

## 分散の加法性を視覚的に理解する（その3）

Sampo Suzuki, CC 4.0 BY-NC-SA

2021-05-31

### はじめに

分散の加法性を視覚的に理解する（その2）において、データが独立であれば分散の加法性がなりたつことがわかりました。では、同一正規分布から取り出した二つの値の平均値の場合はどうなるか確認してみます。

### 同一データからサンプリングした二つの値を平均した場合

最初に以下の処理を行う関数を定義します。

- データを乱数生成する<sup>1</sup>
- 乱数生成したデータをランダムサンプリングする
- 作成したデータの統計量を求める
- 無相関検定の結果と統計量をデータフレームにまとめる

<sup>1</sup> 今回は `rnorm()` 関数による分散が 100 となる正規分布

```
1 f2 <- function(i = NA, n = 5000000) {  
2   # データを乱数生成する  
3   x <- rnorm(n = n, mean = 10, sd = 10)  
4   # 乱数生成したデータから二つのデータを取り出す  
5   a <- sample(x, n, replace = TRUE)  
6   b <- sample(x, n, replace = TRUE)  
7   num <- 2  
8   # 統計量を求める  
9   df <- data.frame(no = i,  
10                    var.x = var(x),  
11                    var.a = var(a), var.b = var(b),  
12                    var.ab = var((a + b) / num), var.sum = (var(a / num) + var(b / num)),  
13                    cov = cov(a / num, b / num),  
14                    cov2 = cov(a / num, b / num) * 2)  
15   # 無相関の検定結果と統計量をデータフレームにまとめる  
16   df <- cor.test(a, b) %>% broom::tidy() %>% dplyr::bind_cols(df)  
17   return(df)  
18 }
```

Table 1: 二つのサンプルを平均した場合の分散

No	相関係数	p 値	母集団	標本 a	標本 b	加法 1	加法 2	差異	母集団比	cov2
1	0.000	0.958	100.008	99.915	99.990	49.977	49.976	0.001	0.500	0.001
2	0.000	0.692	100.066	100.155	100.019	50.035	50.044	-0.009	0.500	-0.009
3	0.000	0.697	99.988	99.980	99.813	49.957	49.948	0.009	0.500	0.009
4	0.000	0.583	99.904	99.873	99.799	49.906	49.918	-0.012	0.500	-0.012
5	0.000	0.651	100.058	100.117	100.088	50.041	50.051	-0.010	0.500	-0.010
6	0.000	0.823	99.950	99.921	99.959	49.975	49.970	0.005	0.500	0.005
7	0.000	0.959	100.032	99.968	100.008	49.993	49.994	-0.001	0.500	-0.001
8	0.000	0.442	100.037	100.052	99.891	50.003	49.986	0.017	0.500	0.017
9	0.000	0.625	100.050	100.107	100.039	50.047	50.037	0.011	0.500	0.011
10	0.001	0.150	99.827	99.890	99.781	49.950	49.918	0.032	0.500	0.032
11	0.000	0.696	99.990	100.017	99.971	50.006	49.997	0.009	0.500	0.009
12	0.000	0.458	100.031	100.015	99.950	49.975	49.991	-0.017	0.500	-0.017
13	0.001	0.093	100.024	100.043	100.025	50.055	50.017	0.038	0.500	0.038
14	0.000	0.478	99.903	99.828	99.876	49.942	49.926	0.016	0.500	0.016
15	0.000	0.330	99.931	100.030	99.943	50.015	49.993	0.022	0.500	0.022
16	0.000	0.578	100.132	100.030	100.146	50.057	50.044	0.012	0.500	0.012
17	-0.001	0.187	99.952	99.939	99.856	49.920	49.949	-0.029	0.499	-0.029
18	0.000	0.346	100.091	100.099	100.080	50.024	50.045	-0.021	0.500	-0.021
19	0.000	0.684	100.009	100.065	99.993	50.006	50.015	-0.009	0.500	-0.009
20	0.000	0.902	99.876	99.959	99.979	49.987	49.984	0.003	0.500	0.003
21	0.000	0.605	99.962	99.812	99.997	49.964	49.952	0.012	0.500	0.012
22	0.000	0.353	99.955	99.921	99.916	49.980	49.959	0.021	0.500	0.021
23	0.000	0.516	100.086	100.092	100.106	50.035	50.049	-0.015	0.500	-0.015
24	0.000	0.701	99.945	99.970	100.001	49.984	49.993	-0.009	0.500	-0.009
25	0.000	0.987	99.943	100.038	100.016	50.014	50.014	0.000	0.500	0.000
26	0.000	0.802	99.968	100.007	99.949	49.995	49.989	0.006	0.500	0.006
27	0.000	0.527	99.980	99.860	99.937	49.935	49.949	-0.014	0.499	-0.014
28	0.000	0.530	100.023	100.081	99.938	50.019	50.005	0.014	0.500	0.014
29	0.000	0.846	99.972	99.992	99.900	49.977	49.973	0.004	0.500	0.004
30	-0.001	0.094	100.079	99.953	100.198	50.000	50.038	-0.037	0.500	-0.037

Table 2: 二つのサンプルが独立でない場合

No	相関係数	p 値	母集団	標本 a	標本 b	加法 1	加法 2	差異	母集団比	cov2
----	------	-----	-----	------	------	------	------	----	------	------

$$\text{加法 1} = \text{var}\left(\frac{a+b}{2}\right), \text{加法 2} = \text{var}\left(\frac{a}{2}\right) + \text{var}\left(\frac{b}{2}\right)$$

## 同一データからサンプリングした三つの値を平均した場合

最初に以下の処理を行う関数を定義します。

- データを乱数生成する<sup>2</sup>
- 乱数生成したデータをランダムサンプリングする
- 作成したデータの統計量を求める
- 無相関検定の結果と統計量をデータフレームにまとめる

<sup>2</sup>今回は `rnorm()` 関数による分散が 100 となる正規分布

```

1  f3 <- function(i = NA, n = 5000000) {
2    # データを乱数生成する
3    x <- rnorm(n = n, mean = 10, sd = 10)
4    # 乱数生成したデータから三つのデータを取り出す
5    a <- sample(x, n, replace = TRUE)
6    b <- sample(x, n, replace = TRUE)
7    c <- sample(x, n, replace = TRUE)
8    num <- 3
9    # 統計量を求める
10   df <- data.frame(no = i,
11                   var.x = var(x),
12                   var.a = var(a), var.b = var(b), var.c = var(c),
13                   var.abc = var((a + b + c) / num),
14                   var.sum = (var(a / num) + var(b / num) + var(c / num)),
15                   cov.ab = cov(a, b), cov.ac = cov(a, c), cov.bc = cov(b, c),
16                   cov2.ab = cov(a, b) * 2, cov2.ac = cov(a, c) * 2, cov2.bc = cov(b, c) * 2)
17   # 無相関の検定結果と統計量をデータフレームにまとめる
18   df <- cor.test(a, b) %>% broom::tidy() %>% dplyr::bind_cols(df)
19   df <- cor.test(a, c) %>% broom::tidy() %>% dplyr::bind_cols(df)
20   df <- cor.test(b, c) %>% broom::tidy() %>% dplyr::bind_cols(df)
21   return(df)
22 }
```

Table 3: 三つのサンプルを平均した場合の分散

No	母集団	標本 a	標本 b	標本 c	加法 1	加法 2	差異	母集団比
1	99.949	99.863	99.848	99.874	33.282	33.287	-0.005	0.333
2	99.996	100.092	100.031	100.105	33.352	33.359	-0.007	0.334
4	100.078	100.144	100.110	100.083	33.386	33.371	0.015	0.334
5	99.990	99.949	99.943	99.964	33.315	33.317	-0.002	0.333
7	99.852	99.756	99.913	99.804	33.270	33.275	-0.005	0.333
8	100.019	100.029	100.101	100.023	33.329	33.350	-0.022	0.333
9	99.999	99.906	100.056	99.913	33.313	33.319	-0.007	0.333
10	99.954	100.010	99.949	99.943	33.332	33.322	0.010	0.333
11	100.024	100.081	100.010	100.036	33.371	33.347	0.024	0.334
12	100.055	100.112	100.129	99.978	33.342	33.358	-0.016	0.333
13	99.917	99.934	99.936	99.908	33.306	33.309	-0.002	0.333
14	100.016	100.126	100.020	99.989	33.340	33.348	-0.009	0.333
15	99.983	100.065	99.996	99.867	33.326	33.325	0.001	0.333
17	99.961	100.048	99.992	99.983	33.341	33.336	0.005	0.334
18	100.026	99.995	99.948	99.957	33.305	33.322	-0.017	0.333
19	99.981	99.989	99.982	99.957	33.329	33.325	0.004	0.333
21	99.987	99.970	99.934	100.030	33.330	33.326	0.004	0.333
22	99.968	100.030	99.999	99.942	33.342	33.330	0.012	0.334
23	100.006	100.023	99.971	100.033	33.356	33.336	0.020	0.334
24	100.036	100.000	100.044	100.069	33.329	33.346	-0.017	0.333
25	99.975	100.009	100.045	100.046	33.327	33.344	-0.018	0.333
26	100.016	100.101	100.009	99.967	33.320	33.342	-0.022	0.333
27	99.908	99.906	99.897	99.810	33.271	33.290	-0.020	0.333
28	100.037	100.055	100.152	100.043	33.363	33.361	0.002	0.334

Table 4: 三つのサンプルのどれかが独立でない場合

No	母集団	標本 a	標本 b	標本 c	加法 1	加法 2	差異	母集団比
3	100.089	100.083	100.201	100.178	33.369	33.385	-0.016	0.333
6	99.892	99.938	99.905	99.854	33.289	33.300	-0.010	0.333
16	99.893	100.003	99.848	99.967	33.337	33.313	0.024	0.334
20	99.978	99.977	100.064	99.981	33.310	33.336	-0.026	0.333
29	99.982	100.030	99.966	99.899	33.347	33.322	0.025	0.334
30	100.067	100.050	100.056	99.998	33.365	33.345	0.020	0.333

$$\text{加法 1} = \text{var}\left(\frac{a+b+c}{3}\right), \text{ 加法 2} = \text{var}\left(\frac{a}{3}\right) + \text{var}\left(\frac{b}{3}\right) + \text{var}\left(\frac{c}{3}\right)$$

## まとめ

データが独立であれば分散の加法性が成り立っており、 $n$  個の平均をとった場合、分散が  $\frac{1}{n}$  になることが予想できます。

### *About handout style*

The Tufte handout style is a style that Edward Tufte uses in his books and handouts. Tufte's style is known for its extensive use of sidenotes, tight integration of graphics with text, and well-set typography. This style has been implemented in LaTeX and HTML/CSS<sup>3</sup>, respectively.

<sup>3</sup> See Github repositories [tufte-latex](#) and [tufte-css](#)