分散の加法性を視覚的に理解する

Sampo Suzuki, CC 4.0 BY-NC-SA

2021-05-30

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。ここでは処理の都合上、二つをデータフレームにまとめてあります。

[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
$N(\mu_a, \sigma_a^2)$	$\mu_a = 10$	$\sigma_a = 10$	
$N(\mu_b, \sigma_b^2)$	$\mu_b = 30$	$\sigma_b = 10$	

```r
x <- data.frame(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	標準偏差	備考
$N(\mu_a, \sigma_a^2)$	10.002356	10.0000116	
$N(\mu_b, \sigma_b^2)$	30.0097476	9.9981689	

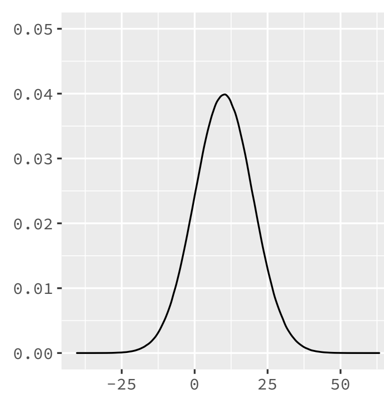

Figure 1: $N(\mu_a, \sigma_a^2)$ の分布

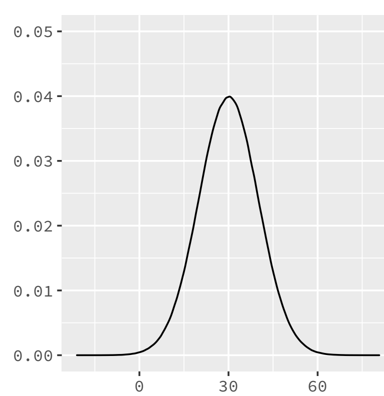

Figure 2: $N(\mu_b, \sigma_b^2)$ の分布

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。

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

```
##  [1] 48.485594 28.823031 23.248879 28.408149 45.032924  8.936533 49.494164
##  [8] 55.556972 25.880179 27.429343 50.647700 54.997154 49.599173 38.139119
## [15]  9.302771  4.988818 25.199294 60.100116 31.650379 49.444355 21.636089
## [22] 41.094214 51.652768 44.234324 37.941525 34.807897 40.068443 43.701257
## [29] 48.249209 26.698325 59.878860 21.431764 42.874170 25.675688 40.374510
## [36] 36.332160 63.247511 38.474428 40.171576 29.171890 37.466399 52.377433
## [43] 21.579113 60.662584 29.418917 34.092421 63.260066 25.692227 36.561038
## [50] 39.142159
```

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。

正規分布	平均	分散	備考
$N(\mu_a,\sigma_a^2)$	10.002356	100.0002314	元の分布
$N(\mu_b,\sigma_b^2)$	30.0097476	99.9633815	元の分布
$N(\mu_a+\mu_b,\sigma_a^2+\sigma_b^2))$	40.0121036	199.9636129	分散の加法性
$N(\mu_c,\sigma_c^2)$	40.006782	200.0706021	

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。

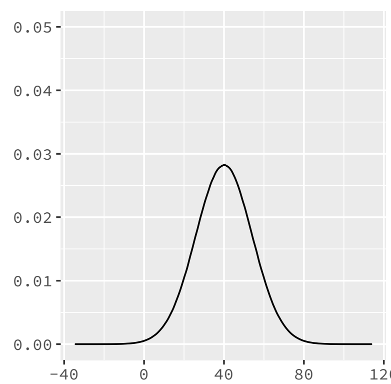

Figure 3: $N(\mu_c,\sigma_c^2)$ の分布

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。

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

```
##  [1]  17.837237013   8.482900753  13.348557174   0.833218091 -11.271911843
##  [6]  11.810127441   0.034802274  16.492975113   3.002560323  -0.851511683
## [11]  22.135161664  -0.166477804   6.147637424  18.406966680   5.564994415
## [16]   8.767292828  15.026947179  32.566077791  15.767011513   6.894729304
## [21]   4.868588909   3.476618059  15.815670741  13.958529606  23.009358797
## [26]  22.407201982  12.487705402  16.978239007   7.928895111  11.703646000
## [31]   5.151864197  14.157120260  11.125565769  33.486652923  14.756405203
## [36]  19.003536543  -1.253908520   5.586895935  -9.414605355   4.300161239
## [41]   5.665699495  -0.006142333   0.646850683  -4.919874157  18.397642717
## [46]   1.134203302  11.515176448   7.944299736   8.035612760  18.384972007
```

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。

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

```
##  [1]   7.117047  40.197022   4.025483  -8.412842  12.103898  18.963426
##  [7]  26.227223  15.080941  12.571927 -13.045191  35.568689  20.780823
## [13]  35.835295  30.910565   5.976138  21.724379   7.062453   6.573209
## [19]  22.820893  37.885613  31.634038  28.631866  16.196499  10.280282
## [25]  23.547313  22.598801  14.656059  18.291213  13.407883   6.318429
## [31]   4.846178  26.505695  26.027938  -4.479397  27.869640  23.034868
## [37]   9.151382  35.418398  36.657953  17.794948  -3.344809  16.736690
## [43]   6.265948  38.804198  27.958440   8.895961  31.850665  10.514966
## [49]  -5.838099  13.289155
```

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

正規分布	平均	分散	備考
$N(\mu_e, \sigma_e^2)$	$2\mu_d$	$2\sigma_d^2$	

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

正規分布	平均	分散	備考
$N(\mu_d, \sigma_d^2)$	9.9940704	99.95931	元の分布
$N(2\mu_d, 2\sigma_d^2)$	19.9881408	199.91862	分散の加法性
$N(\mu_e, \sigma_e^2)$	19.9832408	199.8265825	

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。

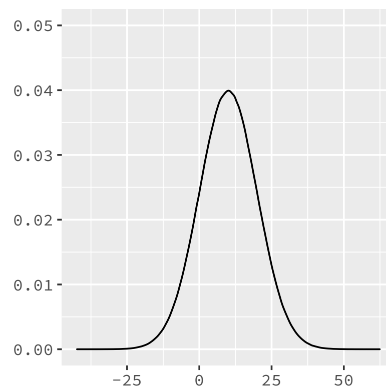

Figure 4: $N(\mu_d, \sigma_d^2)$ の分布

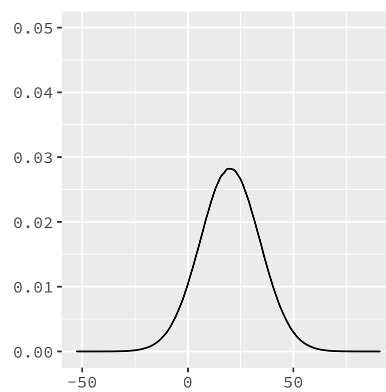

Figure 5: $N(\mu_e, \sigma_e^2)$ の分布

同一の正規分布から取り出した値を平均した場合

最後に同一の正規分布 $N(\mu_d, \sigma_d^2)$ 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。

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

```
##  [1] 11.480955  8.143817  2.679966 17.184109 11.837555  6.839873 10.910810
##  [8]  7.232735 10.055303 18.541158  7.609310  9.367446 18.255863  9.245864
## [15] 14.593417  3.846177 16.133046  5.182148 15.567952  8.662484
```

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。

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

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

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

正規分布	平均	分散	標準偏差	備考
$N(\mu_d, \sigma_d^2)$	9.9940704	99.95931	9.9979653	元の分布
$N(\mu_d, \frac{\sigma_d^2}{2})$	9.9940704	49.979655	7.0696291	分散の加法性
$N(\mu_f, \sigma_f^2)$	9.9923948	50.0161812	7.0722119	

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。

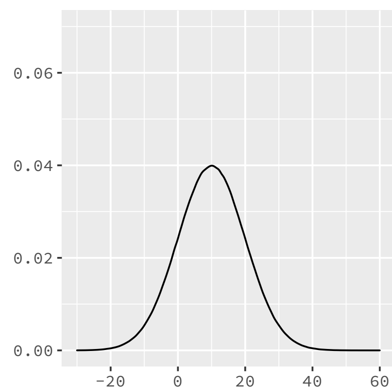

Figure 6: $N(\mu_d, \sigma_d^2)$ の分布

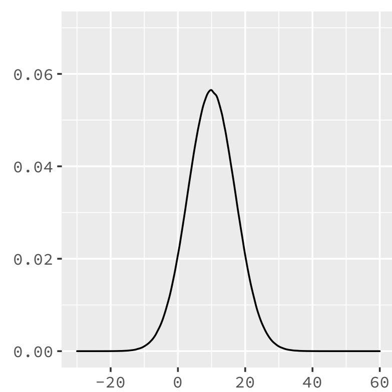

Figure 7: $N(\mu_f, \sigma_f^2)$ の分布

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.