分散の加法性を視覚的に理解する

Sampo Suzuki,

2021-06-01

Introduction

2021 年度データ分析勉強会のテキストである『統計解析のはなし』[大平, 2006] の「標本が 2 つになれば」（P26〜）には分散の加法性の話が出てきます。分散の加法性は理解できるようでいて、理解できていないので、**R** を使って分散の加法性を可視化しながら説明してみます。

以降、平均値 μ、標準偏差 σ、分散 σ^2 である正規分布を $N(\mu, \sigma^2)$ と表記します。

加法性を可視化する

以下の平均値と標準偏差を持つ二つの正規分布を `rnorm()` 関数による正規分布乱数を用いて作成[1]します。

[1] n $= 5 \times 10^6$ 個の値を作成しています

Table 1: 二つの正規分布

正規分布	平均	標準偏差	備考
$N(\mu_a, \sigma_a^2)$	$\mu_a = 10$	$\sigma_a = 10$	
$N(\mu_b, \sigma_b^2)$	$\mu_b = 30$	$\sigma_b = 10$	

```
a <- rnorm(n, mean = 10, sd = 10)
b <- rnorm(n, mean = 30, sd = 10)
```

Table 2: 二つの正規分布の要約統計量

正規分布	平均	分散	標準偏差	備考
$N(\mu_a, \sigma_a^2)$	9.9985888	99.9977663	9.9998883	
$N(\mu_b, \sigma_b^2)$	29.9965583	99.9706305	9.9985314	

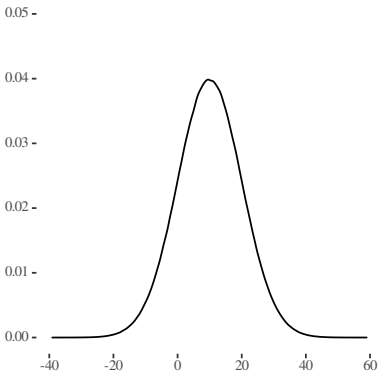

Figure 1: $N(\mu_a, \sigma_a^2)$ の分布

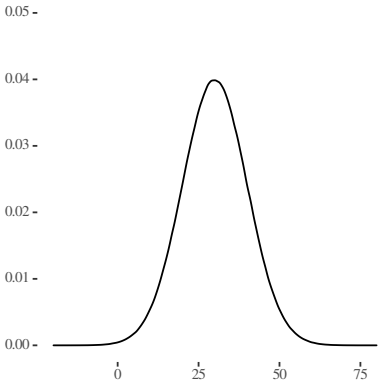

Figure 2: $N(\mu_b, \sigma_b^2)$ の分布

この二つの正規分布 $N(\mu_a, \sigma_a^2)$ と $N(\mu_b, \sigma_b^2)$ からランダムサンプリングにより一つずづ値を取り出して加算します。すなわち

$$N(\mu_a, \sigma_a^2) \text{ から取り出した値} + N(\mu_b, \sigma_b^2) \text{ から取り出した値}$$

という新しい値を作成します。取り出した値は元に戻し、同様の取り出し、加算を繰り返すと以下のようなデータが作成できます。ここではスペースの都合で先頭から限定して表示しています。

```
c <- c(sample(a, n, replace = TRUE) + sample(b, n, replace = TRUE))
head(c, 50)
```

```
##  [1] 55.599241  5.237594 41.600298 23.359751 37.085258 44.437978  8.561242
```

```
##  [8] 20.376646 36.391691 50.055509 39.175702 23.941961 30.602629 33.448960
## [15] 32.419665 49.567928 39.905698 59.922084 63.553437  3.195873 25.325148
## [22] 38.901329 41.377471 57.117994 53.376228 63.967994 36.382754 61.459881
## [29] 35.738357 54.465682 73.874611 26.561265 55.829688 43.607332 56.616833
## [36] 51.793272 21.109510 40.752398 30.066753 63.453079 44.989330 52.826389
## [43] 34.170679 42.629914 30.266823 34.935418 55.543318 61.329707 63.234686
## [50] 67.256212
```

分散の加法性により上記のデータは $N(\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2))$ という正規分布になるはずですが実際はどうでしょう。各正規分布の平均値と分散を比較します。

Table 3: 各分布の要約統計量

正規分布	平均	分散	備考
$N(\mu_a, \sigma_a^2)$	9.9985888	99.9977663	元の分布
$N(\mu_b, \sigma_b^2)$	29.9965583	99.9706305	元の分布
$N(\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2))$	39.9951472	199.9683968	分散の加法性
$N(\mu_c, \sigma_c^2)$	40.0018296	199.9790041	実際の分布

このように確かに分散の加法性が成り立っており、正規分布 $N(\mu_a, \sigma_a^2)$ や $N(\mu_b, \sigma_b^2)$ より横に広がった正規分布になっていることが分かります。

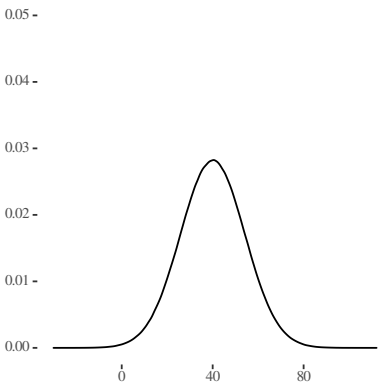

Figure 3: $N(\mu_c, \sigma_c^2)$ の分布

同一の正規分布から取り出し値を加算した場合

次に二つの正規分布 $N(\mu_a, \sigma_a^2)$ と $N(\mu_b, \sigma_b^2)$ がまったく等しいと仮定します。つまり

$$\mu_a = \mu_b = \mu_d$$

$$\sigma_a = \sigma_b = \sigma_d$$

という正規分布 $N(\mu_d, \sigma_d^2)$ を作成します。

```
d <- rnorm(n, mean = 10, sd = 10)
head(d, 50)
```

```
##  [1]  20.6262427  18.7804457   2.4013990   4.0130979  11.3286222   9.8768041
##  [7]   4.7568502   9.9090505  -4.0445735  16.5491671  20.6799970  12.4881164
## [13]   4.0029392  10.0473607  14.8913675  18.7336644  21.2705317  13.9545623
## [19]  14.4062801  12.8104317  26.6215824  29.2233908  24.3489912  -9.0156795
## [25]   8.8831009  -7.5751283  29.5895387   5.3455020  13.8686233  12.1810300
## [31]  -4.4800935   8.5103878  15.2268517  -5.5262560  11.7653608   8.7079841
## [37]  15.8605852   8.0033821  27.2542045 -12.9758471  16.5038052  10.0209208
## [43]  20.5733627  39.4137038  14.7399978  -1.0212168  -0.3178018  15.1327986
## [49]  15.6217370  22.2421247
```

この正規分布 $N(\mu_d, \sigma_d^2)$ から先程と同様にランダムサンプリングにより一つずづ値を取り出して加算しますが、今回は同一正規分布 $N(\mu_d, \sigma_d^2)$ ですので、二つ取り出します。取り出した値は元の正規分布に戻し同様の操作を繰り返します。

```
e <- c(sample(d, n, replace = TRUE) + sample(d, n, replace = TRUE))
head(e, 50)
```

```
##  [1] 37.037753 38.172461 19.844289 24.561710 34.982592 -9.961573 28.119099
##  [8] -5.797978 36.002053 39.808280 21.175291 41.142078 47.242403 17.248593
## [15] 14.811112 29.322576 40.337956 20.577689 -9.489254 30.694281 36.712260
## [22] 40.807344 27.422842 24.136736 37.564558  6.766611  3.815439 46.948949
## [29] 23.909628 42.335233 38.518215  8.076119 -3.642567 32.897224 31.090545
## [36]  8.275585 15.385796 12.435561 24.388711 19.495090 42.531706 33.668978
## [43] 24.224242 37.670191 31.613863 14.885696 36.773785 27.708292 30.518625
## [50] 26.337926
```

分散の加法性により以下が成り立ちます。

$$N(\mu_d + \mu_d, \sigma_d^2 + \sigma_d^2) = N(2\mu_d, 2\sigma_d^2)$$

つまり、正規分布 $N(\mu_d, \sigma_d^2)$ から取り出した二つの値の和である正規分布 $N(\mu_e, \sigma_e^2)$ は

Table 4: 加法性による要約統計量

正規分布	平均	分散	標準偏差	備考
$N(\mu_e, \sigma_e^2)$	$2\mu_d$	$2\sigma_d^2$	$\sqrt{2\sigma_d^2} = \sqrt{2}\sigma_d$	

という正規分布をすることになります。加法性と実際の正規分布を比べてみると

Table 5: 各分布の要約統計量

正規分布	平均	分散	備考
$N(\mu_d, \sigma_d^2)$	10.0045189	100.1240361	元の分布
$N(2\mu_d, 2\sigma_d^2)$	20.0090379	200.2480722	分散の加法性
$N(\mu_e, \sigma_e^2)$	19.9950719	200.0998858	実際の分布

となり、同一正規分布の場合でも分散の加法性が成り立っていることが分かります。

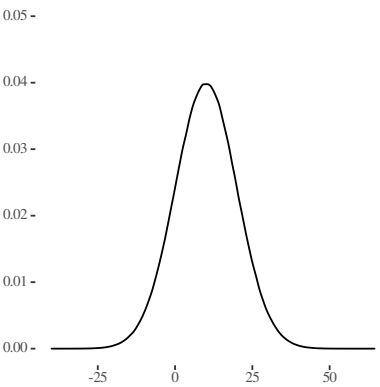

Figure 4: $N(\mu_d, \sigma_d^2)$ の分布

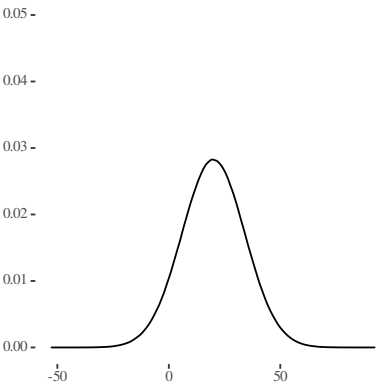

Figure 5: $N(\mu_e, \sigma_e^2)$ の分布

同一の正規分布から取り出した値を平均した場合

同一の正規分布 $N(\mu_d, \sigma_d^2)$ から取り出した二つの値の**平均値の分布**を考えてみます。「二つの値の平均値の平均値」とは、正規分布 $N(\mu_d, \sigma_d^2)$ から、ランダムサンプリングで二つの値を取り出して、その平均値を取るということです。取り出した値は元の正規分布へ戻し、同様の操作を繰り返します。

```
f <- c((sample(d, n, replace = TRUE) + sample(d, n, replace = TRUE)) / 2)
head(f, 20)
```

```
##  [1] 11.872711 22.680748  6.265354  7.445902 18.570496  7.158274 12.881660
##  [8] 18.359528 11.059736 16.541349  4.429276 -3.217499  1.008533  8.969713
## [15] 12.522954 16.699106  8.177954  9.148571 12.510590  4.637624
```

この正規分布正規分布 $N(\mu_f, \sigma_f^2)$ は、二つの値の平均値、つまり二つの値を半分に割った値ですので正規分布 $N(2\mu_d, 2\sigma_d^2)$ のすべての値を半分にした正規分布になると予想できます。

$$\text{「二つの標本の平均値」の平均値} = \frac{2\mu_d}{2} = \mu_d$$

$$\text{「二つの標本の平均値」の標準偏差} = \frac{\sqrt{2}\sigma_d}{2} = \frac{\sigma_d}{\sqrt{2}}$$

$$\text{「二つの標本の平均値」の分散} = (\frac{\sigma_d}{\sqrt{2}})^2 = \frac{\sigma_d^2}{2}$$

Table 6: 各分布の要約統計量

正規分布	平均	分散	標準偏差	備考
$N(\mu_d, \sigma_d^2)$	10.0045189	100.1240361	10.0061999	元の分布
$N(\mu_d, \frac{\sigma_d^2}{2})$	10.0045189	50.0620181	7.0754518	分散の加法性
$N(\mu_f, \sigma_f^2)$	9.9998516	50.0590979	7.0752454	実際の分布

このように元の分布よりも鋭い分布になっていることがわかります。

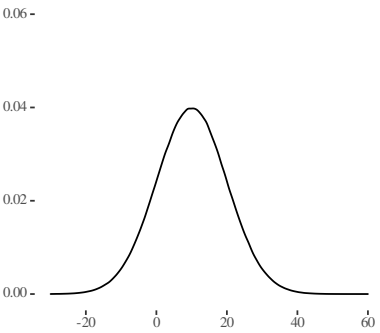

Figure 6: $N(\mu_d, \sigma_d^2)$ の分布

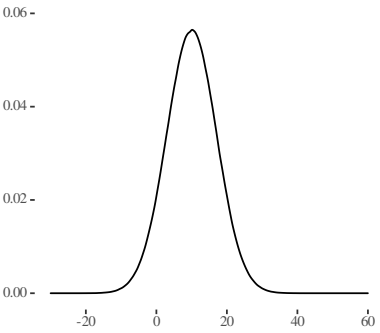

Figure 7: $N(\mu_f, \sigma_f^2)$ の分布

三つ値の平均値の場合

次に同一の正規分布 $N(\mu_d, \sigma_d^2)$ から取り出した三つの値の**平均値の分布**を考えてみます。

```
g <- c((sample(d, n, replace = TRUE) + sample(d, n, replace = TRUE)
        + sample(d, n, replace = TRUE)) / 3)
head(g, 20)
```

```
##  [1] 10.898203 -1.444164 12.990511 16.287384  4.366764  9.781515  6.455512
##  [8]  8.475872  7.239857 10.628173  6.946945  5.182297 -0.421174 17.849796
## [15]  5.557278 20.180071 16.470020  9.177506 -4.355197 22.314048
```

Table 7: 各分布の要約統計量

正規分布	平均	分散	標準偏差	備考
$N(\mu_d, \sigma_d^2)$	10.0045189	100.1240361	10.0061999	元の分布
$N(\mu_g, \sigma_g^2)$	10.0074698	33.3804254	5.7775795	実際の分布
比率	1.000295	0.3333907	0.5774	元の分布に対する比率

標準偏差の比率（0.5774）は、$\frac{1}{\sqrt{3}} = 0.5773503$ とほぼ等しいことが分かります。これより

$$N(\mu_g, \sigma_g^2) = N(\mu_d, \frac{\sigma_d^2}{3})$$

となることがわかります。

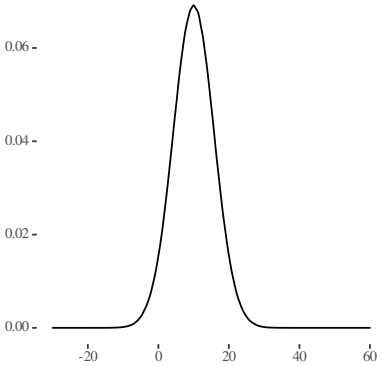

Figure 8: $N(\mu_g, \sigma_g^2)$ の分布

一般化すると

同一正規分布 $N(\mu, \sigma^2)$ から取り出した n 個の値の**平均値の分布**$N(\mu, \sigma_n^2)$ は

$$N(\mu_n, \sigma_n^2) = N(\mu, \frac{\sigma^2}{n})$$

であり、平均は変わらず標準偏差が $\frac{\sigma}{\sqrt{n}}$ となります。

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 = 0.46846, df = 4999998, p-value = 0.6395
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0006670199  0.0010860252
## sample estimates:
##          cor
## 0.0002095028
```

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[2], respectively.

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

References

平大平. 『統計解析のはなし』. 日科技連出版, 改訂版 edition, 2006. URL https://www.juse-p.co.jp/products/view/196. ISBN 978-4-8171-8028-5.