

## rd-6. 相関, 相関係数

### データサイエンス演習 (R システムを使用)

<https://www.kkaneko.jp/de/rd/index.html>

金子邦彦



- **相関**は、2つの変数の間に関連性があるかを示す
- **相関がある**場合、一方が変化すると、もう一方も変化  
する傾向にある

## 【相関ありの場合】

- Xが増えると、Yが増える傾向がある（**正の相関**）  
勉強時間が増えると、得点上がる
- Xが増えると、Yが減る傾向がある（**負の相関**）  
ガソリン代が上がると、車の利用が減る

## 【相関なしの場合】

XとYに関係がない

足のサイズと勉強時間に関係がない

# 相関係数



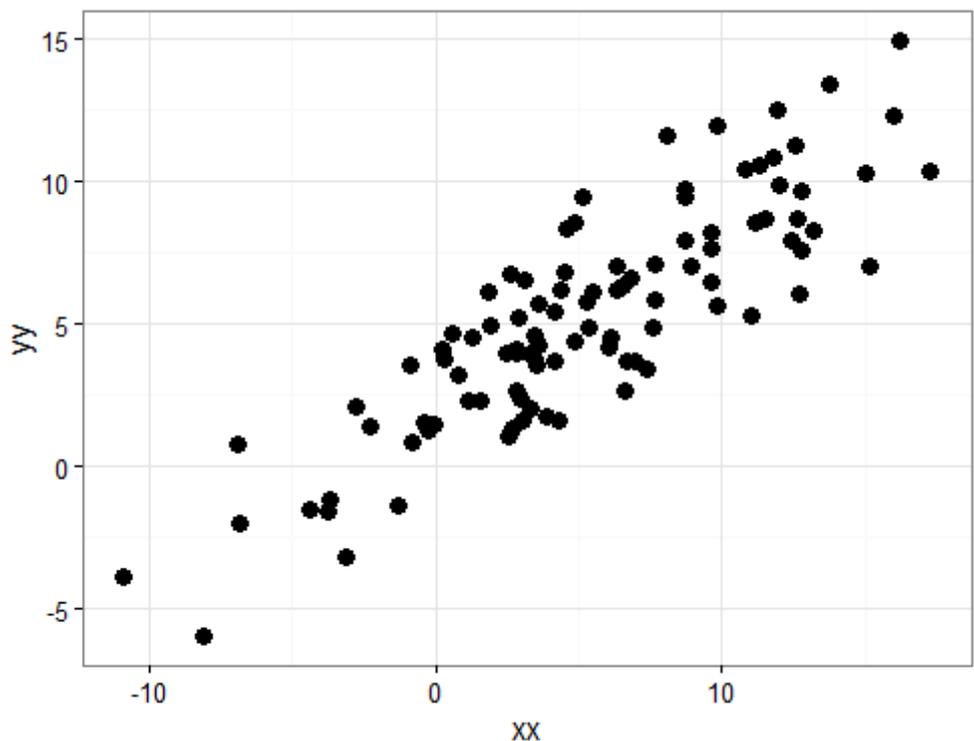
- **相関係数**は、**相関**を算出した**数値**。範囲は**-1から1まで**
- 相関係数を算出することで、変数間の関係の分析ができる

**1に近い値**： **相関**あり。正の相関

**0に近い値**： **相関なし**

**-1に近い値**： **相関**なし。負の相関

# 正の相関



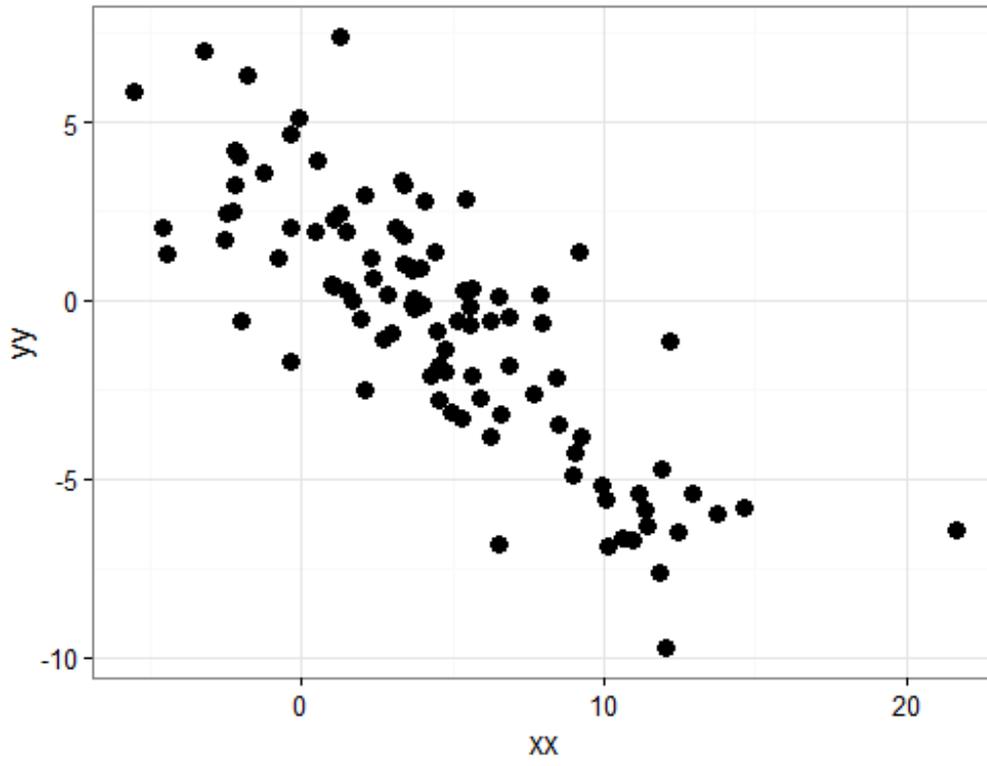
2つの変数  $xx$ ,  $yy$  に相関がある

$xx$  の値が増えると  
 $yy$  の値が増える傾向がある  
(正の相関)

相関係数の算出結果

0.8620027 (1 に近い値)

# 負の相関



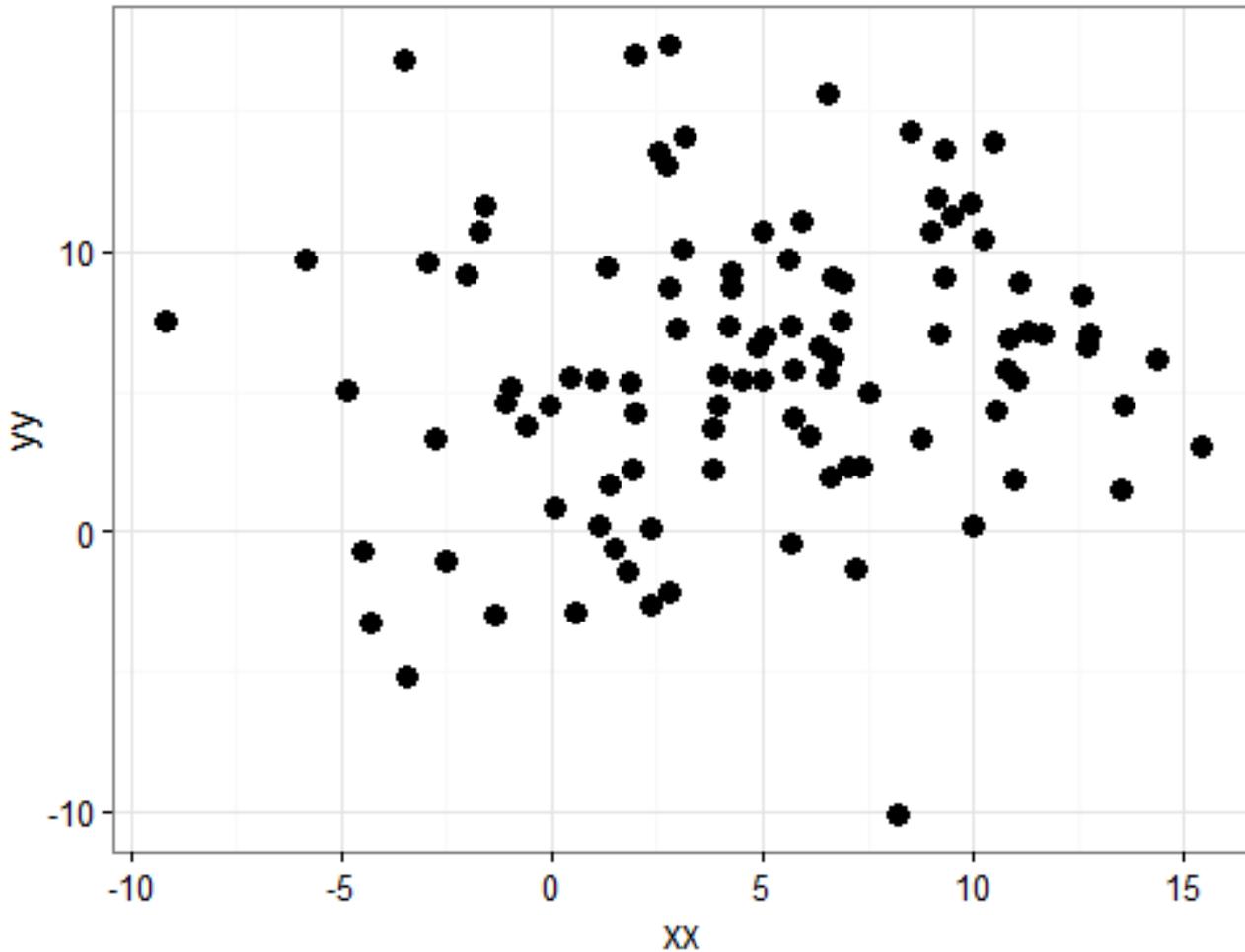
2つの変数  $xx$ ,  $yy$  に  
相関がある

$xx$  の値が増えると  
 $yy$  の値が減る傾向がある  
(負の相関)

相関係数の算出結果

**-0.8502535 (-1 に近い値)**

# 相関なし



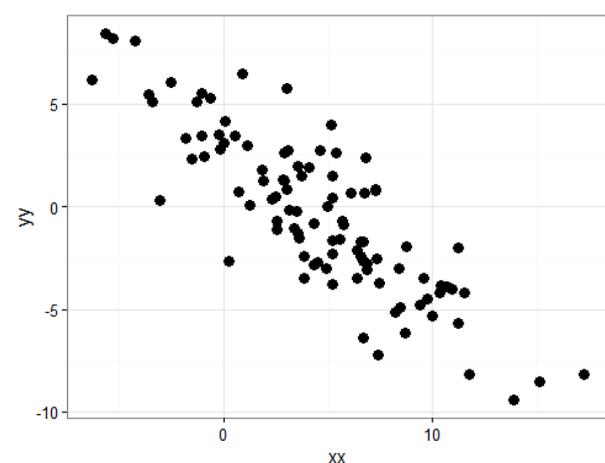
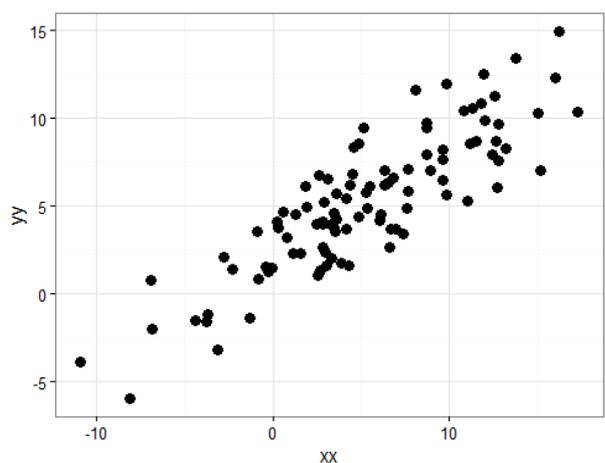
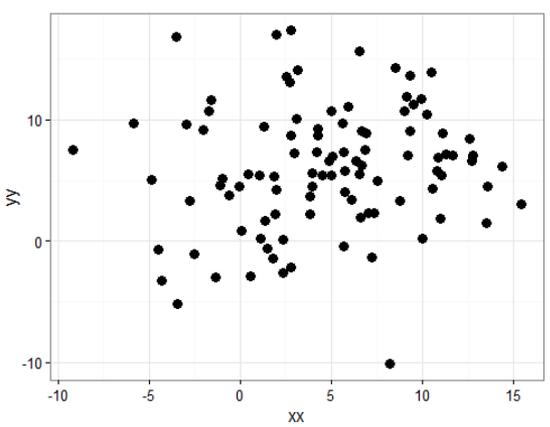
相関係数の算出結果

0.1252164 (0 に近い値)

# 相関係数のまとめ



- 1に近い値： 相関あり。 正の相関
- 0に近い値： 相関なし
- -1に近い値： 相関なし。 負の相関



0.1252164 (0に近い値)

相関なし

0.8620027 (1に近い値)

正の相関

-0.8502535 (-1に近い値)

負の相関

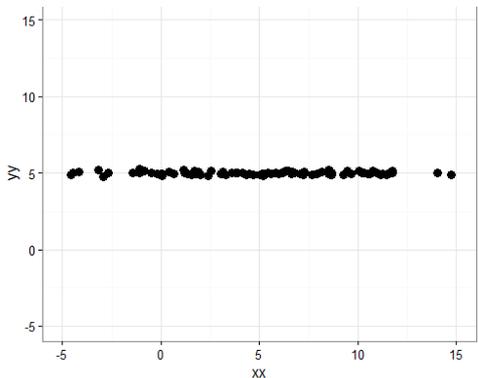
## 2つの量の間の関係性の分析

- 広告を増やすと，売上高が増えそうか
- 相関が高い複数の金融商品を扱うと，リスクが高いか
- 遺伝子と疾患に関係がありそうか

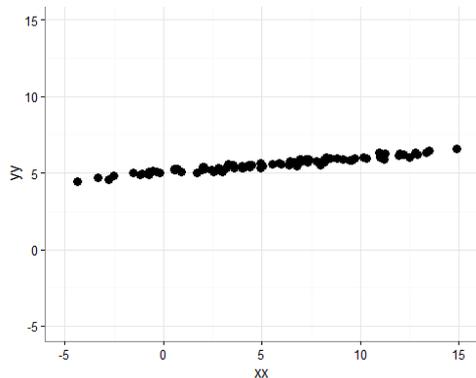
# 相関係数の性質



「相関の強弱」の尺度である。「傾き」ではない

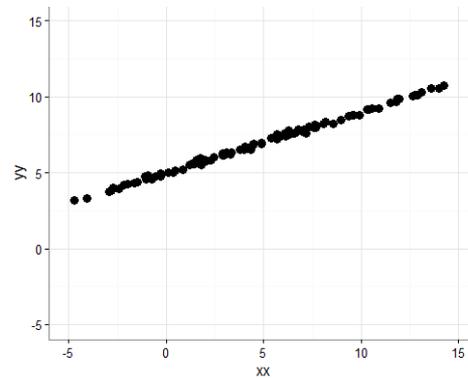


```
> cor(d9$xx, d9$yy)
[1] -0.06027409
```



```
> cor(d10$xx, d10$yy)
[1] 0.9743955
```

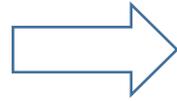
1 に近い値



```
> cor(d11$xx, d11$yy)
[1] 0.998944
```

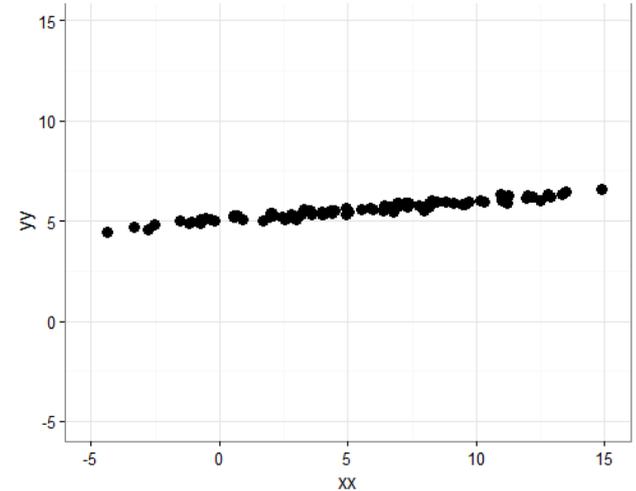
1 に近い値

# 合成データからランダムに100個選び標本を作る



サイズ **100**  
の標本を2セット

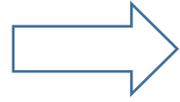
タイプ : 数値 (整数化しない)  
サイズ : **100,000**



```
x2 <- rnorm(100000, mean=5, sd=5)
y2 <- rnorm(100000, mean=5, sd=0.1)
d10 <- data.frame( xx=x2[floor( runif(100, 1, 100000+1) )],
  yy=y2[floor( runif(100, 1, 100000+1) )] )
d10$yy <- 0.1 * d10$xx + d10$yy
library(ggplot2)
ggplot(d10, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d10$xx, d10$yy)
```

合成データに,  
正の相関関係をもたせる

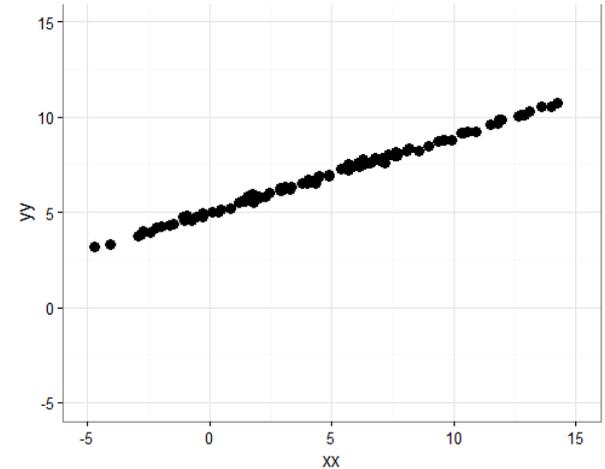
# 合成データからランダムに100個選び標本を作る



サイズ **100**  
の標本を2セット

タイプ：数値（整数化しない）

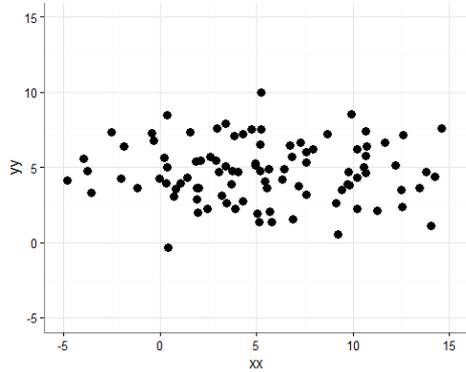
サイズ：**100,000**



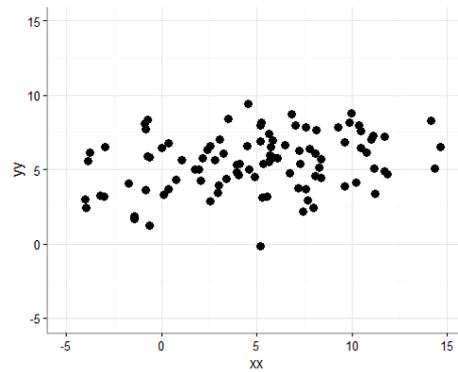
```
x2 <- rnorm(100000, mean=5, sd=5)
y2 <- rnorm(100000, mean=5, sd=0.1)
d11 <- data.frame( xx=x2[floor( runif(100, 1, 100000+1) )],
  yy=y2[floor( runif(100, 1, 100000+1) )] )
d11$yy <- 0.4 * d11$xx + d11$yy
library(ggplot2)
ggplot(d11, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d11$xx, d11$yy)
```

合成データに、  
正の相関関係をもたせる

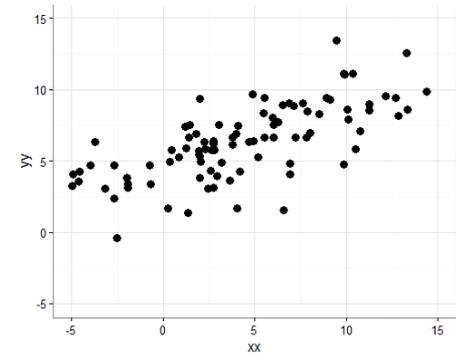
# 相関係数の例



```
> cor(d12$xx, d12$yy)
[1] -0.02723688
>
```



```
> cor(d13$xx, d13$yy)
[1] 0.2808435
>
```



```
> cor(d14$xx, d14$yy)
[1] 0.7268933
>
```

# おわりに



- 相関がある場合、一方が変化すると、もう一方も変化する傾向にある
  - 1に近い値： 相関あり。 正の相関
  - 0に近い値： 相関なし
  - -1に近い値： 相関なし。 負の相関
- 3つ以上の変数があるとき、相関係数は多数求まる  
変数 A, B, C に対して
  - A と B の相関係数,
  - B と C の相関係数,
  - C と A の相関係数