分散の加法性を視覚的に理解する（その３）

Sampo Suzuki,

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、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
1	0.000	0.395	100.074	100.056	100.066	50.012	50.031	-0.019	0.500	-0.019
2	0.000	0.332	100.058	100.045	100.061	50.005	50.027	-0.022	0.500	-0.022
3	0.000	0.485	100.059	100.058	100.064	50.015	50.031	-0.016	0.500	-0.016
4	0.000	0.364	100.030	99.963	100.053	49.984	50.004	-0.020	0.500	-0.020
5	0.000	0.949	100.014	100.019	100.069	50.023	50.022	0.001	0.500	0.001
6	0.000	0.639	99.895	99.933	99.970	49.986	49.976	0.010	0.500	0.010
7	0.000	0.994	99.934	99.865	99.870	49.934	49.934	0.000	0.500	0.000
8	0.000	0.431	99.948	99.979	100.001	49.977	49.995	-0.018	0.500	-0.018
9	0.000	0.961	99.993	100.017	100.042	50.016	50.015	0.001	0.500	0.001
11	0.000	0.991	100.019	100.064	100.073	50.035	50.034	0.000	0.500	0.000
12	0.000	0.821	100.083	100.069	99.990	50.020	50.015	0.005	0.500	0.005
13	0.000	0.826	100.007	100.070	99.888	49.994	49.990	0.005	0.500	0.005
14	0.000	0.364	99.939	99.898	99.928	49.977	49.957	0.020	0.500	0.020
15	0.000	0.291	99.939	100.017	99.932	50.011	49.987	0.024	0.500	0.024
16	0.000	0.521	100.010	100.006	99.896	49.961	49.975	-0.014	0.500	-0.014
17	0.000	0.440	100.105	100.200	100.151	50.105	50.088	0.017	0.501	0.017
18	0.001	0.247	99.980	99.962	99.954	50.005	49.979	0.026	0.500	0.026
19	0.000	0.431	99.861	99.779	99.886	49.934	49.916	0.018	0.500	0.018
21	0.000	0.983	100.056	100.062	99.929	49.998	49.998	0.000	0.500	0.000
22	0.001	0.097	99.925	99.960	99.993	50.025	49.988	0.037	0.501	0.037
23	0.001	0.233	100.047	99.959	100.112	50.044	50.018	0.027	0.500	0.027
24	0.000	0.796	99.992	99.933	99.934	49.961	49.967	-0.006	0.500	-0.006
25	0.000	0.525	99.995	100.054	100.013	50.031	50.017	0.014	0.500	0.014
26	0.000	0.592	100.027	100.090	100.090	50.057	50.045	0.012	0.500	0.012
27	0.000	0.973	100.100	100.014	100.156	50.042	50.042	-0.001	0.500	-0.001
28	0.000	0.551	99.940	99.945	99.939	49.984	49.971	0.013	0.500	0.013
29	0.000	0.269	100.063	99.982	100.096	49.995	50.020	-0.025	0.500	-0.025
30	0.001	0.087	100.053	100.078	100.084	50.079	50.041	0.038	0.501	0.038

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
10	-0.001	0.047	99.989	100.017	99.979	49.955	49.999	-0.044	0.500	-0.044
20	0.001	0.005	99.852	99.918	99.798	49.992	49.929	0.063	0.501	0.063

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

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	99.971	99.918	99.795	100.021	33.306	33.304	0.002	0.333
2	99.948	99.991	99.882	99.918	33.314	33.310	0.004	0.333
3	100.094	100.125	100.078	99.943	33.364	33.349	0.014	0.333
4	100.023	99.896	100.007	99.991	33.323	33.321	0.001	0.333
5	100.115	100.039	100.045	100.165	33.354	33.361	-0.007	0.333
6	99.867	99.818	99.964	99.820	33.289	33.289	0.000	0.333
7	100.060	100.130	100.148	99.985	33.362	33.363	0.000	0.333
8	100.050	100.057	100.038	99.961	33.354	33.340	0.014	0.333
9	99.876	99.807	100.023	99.796	33.302	33.292	0.011	0.333
10	100.021	100.132	100.036	99.849	33.330	33.335	-0.006	0.333
11	99.957	100.080	99.937	99.998	33.309	33.335	-0.026	0.333
12	100.065	100.031	100.047	100.070	33.325	33.350	-0.024	0.333
13	100.089	100.043	100.099	100.026	33.343	33.352	-0.010	0.333
14	99.950	99.912	100.009	99.819	33.296	33.304	-0.008	0.333
16	100.036	100.070	100.099	100.116	33.361	33.365	-0.004	0.333
17	100.008	99.917	100.020	100.002	33.334	33.326	0.008	0.333
18	99.989	100.040	100.072	99.922	33.341	33.337	0.004	0.333
19	100.005	100.075	100.074	99.926	33.327	33.342	-0.015	0.333
20	99.911	99.922	99.890	99.928	33.326	33.304	0.022	0.334
21	100.136	100.092	100.150	100.213	33.345	33.384	-0.039	0.333
22	99.998	100.012	100.039	100.022	33.344	33.341	0.002	0.333
23	99.950	100.032	99.954	100.007	33.308	33.333	-0.024	0.333
24	100.007	100.007	99.972	100.026	33.339	33.334	0.006	0.333
25	99.999	99.903	99.963	99.905	33.304	33.308	-0.004	0.333
26	99.930	99.946	99.945	99.907	33.280	33.311	-0.031	0.333
27	99.996	100.135	100.070	100.008	33.368	33.357	0.012	0.334
28	100.022	100.058	100.055	100.078	33.336	33.355	-0.019	0.333
29	100.032	100.117	100.140	99.993	33.383	33.361	0.021	0.334
30	100.003	100.026	100.078	100.097	33.344	33.356	-0.011	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
15	99.927	100.091	99.81	99.954	33.293	33.317	-0.024	0.333

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

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