分散の加法性を視覚的に理解する

Sampo Suzuki, CC 4.0 BY-NC-SA

2021-05-31

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 = 5×10^6 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
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
$N(\mu_a, \sigma_a^2)$	10.0018845	99.994408	9.9997204	
$N(\mu_b, \sigma_b^2)$	29.9968633	100.0170421	10.0008521	

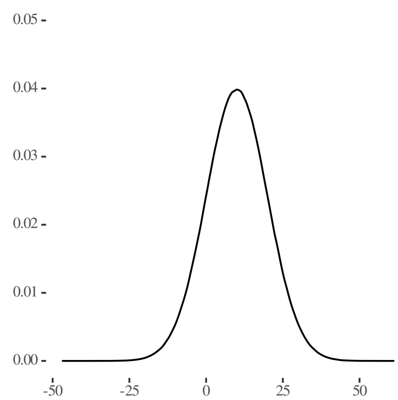

Figure 1: $N(\mu_a, \sigma_a^2)$ の分布

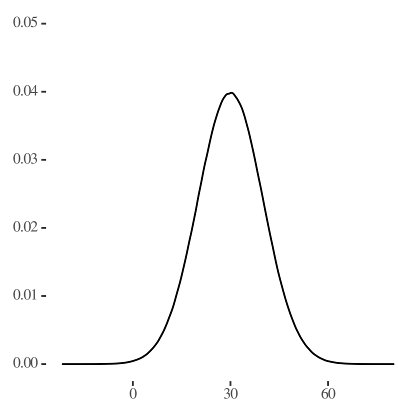

Figure 2: $N(\mu_b, \sigma_b^2)$ の分布

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

```
##  [1] 24.667622 31.612037 47.536352 66.916889 33.749761 36.649172 54.290921
##  [8] 77.884441 44.691841 40.621955 43.838847 16.358541 34.144425 38.581142
## [15] 42.666889 40.791640 37.229227 -7.237897 25.148347 61.800948 61.526631
## [22] 60.641850 40.064745 58.387025 33.583883 57.739139 39.467519 41.188832
## [29] 64.905717 30.787344 36.475048 36.619361 44.598386 47.886578 14.004975
## [36] 66.057824 46.678635 21.642013 44.987430 36.928684 32.562940 57.779482
## [43] 11.197685 38.806596 26.628165 59.812965 47.331169 58.127789 24.942328
## [50] 59.868045
```

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
$N(\mu_a, \sigma_a^2)$	10.0018845	99.994408	元の分布
$N(\mu_b, \sigma_b^2)$	29.9968633	100.0170421	元の分布
$N(\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2))$	39.9987478	200.0114501	分散の加法性
$N(\mu_c, \sigma_c^2)$	40.0002214	199.8853603	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。

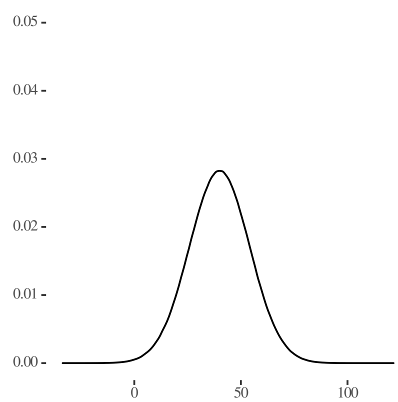

Figure 3: $N(\mu_c, \sigma_c^2)$ の分布

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。

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

```
##  [1] 12.034606 -1.214004 13.776530 16.143755  3.106183  6.182120 23.451810
##  [8] -6.150912 15.639383  7.830779  8.198381 -1.305731 18.704214 10.084423
## [15] 14.747408 -2.926950 10.300094 -4.751576  8.199263  7.273566  6.745570
## [22] -0.311655 22.643916 16.646436 24.806921  7.657171 15.581767 39.094705
## [29] -1.861820  5.465032  8.717058  8.975608 -7.211283 10.600658 12.367784
## [36] -3.024190 17.559373 22.125600  8.687596  7.064100  9.843990 20.489812
## [43] 13.497414 14.323515 15.146395 11.668585  4.629327  3.715618 11.520781
## [50] 11.972610
```

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] 20.8336628 14.4599047  3.9145979 14.5565685 -0.9508145 14.9970491
##  [7] -6.9557951 28.2038815 10.5317038 -0.9593997 25.6304693  1.6113688
## [13]  6.3702246 35.1270067 28.1501076 46.6483243 22.0890673 10.7547139
## [19] 26.2734756 29.6064915  0.5089567  2.1194499  8.3471013 28.5236195
## [25] 13.1189476  3.0607473 17.9480463 30.3358709  9.8321002 32.7045648
## [31]  5.5616795 -0.7902418 18.7189952 -2.3806151  9.6897299  7.1552882
## [37] 21.8980739 17.3509717 13.9177152 27.6431257 44.5932552 48.9292005
## [43] 28.4559518 40.3104493 31.0517835 14.9568908  8.5540292 10.3550870
## [49] 35.1732325 23.2839825
```

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
$N(\mu_d, \sigma_d^2)$	9.9972166	99.9558471	元の分布
$N(2\mu_d, 2\sigma_d^2)$	19.9944332	199.9116942	分散の加法性
$N(\mu_e, \sigma_e^2)$	19.9947345	199.9341886	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。

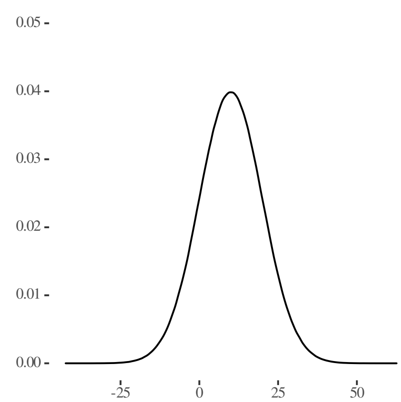

Figure 4: $N(\mu_d, \sigma_d^2)$ の分布

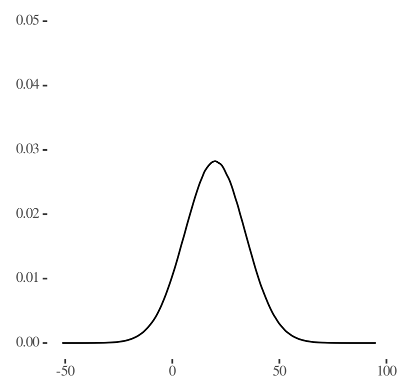

Figure 5: $N(\mu_e, \sigma_e^2)$ の分布

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]  6.652945  9.960451  7.072505 12.127334  7.176384  8.525928 13.713095
##  [8]  5.759468  3.484226 21.679843 24.694319  8.824793  7.736349  7.051991
## [15]  7.460617  5.019521  9.915649  7.197053 12.885202 17.102998
```

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。

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

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

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

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
$N(\mu_d, \sigma_d^2)$	9.9972166	99.9558471	9.9977921	元の分布
$N(\mu_d, \frac{\sigma_d^2}{2})$	9.9972166	49.9779235	7.0695066	分散の加法性
$N(\mu_f, \sigma_f^2)$	9.9982443	49.9502678	7.0675503	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。

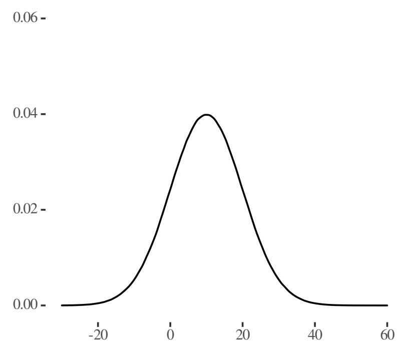

Figure 6: $N(\mu_d, \sigma_d^2)$ の分布

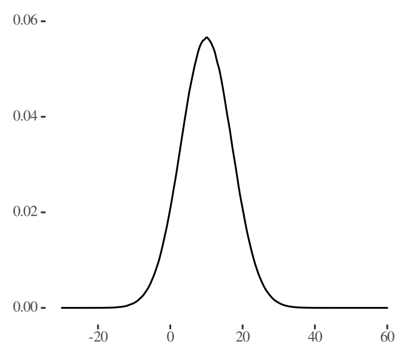

Figure 7: $N(\mu_f, \sigma_f^2)$ の分布

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]  7.1260897 18.2776983  8.8753585  7.7287394  8.9023006  8.2626464
##  [7] 10.4617740 14.2212462  5.8337332  4.1211284  5.7325064  8.9405417
## [13]  0.4938622  0.1307224  8.7384554 17.1890772  3.8833432 19.4887704
## [19] 13.5790915  5.5292824
```

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
$N(\mu_d, \sigma_d^2)$	9.9972166	99.9558471	9.9977921	元の分布
$N(\mu_g, \sigma_g^2)$	9.9998233	33.3572339	5.7755722	実際の分布
比率	1.0002607	0.3337197	0.5776848	元の分布に対する比率

標準偏差の比率（0.5776848）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。

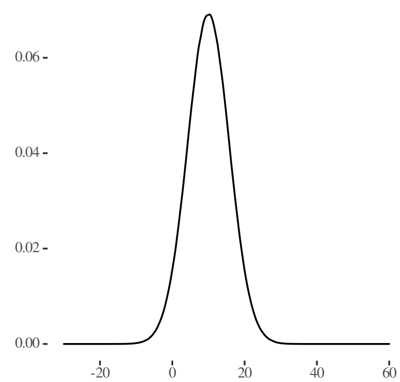

Figure 8: $N(\mu_g, \sigma_g^2)$ の分布

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 = 0.46867, df = 4999998, p-value = 0.6393
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0006669293  0.0010861158
## sample estimates:
##          cor
## 0.0002095934
```

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.