=floor(X*6)+1;
>> mean(Y*1000)
ans =   3445.2
```

$$E[X] \approx \frac{1}{N} \sum_{n=1}^{N} X_n$$

`floor`: 床関数
実数xに対して，x以下の最大の整数．
受験ではガウス記号でおなじみ．
仲間に天井関数`ceil` がある．

TOHOKU
UNIVERSITY

二項分布

二項分布（binomial distribution）

例: コインを n 回投げて表が k 回出る確率（ただし，表の出る確率は p とする）

$$P[X=k] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for } k = 0, 1, \ldots, n$$

二項係数

$$ {}_nC_k = \frac{n!}{k!(n-k)!} = \frac{n \times (n-1) \times \ldots \times (n-k+1)}{k \times (k-1) \times \ldots \times 1}$$

▶ `nchoosek(n,k)`

この分布を二項分布とよび，$B(n, p)$ と表現する

期待値: $E[X] = np$

統計関数のインストール＆ロード

次のページから統計関数を使用するのでインストール

```
>> pkg install -forge statistics
```

初回のみ必要.

```
>> pkg load statistics
```

こちらはOctaveを立ち上げ直したらその都度.

Octaveでの二項分布の関数

二項分布の確率密度関数 `binopdf(k,n,p)`

Probability density function

全試行回数 n，1試行当たりの「真」の確率 p，全ての「真」の回数 k

例: 1/8のくじを10回引いて，2回あたりが出る▶ `binopdf(2,10,1/8)`

二項分布の積算分布関数 `binocdf(k,n,p)`

Cumulative distribution function

全試行回数 n，1試行当たりの「真」の確率 p，全ての「真」の回数 0 から k である

例: 1/8のくじを10回引いて，2回以上あたりが出る▶ `1-binocdf(1,10,1/8)`

定義上

`binocdf(k,n,p)` 同値 `sum(binopdf([0:k],n,p))`

モンテカルロ法による二項分布

例: $B(10, 0.4)$ に従う変数X

```
>> X=rand(1,10)<0.4
X =
   0   0   0   1   0   1   1   1   1   0
>> sum(X)
ans = 5
```

真 ▸ 1
偽 ▸ 0

10回の試行で確率0.4の事象が起こる回数
この計算では，

偽	偽	偽	真	偽	真	真	真	真	偽

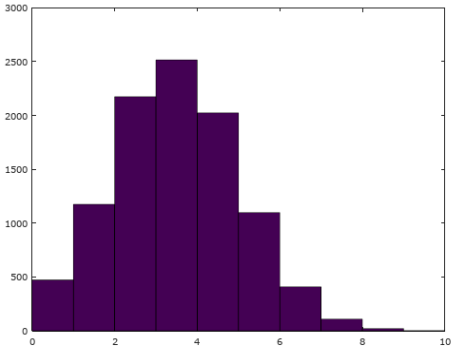
となり，5回起こったことを再現している．

これを踏まえて，
10回のセットを10000回行ったという
モンテカルロ法

例: モンテカルロ法による分布の生成

```
>> Y=sum(rand(10,10000)<0.4);
>> hist(Y,10);
>> mean(Y)
ans =  4.0098
```

列	1	2	…	10000
	偽	偽		偽
	偽	真		真
	偽	偽		真
	真	真		真
	真	偽	…	偽
	真	真		偽
	偽	偽		偽
	真	偽		真
	偽	真		偽
	偽	偽		偽
sum	4	4		4



比較演算の効率的な計算

例えばさいころを1万回投げる実験の模擬として．．．

```
X = randi([1,6],10000,1);
```

整数乱数作成関数　1から6まで　100000×1 行列

```
sum(X == 1:6);
```

列 ベクトル　**行** ベクトル　を一気に比較できる

（行と列が一致すれば1）

X	1:6 1	2	3	4	5	6
1	1	0	0	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
3	0	0	1	0	0	0
2	0	1	0	0	0	0
1	1	0	0	0	1	0
6	0	0	0	0	0	1
⋮						
2	0	1	0	0	0	0
3	0	0	1	0	0	0
5	0	0	0	0	1	0
sumにより合計が出る▸	1661	1674	1673	1661	1669	1677

二項分布の例

条件:
5枚のカードからランダムに1枚取り出し，マークを当てるゲームで，
10回のうち6回正解を出した場合，自分は超能力者だといえるか？

ゼナー・カード
（Zener cards）

考え方:
一般人であれば1回の試行でマークを当てる確率は1/5 すなわち0.2である.

計算は二項分布$B(10,p)$であるので
当てる回数を0～10回まで列挙すると▸

```
>> [[0:10]' binopdf([0:10]',10,0.2)]
ans =

    0.00000    0.10737
    1.00000    0.26844
    2.00000    0.30199
    3.00000    0.20133
    4.00000    0.08808
    5.00000    0.02642
    6.00000    0.00551
    7.00000    0.00079
    8.00000    0.00007
    9.00000    0.00000
   10.00000    0.00000
```

かなり珍しい事態である.

あなたは
超能力者かもしれないね！

TOHOKU UNIVERSITY

ポアソン分布

ポアソン分布（Poisson distribution）

例: 所定の時間 τ に平均 λ 回発生する事象が τ 内に k 回その事象が起こる確率

期待値: λ

二項分布との違い ▸ 連続時間なので，明確な試行回数 n が分からない．

▶ 見えざる手による無限回のくじびきをイメージ

時間 τ 中に n 回くじを引く

当たりの出る確率 $p = \dfrac{\lambda}{n}$

回数 n を大きくした分，

確率 p を小さくして，

期待値 λ を一定に保つ

p = 1/2 のくじを2回引けば，1回は当たると期待される．

p = 1/3 のくじを3回引けば，1回は当たると期待される．

p = 1/100 のくじを100回引けば，1回は当たると期待される．

p = 1/1000000000 のくじを1000000000回引けば，1回は当たると期待される．

このイメージで二項分布の極限を考える

二項分布の極限としてのポアソン分布

式の変形

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \frac{n!}{k!\,(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

◂ pに代入 $p=\frac{\lambda}{n}$

$$= \frac{\lambda^k}{k!}\frac{n\times(n-1)\times\ldots\times(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}$$

$$\xrightarrow[n\to\infty]{} \frac{\lambda^k}{k!}e^{-\lambda}$$

時の流れの中に手を突っ込み，コンスタントにくじを引き続ける．　時間

もっと区切る

箱の大きさが1回の試行の期待値 ▸ 1回当たりの期待値は減ってもその総和は同じ

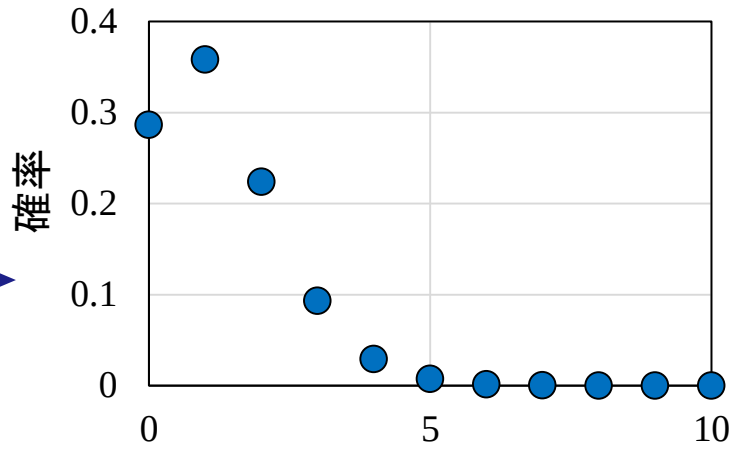
ポアソン分布の実例

例: 1時間に平均5通のemailを受ける人が次の15分で受け取るメール数

$\lambda = 5/(60/15)$

$= 1.25$（15分だと平均1.25通）

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$



k: 15分での受信件数

期待値: $E[X] = \lambda$

ポアソン分布に従う乱数を使うシミュレーション

λ

ポアソン分布に従う乱数　randp(l,m,n)

$m \times n$行列
1個省略すると
正方行列

例: 1時間に平均5通のemailを受ける人が次の15分で受け取るメール数

```
>> randp(5/60*15,1,10)
ans =
    4     1     1     0     2     1     1     1     2     3
```

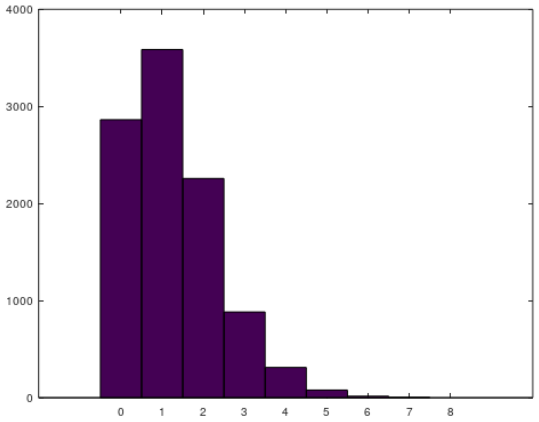
1行10列

ある時4通　ある時1通　1通　0通　2通　1通　1通　1通　2通　3通

というのを模擬している

```
>> hist(randp(5/60*15,1,100000),0:8)
```

1行100000列



Octaveでのポアソン分布の関数

ポアソン分布の確率密度関数 `poisspdf(k,l)`

Probability density function

平均 l 回の現象が k 回起こる確率

例: emailの1日平均受信数 20件で，2日で45件の確率 ▸ `poisspdf(45,40)`

2日なら平均40通

ポアソン分布の積算分布関数 `poisscdf(k,l)`

Cumulative distribution function

平均 l 回の現象が 0 から k 回起こる確率

例: emailが1日平均受信数 20件で，3日で50件以下の確率 ▸ `poisscdf(50,60)`

定義上

`poisscdf(k,l)` 同値 `sum(poisspdf([0:k],l))`

二項分布とポアソン分布のまとめ

	試行	発生	1回の試行による発生の確率	期待値
二項分布	離散的（数えられる）	離散的（数えられる）	p	np
ポアソン分布	連続的（数えられない）	離散的（数えられる）	「1回」を定義できない	λ

確率の小さな事象 ▸ ポアソン分布で近似可能.

確率と期待値を混同しないように！

混同の例:「あたりの確率が1/256ということは256回引けば1回は当たる，ってことだよね？」

▶ 当たりません．むしろ37%くらいまったく当たらないことがあり得ます．

ポアソン過程と指数分布

λ: 単位時間当たりの平均 と取る
- ある基準時刻0 から t までの回数の期待値は λt となり，式は，

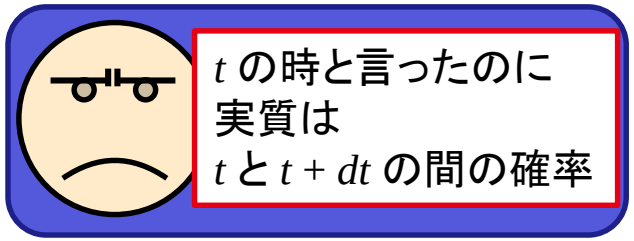
$$P(N_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

ポアソン分布の λ を λt に置き換えるだけ．
これがポアソン過程．

待ち時間に注目した場合：
ポアソン過程に従うような事象が1回発生したのち，次の1回が t 後に起こる確率 f について

$[0, t]$では発生せず，$[t, t + \mathrm{d}t]$に1回以上発生する確率を考えればよいので，

$$f(t)\mathrm{d}t = \underbrace{\left(1 - e^{-\lambda(t+\mathrm{d}t)}\right)}_{[0,\, t+\mathrm{d}t]\text{に1回以上発生する確率}} - \underbrace{\left(1 - e^{-\lambda t}\right)}_{[0,\, t]\text{に1回以上発生する確率}}$$



$$f(t) = -\frac{\left(e^{-\lambda(t+\mathrm{d}t)} - e^{-\lambda t}\right)}{dt}$$

指数関数の微分の定義

$$f(t) = \lambda e^{-\lambda t}$$

指数分布

確率密度関数と確率質量関数

二項分布やポアソン分布：発生するイベントの回数が数えられる

1回起こる確率，2回起こる確率が定義できる.

確率質量関数

指数分布：発生するイベント回数ではなくタイミングを表しており，「数える」ものではない

たとえば「1秒」ぴったりの確率は定義できない
（したとしても微小時間なので限りなく0に近い）

幅を伴って積分によって具体的な確率を考える.

確率密度関数

なので，確率密度関数は点の値が1を超えることがあり得る.

TOHOKU
UNIVERSITY

課題

Exercise 9.1

あるコンビニではお昼の12時～13時に平均100人の来店がある.
ある10分間に来店者数がX人以下となる確率を
モンテカルロ法と解析的な手法の両方で求めよ.

モンテカルロ法: 乱数を使う手法

解析的手法: 数式から求まる手法

ここでXを回答者の学籍番号4ケタの各桁の合計とする.

つまり，学籍番号B〇TB1357　の場合，

X=1+3+5+7=16　として計算せよ.

Exercise 9.2

あたりの確率が1/1024の電子くじ（ガチャ）をZ回引いた場合，
あたりが計0回，計1回，…，計10回である確率を
二項分布の理論的解とモンテカルロ法での計算の求め，
双方をグラフで示せ．

ここでZを回答者の学籍番号とする．

モンテカルロ法のヒント：
あたりが 1/1024 のくじをZ回引くことを 1セット として行or列を作り，
それを列or行方向に重ねることで 複数セット 行うことを模擬することで分布を作る．

Exercise 9.3

あたりの確率が1/1024の電子くじ（ガチャ）をZ回引いた場合，
あたりが計0回，計1回，…，計10回である確率を
ポアソン分布で近似した場合の理論的解と
ポアソン乱数を用いたモンテカルロ法の計算の求め，
双方をグラフで示せ．

ここでZを回答者の学籍番号とする．

ヒント：
「くじをZ回引くこと」を「1セット」とした場合に，
その1セット内に何回あたりがあるか近似分布を示すのがポアソン分布．
その1セット内のあたりの回数を模擬するのがポアソン乱数．

TOHOKU
UNIVERSITY

Appendix

標本分散の期待値(1)

母集団（要素数N）から要素数nの標本を抜き出す

- 標本の選び方の数は下の式

$$\binom{N}{n} = {}_N C_n = \frac{N!}{(N-n)!\,n!} = M$$ とりあえずMと置く.

以下，母集団の要素を意識する場合は$\{x_i\}$と表記し，

ある標本jの要素であることを意識する場合は，$\{x_{jk}\}$と表記する.

（母集団の要素に$1, 2, \ldots N$と番号を振り，
グループjに属する要素にも別途$1, 2, \ldots, n$と番号を振る.
当然，$\{x_{jk}\} \subset \{x_i\}$であり，$\{x_{jk}\} \cap \{x_{j'k}\}$が$\emptyset$でない場合がある.）

標本分散の期待値(2)

$$\overline{s^2} = \frac{1}{M}\sum_{j=1}^{M} s_j^2$$

標本分散の期待値を式化
（M個あるグループの標本分散を全部して平均）

$$= \frac{1}{M}\sum_{j=1}^{M}\left\{\frac{1}{n}\sum_{k=1}^{n}(x_{jk}-\overline{x_j})^2\right\}$$

定義式

定義

$$\overline{x_j} = \frac{1}{n}\sum_{k=1}^{n} x_{jk}$$

$$= \frac{1}{M}\sum_{j=1}^{M}\left\{\frac{1}{n}\sum_{k=1}^{n}(x_{jk})^2-\overline{x_j}^2\right\}$$

公式

ある要素 x_i が含まれるグループ数を考えると，
$(N-1)$から$(n-1)$を取り出す組合せであるから

$$\binom{N-1}{n-1} = \frac{n}{N}M$$

すべての要素にとって同様なので，

$$\sum_{j=1}^{M}\left\{\sum_{k=1}^{n}(x_{jk})^2\right\} = \frac{n}{N}M\sum_{i=1}^{N}(x_i)^2$$

となり

標本分散の期待値(3)

$$\bar{s}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i)^2 - \frac{1}{M}\sum_{j=1}^{M}\bar{x}_j^2$$ となる.

再掲

$$\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{jk}$$

ある要素 x_α と x_β がともに含まれるグループ数を考えると,
$(N-2)$ から $(n-2)$ を取り出す組合せであるから

$$\frac{1}{M}\sum_{j=1}^{M}\bar{x}_j^2 = \frac{1}{M}\frac{1}{n^2}\left\{\frac{n}{N}M\sum_{i=1}^{N}(x_i)^2 + \frac{n}{N}\frac{n-1}{N-1}M\sum_{\substack{\alpha=[1,N]\\ \beta=[1,N]\\ \alpha\neq\beta}}(x_\alpha x_\beta)\right\}$$

$$N^2\mu^2 = \left\{\sum_{\substack{\alpha=[1,N]\\ \beta=[1,N]}}(x_\alpha x_\beta)\right\} = \left\{\sum_{i=1}^{N}(x_i)^2 + \sum_{\substack{\alpha=[1,N]\\ \beta=[1,N]\\ \alpha\neq\beta}}(x_\alpha x_\beta)\right\}$$ を用いて変形すると

標本分散の期待値(4)

$$\frac{1}{M}\sum_{j=1}^{M}\bar{x}_j^2 = \frac{1}{nN}\frac{N-n}{N-1}\sum_{i=1}^{N}(x_i)^2 + \frac{N}{n}\frac{n-1}{N-1}\mu^2$$

が得られる.
これを代入して,

$$\bar{s}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i)^2 - \frac{1}{M}\sum_{j=1}^{M}\bar{x}_j^2$$

$$= \frac{n-1}{n}\frac{N}{N-1}\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i)^2 - \mu^2\right\}$$

$$= \frac{n-1}{n}\frac{N}{N-1}\sigma^2$$

N は 自然現象であれば非常に巨大な数であるし,
通常非常に大きな数であるので約分できる.

結局, 分母の n は標本分散を求める際に用いたものがそのまま出てきているだけなので,
これを $(n-1)$ に置き換えたほうが母分散に近づける. ということで不偏分散が使われる.