# 相関関係 (散布図)

Code

**correlation**

## 相関と散布図

２次元 data と散布図

- data 同士の関係を図で表す

| 日 | ビールの販売数 | 気温 |
|---|---|---|
| 1 | 64 | 26.6 |
| 2 | 53 | 22.4 |
| 3 | 58 | 24.4 |

## 散布図

- 2変数の **相関関係** がわかる

相関の出発点は　→　**散布図**

## 散布図と相関

- 散布図から **5つの相関** の種類分けが出来る
    - **強い相関**
        - 右肩上がり
        - 片方が上がれば, もう一方も上がる
        - data が直線のように密集している
    - **弱い相関**
        - 右肩上がり
        - 片方が上がれば, もう一方も上がる
        - data がまばら
    - **無相関**
        - 全体的に散らばっている(円形の様に)
        - 特に x軸, y軸との関係はみられない
    - **強い負の相関**
        - 右肩下がり
        - 片方が上がれば, 片方が下がる
        - 直線のように密集している
    - **弱い負の相関**
        - 右肩下がり
        - 片方が上がれば, 片方が下がる
        - data がまばらに

## 相関の記述

Hide

```
cd = cars
cd
```

| speed<br><dbl> | dist<br><dbl> |
|---|---|
| 4 | 2 |
| 4 | 10 |
| 7 | 4 |

| speed<br><dbl> | dist<br><dbl> |
|---|---|
| 7 | 22 |
| 8 | 16 |
| 9 | 10 |
| 10 | 18 |
| 10 | 26 |
| 10 | 34 |
| 11 | 17 |

1-10 of 50 rows Previous **1** 2 3 4 5 Next

Hide

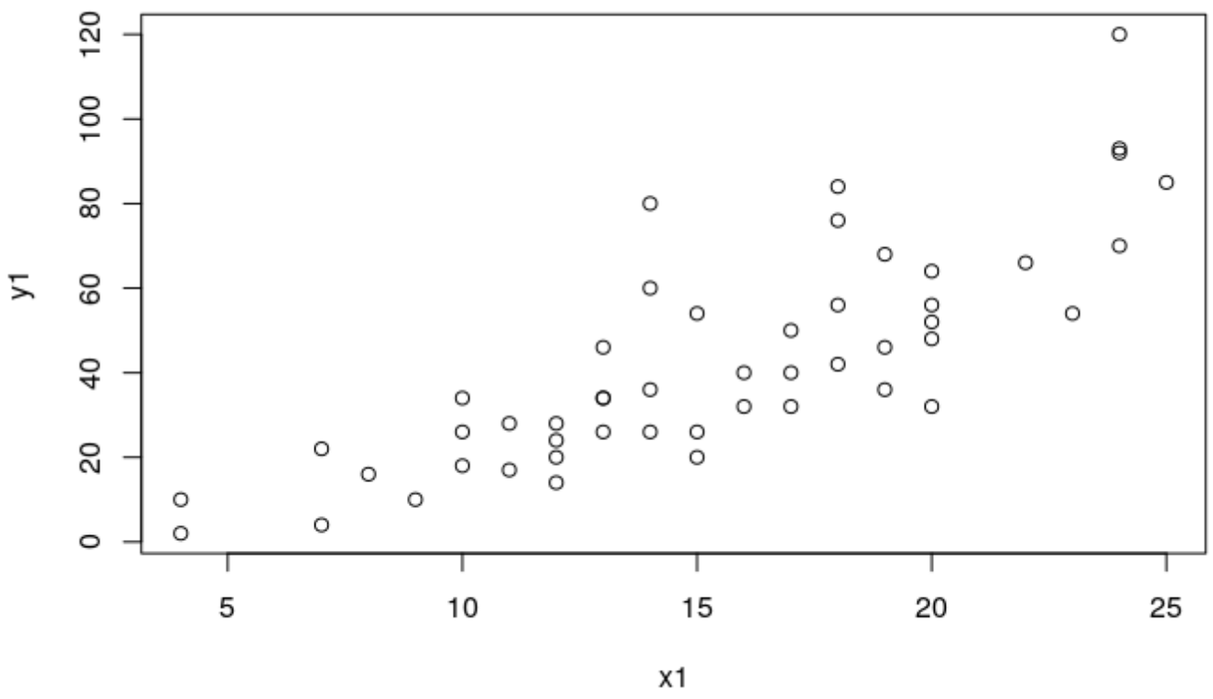
```
x1 <- cd$speed
y1 <- cd$dist
```

# 散布図 plot

- **speed(速度)** : x1軸
- **dist(制動距離)** : y1軸
    - **相関関係**　があることが窺える

Hide

```
plot(x1, y1)
```

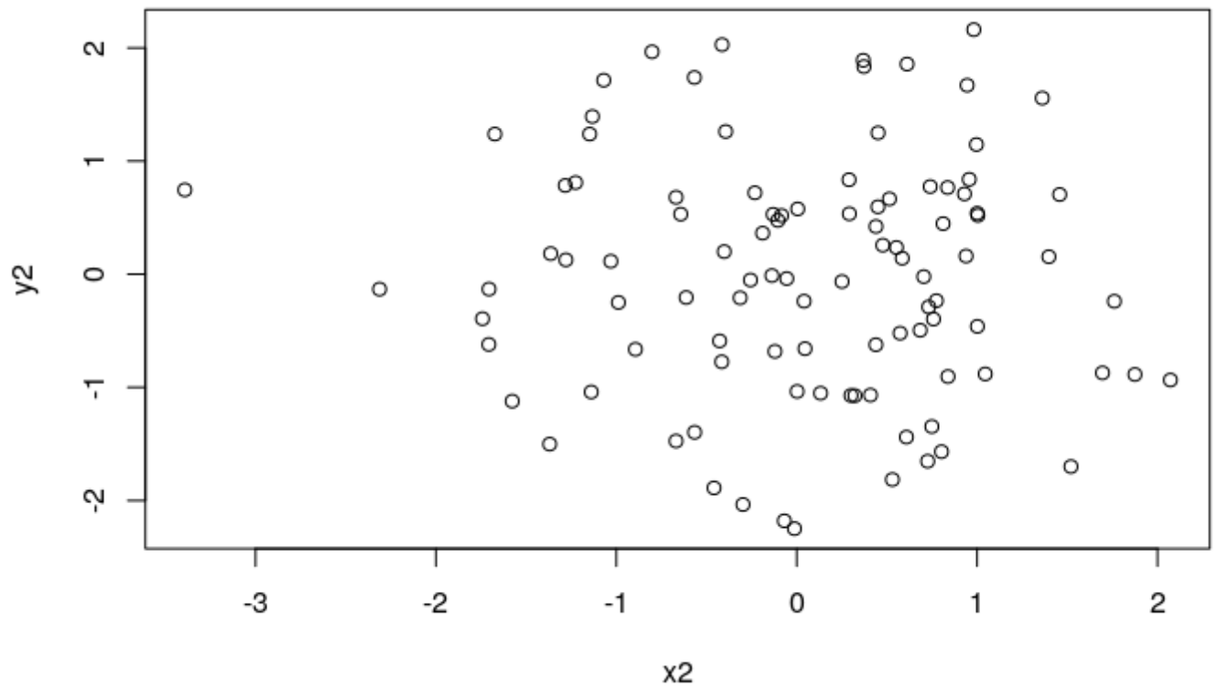



# 無相関

- 乱数を発生させた為に data に **相関関係が窺えない**

Hide

```
x2 <- rnorm(100, mean = 0, sd =1)
y2 <- rnorm(100, mean = 0, sd =1)
plot(x2, y2)
```




# 相関係数

2つの変数の強さを -1 ~ 1 の数値で示す統計量

n 個の変数 x と変数 y がある時の相関係数は r

↓

- **標準偏差　＝　**$\sigma$
- **変数　＝　**$n$**個**
- **変数　＝　**$x, y$

$$r \quad = \quad \frac{\text{共分散}}{x\sigma \times y\sigma}$$

# 公式

$$r_{xy} \quad = \quad \frac{\sum(x_i - \bar{x})(y - \bar{y})/n}{\sqrt{\sum(x_i - \bar{x})^2/n}\sqrt{\sum(y_i - \bar{y})^2/n}}$$

$$= \quad \frac{\sum(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

**散布図**

+

**相関係数**

↓

セットでみることでより data の結びつきがわかる

# 相関係数を求める

- 相関係数は　**-1**　~　**1**　の間で表す
  - ー(マイナス)は **負の相関**
  - ＋(プラス)は **正の相関**

## speed(速さ) と dist（制動距離）相関係数

- **相関係数** : 0.8068949
    - 結構強めの相関があることが窺える

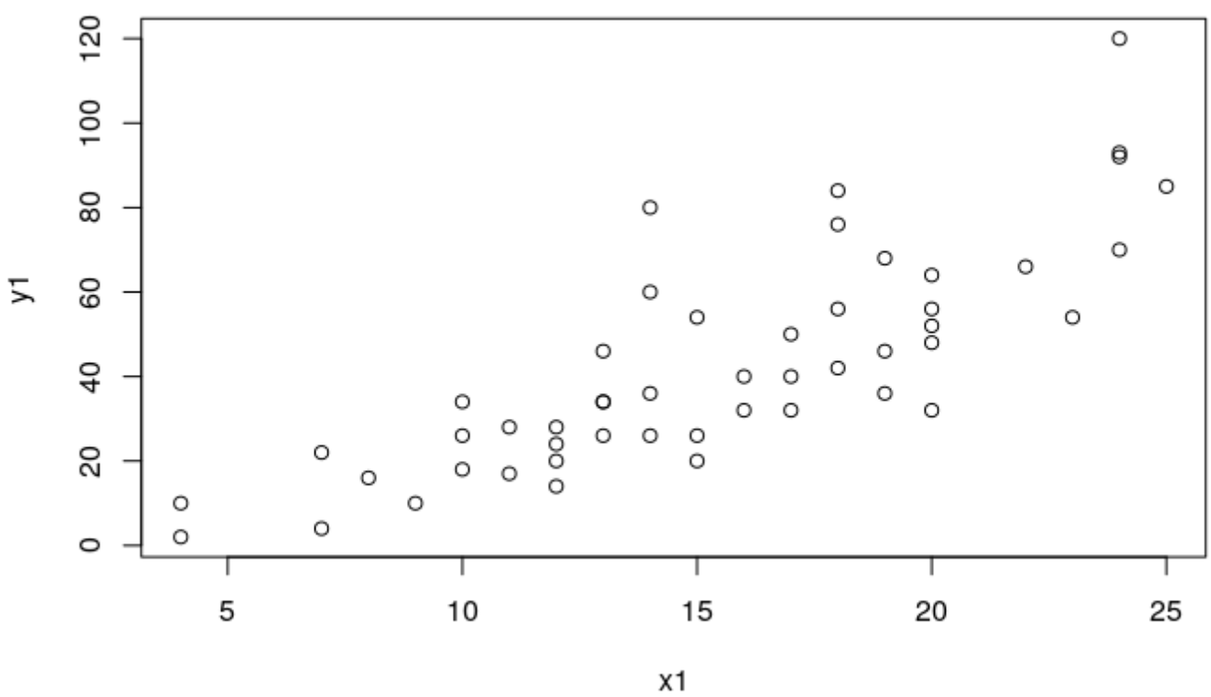
Hide

```
cor(x1, y1)
```

```
[1] 0.8068949
```

Hide

```
plot(x1, y1)
```




# 無相関の相関係数

- **相関係数** : -0.0596985
    - 相関関係は全く窺えない

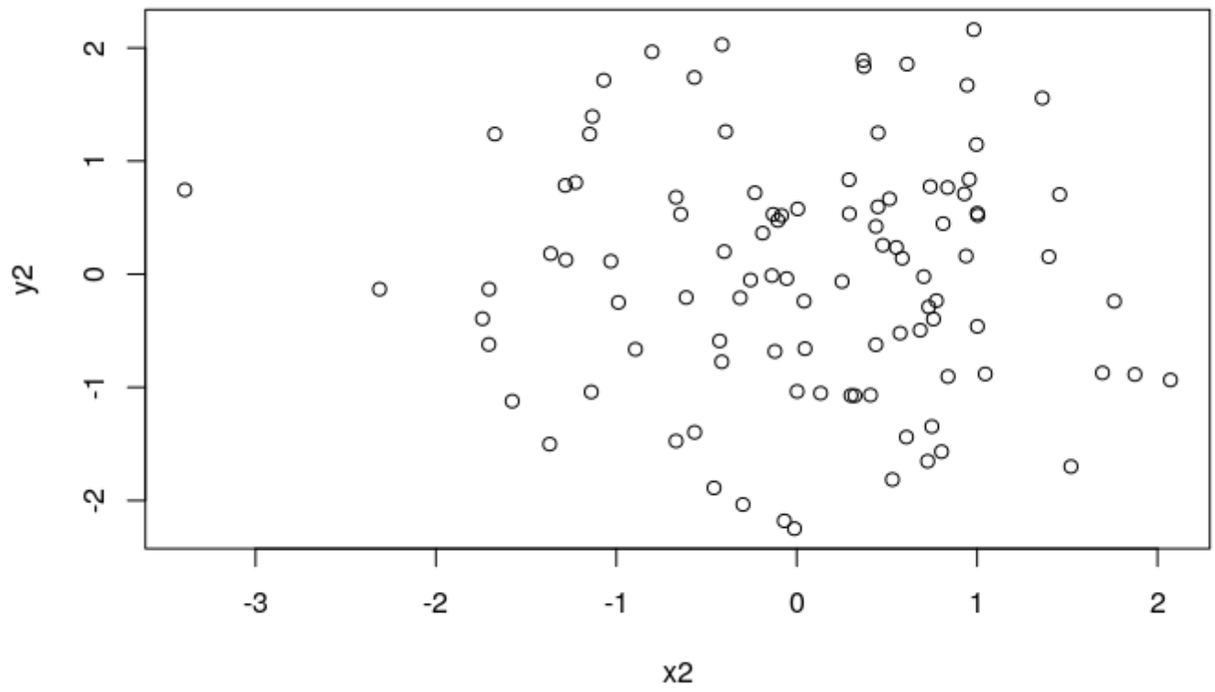
Hide

```
cor(x2, y2)
```

```
[1] -0.05969848
```

Hide

```
plot(x2, y2)
```




# 相関係数２

## - 相関係数の注意点１ -

- **一般的な相関の強さと相関係数の値**

| 値 | 相関 |
|---|---|
| 0 < r ≦ ±0.2 | ほとんど相関無し |
| ±0.2 ≦ r ≦ ±0.4 | 弱い相関無し |
| ±0.4 ≦ r ≦ ±0.7 | 相関有り |
| ±0.7 ≦ r ≦ ±1.0 | 強い相関有り |

同じ相関係数でも　**data数**，　**外れ値**　でだいぶ印象が変わる

↓

**必ず散布図とセットで確認すること**

## - 相関係数の注意点２ -

- 相関関係と因果関係

**相関関係　≠　因果関係**

↓

相関関係があるからといって **必ず因果関係があるわけではない**

### 因果関係が成立する条件

1. **相関関係がある**
2. **時間的順序**
3. **第3因子が存在しない**

### グラフで確認

- **sample数** : 30

- **相関係数** : 0
- **正規分布に従うdata** : x, e

Hide

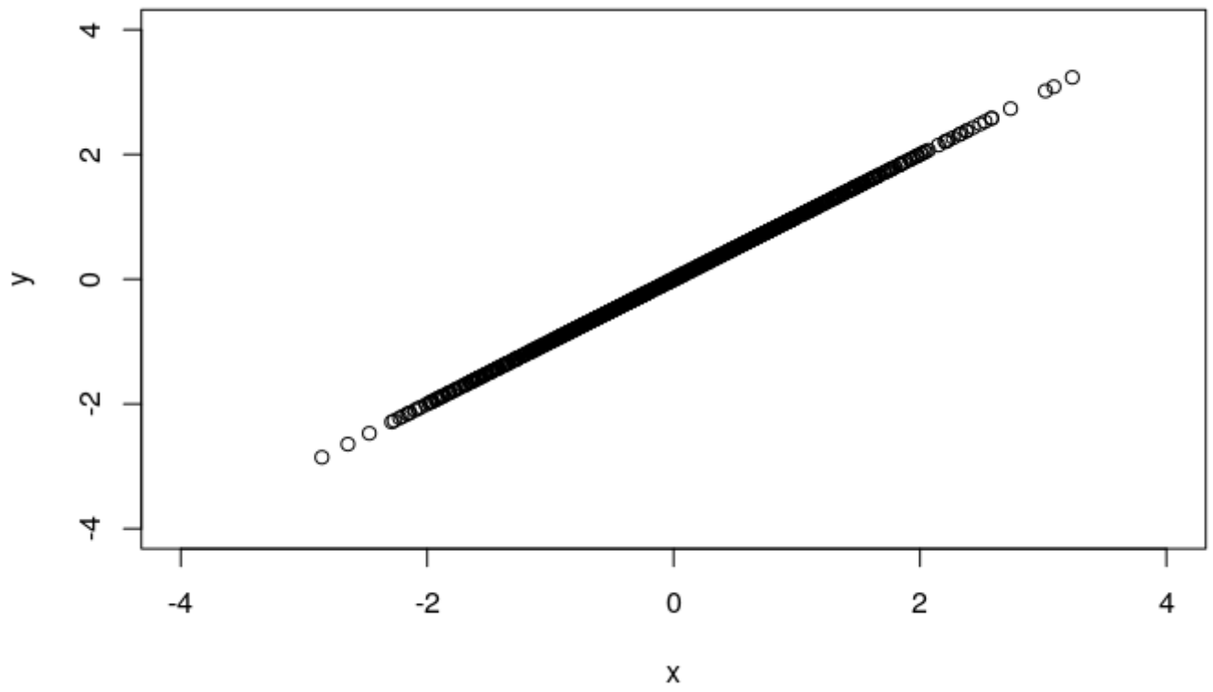
```
n <- 1000
r <- 1

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e
```

相関係数が **1** となるグラフ

Hide

```
plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

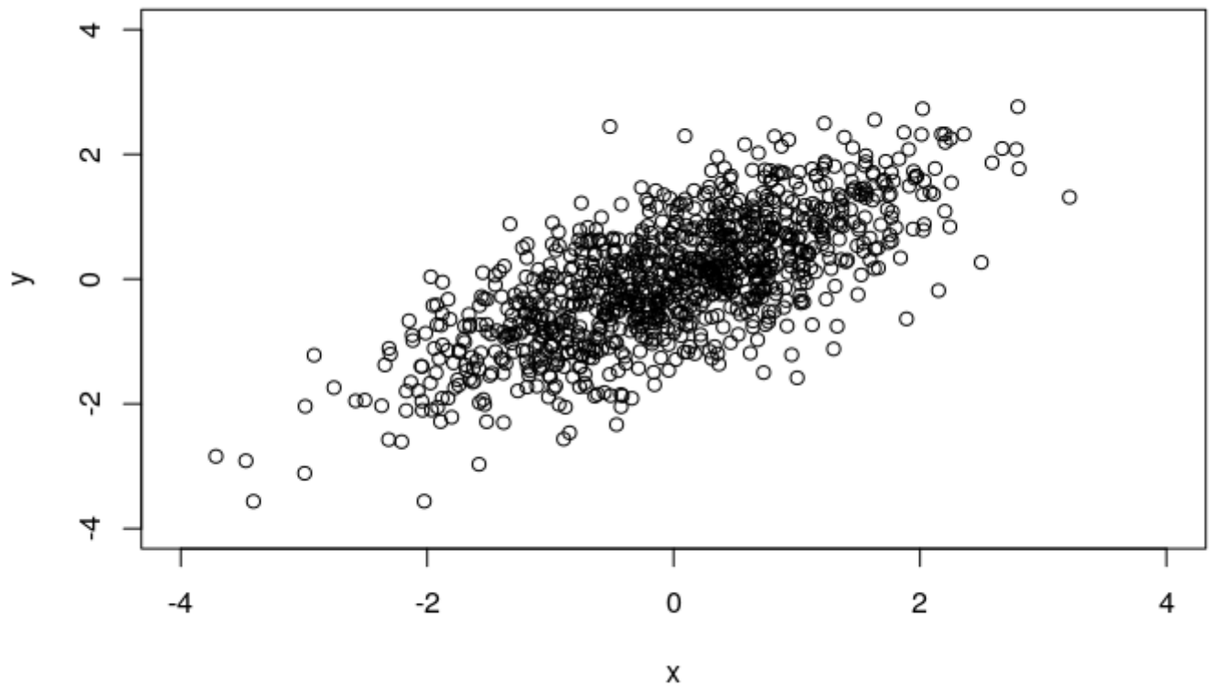
```
[1] 1
```

相関係数が **0.7** となるグラフ

Hide

```
n <- 1000
r <- 0.7

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e
plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

```
[1] 0.7129333
```

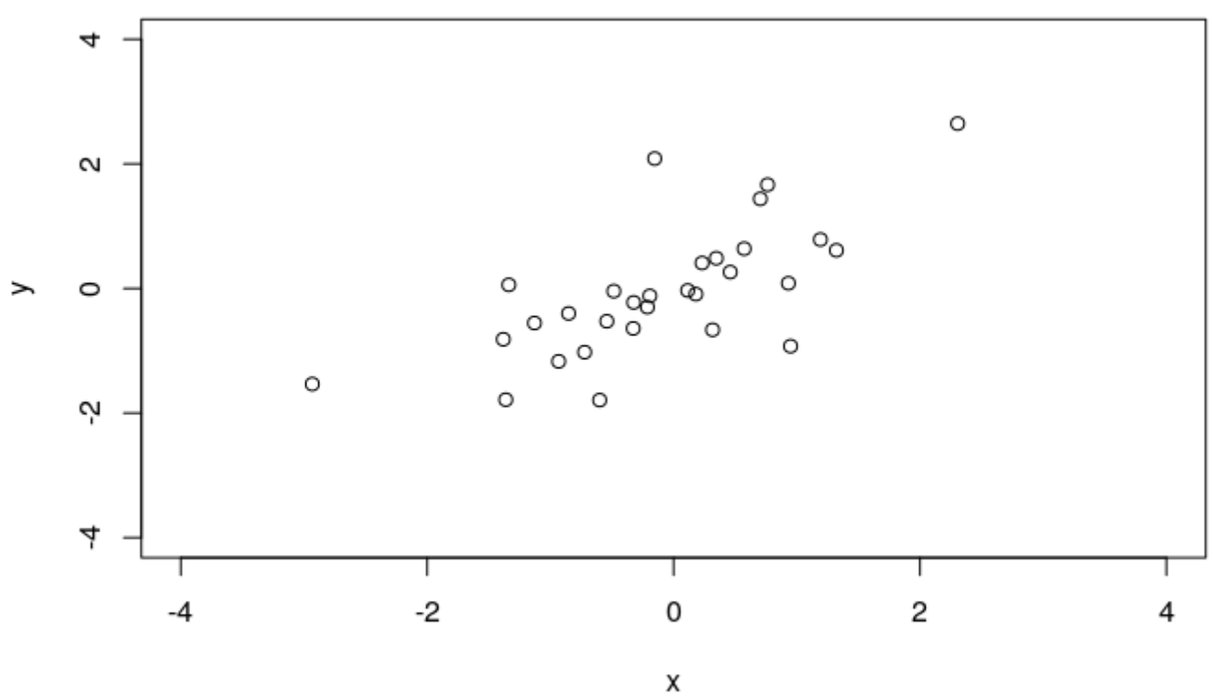
相関係数が **0.7** となるグラフ

- **数を減らした場合**

Hide

```
n <- 30
r <- 0.7

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e
plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```



Hide

```
cor(x, y)
```

```
[1] 0.6948065
```

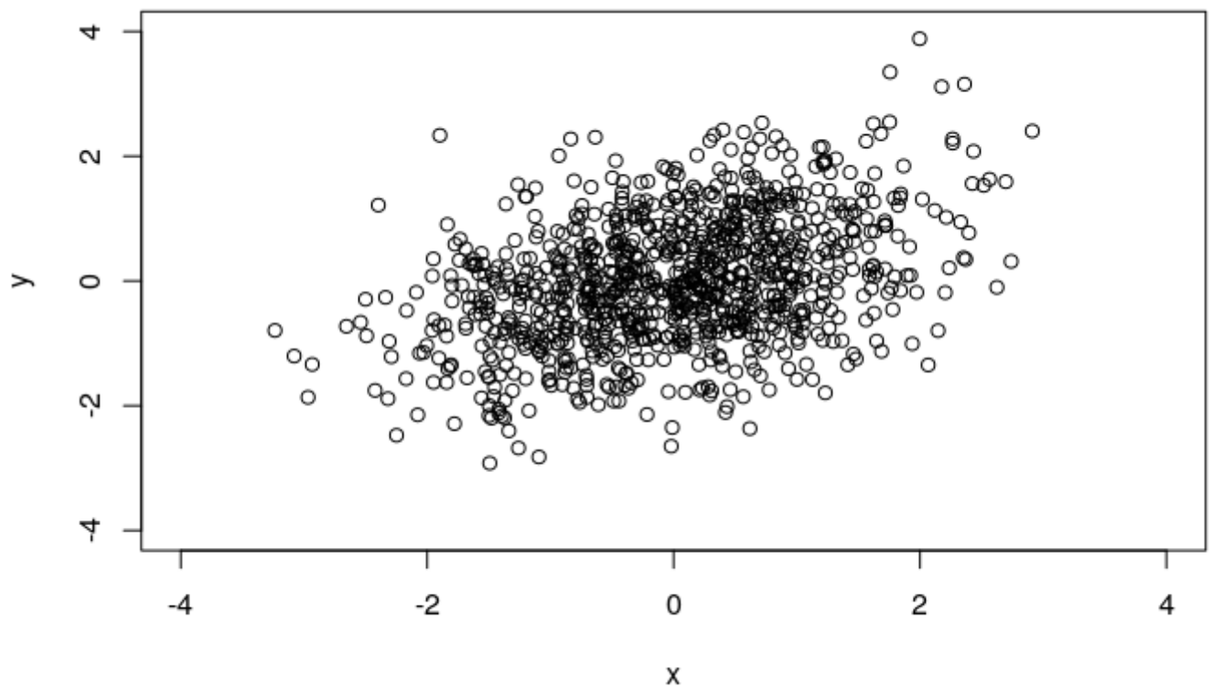
相関係数が **0.4** となるグラフ

Hide

```
n <- 1000
r <- 0.4

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e

plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

```
[1] 0.3957364
```

相関係数が **0.4** となるグラフ

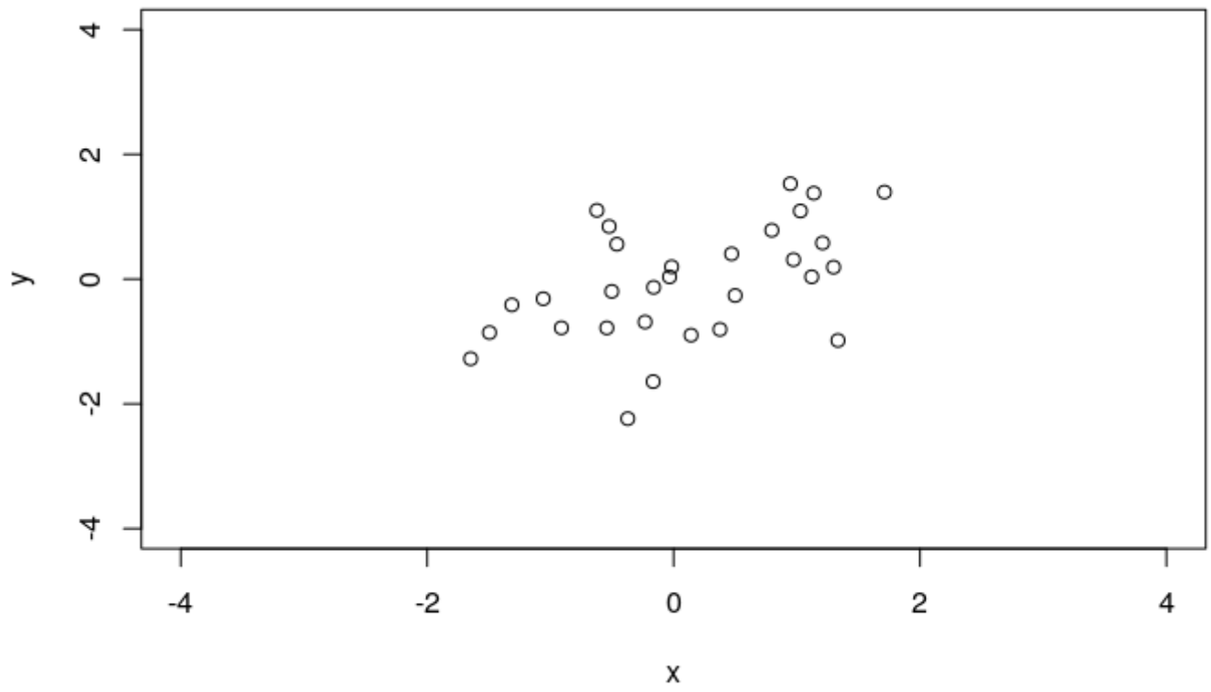
- **数を減らした場合**

Hide

```
n <- 30
r <- 0.4

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e

plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

```
[1] 0.4940882
```

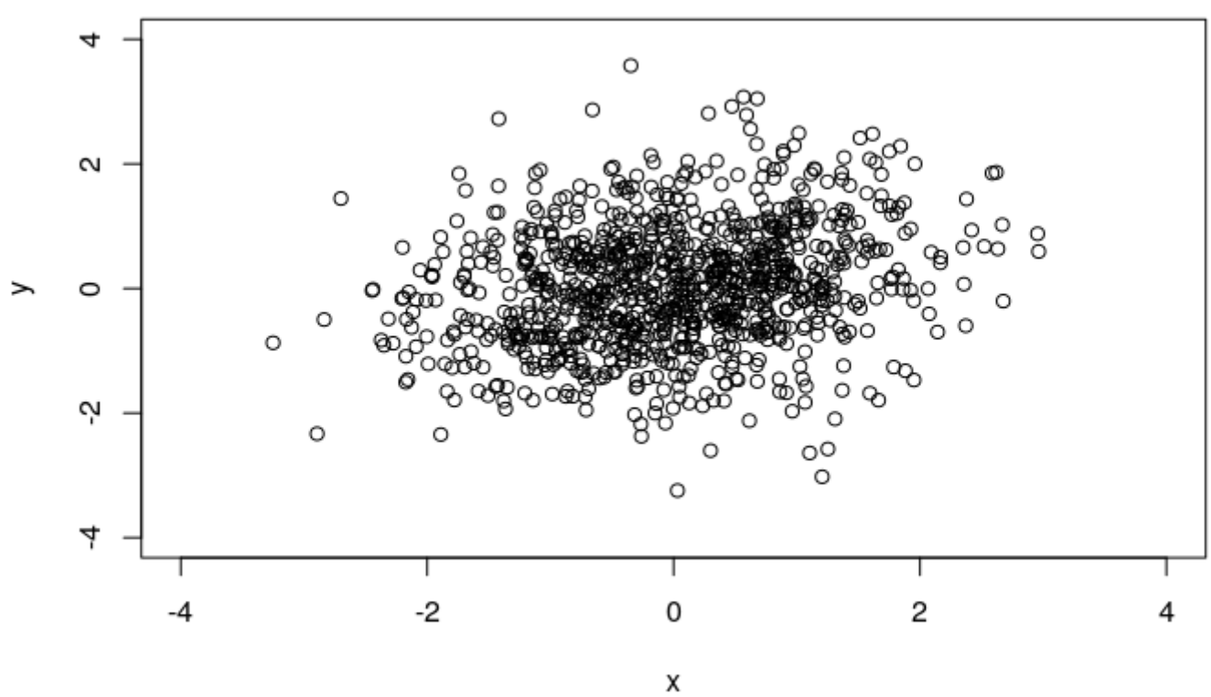
相関係数が **0.2** となるグラフ

Hide

```
n <- 1000
r <- 0.2

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e

plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

```
[1] 0.2260471
```

相関係数が **0.2** となるグラフ

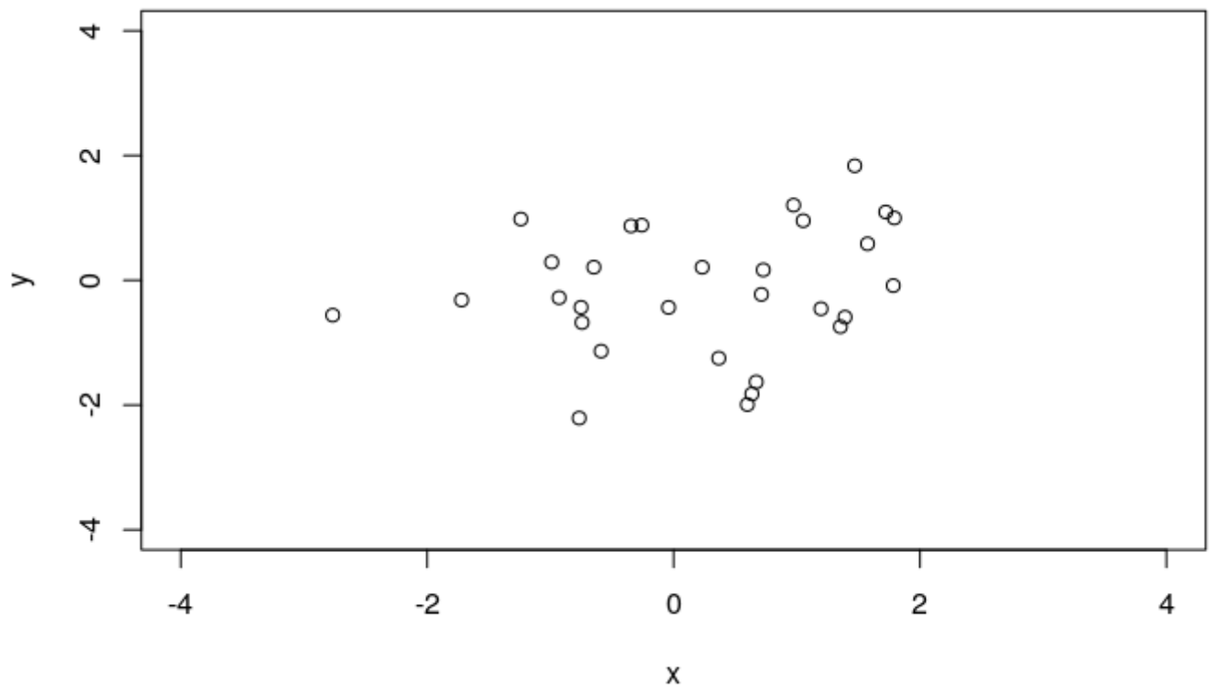
- **数を減らした場合**

Hide

```
n <- 30
r <- 0.2

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e

plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

```
[1] 0.2155738
```

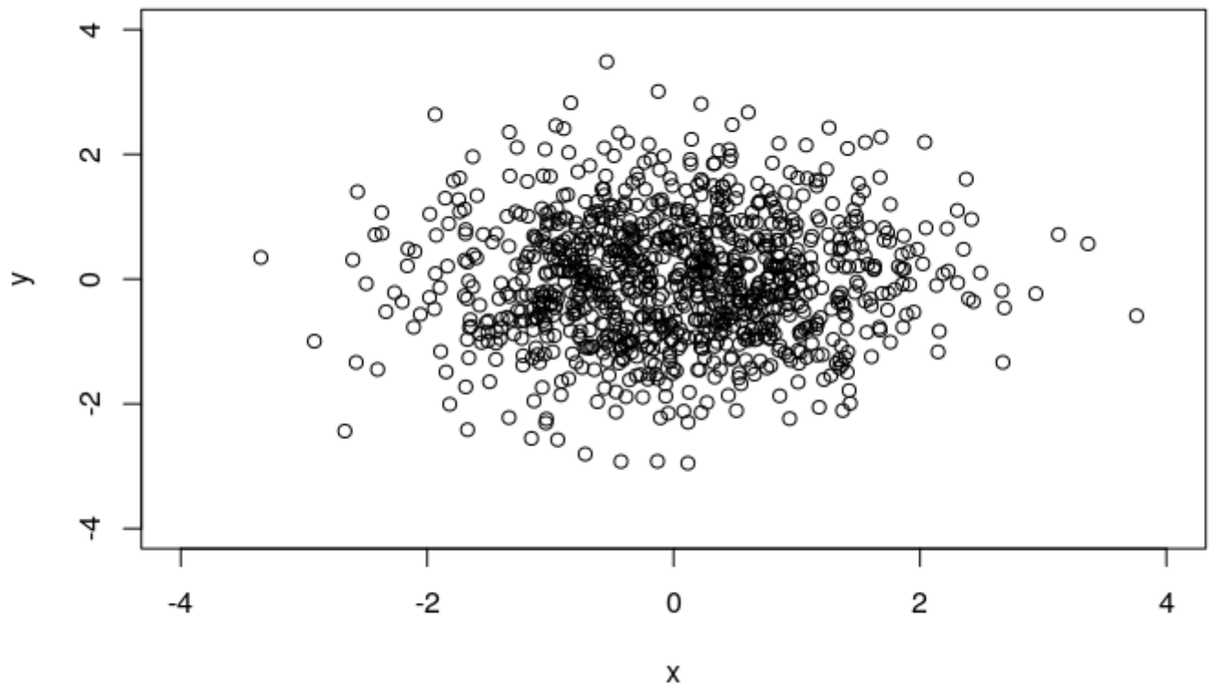
相関係数が **0** となるグラフ

Hide

```
n <- 1000
r <- 0

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e

plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```




Hide

```
cor(x, y)
```

```
[1] 0.03727087
```

相関係数が **0** となるグラフ

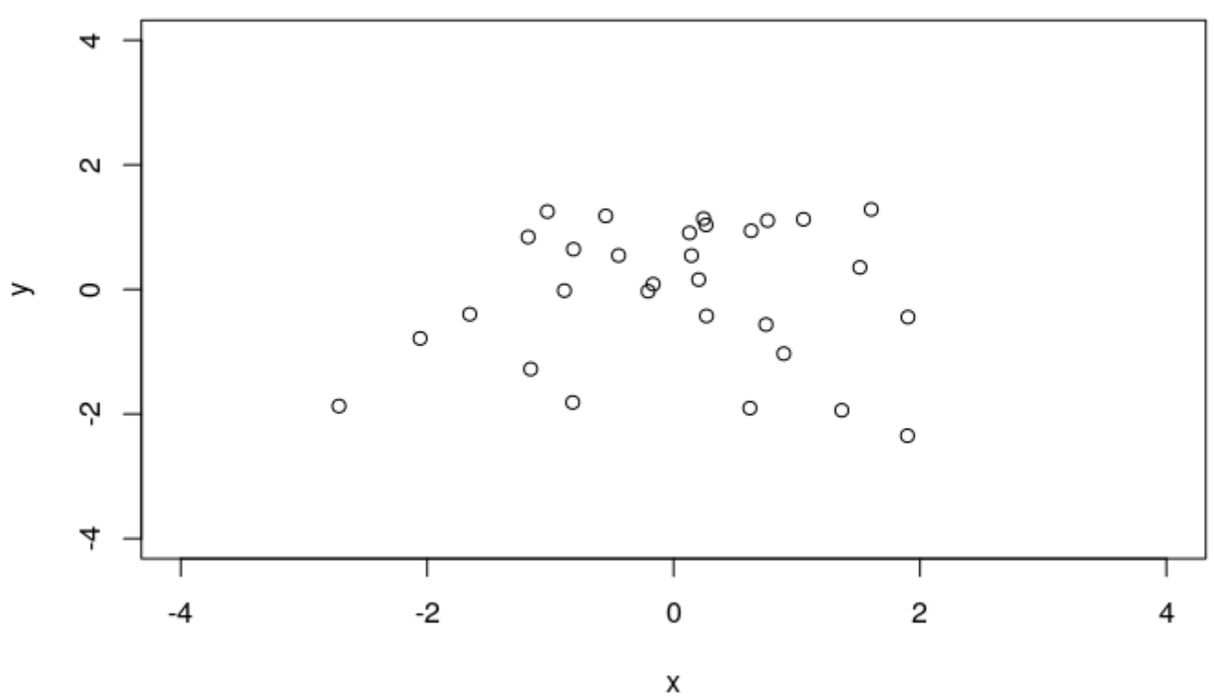
- **数を減らした場合**

Hide

```
n <- 30
r <- 0

x <- rnorm(n, mean = 0, sd = 1)
e <- rnorm(n, mean = 0, sd = 1)
y <- r*x + sqrt(1-r^2)*e

plot(x, y, xlim = c(-4, 4), ylim = c(-4, 4))
```



Hide

```
cor(x, y)
```

```
[1] 0.05498187
```