分散の加法性を視覚的に理解する（その３）

Sampo Suzuki,

2021-06-01

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、および、三つの値の平均値の場合はどうなるか、その２と同様の手段で確認してみます。

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[1]
- 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

[1] 今回は `rnorm()` 関数による分散が 100 となる正規分布

```
f2 <- function(i = NA, n = 5000000) {
  # データを乱数生成する
  x <- rnorm(n = n, mean = 10, sd = 10)
  # 乱数生成したデータから二つのデータを取り出す
  a <- sample(x, n, replace = TRUE)
  b <- sample(x, n, replace = TRUE)
  num <- 2
  # 統計量を求める
  df <- data.frame(no = i,
                   var.x = var(x),
                   var.a = var(a), var.b = var(b),
                   var.ab = var((a + b) / num), var.sum = (var(a / num) + var(b / num)),
                   cov = cov(a / num, b / num ),
                   cov2 = cov(a / num, b / num) * 2)
  # 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
  df <- cor.test(a, b) %>% broom::tidy() %>% dplyr::bind_cols(df)
  return(df)
}
```

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
2	0.000	0.358	99.870	99.909	99.957	49.987	49.967	0.021	0.501	0.021
4	0.000	0.842	100.079	100.129	100.043	50.038	50.043	-0.004	0.500	-0.004
5	-0.001	0.178	100.086	100.129	100.016	50.006	50.036	-0.030	0.500	-0.030
6	0.000	0.322	100.059	100.070	100.064	50.011	50.034	-0.022	0.500	-0.022
7	0.000	0.551	99.974	99.993	99.896	49.959	49.972	-0.013	0.500	-0.013
8	0.000	0.684	100.021	99.948	100.013	49.999	49.990	0.009	0.500	0.009
9	0.000	0.604	100.022	100.032	99.981	49.991	50.003	-0.012	0.500	-0.012
10	0.001	0.197	99.869	99.844	99.867	49.957	49.928	0.029	0.500	0.029
11	0.000	0.628	100.059	100.149	99.999	50.048	50.037	0.011	0.500	0.011
12	0.000	0.570	99.988	99.986	99.940	49.994	49.982	0.013	0.500	0.013
13	0.000	0.783	99.986	100.023	99.891	49.972	49.979	-0.006	0.500	-0.006
16	0.000	0.555	100.025	100.083	99.997	50.007	50.020	-0.013	0.500	-0.013
17	0.000	0.718	100.092	100.200	100.107	50.069	50.077	-0.008	0.500	-0.008
18	0.000	0.493	99.936	99.861	100.044	49.992	49.976	0.015	0.500	0.015
19	0.000	0.613	100.071	99.955	100.036	49.986	49.998	-0.011	0.500	-0.011
20	0.000	0.383	99.987	99.966	99.947	49.998	49.978	0.020	0.500	0.020
21	0.000	0.401	100.018	100.084	100.004	50.003	50.022	-0.019	0.500	-0.019
22	0.000	0.884	100.019	99.932	100.023	49.992	49.989	0.003	0.500	0.003
23	0.000	0.709	100.010	100.011	100.098	50.036	50.027	0.008	0.500	0.008
24	0.000	0.452	100.051	99.943	100.129	50.035	50.018	0.017	0.500	0.017
25	0.000	0.673	100.014	100.075	100.003	50.010	50.019	-0.009	0.500	-0.009
26	0.000	0.416	100.010	99.966	99.993	49.972	49.990	-0.018	0.500	-0.018
27	0.000	0.744	100.028	100.095	100.036	50.040	50.033	0.007	0.500	0.007
28	-0.001	0.180	99.916	99.992	99.932	49.951	49.981	-0.030	0.500	-0.030
29	0.001	0.169	99.979	100.079	99.971	50.043	50.012	0.031	0.501	0.031
30	0.001	0.185	100.156	100.189	100.185	50.123	50.094	0.030	0.500	0.030

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
1	-0.001	0.045	100.011	100.128	100.048	49.999	50.044	-0.045	0.500	-0.045
3	0.001	0.033	100.040	100.097	100.066	50.088	50.041	0.048	0.501	0.048
14	-0.001	0.013	100.070	100.096	100.076	49.987	50.043	-0.056	0.500	-0.056
15	0.001	0.049	100.106	100.067	100.094	50.084	50.040	0.044	0.500	0.044

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

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[2]
- 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

[2] 今回は rnorm() 関数による分散が 100 となる正規分布

```
f3 <- function(i = NA, n = 5000000) {
  # データを乱数生成する
  x <- rnorm(n = n, mean = 10, sd = 10)
  # 乱数生成したデータから三つのデータを取り出す
  a <- sample(x, n, replace = TRUE)
  b <- sample(x, n, replace = TRUE)
  c <- sample(x, n, replace = TRUE)
  num <- 3
  # 統計量を求める
  df <- data.frame(no = i,
                   var.x = var(x),
                   var.a = var(a), var.b = var(b), var.c = var(c),
                   var.abc = var((a + b + c) / num),
                   var.sum = (var(a / num) + var(b / num) + var(c / num)),
                   cov.ab = cov(a, b), cov.ac = cov(a, c), cov.bc = cov(b, c),
                   cov2.ab = cov(a, b) * 2, cov2.ac = cov(a, c) * 2, cov2.bc = cov(b, c) * 2)
  # 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
  df <- cor.test(a, b) %>% broom::tidy() %>% dplyr::bind_cols(df)
  df <- cor.test(a, c) %>% broom::tidy() %>% dplyr::bind_cols(df)
  df <- cor.test(b, c) %>% broom::tidy() %>% dplyr::bind_cols(df)
  return(df)
}
```

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	99.945	99.934	100.023	99.931	33.305	33.321	-0.016	0.333
3	99.973	99.938	99.870	99.907	33.306	33.302	0.005	0.333
4	100.020	99.998	99.930	100.067	33.307	33.333	-0.026	0.333
5	99.947	99.905	100.021	99.909	33.319	33.315	0.004	0.333
6	99.966	99.965	99.928	99.928	33.329	33.314	0.016	0.333
7	100.002	100.063	100.001	99.956	33.351	33.335	0.016	0.334
8	100.004	100.050	99.928	100.014	33.319	33.332	-0.013	0.333
9	100.079	100.021	100.057	100.095	33.365	33.353	0.013	0.333
10	100.228	100.285	100.336	100.261	33.418	33.431	-0.013	0.333
11	100.006	100.048	100.064	100.030	33.355	33.349	0.006	0.334
12	99.858	99.744	99.905	99.867	33.286	33.280	0.006	0.333
13	99.980	99.878	99.915	99.958	33.308	33.306	0.003	0.333
14	100.110	100.019	100.165	100.043	33.368	33.359	0.009	0.333
15	100.095	100.124	100.090	100.110	33.376	33.369	0.007	0.333
16	99.968	99.979	99.984	99.981	33.351	33.327	0.024	0.334
17	100.057	100.078	100.166	100.171	33.373	33.379	-0.007	0.334
18	99.897	99.885	99.888	100.061	33.304	33.315	-0.011	0.333
19	99.856	99.790	99.856	99.870	33.281	33.280	0.002	0.333
20	99.971	99.955	100.086	99.948	33.354	33.332	0.022	0.334
22	99.895	99.792	99.833	99.839	33.289	33.274	0.015	0.333
25	100.140	99.969	100.260	100.097	33.389	33.370	0.020	0.333
26	99.920	99.968	99.884	99.923	33.307	33.308	-0.002	0.333
27	100.025	99.991	100.153	99.970	33.352	33.346	0.006	0.333
28	100.000	99.940	100.028	100.013	33.351	33.331	0.020	0.334
29	99.960	100.023	99.911	99.962	33.321	33.322	-0.001	0.333
30	99.999	100.052	100.067	99.986	33.340	33.345	-0.005	0.333

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
2	99.968	100.051	99.987	100.012	33.292	33.339	-0.047	0.333
21	99.966	99.894	100.034	99.932	33.297	33.318	-0.021	0.333
23	100.047	99.977	99.869	100.056	33.289	33.322	-0.034	0.333
24	99.930	99.913	99.977	99.899	33.343	33.310	0.033	0.334

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

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[3], respectively.

[3] See Github repositories tufte-latex and tufte-css