分散の加法性を視覚的に理解する（その２）

Sampo Suzuki,

2021-05-31

はじめに（小室先生のアドバイスから）

分散の加法性が成り立つには「データが独立」であるという前提条件があります。乱数生成した二つのデータが本当に独立なのかを確認すると共に分散の加法性も確認してみます。

関数の定義

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[2]
- 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

[2] `sampling = TRUE` の場合のみ

```
f <- function(i = NA, sampling = FALSE, n = 5000000) {
  # データを乱数生成する
  x <- rnorm(n = n, mean = 10, sd = 10)
  y <- rnorm(n = n, mean = 30, sd = 10)
  # 乱数生成したデータからサンプリングする場合
  if (sampling == TRUE) {
    x <- sample(x, n, replace = TRUE)
    y <- sample(y, n, replace = TRUE)
  }
  # 統計量を求める
  df <- data.frame(no = i, var.x = var(x), var.y = var(y),
                   var.xy = var(x + y), var.sum = var(x) + var(y),
                   cov = cov(x, y), cov2 = cov(x, y) * 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(x, y) %>% broom::tidy() %>% dplyr::bind_cols(df)
  return(df)
}
```

この関数を `for` ループで一定回数繰り返し、その結果をデータフレームにまとめ、分散がどのようになっているかを比較します。

Table 1: 変数の意味

変数名	その意味	備考
var.x	データ x の分散	
var.y	データ y の分散	
var.xy	データ x と y を加算したものの分散（var(x + y)）	加法 1
var.sum	データ x, y の分散を加算したもの（var(x) + var(y)）	加法 2
var.diff	var.xy から var.sum を減算したもの	加法 1 と加法 2 の差異
cov2	データ x, y の共分散の 2 倍数	
cov	データ x, y の共分散	計算のみで未出力

乱数生成したデータの場合

Table 2: 乱数生成した二つのデータの分散

No	相関係数	p 値	標本 x	標本 y	加法 1	加法 2	差異	cov2
1	0.000	0.433	100.017	100.043	200.131	200.060	0.070	0.070
2	-0.001	0.153	100.001	100.036	199.910	200.037	-0.128	-0.128
3	0.001	0.148	100.006	99.947	200.082	199.953	0.129	0.129
4	0.001	0.223	99.970	99.980	200.059	199.950	0.109	0.109
5	0.000	0.752	100.022	100.020	200.070	200.042	0.028	0.028
6	0.000	0.784	99.958	100.039	199.972	199.996	-0.025	-0.025
7	0.000	0.318	99.974	100.003	199.888	199.978	-0.089	-0.089
8	0.000	0.883	100.063	99.914	199.990	199.977	0.013	0.013
9	0.000	0.987	99.960	99.983	199.942	199.944	-0.001	-0.001
10	0.000	0.383	100.027	100.055	200.003	200.081	-0.078	-0.078
11	0.001	0.207	99.969	100.202	200.283	200.170	0.113	0.113
12	0.000	0.748	100.041	100.001	200.071	200.043	0.029	0.029
13	0.000	0.490	99.995	100.105	200.038	200.100	-0.062	-0.062
14	0.000	0.857	99.945	99.817	199.746	199.762	-0.016	-0.016
15	0.001	0.198	99.994	99.923	200.032	199.917	0.115	0.115
16	0.000	0.994	100.101	99.945	200.045	200.045	-0.001	-0.001
17	0.000	0.874	99.973	99.875	199.834	199.848	-0.014	-0.014
18	-0.001	0.102	99.948	100.033	199.836	199.982	-0.146	-0.146
19	0.000	0.868	100.034	99.878	199.927	199.912	0.015	0.015
20	0.000	0.913	100.019	99.952	199.962	199.971	-0.010	-0.010
21	0.000	0.643	99.996	100.083	200.121	200.080	0.042	0.042
22	0.000	0.640	100.032	99.980	199.969	200.011	-0.042	-0.042
24	-0.001	0.130	99.927	100.098	199.890	200.025	-0.136	-0.136
25	0.000	0.359	99.976	100.001	200.059	199.977	0.082	0.082
26	0.000	0.886	99.957	100.046	199.990	200.003	-0.013	-0.013
27	0.001	0.243	99.919	100.035	200.058	199.954	0.104	0.104
28	0.000	0.443	99.895	99.995	199.959	199.891	0.069	0.069
29	0.001	0.162	99.947	99.992	200.064	199.939	0.125	0.125
30	0.000	0.763	100.104	99.950	200.081	200.054	0.027	0.027

Table 3: 乱数生成した二つのデータが独立でない場合

No	相関係数	p 値	標本 x	標本 y	加法 1	加法 2	差異	cov2
23	-0.001	0.023	99.881	99.916	199.595	199.797	-0.203	-0.203

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

乱数生成したデータをランダムサンプリングした場合

Table 4: ランダムサンプリングしたデータの分散

No	相関係数	p 値	標本 x	標本 y	加法 1	加法 2	差異	cov2
2	0.000	0.687	100.070	100.010	200.117	200.080	0.036	0.036
3	-0.001	0.064	99.868	99.977	199.679	199.845	-0.166	-0.166
4	-0.001	0.246	100.041	100.012	199.949	200.053	-0.104	-0.104
5	0.001	0.153	100.056	100.104	200.287	200.159	0.128	0.128
6	0.000	0.612	100.070	99.982	200.097	200.051	0.045	0.045
7	0.000	0.840	100.023	99.714	199.755	199.737	0.018	0.018
8	0.000	0.783	100.093	100.187	200.255	200.280	-0.025	-0.025
9	0.000	0.277	100.125	100.047	200.270	200.173	0.097	0.097
10	0.000	0.845	99.834	99.906	199.757	199.740	0.017	0.017
11	0.000	0.938	99.785	100.043	199.821	199.828	-0.007	-0.007
12	0.000	0.869	100.098	100.083	200.196	200.181	0.015	0.015
13	0.000	0.842	100.024	100.052	200.093	200.076	0.018	0.018
14	0.001	0.128	99.996	100.127	200.259	200.123	0.136	0.136
15	0.000	0.592	100.025	100.037	200.110	200.062	0.048	0.048
16	0.001	0.057	99.988	99.997	200.156	199.985	0.171	0.171
17	0.000	0.334	99.882	99.946	199.914	199.828	0.086	0.086
18	0.000	0.570	100.050	100.221	200.220	200.271	-0.051	-0.051
19	0.000	0.781	99.904	100.166	200.045	200.070	-0.025	-0.025
20	0.000	0.822	99.963	100.032	200.015	199.995	0.020	0.020
21	0.000	0.675	100.114	99.855	200.006	199.969	0.037	0.037
22	0.001	0.129	99.965	99.976	200.077	199.941	0.136	0.136
23	0.000	0.369	100.031	100.084	200.195	200.115	0.080	0.080
24	0.001	0.113	100.013	100.093	200.248	200.106	0.142	0.142
25	0.000	0.808	100.039	99.916	199.933	199.955	-0.022	-0.022
26	0.001	0.174	99.932	99.828	199.882	199.760	0.121	0.121
27	0.000	0.333	100.023	100.064	200.173	200.086	0.087	0.087
28	0.000	0.355	100.064	99.856	200.003	199.920	0.083	0.083
29	0.000	0.959	100.056	100.010	200.061	200.066	-0.005	-0.005
30	-0.001	0.138	100.009	99.993	199.870	200.002	-0.133	-0.133

Table 5: ランダムサンプリングしたデータが独立でない場合

No	相関係数	p 値	標本 x	標本 y	加法 1	加法 2	差異	cov2
1	-0.002	0.001	100.166	99.961	199.817	200.126	-0.309	-0.309

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

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いることがわかります。データが独立とは言い難い無相関の検定が成功するケース（95% 信頼区間に 0 が入らない）では、分散の差（共分散の 2 倍数）が一桁大きいので加法性が成り立っているとは言い難いように言えますがこのケースでは数値だけを見ている限り差はよくわかりません。

cor.test() 関数について

`cor.test()` 関数は無相関の検定を行う関数です。対立仮説（H_1）は下記の出力の通り「true correlation is **not** equal to 0（相関係数はゼロではない）」ですので、帰無仮説（H_0）は「相関係数はゼロである（相関はない）」となります。有意水準 α で検定が失敗すれば（帰無仮説が棄却されない、$p \geqq \alpha$ である）帰無仮説が採択されますので相関係数はゼロ（データ間には相関がない）と考えられます。

```
##
##  Pearson's product-moment correlation
##
## data:  rnorm(n) and rnorm(n)
## t = -0.18054, df = 4999998, p-value = 0.8567
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0009572605  0.0007957847
## sample estimates:
##           cor
## -8.073797e-05
```

Appendix

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