なぜ不偏分散は N-1 で割るのか *

Kosugitti の研究ノート

2019 年 8 月 21 日

1　母集団と標本

1.1　サンプリング

推測統計学の基本は，全体の情報を知るために，全ての要素をチェックしていくのは大変だから，少ないサンプルをもとに全体的特徴を推し量ろう，という考え方である。選挙でどこの政党がもっとも票を集めるのか，を知るために，日本国民全員に聞いていくのでは選挙を一度やる苦労と変わりがない。そこで小数の電話調査などで，だいたいの当たりをつけるのである。

このとき，全体のことを**母集団**といい，(すでに出てきたが) そこから集められる少数のデータのことを**サンプル**あるいは**標本**という。

問題は，サンプルが母集団の特徴をきちんと反映しているかどうか，である。母集団からサンプルを集める方法を**サンプリング**という。普通，サンプリングはランダム・サンプリング (無作為抽出) がなされる。ランダムとは，無作為，つまり調査者の意図が入っていないということである。調査者が自分に都合の良いようなサンプルを集めたら (例えば自民党員ばかりに支持政党調査を行ったら)，母集団をうまく反映しない結果が出るのは容易に想像できるだろう。

サンプリングの方法については，様々なものがあるので他書に譲る。

1.2　なぜサンプルで正しいことがわかるか

1.2.1　経験的な説明

さて，では母集団からサンプルを取り出すと，何が分かるのだろうか。例を挙げてみてみよう。

$N(50, 10)$ の正規分布に従う乱数を用いて，10,000 個のデータを作ってみる。10,000 人の学生さんによるテストの点数だとでも思ってくれればよい。このようなシミュレーションは，R の簡単なコードで実行することができる。

```
> library(tidyverse)
> set.seed(20190820)
> # サイズを決定
> N <- 10000
> # 正規乱数の発生
```

* written on 2008.02.06 / revised on 2015.02.26 / Last updated on 2019.08.20

```
> x <- round(rnorm(N, 50, 10)) %>% matrix(ncol=10)
> # 平均値の計算
> mean(x)
[1] 49.8459
```

10,000 サンプルの平均は 49.8459 である。理論的には 50 であってほしいのだが，今回の標本ではこの数字になった。まあほぼ 50 なので，十分近似していると考えよう。そして，以降はこの 10,000 の数字が母集団だと考えることにする。母平均は 49.8459 であり，母分散は 99.88015 である[*1]。これもその平方根を取ると 9.99 なので，ほぼ 10 と理論値に十分近似していると言えるだろう。

```
> sampleVar <-function(x){
+     mean((x - mean(x))^2)
+ }
> sampleVar(x)
[1] 99.88015
```

さてこの 100 人のデータから，10 人分ずつサンプルを取ったとしよう。今回は生成した乱数を 10 列の行列サイズにしてあるので (`%>% matrix(ncol=10` の箇所)，各列が母集団からのサンプリングされた標本であると考えて欲しい。

この平均値，すなわち各サンプルの平均なので標本平均だが，これを算出し，その最大値と最小値を計算してみた。最大値は 60.0，最小値は 38.5 である。母平均 49.8459 の集団から，10 人無作為にとってきたとしても，その平均値は 60 にもなり得るし 38 にもなり得るのだ。これは結構なブレかたである。

```
> apply(x,1,mean) %>% as.data.frame %>% summarise(max=max(.),min=min(.))
  max  min
1  60 38.5
```

サンプルを繰り返せば，母集団の情報に近づいていく－これがサンプリングをするメリットである。特に，母集団がやたらと大きい場合は，母集団の悉皆調査が実際的に不可能であったとしても，サンプルを数回取り出すことぐらいなら，何とか実現可能な範囲内にあるはずだからである。

1.2.2 理論的な説明

上の例を，理論的にきちんとフォローしておこう。

サンプルを取ってくると，その平均値 (標本平均) は簡単に求められる。すなわち，

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \tag{1}$$

同様に，サンプルの分散 (標本分散) も次式で得られる。

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

[*1] R の var 関数は $N-1$ で割る不偏分散を計算するので，今回は N で割る標本分散を計算する関数を自作して算出した

これは以下のように変形できる。通常の計算はこちらの方が楽だろう。

$$\begin{aligned} s^2 &= \frac{1}{n}\{(x_i^2 - 2x_1\bar{x} + \bar{x}^2) + \cdots + (x_n^2 - 2x_n\bar{x} + \bar{x}^2)\} \\ &= \frac{1}{n}\{(x_1^2 + x_2^2 + \cdots + x_n^2) - 2\bar{x}(x_1 + x_2 + \cdots + x_n) + n\bar{x}^2\} \\ &= \frac{1}{n}\{\sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\ &= \frac{1}{n}\sum x_i^2 - \bar{x}^2 \end{aligned} \tag{3}$$

余談だが，共分散も同様に

$$s_{xy} = \frac{1}{N}\sum x_i y_i - \bar{x}\bar{y} \tag{4}$$

で求められる。

さて，推測統計においては，サンプル x は母集団分布に従う確率変数 X が x という形に具現化したもの，と考える。ここで，標本の大きさが n の時，その期待値は

$$E[\bar{X}] = E[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)] = \frac{1}{n}\{E[X_1] + E[X_2] + \cdots + E[X_n]\}$$

であるから，

$$E[\bar{X}] = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu \tag{5}$$

である。これは，標本平均の期待値が母平均に等しいことを意味しており，サンプルを何度も取り，その平均値の平均値は母平均に等しくなることが理論的に示される。

ところで，後々必要になってくるから，ここで標本を繰り返したときの，平均値の散らばりについて考えておく。標本平均 $\bar{X}$ の散らばりの期待値だから，

$$E[(\bar{X} - \mu)^2] = E\left[\left\{\frac{1}{n}(X_1 + X_2 + \cdots + X_n - n\mu)\right\}^2\right]$$

$$= E\left[\left\{\frac{1}{n}\{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)\}\right\}^2\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} E[(X_i - \mu)^2]$$

$$= \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n\text{ 個}})$$

となり，

$$= \frac{1}{n}\sigma^2 \tag{6}$$

であることがわかる。

では次に標本分散を見てみよう。今回の母分散は 99.88015 なのであった。先ほどと同じように，各列を n=10 のサンプルだと考えて計算してみる。その上で，それの平均を出してみると，89.82037 となった。今回は、流石に近似しているとは言えないレベルのようだ。ただのサンプリングミスだろうか？

```
> apply(x,1,sampleVar) %>% mean
[1] 89.82037
```

では元の式に戻って，考え直してみよう。標本の大きさが n のとき，標本分散 S^2 の期待値を母分散で表したい。

$$S^2 = \frac{1}{n}\{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\}$$

で，$\sigma^2 = E[(X-\mu)^2]$ である。なんとかして $X - \mu$ を式の中に入れたいので，

$$S^2 = \frac{1}{n} = \{(X_1 - \mu - \bar{X} + \mu)^2 + \cdots + (X_n - \mu - \bar{X} + \mu)^2\}$$

と置こう。これを展開すると，

$$= \frac{1}{n}\{\{(X_1 - \mu) - (\bar{X} - \mu)\}^2 + \cdots + \{(X_n - \mu) - (\bar{X} - \mu)\}^2\}$$

$$= \frac{1}{n}\{(X_1 - \mu)^2 - 2(X_1 - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 + \cdots + (X_n - \mu)^2 - 2(X_n - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2$$

ここで第二項は，

$$-2\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)(\bar{X} - \mu) = -2(\bar{X} - \mu)\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)$$

$$= -2(\bar{X} - \mu)(\bar{X} - \mu)$$

$$= -2(\bar{X} - \mu)^2$$

だから，元の式は

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)^2 + (\bar{X} - \mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2$$

となる。これの期待値，$E[S^2]$ は

$$E[S^2] = \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2]$$

$$= \sigma^2 - E[(\bar{X} - \mu)^2]$$

である。これの第二項は，標本平均の分散であり，σ^2/n で得られるのだったから (式 6)，$E[S^2]$ は，

$$E[S^2] = \sigma^2 - \frac{1}{n}\sigma^2$$

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

となる。さて，標本分散の期待値が $\frac{n-1}{n}\sigma^2$ であるから，$\frac{1}{n}\sum(X_i - \bar{X})^2$ で求めた標本分散とずれていることがわかるだろう。標本分散は

$$\frac{1}{n}\sum(X_i - \bar{X})^2 \times \frac{n}{n-1} = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

で求めるべきであり，このようにして求める分散のことを特に**不偏分散**という。

分散だけ $n-1$ で割るのは，どうも不公平な感じがする，という方がいるかもしれない。試しに，実際のデータで見てみよう。先ほどの数値例では，母分散が 99.88015 であるのに，標本分散の平均は 89.82037 なのであった。では，不偏分散の平均値を取ってみよう。

```
> apply(x,1,var) %>% mean
[1] 99.80041
```

このように，サンプルサイズ 10 の各標本から不偏分散を計算し，その平均を出すと 99.80041 となった。これならば母分散の近似値として十分だろう。

このことは，サンプル数を徐々に増やしていけばもっとわかりやすい。図 1.2.2 は，サンプル数を増やしていったときの標本分散と不偏分散の平均値の推移である[*2]。

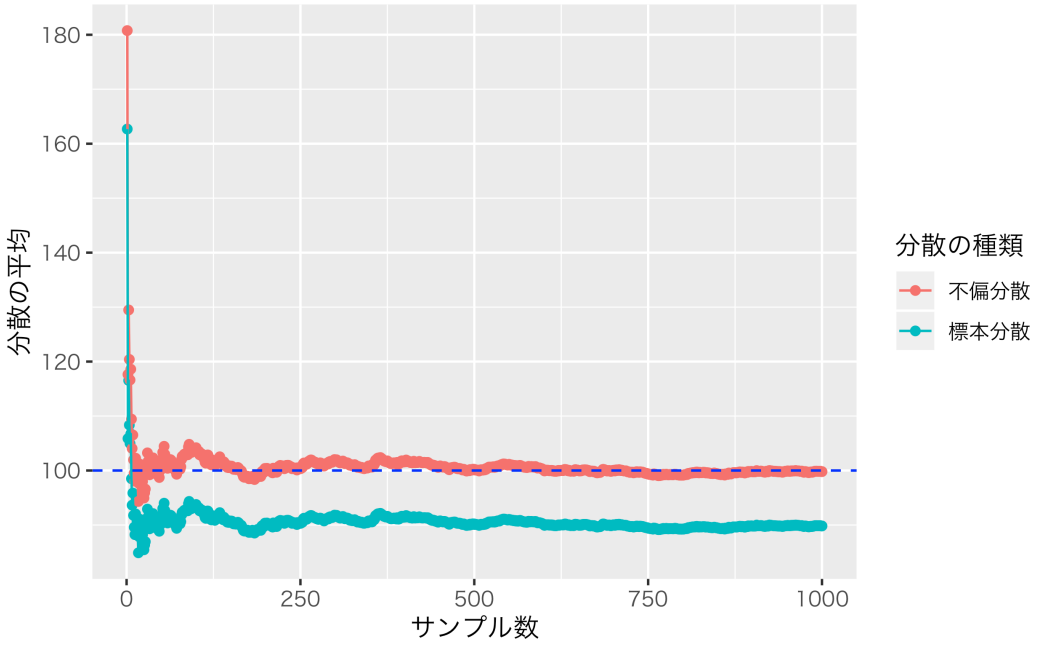

図 1　標本分散と不偏分散の期待値

図 1.2.2 にあるように，標本分散は母分散 100 から外れたところを推移し，決して母分散に近づくことはない。最終的に，標本分散の期待値は 89.82，不偏分散の期待値は 99.80 と，後者がより母分散に近いことは明らかだ。

以上のことから，頻度主義的な分析を行う際の分散の推定値としては，以後 $\frac{1}{n-1}$ で除したものを用いる。

[*2] この図を書くコードは付録を参照のこと

1.3 必要なサンプル数の求め方

1.4 推定のひみつ

さて，サンプリングからいくつかの基本的なカラクリが見えてくる。

まず重要なのは，サンプルの平均値 $\bar{x}$ の期待値は，母集団の平均値 $\bar{X}$ と一致するということ (式 5)。繰り返しになるが，何度もサンプルを繰り返して，サンプル平均の平均をとると，母平均になるということ。

次に，サンプルの平均値の分散は，母分散を標本の大きさで割ったものに等しくなるということ (式 6)。この第二の特徴からは，もうひとつ重要なことがわかる。サンプル平均の標準偏差，つまり何度も繰り返して取られるサンプルの，平均の散らばり具合は，$\sqrt{\sigma/n}$ に一致するとのこと。この式の分母に入っているのは，サンプルの大きさ n(の平方根) である。分母が大きくなればなるほど，ここで得られる値は小さくなる。つまり，サンプルのサイズが大きくなればなるほど，精度の高い母平均の推定が出来ることを意味している。これも一緒に頭の中に入れておこう。

第三に，サンプルの平均値 $\bar{x}$ は，母集団の平均値 $\bar{X}$ を中心にした正規分布に従うことがわかっている。これは中心極限定理と呼ばれる特徴なのだが，その詳細については本稿の範囲を超えるので割愛する。

ともあれ，これら三つの特徴から，サンプルの平均値 $\bar{x}$ がわかれば，母平均は $N(\bar{x}, \sigma^2/n)$ の正規分布のどこかにあることがわかる。正規分布の確率密度関数は，理論的に明らかであるから，$\pm 1\sigma$ の範囲に全体の何 % が含まれるか，ということが算出できる。例えば，$\bar{x} \pm \sqrt{\sigma/n}$ の範囲内に母平均が入る確率が，67.3% あるということだ。

この辺りに推測統計学のカラクリが潜んでいる。

母集団から，サンプルを取ってきたとする。サンプル平均 $\bar{x}$ とその標準偏差 σ が得られる。さて，ここで母平均を求めたいとする。サンプルを何度も取り，その平均値をどんどん均していけば，上の第一定理により母平均に一致するはずだ。でもサンプルを何度も取るのは大変。だから，一回のサンプルを信じて，サンプル平均 $\bar{x}$ が母平均と一緒だと考えよう (これが推測の第一歩。手抜きの第一歩でもある)。だけどさすがに，一回のサンプル平均が母平均とぴったり一致するなんてコトはありそうにない。だから，あんまり「これ (サンプル平均) が母平均なんです」と言い切るのはよろしくない。そこで，第二の特徴を考える。母分散がわかっていれば，$\sqrt{\sigma/n}$ の散らばりをもつ正規分布に従うのだから，「一回のサンプル平均 $\bar{x}$ の，周囲 $\bar{x} \pm \sqrt{\sigma/n}$ の範囲内に母平均が入る確率が，67.3% ある。$\pm 2\sqrt{\sigma/n}$ の範囲内には 95.3% の確率で入ってくる。」と言えるだろう。しかし，ここにも問題がある。普通，母分散なんてわからないのだ。だから，その時はサンプルの分散 S^2 が母分散の推定値だと考えて (推測の第二歩，さらに足下が怪しくなる)，サンプルの分散を使いながら，母平均の入りそうな範囲を確率と共に推定する。[*3]

1.5 有限母集団からのサンプリング

ところで，サンプルの分散を，そのまま推定値として使うのには，ちょっと問題がある。今までのサンプルと母集団の関係式 (式 6 など) は，母集団が無限であることを前提とした式になっている。しかし，社会調査のシーンで使われる母集団は有限 (人類全体とか) であるから，有限を前提とした式になるように，修正が必要である。少し横道に逸れるが，以下にそのプロセスを辿ってみよう。

[*3] このように，推測の幅を持たせることを**区間推定**といい，そうでない**点推定**とは区別する。

まず，サンプル平均の期待値の算出について[*4]。

母集団の要素が $X_1, X_2, \cdots, X_n$ である (有限ですね) とき，ここから大きさ n のサンプルを取り出す取り出し方は，${}_N C_n$ だから $N!/(N-n)!n!$ 通りある。この中から，たまたま X_1 がサンプルされて取り出された，としよう。このような取り出され方は何回あるだろうか？

n 個のサンプルのうち，ひとつが X_1 で，残り $n-1$ 個は X_1 以外の何であっても良い。X_1 以外を取り出すのは，${}_{N-1}C_{n-1}$ 通りである。このことに注意した上で，サンプル平均の期待値を求めてみよう。全ての平均値の平均値だから，

$$E(\bar{x}) = \frac{1}{{}_N C_n}\left\{\frac{1}{n}(X_1 + X_2 + \cdots + X_n) + \frac{1}{n}(X_1 + X_2 + \cdots + X_{n-1} + X_{n+1}) \right.$$
$$\left. + \cdots + \frac{1}{n}(X_{N-n+1} + X_{N-n+2} + \cdots + X_N)\right\}$$

右辺第二の項は，母集団の X_n 番目の要素がないサンプルである。それ以降の項は同様に，n 個のサンプルにおいて，欠けている要素がひとつずつズレながら，X_N 番目の要素まで続いている。この中で，X_1 が入っている項は ${}_{N-1}C_{n-1}$ 個。$X_2, X_3, \cdots, X_N$ もそれぞれ，${}_{N-1}C_{n-1}$ 個入っている。とすると，この式の {} の内側は，${}_{N-1}C_{n-1}$ 個の X_i を全部足して，n で割っていることになる。つまり，右辺は

$$= \frac{1}{n}\frac{1}{{}_N C_n}{}_{N-1}C_{n-1}(X_1 + X_2 + \cdots + X_N)$$

である。これを紐解くと，

$$= \frac{1}{n}\frac{1}{N!/(N-n)!n!}\frac{(N-1)!}{((N-1)-(n-1))!(n-1)!}(X_1 + X_2 + \cdots + X_N)$$
$$= \frac{1}{n}\frac{n!(N-1)!(N-n)!}{N!(N-1)!(N-n)!}(X_1 + X_2 + \cdots + X_N)$$
$$= \frac{1}{n}\frac{n}{N}(X_1 + X_2 + \cdots + X_N)$$
$$= \frac{1}{N}(X_1 + X_2 + \cdots + X_N)$$
$$= \bar{X}$$

となる。有限であれ，無限であれ，サンプル平均の期待値は母平均に一致するというわけだ。t ところがサンプル平均の分散の場合は少し話が違ってくる。サンプル平均の分散の期待値，$E(\bar{x}^2)$ はどうなるかというと，

$$E(\bar{x}^2) = \frac{1}{{}_N C_n}\left[\left\{\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - E(\bar{x})\right\}^2\right.$$
$$+ \left\{\frac{1}{n}(X_1 + X_2 + \cdots + X_{n+1}) - E(\bar{x})\right\}^2 + \cdots$$
$$\left. + \left\{\frac{1}{n}(X_{N-n+1} + X_{N-n+2} + \cdots + X_N) - E(\bar{x})\right\}^2\right]$$

[*4] 母分散の推定値についての項目なのだが，計算の途中でサンプル平均が関係してくるので，まずそれをやっつけておくのです。

となる。

これをこのまま計算すると非常に面倒なので，$(a-b)^2 = a^2 - 2ab + b^2$ の公式に従って分解してみよう。もちろん a に当たるのがサンプルで，b に当たるのが $E(\bar{x})$ である。

第一の項はこんな感じである。

$$\text{第一項} = \frac{1}{n^2}\left\{(X_1 + X_2 + \cdots + X_n)^2 + (X_1 + X_2 + \cdots + X_{n-1} + X_{n+1})^2 + \cdots + (X_{N-n+1} + \cdots + X_N)^2\right\}$$

. 第二項は，

$$\text{第二項} = -2E(\bar{x})\left\{\frac{1}{n}(X_1 + X_2 + \cdots + X_n) + \frac{1}{n}(X_1 + X_2 + \cdots + X_{n-1} + X_{n+1}) + \cdots + \frac{1}{n}(X_{N-n+1} + \cdots + X_N)\right\}$$

. 第三項は，

$$\text{第三項} = [{}_NC_n \{E(\bar{x})\}^2] = \{E(\bar{x})\}^2$$

. これら全てに $\frac{1}{{}_NC_n}$ がかかっていることを忘れずに。

さてまず第一項。これは () 内の n 個の要素を足して，二乗するのだが，記号だけでやるので得てしてわけが分からなくなる。でも振り落とされずについて来て下さい。

簡単な例から考えてみよう。$(a+b+c+d)^2 = a^2 + b^2 + c^2 + d^2 + 2ab + 2ac + 2ad + 2bc + 2bd + 2cd$ である。ということは，第一項のなかの最初の項からは，$(X_1^2 + X_2^2 + \cdots + X_n^2 + X_1X_2 + X_1X_3 + \cdots + X_{n-1}X_n)$ が得られることになる。つまり X_i^2 と，X_jX_k の組み合わせがつらつらと。特に後者は，$j < k$ という制限を付けると，$2\sum_j \sum_k X_jX_k$ と書ける。これが $(X_{N-n+1} + X_{N-n+2} + \cdots + X_N)$ の項まで同様に続くわけだ。では結局どれだけの数が含まれるのだろうか。

X_i^2 の項は，${}_NC_n$ 個ある中括弧 {} の中に，全部で $n \cdot {}_NC_n$ 個入っている。さらに，$X_1, X_2, \cdots, X_n$ のどの項も同じ数だけ入っているはずだから，X_1^2 は $n \cdot {}_NC_n/N$ 個あるはずだ。X_2 や X_3 など，全ての項が同じだけあるはずなので，ここは $n \cdot {}_NC_n/N(X_1^2 + X_2^2 + \cdots + X_n^2)$ と書けるだろう。

では X_jX_k の項はどうなるか。これはともかく，二つの項の組み合わせである。N 個の中から 2 つの項を取り出す取り出し方は，${}_NC_2$ 通り。ひとつのサンプルから 2 つの項を取り出す取り出し方は，${}_nC_2$ 通りある。全てのサンプリング方法，${}_NC_n$ 通りの中には，${}_nC_2/{}_NC_2$ の割合で X_jX_k が入っているに違いない。ということは全部で $2\frac{{}_nC_2{}_NC_n}{{}_NC_2}$ 個あるはずなのだ。

さて少し式の展開をしてみよう。第一項を更に展開すると以下のようになる。

$$\text{第一項} = \frac{1}{{}_NC_n}\left[\frac{1}{n^2}\left\{(X_1 + X_2 + \cdots + X_n)^2 + (X_1 + X_2 + \cdots + X_{n-1} + X_{n+1})^2 + \cdots + (X_{N-n+1} + \cdots + X_N)^2\right\}\right]$$

.

$$= \frac{1}{{}_NC_n}\left[\frac{1}{n^2}\left\{\frac{n{}_NC_n}{N}\sum X_i^2 + 2\frac{{}_nC_2{}_NC_n}{{}_NC_2}\sum_j\sum_k X_jX_k\right\}\right]$$

$$= \frac{1}{Nn}\sum X_i^2 + 2\frac{1}{n^2}\frac{{}_nC_2}{{}_NC_2}\sum\sum X_jX_k$$

$$= \frac{1}{Nn}\sum X_i^2 + 2\frac{1}{n^2}\frac{\frac{n!}{(n-2)!2!}}{\frac{N!}{(N-2)!2!}}\sum\sum X_jX_k$$

ここで後者の項が少し面倒になってきたので，整理しよう。

$$\frac{1}{n^2}\frac{{}_nC_2}{{}_NC_2} = \frac{1}{n^2}\frac{\frac{n!}{(n-2)!2!}}{\frac{N!}{(N-2)!2!}} = \frac{1}{n^2}\frac{n!(N-2)!2!}{(n-2)!2!N!}$$

$$= \frac{1}{n^2} \frac{(n \times n-1 \times n-2 \times \cdots \times 1) \times (N-2 \times N-3 \times \cdots 1) \times 2 \times 1}{(n-2 \times n-3 \times \cdots 1) \times 2 \times 1 \times (N \times N-1 \times N-2 \times \cdots 1)}$$

分子と分母に同じものがあるので，それらはキャンセルしあって，

$$= \frac{n-1}{Nn(N-1)}$$

これだけになる。さあこれを元の式に組み込もう。

$$\text{第一項} = \frac{1}{Nn} \sum X_i^2 + 2\frac{n-1}{Nn(N-1)} \sum\sum X_j X_k$$

やっと見やすい形になった。しかしこれはまだ一つ目，第二項もやっつけよう。第二項は全部書くと，

$$-\frac{1}{{}_N C_n} 2E(\bar{x}) \left\{ \frac{1}{n}(X_1 + X_2 + \cdots + X_n) + \cdots + \frac{1}{n}(X_{N-n+1} + X_{N-n+2} + \cdots + X_N) \right\}$$

となる項である。ここで，$2E(\bar{x})$ を除いた箇所，つまり

$$\frac{1}{{}_N C_n} \left\{ \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \cdots \right\}$$

は，結局のところ $E(\bar{x})$ の式と変わらないわけである。つまり，第二項は $= -2E(\bar{x})E(\bar{x})$ である。

第二項と第三項を併せて考えてみよう。

$$\text{第二項} + \text{第三項} = -2E(\bar{x})E(\bar{x}) + \{E(\bar{x})\}^2 = -\{E(\bar{x})\}^2 = -\bar{X}^2$$

$$= -\left\{ \frac{1}{N}(X_1 + X_2 + \cdots + X_N) \right\}^2$$

$$= -\frac{1}{N^2} \left\{ (X_1^2 + X_2^2 + \cdots + X_N^2) + 2(X_1X_2 + X_1X_3 + \cdots + X_{N-1}X_N) \right\}$$

$$= -\frac{1}{N^2} \sum X_i^2 - 2\frac{1}{N^2} \sum\sum X_j X_k$$

さてこれで，もともとの式に戻ってみると，

$$E(\bar{x}^2) = \frac{1}{Nn} \sum X_i^2 + 2\frac{n-1}{Nn(N-1)} \sum\sum X_j X_k - \frac{1}{N^2} \sum X_i^2 - 2\frac{1}{N^2} \sum\sum X_j X_k$$

$$= \left(\frac{1}{Nn} - \frac{1}{N^2}\right) \sum X_i^2 + 2\left\{ \frac{n-1}{Nn(N-1)} - \frac{1}{N^2} \right\} \sum\sum X_j X_k$$

$$= \frac{N-n}{N^2 n} \sum X_i^2 + 2\frac{n-N}{N^2 n(N-1)} \sum\sum X_j X_k$$

$$= \frac{N-n}{n(N-1)} \left\{ \frac{N-1}{N^2} \sum X_i^2 - \frac{2}{N^2} \sum\sum X_j X_k \right\}$$

$$= \frac{N-n}{n(N-1)} \left[\frac{N}{N^2} \sum X_i^2 - \frac{1}{N^2} \sum X_i^2 - \frac{2}{N^2} \sum\sum X_j X_k \right]$$

$$= \frac{N-n}{n(N-1)} \left[\frac{N}{N^2} \sum X_i^2 - \frac{1}{N^2} \left\{ \sum X_i^2 + 2\sum\sum X_j X_k \right\} \right]$$

$$= \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum X_i^2 - \frac{1}{N^2} \left\{ \sum X_i \right\}^2 \right]$$

$$= \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum X_i^2 - \left\{ \frac{1}{N} \sum X_i \right\}^2 \right]$$

$$= \frac{N-n}{n(N-1)} \left\{ \frac{1}{N} \sum X_i^2 - \bar{X}^2 \right\}$$

さてここで，最後に残った $1/N \sum X_i^2 - \bar{X}^2$ は分散の公式である。ここから以下の関係が得られる。

$$E(\bar{x}^2) = \frac{N-n}{N-1} \frac{\sigma^2}{n} \tag{7}$$

つまり，有限母集団を対象とする場合は，有限修正項 $N-n/N-1$ をつけないと無限母集団の値と一致しないことがわかる。もっとも，多くのサンプリング調査の場合は，n に比べて N が非常に大きい (10 万人とか，500 万人とか) ため，$N-n/N-1$ がほとんど 1.0 に近くなるから，標本誤差に大きな影響を与えることはない。

2 付録

描画を含む本稿のデータを算出するための全コード

```
library(tidyverse)
set.seed(20190820)
# サイズを決定
N <- 10000
# 正規乱数の発生
x <- round(rnorm(N, 50, 10)) %>% matrix(ncol=10)
# 平均値の計算
mean(x)

# 行平均
apply(x,1,mean)
# 行平均の最大値・最小値
apply(x,1,mean) %>% as.data.frame %>% summarise(max=max(.),min=min(.))

# 行平均の平均
apply(x,1,mean) %>% mean

# 標本分散を算出する関数
sampleVar <-function(x){
    mean((x - mean(x))^2)
}

# 今回のサンプル全ての標本分散（母分散）
sampleVar(x)
```

```
# 行毎の標本分散の平均
apply(x,1,sampleVar) %>% mean
# 行毎の不偏分散の平均
apply(x,1,var) %>% mean

### 描画
# 二種類の分散をデータセットに
varSet <- data.frame(VAR = apply(x,1,sampleVar), N_1Var = apply(x,1,var))
# 新しい変数を作成（中身は NA）
varSet$mVar <- NA
varSet$mN_1Var <- NA
# 第 n 行目までの分散の平均を計算して代入していく
for(n in 1:NROW(varSet)){
    varSet$mVar[n] <- mean(varSet$VAR[1:n])
    varSet$mN_1Var[n] <- mean(varSet$N_1Var[1:n])
}
# 描画
#          変数の選択
varSet %>% dplyr::select(mVar,mN_1Var) %>%
# サンプル数を変数として持つ
    dplyr::mutate(nSamp=rownames(.)) %>%
# データを縦長に
    tidyr::gather(key,val,-nSamp) %>%
# データの型を整える
    dplyr::mutate(N=as.numeric(nSamp),key=as.factor(key)) %>%
# ggplot!
    ggplot(aes(x=N,y=val,color=key))+geom_point()+geom_line() +
    geom_hline(yintercept = 100,color="blue",lty=2) +
    xlab("サンプル数") + ylab("分散の平均") +
    theme_set(theme_grey(base_family = "HiraKakuProN-W3")) +
    scale_color_hue(name = "分散の種類", labels = c(mVar = "標本分散", mN_1Var ="不偏分
散")) -> p
# 描画の保存
ggsave(plot=p,file = "sample3.png", device = "png",
dpi = 400, width = 16, height = 10,units="cm")
```