

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 × 106` 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
$N(\mu_a, \sigma_a^2)$	$\mu_a = 10$	$\sigma_a = 10$	
$N(\mu_b, \sigma_b^2)$	$\mu_b = 30$	$\sigma_b = 10$	

```
1 x <- data.frame(a = rnorm(n, mean = 10, sd = 10),  
2                   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
$N(\mu_a, \sigma_a^2)$	9.9951475	9.9959159	
$N(\mu_b, \sigma_b^2)$	29.9989783	10.0051801	

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つ値を取り出して加算します。すなわち

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

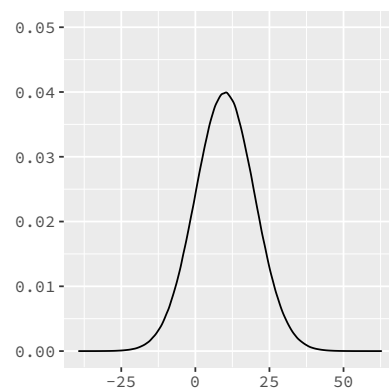


Figure 1: $N(\mu_a, \sigma_a^2)$ の分布

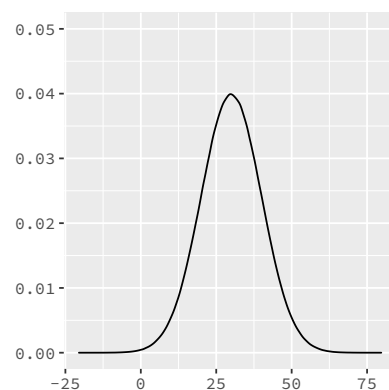


Figure 2: $N(\mu_b, \sigma_b^2)$ の分布

```

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

## [1] 57.33431 13.86405 21.24315 21.52298 60.14019 64.89915 39.53705 30.66057
## [9] 35.58985 44.39301 42.75026 24.11251 46.36193 33.31220 35.95976 25.68859
## [17] 46.92637 70.78838 57.07144 70.56954 47.77346 31.30232 26.13623 54.24486
## [25] 33.07926 27.85359 46.79052 56.22061 44.89961 42.29841 40.51847 48.20768
## [33] 61.10274 53.15060 39.06615 45.27197 12.96846 46.94400 37.67426 49.17930
## [41] 55.30332 43.38246 37.50048 16.14157 53.31532 36.56948 36.82211 35.10256
## [49] 53.73004 39.41396

```

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
$N(\mu_a, \sigma_a^2)$	9.9951475	99.9183346	元の分布
$N(\mu_b, \sigma_b^2)$	29.9989783	100.1036293	元の分布
$N(\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2)$	39.9941258	200.0219639	分散の加法性
$N(\mu_c, \sigma_c^2)$	39.9905755	200.0174879	

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。

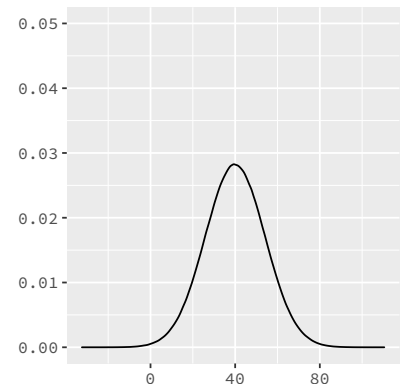


Figure 3: $N(\mu_c, \sigma_c^2)$ の分布

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。

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

## [1]  2.2233014  1.5846010  9.6193505  6.4874410 -6.7454230  3.7869663
## [7]  7.1527368  8.8026565 15.4411933  9.4532447 32.2050733 -3.9343231
## [13] -0.5036607 16.6210855 12.6110213 16.7782865  3.2212688  2.9894326
## [19]  2.5727131 12.8557626 17.3988821 11.1802089  4.8892306 15.0641506
## [25]  5.3592454 12.9890084 20.9063492  9.4312083  9.8147737 -4.3556116
## [31] 20.9061069  0.3316442  0.5345106 17.6295992 11.4236104 -2.4654854
## [37] 14.2736619  4.6445990  7.4769036  8.6161968  7.1881268  9.9389923
## [43] 13.1300344  0.4984865 23.6710128 21.3089260 20.7042437  3.2072904
## [49] 14.2789681 16.7855235
```

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つ値を取り出して加算しますが、今回は同一正規分布 $N(\mu_d, \sigma_d^2)$ でするので、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。

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

## [1] 27.3004509 29.7200016 23.9848304  3.4979476 13.4630279 31.4064247
## [7]  3.5155251 19.0187285 28.3528654 20.0649361 18.2230424 24.6737595
## [13]  7.3944569 17.7851179 52.5397998 34.1228952 10.0614182 42.4256975
## [19]  6.4739662  2.8623599 39.3945190 18.6478855  1.5683397  6.2463979
## [25] 23.2930933 22.1717747 16.2854189 29.2877362 12.5280934  5.7419326
## [31] 24.0009893 14.8794803 29.7507442  9.4592030 44.8896683 -0.4141265
## [37] 18.6104193  6.7617456 19.7407097 10.2755228 28.0814354 24.1566475
## [43] 27.4349759 19.9207832  6.6807422  9.2693959 20.5320835 38.7083146
## [49] 17.1549492 -1.1296295
```

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
$N(\mu_d, \sigma_d^2)$	10.0026499	99.9740994	元の分布
$N(2\mu_d, 2\sigma_d^2)$	20.0052998	199.9481988	分散の加法性
$N(\mu_e, \sigma_e^2)$	20.0026644	199.9443459	

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。

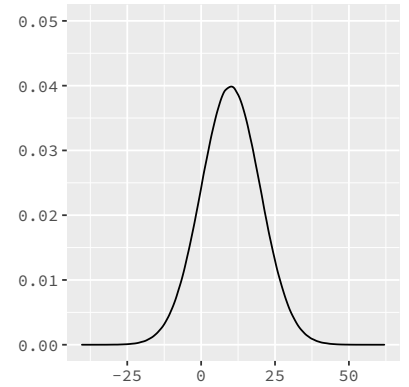


Figure 4: $N(\mu_d, \sigma_d^2)$ の分布

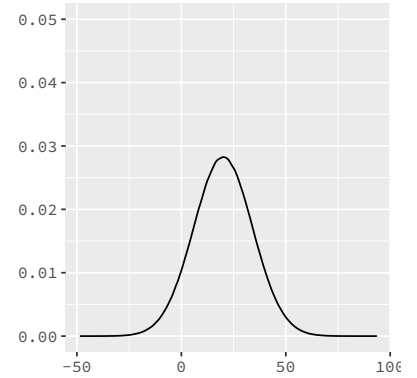


Figure 5: $N(\mu_e, \sigma_e^2)$ の分布

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。

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

## [1]  8.7305884 14.9020097  8.8849919 16.4786882  8.3310684  5.1886257
## [7] 12.4705646 -1.5500875 -0.5659579  9.1376441 11.6390996 18.5703466
## [13]  9.7226476 15.8533947 19.5455885 11.9792291 11.4669204  3.8360299
## [19] 15.0274205 16.0350309
```

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。

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

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

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

正規分布	平均	分散	標準偏差	備考
$N(\mu_d, \sigma_d^2)$	10.0026499	99.9740994	9.9987049	元の分布
$N(\mu_d, \frac{\sigma_d^2}{2})$	10.0026499	49.9870497	7.070152	分散の加法性
$N(\mu_f, \sigma_f^2)$	10.0038582	49.9996318	7.0710418	

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。

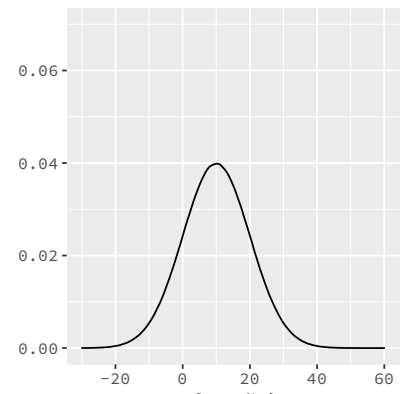


Figure 6: $N(\mu_d, \sigma_d^2)$ の分布

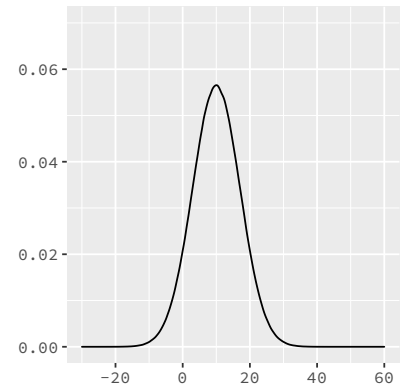


Figure 7: $N(\mu_f, \sigma_f^2)$ の分布

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/CSS², 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.