Rで始めるデータサイエンス①

Ｒを使って 統計量を計算

- Ｒをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　～相関係数
- Ｒでよく使う関数
- 乱数を生成する　～分布関数

クイズ１

A組とB組の違いは？

学生	A組	B組
1	86	70
2	96	90
3	64	80
4	95	81
5	76	76
6	63	81
7	90	82
8	70	80
平均	80.0	80.0

どちらの組も、
平均は80点で同じだが、

学生の理解の度合いに
違いはないか？

違いを定量的に
評価できないか？

クイズ２

理科と英語の関係は？

学生	理科	英語
1	85	65
2	96	78
3	64	80
4	90	85
5	76	91
6	80	77
7	73	85
8	70	73
平均	79.25	79.25
標準偏差	10.67	8.05

理科ができると、
英語もできると
言えるのか？

関係の強さ・弱さを
定量的に評価したい

クイズ３

相関係数の理論値

雑音 x と雑音 y は、
エネルギー（分散）は同じであり
互いに無関係（無相関）である。

x と x+y の相関係数をRで計算したら
0.705 となった。

このことを、理論的に裏付けよう。

クイズ４

２つの乱数の和

① ある雑音に別の雑音が重なると、
エネルギー（分散）は何倍になるか？

② ある雑音の大きさを２倍にすると、
エネルギー（分散）は何倍になるか？

但し、２つの雑音は、互いに無関係（無相関）
かつ、エネルギー（分散）は同じ。

Ｒで始めるデータサイエンス①

Ｒを使って 統計量を計算

- Ｒをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　～相関係数
- Ｒでよく使う関数
- 乱数を生成する　～分布関数

R をダウンロード（1/2）

https://cran.r-project.org/

The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Search

About R
R Homepage
The R Journal

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

base — Binaries for base distribution. This is what you want to **install R for the first time**.

R をダウンロード（2/2）

R-4.2.1 for Windows

CRAN
Mirrors
What's new?
Search

Download R-4.2.1 for Windows (79 megabytes, 64 bit)
README on the Windows binary distribution
New features in this version

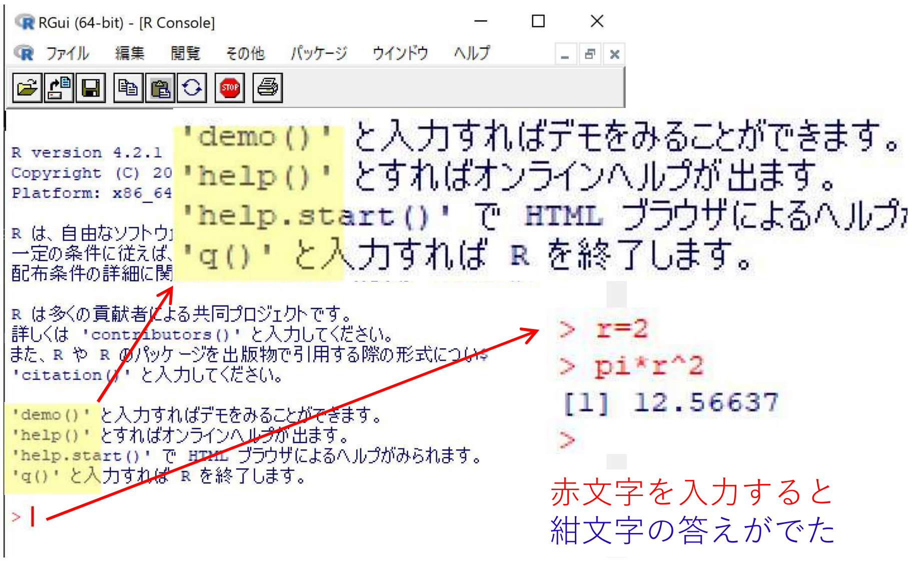
This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT has to be installed manually from here.

R-4.2.1-win.exe（78.7MB）を入手して実行する

R 4.2.1

このアイコンをクリック
するとＲが動く

R を動かしてみる



赤文字を入力すると
紺文字の答えがでた

R で計算してみる

```
> r=2
> pi*r^2
[1] 12.56637
```

☜ R に入力：赤文字
☜ R の出力：紺文字

$r = 2$
πr^2

記号「<-」は代入

```
> r <- c(2, 3, 4)
> pi*r^2
[1] 12.56637 28.27433 50.26548
```

$\boldsymbol{r} = [2 \quad 3 \quad 4]$
$\pi \boldsymbol{r}^2$

```
> pi*r^2 -> a
> mean(a)
[1] 30.36873
```

```
> sum(a)/3
[1] 30.36873
```

$\boldsymbol{a} = \pi \boldsymbol{r}^2$
$E[\boldsymbol{a}]$

統計量を計算してみる

データ x <- c(1, 2 ,3, 4) x_i , i=1,2,3,・・・,n

平均 mean(x)
2.5
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

不偏分散 var(x)
1.666667
$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

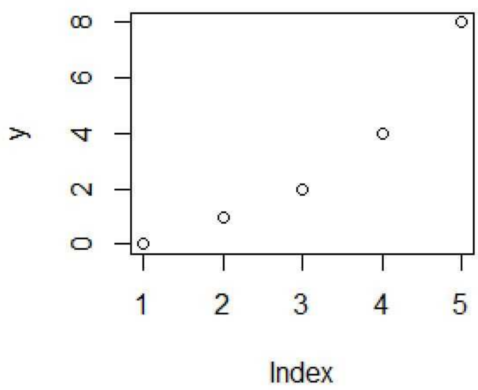
標準偏差 sd(x)
1.290994
$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

プロットしてみる（1/3）

データ y をプロットしたい
$$\boldsymbol{y} = [0 \quad 1 \quad 2 \quad 4 \quad 8]$$

```
y <- c(0,1,2,4,8)
plot(y)
```

R に入力



プロットしてみる（2/3）

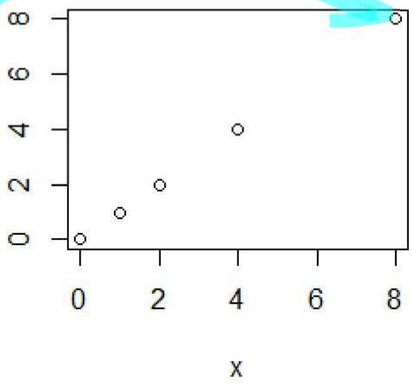
ｘとｙの**散布図**をみたい

$\boldsymbol{x} = [0 \quad 1 \quad 2 \quad 4 \quad 8]$

$\boldsymbol{y} = [0 \quad 1 \quad 2 \quad 4 \quad 8]$

```
x <- c(0,1,2,4,8)
y <- c(0,1,2,4,8)
plot(x,y)
```

Ｒに入力



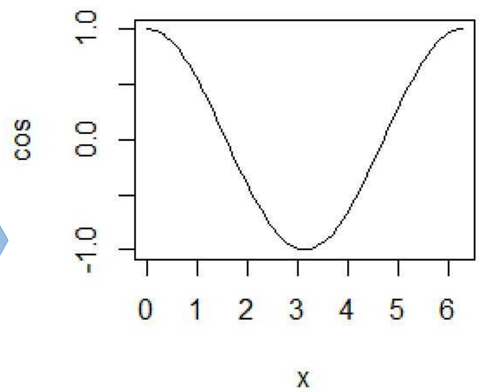
プロットしてみる（3/3）

0から2πまでの範囲で

cos をプロット

```
plot(cos, 0, 2*pi)
```

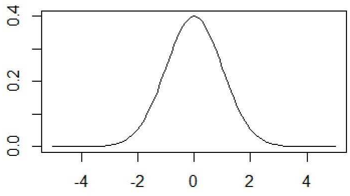
Ｒに入力



関数をかいてみる（1/2）

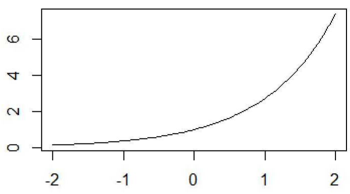
-5 から 5 までの範囲で
正規分布をプロット

curve(dnorm, -5,5)



-2 から 2 までの範囲で
指数関数をプロット

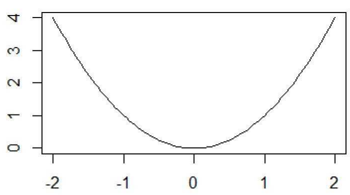
curve(exp(x), -2,2)



関数をかいてみる（2/2）

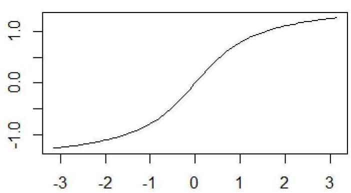
-2 から 2 までの範囲で
べき関数をプロット

curve(x^2, -2,2)



-π から π までの範囲で
逆正接関数をプロット

curve(atan(x), -pi,pi)



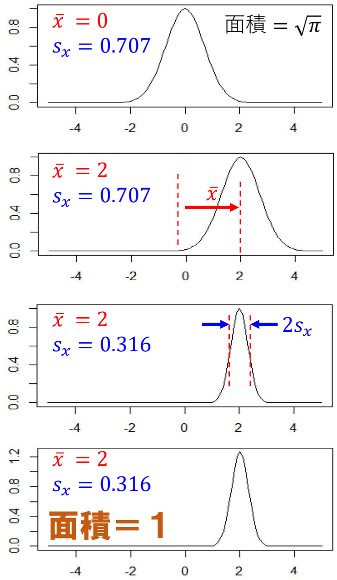
ガウス関数（数式）

$$f(x) = exp(-x^2)$$

$$f(x) = exp(-(x - \bar{x})^2)$$

$$f(x) = exp\left(-\frac{(x - \bar{x})^2}{2s_x^2}\right)$$

$$f(x) = \frac{1}{\sqrt{2\pi s_x^2}} exp\left(-\frac{(x - \bar{x})^2}{2s_x^2}\right)$$



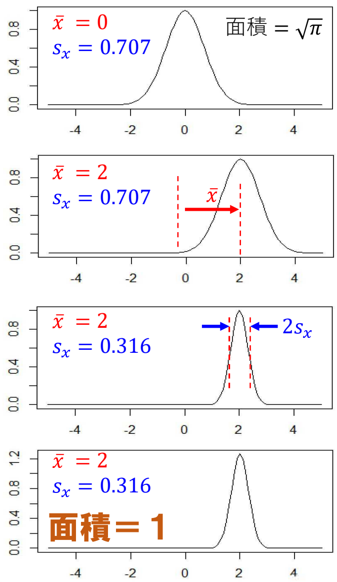
ガウス関数（Ｒのコマンド）

```
> curve(exp(-x^2),-5,5)
```

```
> m=2
> curve(exp(-(x-m)^2),-5,5)
```

```
> s=0.316
> curve(exp(-((x-m)^2)/2/s/s),-5,5)
```

```
> d=sqrt(2*pi*s*s)
> curve(exp(-((x-m)^2)/2/s/s)/d,-5,5)
```



Ｒで始めるデータサイエンス①

Ｒを使って 統計量を計算

- Ｒをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　～相関係数
- Ｒでよく使う関数
- 乱数を生成する　～分布関数

クイズ

A組とB組の違いは？

学生	A組	B組
1	86	70
2	96	90
3	64	80
4	95	81
5	76	76
6	63	81
7	90	82
8	70	80
平均	80.0	80.0

データをＲの配列に代入

```
a <- c(86,96,64,95,76,63,90,70)

b <- c(70,90,80,81,76,81,82,80)

mean(a)
80
mean(b)
80
```

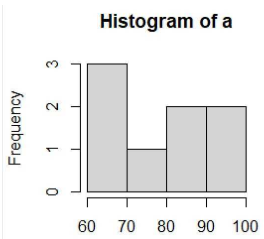
平均を計算したら
同じだった

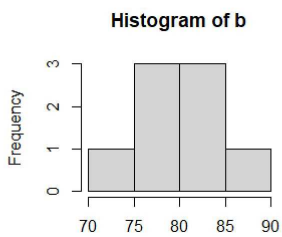
違いは何か？

クイズ

ヒストグラムで違いを調べる

hist(a)　　hist(b)　←Rに入力





←Rの出力

Rに入力→　sd(a)　　sd(b)

Rの出力→　13.5119　　5.6315

ばらつきが大きい

ばらつきが小さい

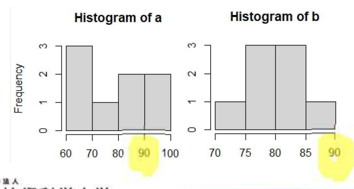
同じ90点でも　価値が違う!?

学生	A組	B組
1	86	70
2	96	**90**
3	64	80
4	95	81
5	76	76
6	63	81
7	**90**	82
8	70	80
平均	80	80
標準偏差	13.19	5.63

偏差値に変換した

学生	A組	B組
1	54.4	32.2
2	61.8	**67.8**
3	38.2	50.0
4	61.1	51.8
5	47.0	42.9
6	37.4	51.8
7	**57.4**	53.6
8	42.6	50.0
平均	**50**	**50**
標準偏差	**10.0**	**10.0**

高価値!?



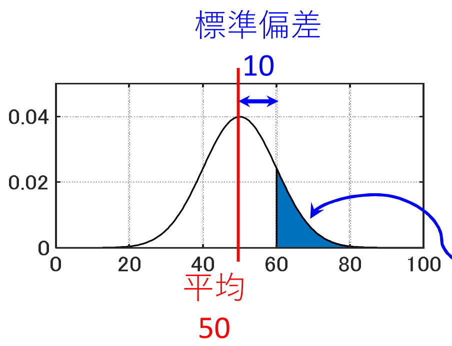

```
aa <- (a-mean(a))/sd(a)*10+50
bb <- (b-mean(b))/sd(b)*10+50
```

標準偏差 と 偏差値

偏差値 $T_i = 50 + \dfrac{x_i - \bar{x}}{s_x} \cdot 10$

データの標準偏差

偏差値の標準偏差

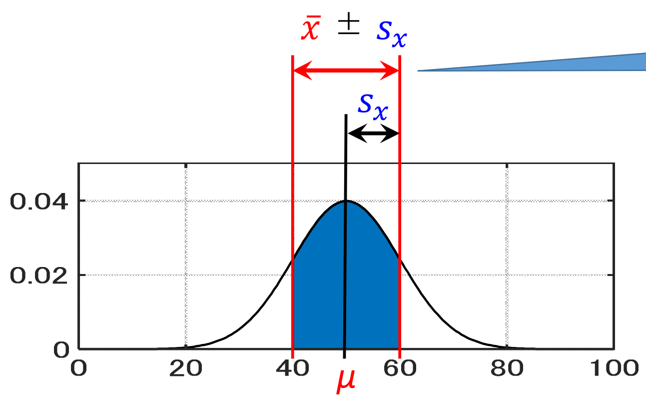


データが、
平均が５０、標準偏差が１０
の正規分布に従うとき、

偏差値60以上は上位15.9%以内

$$\int_{\bar{x}+s_x}^{\infty} f(x)dx = 0.159$$

正規分布 と 標準偏差



この範囲に全体の68.3%が入る

$$\int_{\bar{x}-s_x}^{\bar{x}+s_x} f(x)dx$$

$$= 0.683$$

正規分布

$$f(x) = \frac{1}{\sqrt{2\pi s_x^2}} exp\left(-\frac{(x-\bar{x})^2}{2s_x^2}\right)$$

$\bar{x}$：平均
s_x ：標準偏差

歪度（skewness）

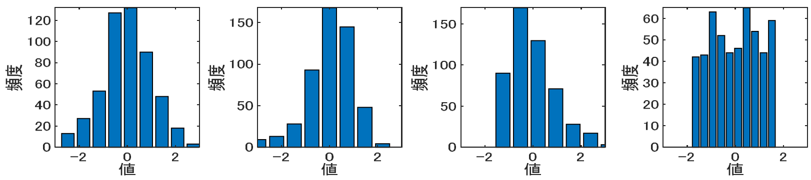
$$w_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}$$

左右対称ならば　0

尖度（kurtosis）

$$k_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4}$$

正規分布ならば　3



歪度 0.0	歪度 -1.1	歪度 1.1	歪度 0.0
尖度 3.2	尖度 5.2	尖度 5.2	尖度 1.8

全て　平均は 0、分散は 1

Ｒで始めるデータサイエンス①

Ｒを使って 統計量を計算

- Ｒをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　　　～相関係数
- Ｒでよく使う関数
- 乱数を生成する　　　　　～分布関数

理科と数学の関係は？

学生	理科	数学
1	85	78
2	96	90
3	64	62
4	90	82
5	76	70
6	80	75
7	73	70
8	70	63
平均	？	？
標準偏差	？	？

Rで平均と標準偏差を計算する

理科

r <- c(85,96,64,90,76,80,73,70)

> mean(r)
79.25
> mean(r)
10.67

数学

s <- c(78,90,62,82,70,75,70,63)

> mean(s)
73.75
> mean(s)
9.51

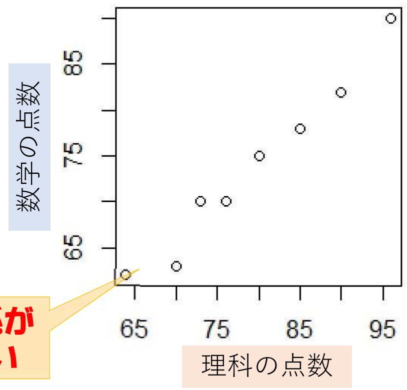
散布図で関係を調べる

学生	理科	数学
1	85	78
2	96	90
3	64	62
4	90	82
5	76	70
6	80	75
7	73	70
8	70	63
平均	79.25	73.75
標準偏差	10.67	9.51

↓Rに入力

plot(r,s)

↓Rの出力



関係が強い

理科ができれば、**数学**もできる

…と言える

クイズ

理科と英語の関係は？

学生	理科	英語
1	85	65
2	96	78
3	64	80
4	90	85
5	76	91
6	80	77
7	73	85
8	70	73
平均	79.25	79.25
標準偏差	10.67	8.05

Rで平均と標準偏差を計算する

理科

r <- c(85,96,64,90,76,80,73,70)

> mean(r)
79.25
> mean(r)
10.67

英語

e <- c(65,78,80,85,91,77,85,73)

> mean(e)
79.25
> mean(e)
8.05

クイズ

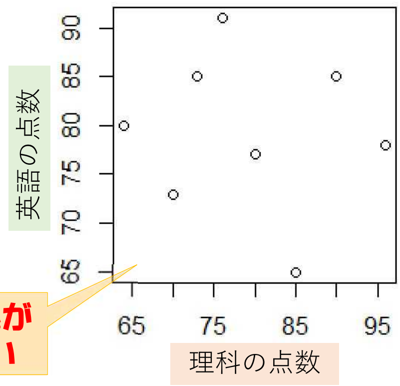
散布図で関係を調べる

学生	理科	英語
1	85	65
2	96	78
3	64	80
4	90	85
5	76	91
6	80	77
7	73	85
8	70	73
平均	79.25	79.25
標準偏差	10.67	8.05

↓Rに入力

plot(r,e)

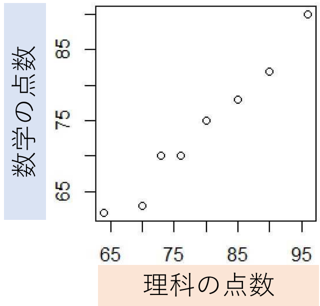
↓Rの出力

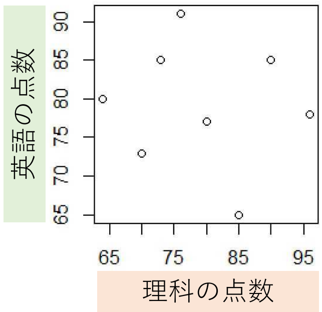


関係が弱い

理科ができれば、**英語**もできる
…と言えない

関係の強さを数字で表す



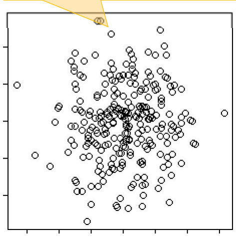


相関係数　> cor(r,s)
0.9855292
関係が強い

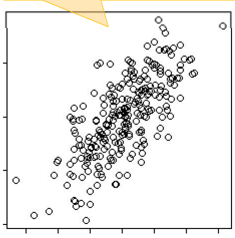
相関係数　> cor(r,e)
－0.117229
関係が弱い

強い関係 → 決定係数が 1

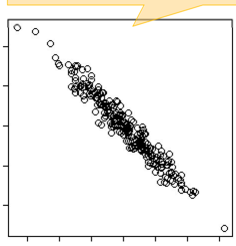
関係が無い



関係がある

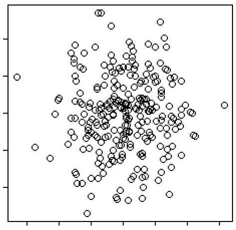


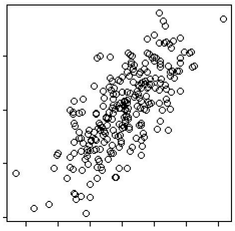
負の関係が強い

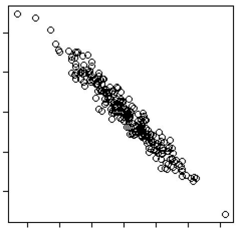


cor (x, y)		cor (x, x+y)	cor (x, y−4*x)
0. 03	相関係数	0. 76	−0. 97
cor (x, y) ^2		cor (x, x+y) ^2	cor (x, y−4*x) ^2
0. 00	決定係数	0. 58	0. 94

データをどのように生成したか







```
x <- rnorm(256, 0, 1)
y <- rnorm(256, 0, 1)
plot(x, y)
```

横軸 $= x_i$
縦軸 $= y_i$
$i = 1,2,\cdots,n$

```
plot(x, x+y)
```

横軸 $= x_i$
縦軸 $= x_i + y_i$

```
plot(x, y-4*x)
```

横軸 $= x_i$
縦軸 $= -4x_i + y_i$

乱数の統計量

```
x <- rnorm(4096, 0, 1)
y <- rnorm(4096, 0, 1)
```

	データ数	平均	分散
x_i	$n = 4096$	$\bar{x} = 0$	$s_x^2 = 1$
y_i	$n = 4096$	$\bar{y} = 0$	$s_y^2 = 1$

$i = 1,2,\cdots,n$

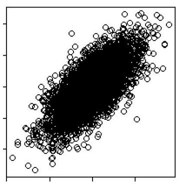
cor(x, y)
0.003

$$\frac{s_{xy}}{s_x s_y} = 0$$

mean(x)
0.01
mean(y)
-0.02

var(x)
1.01
var(y)
1.03

```
plot(x, x+y)
```



cor(x, x+y)
0.705

横軸 $= x_i$
縦軸 $= x_i + y_i$

クイズ

相関係数の理論値

Rで計算した結果、 cor(x, x+y) 0.705 となった。これを理論的に示そう。

x と z = x + y の共分散は

$$s_{xz} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})$$

x と z=x+y の相関係数は

$$r_{xz} = \frac{s_{xz}}{s_x s_z}$$

に

$$\bar{x} = \bar{y} = 0$$
$$s_x = s_y = 1$$
$$s_{xy} = 0$$

を代入してみよう。

☜ cor(x, x+y) **相関係数**

クイズ

相関係数の理論値（解説1/2）

$$s_{xz} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x} + y_i - \bar{y})$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad = \quad s_x^2 + s_{xy}$$

$$s_z^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + y_i - \bar{y})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad = \quad s_x^2 + 2s_{xy} + s_y^2$$

クイズ

相関係数の理論値（解説2/2）

$$s_{xz} = s_x^2 + s_{xy}$$
$$s_z^2 = s_x^2 + 2s_{xy} + s_y^2$$

$$r_{xz} = \frac{s_{xz}}{s_x s_z} = \frac{s_x^2 + s_{xy}}{s_x\sqrt{s_x^2 + 2s_{xy} + s_y^2}}$$

$$s_x = s_y$$
$$s_{xy} = 0$$

$$r_{xz} = \frac{1}{\sqrt{2}} = \mathbf{0.707}$$ ☜ **理論値**

相関係数　cor(x, x+y)

0.705 ☜ **実験値**

R で統計量を計算

データ１　x_i , i=1,2,3,・・・,n

データ２　y_i

↓Rに入力

不偏共分散　$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$　cov(x,y) covariance

相関係数　$r = \frac{s_{xy}}{s_x s_y}$　cor(x,y) correlation

同じ

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

統計量の別計算

> cov(r,s)
100.0714

同じ

> sum((r-mean(r))*(s-mean(s)))/(8-1)
100.0714

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

> cor(x,y)
0.9855292

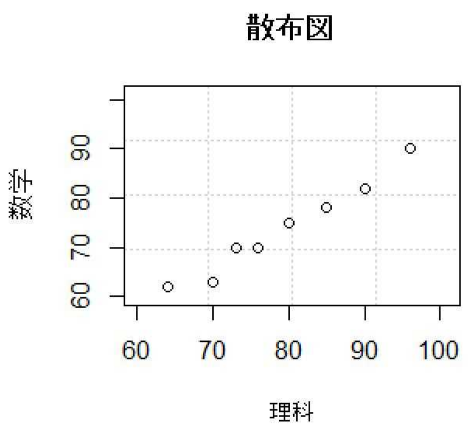
同じ

> cov(r,s)/sd(r)/sd(s)
0.9855292

$$r = \frac{s_{xy}}{s_x s_y}$$

plot の詳細

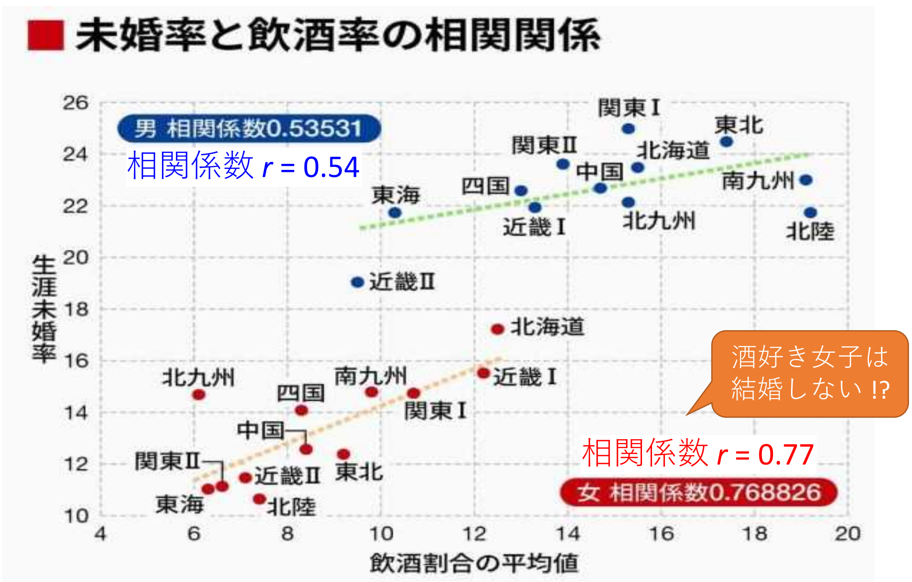
> plot(r,s,xlab="理科",ylab="数学",main="散布図
",panel.first=grid(4,4),xlim=c(60,101) ,ylim=c(60,101))



Rで始めるデータサイエンス①

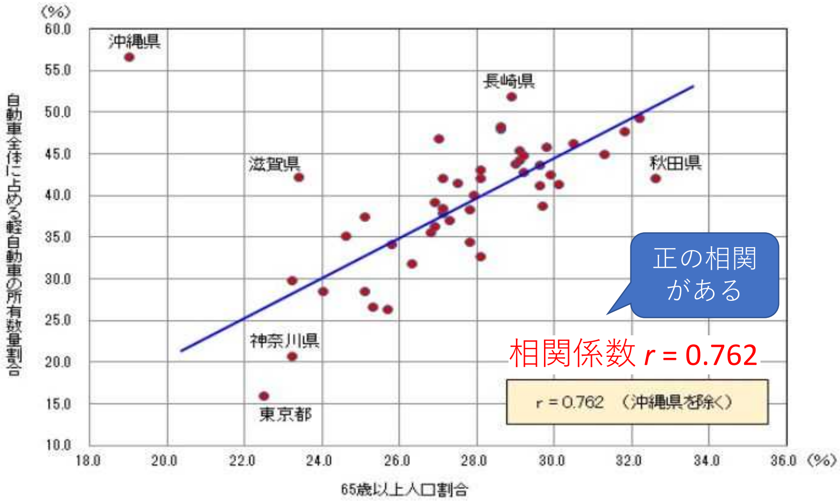
Ｒを使って
統計量を計算

- Ｒをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　　**～疑似相関**
- Ｒでよく使う関数
- 乱数を生成する　　～分布関数



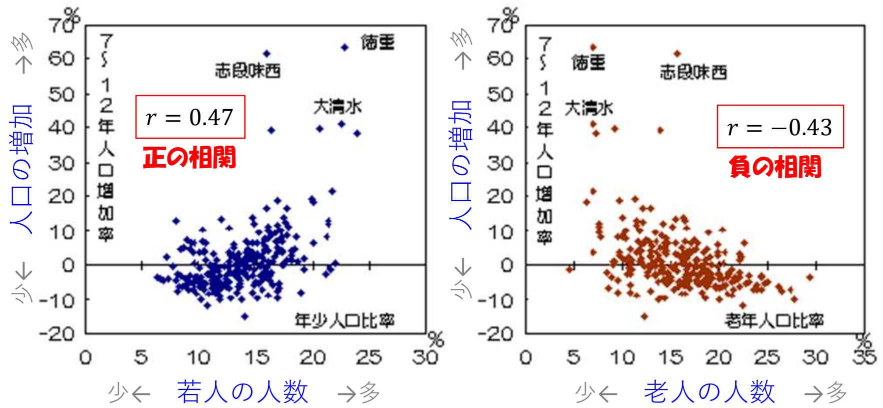
出典：東洋経済ONLINE、2018.10.28の記事　https://toyokeizai.net/articles/-/244802?page=3

「老人の人数」と「軽自動車の台数」



出典：総務省統計局、2018.10.28の記事 https://www.stat.go.jp/info/today/102.html

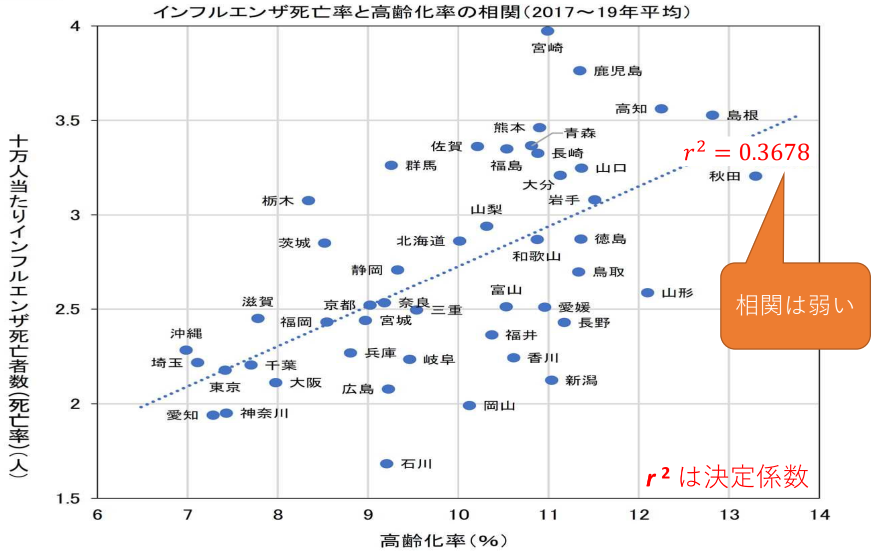
「人口の増加」と「住人の年齢」



出典：名古屋市役所、国勢調査 http://www.city.nagoya.jp/somu/page/0000003971.html

図表5 高齢化の進んだ地域ほどインフル死亡が多く、寒暖差は無関係

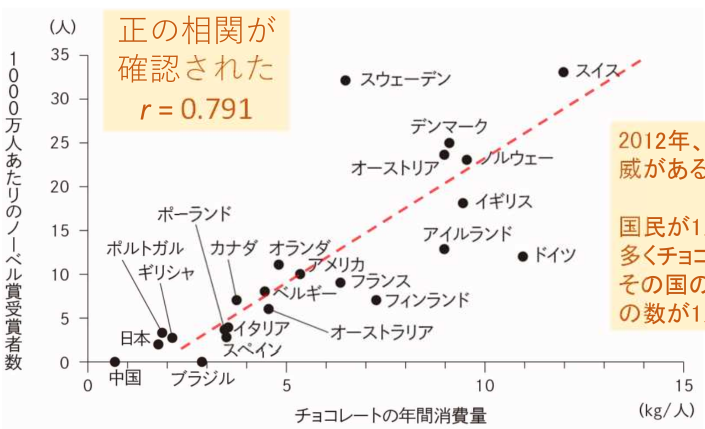
インフルエンザ死亡率と高齢化率の相関（2017〜19年平均）



（注）ここで高齢化率は75歳以上人口比率。使用した人口は2018年推計人口（10月1日日本人人口）
（資料）厚生労働省「人口動態統計」、総務省統計局「推計人口」

出典：PRESIDENT Online、2020.3.6の記事 https://president.jp/articles/-/33466?page=5

チョコレートの消費量が多い国ほど ノーベル賞の受賞者数が多い！？



2012年、臨床医学で最も権威がある雑誌に掲載された

国民が1人あたり年間400g多くチョコレートを摂取するとその国のノーベル賞受賞者の数が1人増える

Messerli, F. H. (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates, The New England Journal of Medicine, 367, 1562-1564.

ダイヤモンド・オンライン： https://diamond.jp/articles/-/124862

疑似相関 と 交絡因子

チョコレートの消費量 —因果関係がある→ ノーベル賞の受賞者数

…とは言えない

裕福な国ほど
消費量が多い

国が裕福なほど
教育費をかける

国のGDP
（交絡因子）

”As we suspected, it turned out that the GDP strongly correlated both with the number of Nobel laureates (r= 0.66; P < 0.001) and chocolate consumption (r= 0.73; P< 0.001)”. Maurage, P., et.al. (The Journal of Nutrition, 2013)

相関がある —必ずしも成立しない→ 因果がある

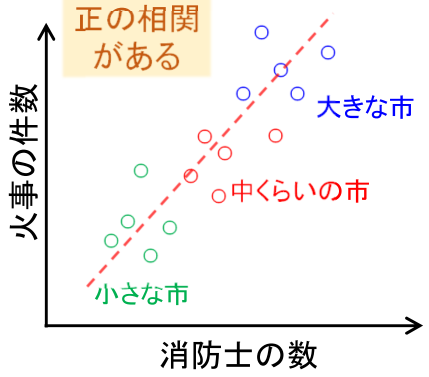
←成立する—

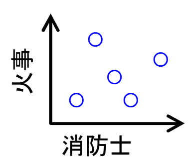
相関がある：
A の値が大きくなると
B の値も大きくなる

因果がある：
事象 A が生じると
事象 B が生じる

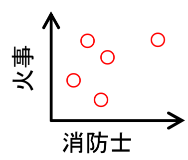
消防士が多いと 火事が多い？

正の相関がある

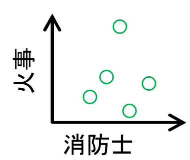




大きな市
相関なし



中くらいの市
相関なし



小さな市
相関なし

市の大きさ（人口の多さ）が
交絡因子となっている

統計に騙されない

年賀状を出す人ほど、高収入
沢山稼ぐと、高血圧になる
早起きすると、お金持ちになる

→「年齢」が交絡因子

アイスが売れると、熱中症が増える
ビールが売れると、水難事故が増える

→「気温」が交絡因子

図書館が多いと、犯罪が多い
消防士が多いと、火事が多い
コンビニが増えると、犯罪も増える

→「人口」が交絡因子

Rで始めるデータサイエンス①

Rを使って統計量を計算

- Rをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　　～相関係数
- Rでよく使う関数
- 乱数を生成する　　～分布関数

ベクトル

```
> x <- 0:5
> x
[1] 0 1 2 3 4 5

> x <- 5:0
> x
[1] 5 4 3 2 1 0

> x <- c(1,2)
> x
[1] 1 2
```

成分を取り出す

```
> x <- 1:5
> x
[1] 1 2 3 4 5

> x[1:3]
[1] 1 2 3

> x[3:5]
[1] 3 4 5
```

```
> seq(0,5,by=2)
[1] 0 2 4

> seq(0,2,by=0.5)
[1] 0.0 0.5 1.0 1.5 2.0

> x <- c(11,20)
> y <- c(31,24,98)
> c(x,y)
[1] 11 20 31 24 98
```

ベクトルを繋げる

行列

```
x <- matrix(1:6, nrow=3)
> x
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

```
y <- matrix(1:6, ncol=3)
> y
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
> x[2,1:2]
[1] 2 5
```

```
> x[2,1]
[1] 2
```

成分を取り出す

転置

```
> t(y)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

対角行列

```
diag(2)
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

行列の演算

```
x <- matrix(1:6, nrow=3)
x
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

```
y <- matrix(1:6, ncol=3)
y
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
u <- diag(2)
u
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

要素ごとの積

```
x * x
     [,1] [,2]
[1,]    1   16
[2,]    4   25
[3,]    9   36
```

行列の積

```
x %*% y
     [,1] [,2] [,3]
[1,]    9   19   29
[2,]   12   26   40
[3,]   15   33   51
```

行列の積

```
x %*% u
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```


行列の結合

```
x <- matrix(1:6, nrow=3)
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

```
y <- matrix(1:6, ncol=3)
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
u <- diag(2)
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
> rbind(x,u)
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
[4,]    1    0
[5,]    0    1
```

```
> cbind(y,u)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    1    0
[2,]    2    4    6    0    1
```

行列の結合（行）

行列の結合（列）

行列の行と列に名前をつける

```
x <- matrix(1:6, nrow=3)
x
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

```
colnames(x) <- c("A","B")
x
     A B
[1,] 1 4
[2,] 2 5
[3,] 3 6
```

列の名前

```
rownames(x) <- c("1st","2nd","3rd")
x
    A B
1st 1 4
2nd 2 5
3rd 3 6
```

行の名前

行列の演算

rowSums(x)	x の各行の総和
solSums(x)	x の各列の総和
rowMeans(x)	x の各行の平均
colMeans(x)	x の各列の平均
t(x)	x を転置
solve(x)	x の逆行列
eigen(x)	x の固有値と固有ベクトル
det(x)	x の行列式

Rで始めるデータサイエンス①

Rを使って
統計量を計算

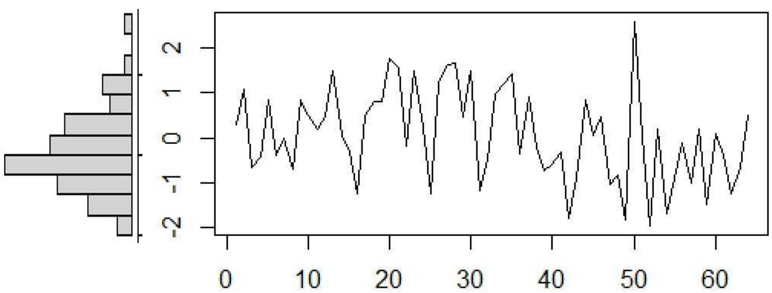
- Rをインストールする
- 平均は同じでも違いはある
- 関係がある？ない？　～相関係数
- Rでよく使う関数
- 乱数を生成する　～分布関数

正規分布する乱数

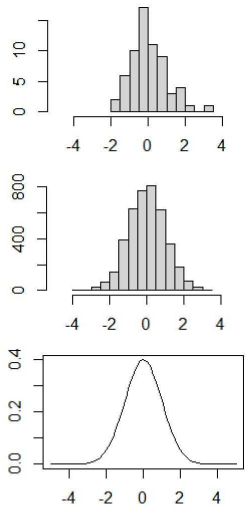
```
m=0
s=1
```
平均
分散

```
x <- rnorm(64, m, s)

plot(x, type="l")
```



正規分布する乱数



m=0　平均
s=1　分散

x <- rnorm(64,m,s)
hist(x,xlim=c(-5,5))

ヒストグラム
データ数64

x <- rnorm(4096,m,s)
hist(x,xlim=c(-5,5))

ヒストグラム
データ数4096

curve(dnorm(x,m,s),-5,5)

正規分布

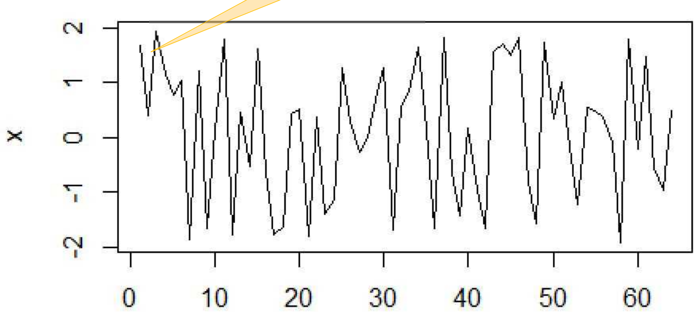
確率密度関数

一様分布する乱数

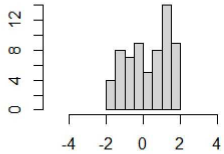
x <- runif(64,-2,2)

plot(x, type="l")

線でつなぐ　小文字のエル

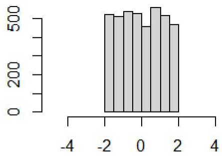


一様分布する乱数



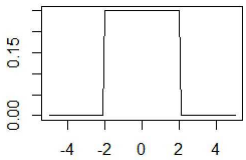
```
x <- runif(64,-2,2)
hist(x,xlim=c(-5,5))
```

ヒストグラム
データ数64



```
x <- runif(4096,-2,2)
hist(x,xlim=c(-5,5))
```

ヒストグラム
データ数4096

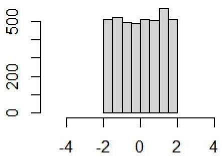


```
curve(dunif(x,-2,2),-5,5)
```

確率密度関数

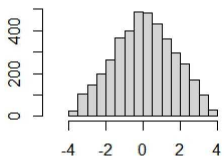
一様分布

一様分布する乱数の和



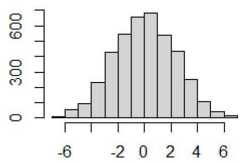
```
x <- runif(4096,-2,2)
hist(x,xlim=c(-5,5))
```

ヒストグラム
１つの乱数



```
x <- runif(4096,-2,2)
y <- runif(4096,-2,2)
hist(x+y,xlim=c(-5,5))
```

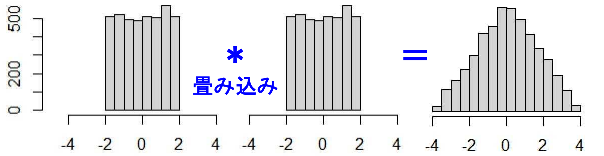
ヒストグラム
２つの乱数の和



```
x <- runif(4096,-2,2)
y <- runif(4096,-2,2)
z <- runif(4096,-2,2)
w <- runif(4096,-2,2)
hist(x+y+z+w,xlim=c(-7,7))
```

ヒストグラム
４つの乱数の和

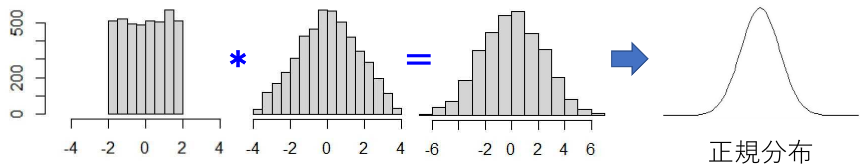
乱数の和の分布



１つの乱数の
ヒストグラム

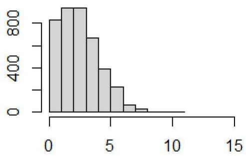
別の乱数の
ヒストグラム

２つの乱数の和
のヒストグラム



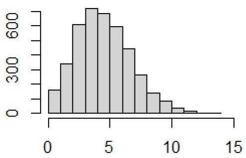
正規分布
に近づく

ポアソン分布する乱数




```
> x <- rpois(4096, lambda=3)
> hist(x, xlim=c(0, 15))
```

ヒストグラム
λ = 3

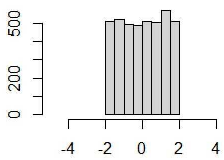



```
> x <- rpois(4096, lambda=5)
> hist(x, xlim=c(0, 15))
```

ヒストグラム
λ = 5

クイズ

２つの乱数の和




```
> x <- runif(4096, -2, 2)
> var(x)
 1.3
> y <- runif(4096, -2, 2)
> var(y)
 1.3
```

乱数 x
（分散＝1.3）

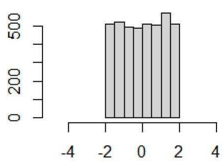
乱数 y
（分散＝1.3）

問題１　乱数 x と乱数 y の和の分散は？

問題２　乱数 x と乱数 x の和の分散は？

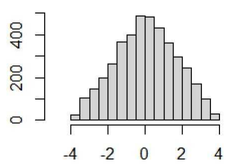
クイズ

Ｒによる実験



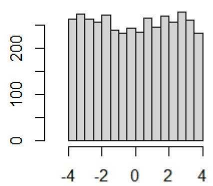

```
> x <- runif(4096, -2, 2)
> var(x)
 1.308826
```

１つの乱数




```
> x <- runif(4096, -2, 2)
> y <- runif(4096, -2, 2)
> var(x+y)
 2.641422
```

問題１
２つの乱数の和




```
> x <- runif(4096, -2, 2)
> var(x+x)
 5.400876
```

問題２
２倍した乱数

クイズ

理論的考察

$$s_{x+y}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + y_i - \bar{y})^2$$

$$s_x^2 = s_y^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad = \quad s_x^2 + 2s_{xy} + s_y^2$$

問題１　乱数ｘと乱数ｙの和の分散は？

xとyが無相関ならs_{xy}=0なので、　$s_{x+y}^2 = 2\ s_x^2$

問題２　乱数ｘと乱数ｘの和の分散は？

xとxは相関が1でありs_{xx}=$s_x{}^2$ なので、　$s_{x+x}^2 = 4\ s_x^2$

初版：　2022年7月

制作：　岩橋政宏
所属：　長岡技術科学大学