

変数と変数の関連性

変数と変数の関連性

- 関連性がある
→ある変数の分布が他の変数の分布の影響を受けている

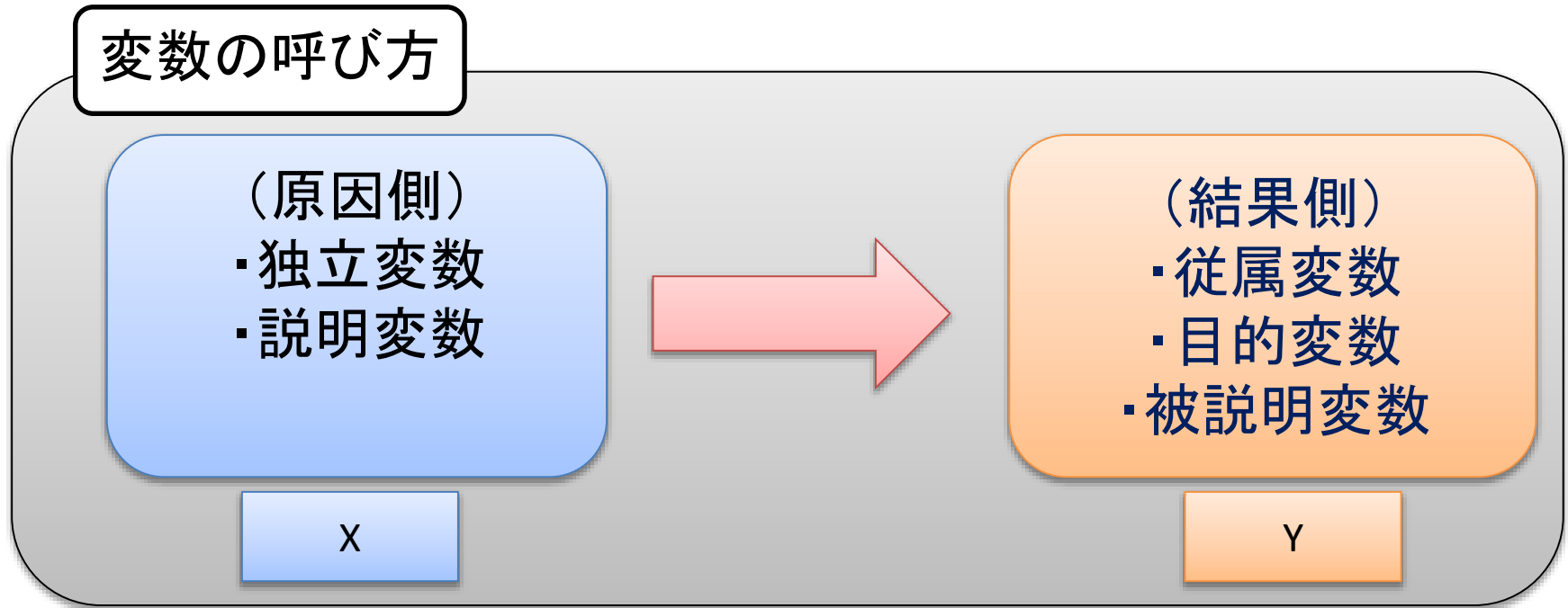
		肺がん	
		発症	発症しない
喫煙	あり	100人	900人
	なし	10人	990人

- 関連性がない
→ある変数の分布が他の変数の影響を受けない

		携帯電話を持っているか	
		持っている	持っていない
性別	男性	900人	100人
	女性	900人	100人

用語などの整理

変数の呼び方



パス図

$X \longrightarrow Y$

XがYの原因

$X \longleftrightarrow Y$

双方向の因果

$X \quad \frown \quad Y$

関連はあるが
関心のない関連性

変数の種類が違えば関連性の調べ方も違う

- 質的変数と質的変数の関連性
→クロス表
- 質的変数と量的変数の関連性
→平均値の比較
- 量的変数と量的変数の比較
→共分散、相関係数

	質的変数	量的変数
質的変数	クロス表	平均値の比較
量的変数	平均値の比較	相関係数 共分散

質的変数と質的変数の関連性を調べる－クロス表

		肺がん		合計
		発生	発生しない	
喫煙	あり	100人	900人	1000人
	なし	10人	990人	1000人
合計		110人	1890人	2000人

行周辺度数

列周辺度数

クロス表－行パーセント

- 各行のセルの度数を行周辺度数で割って100をかけたもの

行周辺度数

		肺がん		合計
		発生	発生しない	
喫煙	あり	10%(100人) $100 \times 1/1000$	90%(900人) $100 \times 900/1000$	1000人
	なし	1%(10人) $100 \times 10/1000$	99%(990人) $100 \times 990/1000$	1000人
合計		110人	1890人	2000人

クロス表－列パーセント

- 各列のセルの度数を列周辺度数で割って100をかけたもの

		肺がん		合計
		発生	発生しない	
喫煙	あり	91%(100人) $100 \times 100 / 110$	48%(900人) $100 \times 900 / 1890$	1000人
	なし	9%(10人) $100 \times 10 / 110$	52%(990人) $100 \times 990 / 1890$	1000人
合計		110人	1890人	2000人

列周辺度数

関連性の指標(1)ーリスク比とオッズ比

- リスク比(相対危険度)

$$\begin{aligned}\text{リスク比} &= \frac{\text{要因がある群(曝露群)のリスク}}{\text{要因がない群(非曝露群)のリスク}} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} \\ &= \frac{(C+D) \times A}{(A+B) \times C}\end{aligned}$$

- オッズ比

$$\begin{aligned}\text{オッズ比} &= \frac{\text{要因がある群(曝露群)のオッズ}}{\text{要因がない群(非曝露群)のオッズ}} = \frac{\frac{A}{B}}{\frac{C}{D}} \\ &= \frac{A \times D}{B \times C}\end{aligned}$$

		イベント	
		発生	発生しない
要因	あり	A	B
	なし	C	D

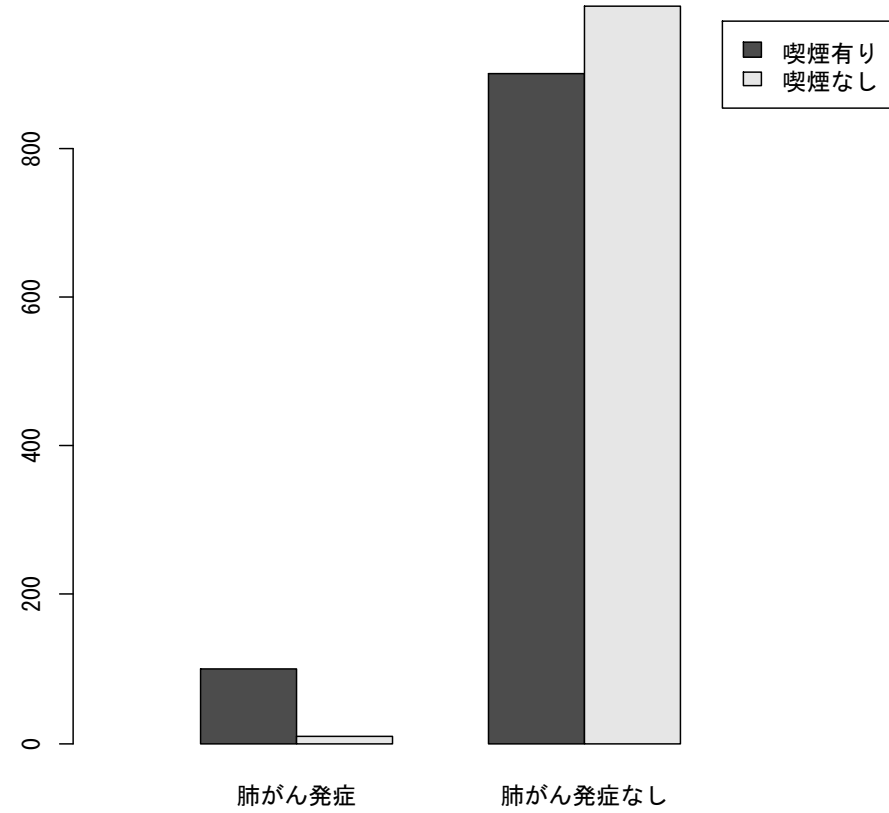
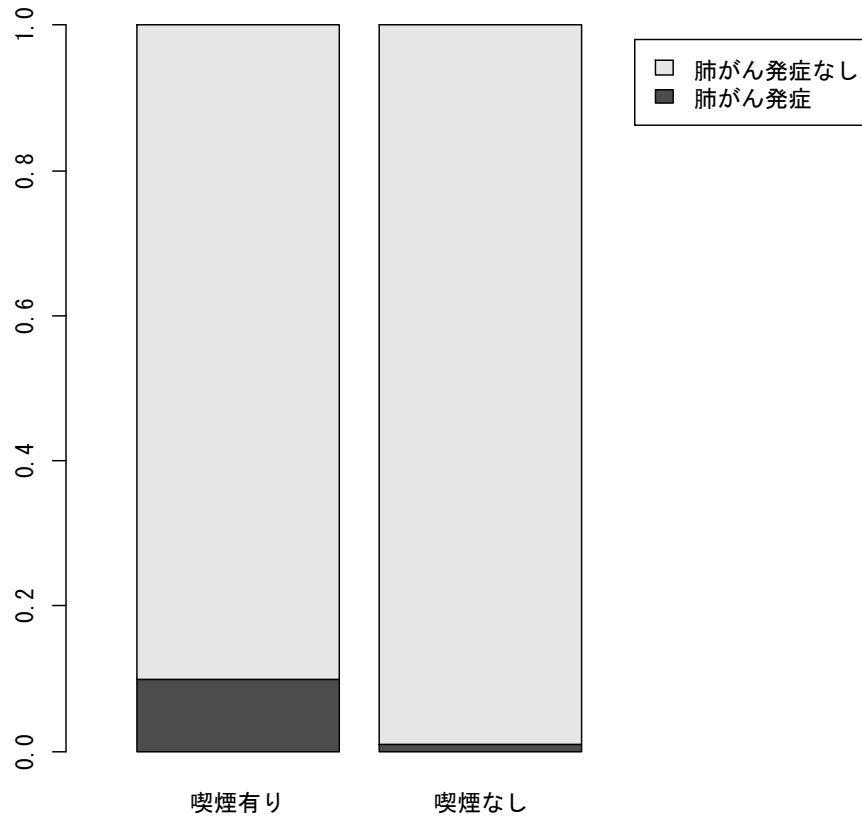
関連性の指標(2)－ファイ係数

- ϕ (ファイ)係数
 - 相関係数(後述)の一種
 - -1から1の間の値をとる
 - 0なら関連なし

$$\phi = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$$

		イベント	
		発生	発生しない
要因	あり	A	B
	なし	C	D

クロス表のグラフ表現



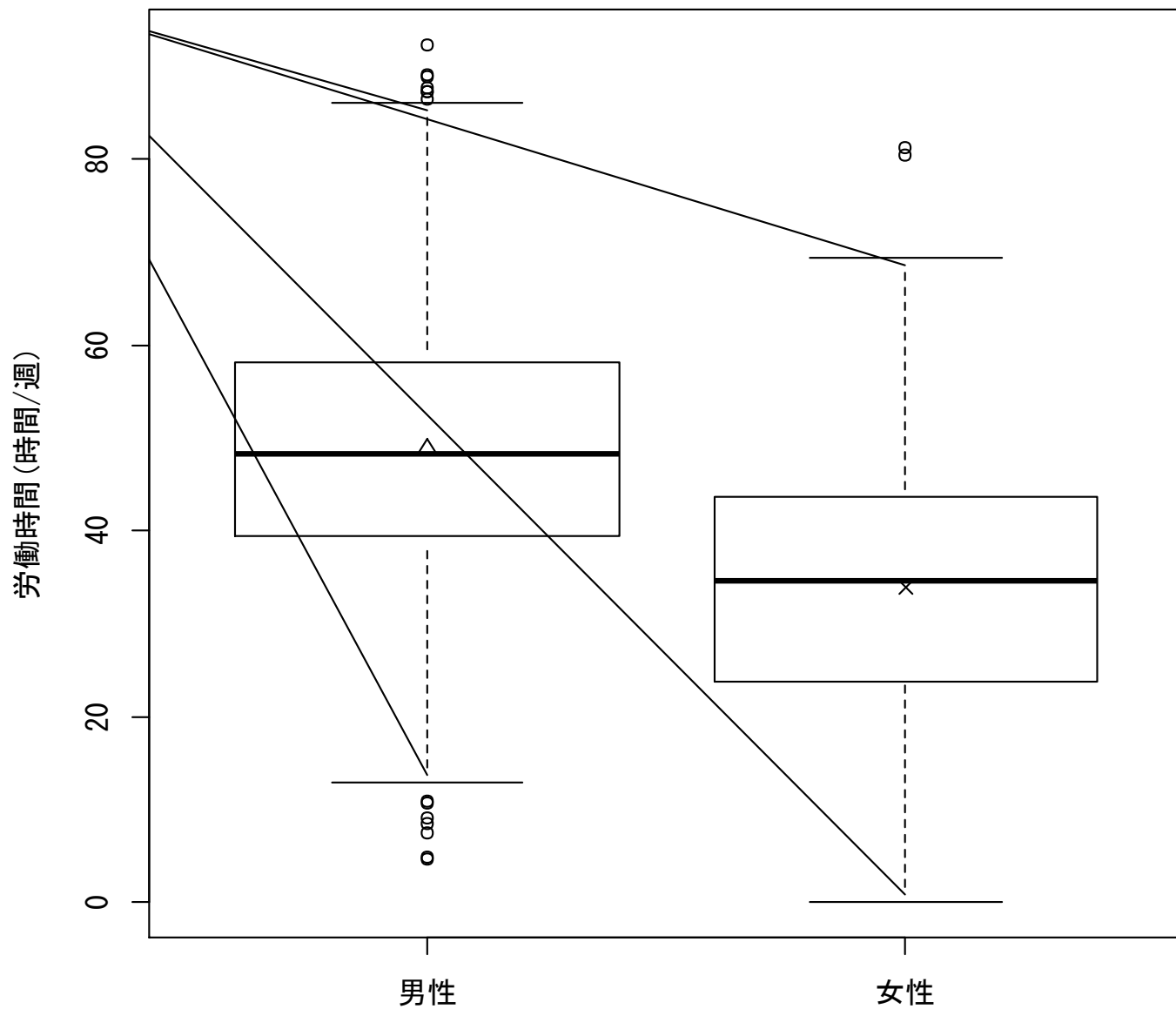
質的変数と量的変数の関連性

- 質的変数のカテゴリーごとに量的変数の平均を出して比べる

男女の1週間の労働時間

性別	平均	標準偏差	度数
男性	48.8	14.1	1144
女性	33.9	14.7	957

グラフにすると



量的変数と量的変数の関連性

- まずグラフにしてみる

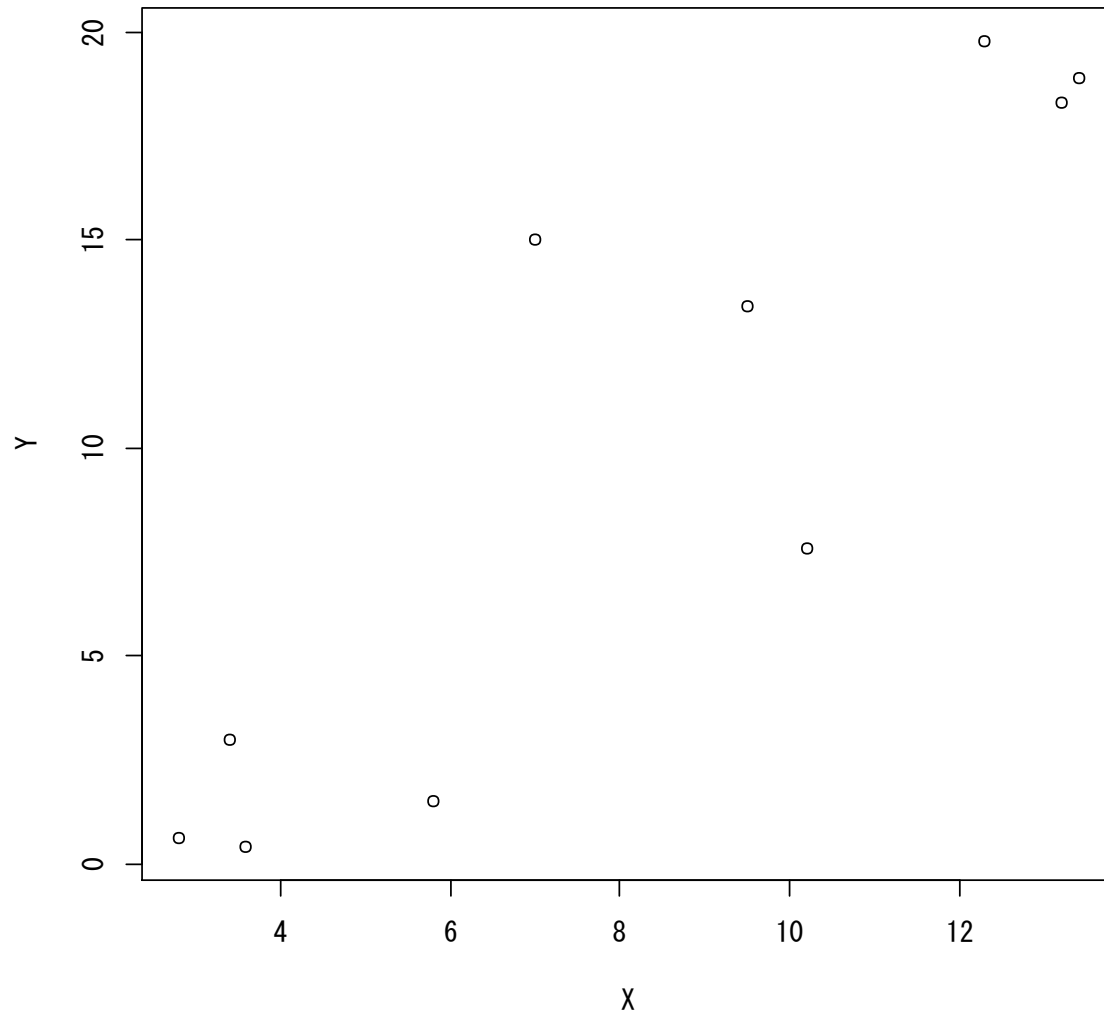
2つの変数の値で決まる座標に点を打っていく

→散布図

i	1	2	3	4	5	6	7	8	9	10	平均
変数 X	2.8	3.4	3.6	5.8	7	9.5	10.2	12.3	13.2	13.4	8.12
変数 Y	0.6	3	0.4	1.5	15	13.4	7.6	19.8	18.3	18.9	9.85
$X_i - (X \text{の平均})$	-5.32	-4.72	-4.52	-2.32	-1.12	1.38	2.08	4.18	5.08	5.28	
$Y_i - (Y \text{の平均})$	-9.25	-6.85	-9.45	-8.35	5.15	3.55	-2.25	9.95	8.45	9.05	

散布図

- XとYの関係をグラフで表すと



量的変数と量的変数の関連性を数値で表す 共分散

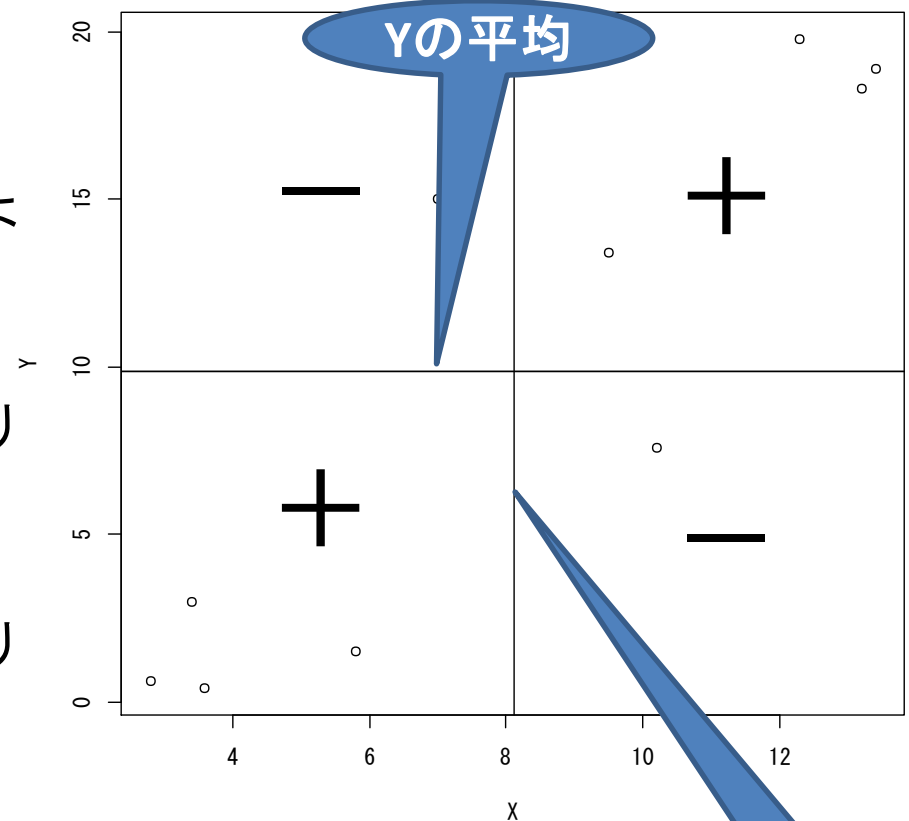
- 共分散
 - 二つの変数の共変動の指標

$$\text{共分散} = \frac{\sum_{i=1}^N (x_i - (x\text{の平均})) \times (y_i - (y\text{の平均}))}{N}$$

i	1	2	3	4	5	6	7	8	9	10	平均
変数 X	2.8	3.4	3.6	5.8	7	9.5	10.2	12.3	13.2	13.4	8.12
変数 Y	0.6	3	0.4	1.5	15	13.4	7.6	19.8	18.3	18.9	9.85
$X_i - (X\text{の平均})$	-5.32	-4.72	-4.52	-2.32	-1.12	1.38	2.08	4.18	5.08	5.28	
$Y_i - (Y\text{の平均})$	-9.25	-6.85	-9.45	-8.35	5.15	3.55	-2.25	9.95	8.45	9.05	

共分散の意味

- XとYの散布図を描く
- Xの平均を通りY軸と平行の線とYの平均を通りX軸に平行の線を引くと・・・
 - (右上)
 - Xの平均よりも高くYの平均よりも高い⇒積の和はプラス
 - (左上)
 - Xの平均より低くYの平均より高い⇒積の和はマイナス
 - (左下)
 - Xの平均より低くYの平均より低い⇒積の和はプラス
 - (右下)
 - Xの平均よりも高くYの平均より低い⇒積の和はマイナス



量的変数と量的変数の関連性を数値で表す ピアソンの積率相関係数

- 共分散は単位やケース数の影響をうけるので比較しづらい
→ 共分散を標準化すれば比較しやすい
- ピアソンの積率相関係数(相関係数)
 - 量的変数Xと量的変数Yの関係の大きさを「係数」という値で表す方法
 - $-1 \sim 1$ の間をとる
 - 0だと関係なし
 - 1にちかいほど、「正」の関係がある
 - Xが大きくなるほどYも大きくなる
 - -1 に近いほど、「負」の関係がある
 - Xが大きくなるほどYが小さくなる

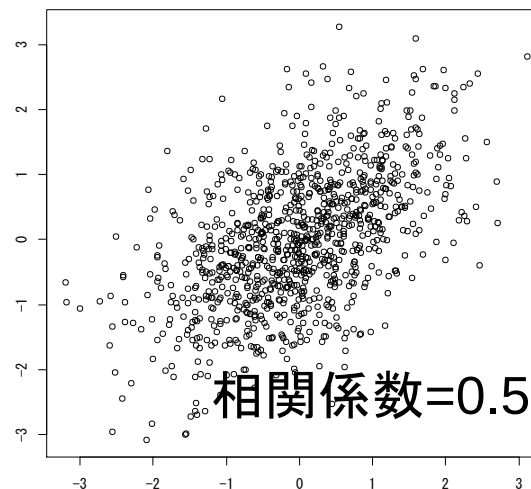
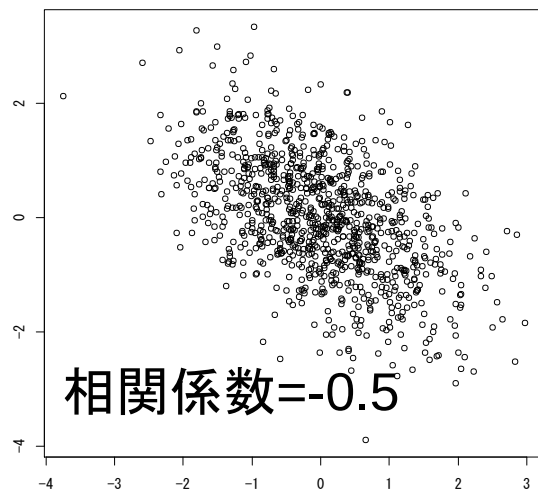
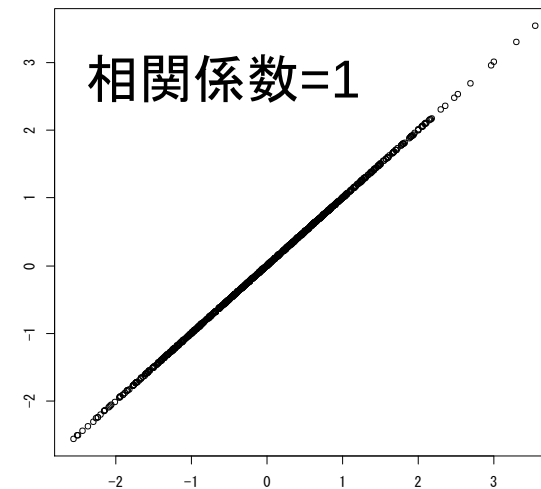
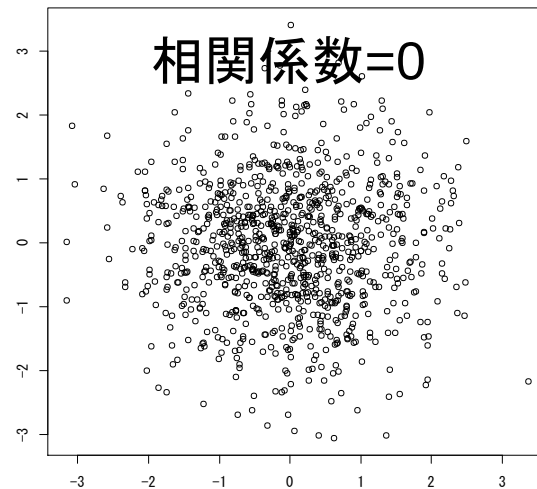
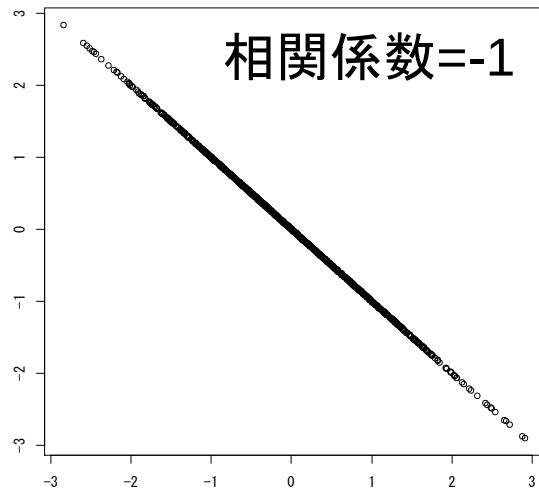
相関係数の計算式

- XとYのピアソンの積率相関係数
 - XとYの共分散を、Xの標準偏差×Yの標準偏差で割って標準化する

$$\text{相関係数} = \frac{X \text{と} Y \text{の共分散}}{X \text{の標準偏差} \times Y \text{の標準偏差}}$$

i	1	2	3	4	5	6	7	8	9	10	標準偏差
変数 Xi	2.8	3.4	3.6	5.8	7	9.5	10.2	12.3	13.2	13.4	3.9
変数 Yi	0.6	3	0.4	1.5	15	13.4	7.6	19.8	18.3	18.9	7.7

相関係数と散布図の形



まとめ

- 質的変数と質的変数の関連性
→ クロス表
- 質的変数と量的変数の関連性
→ 平均値の比較
- 量的変数と量的変数の比較
→ 共分散、相関係数

	質的変数	量的変数
質的変数	クロス表	平均値の比較
量的変数	平均値の比較	相関係数 共分散