

データサイエンスへの誘い

第9回: 複数のデータを比較する

瓜生真也 (デザイン型AI教育研究センター・助教)

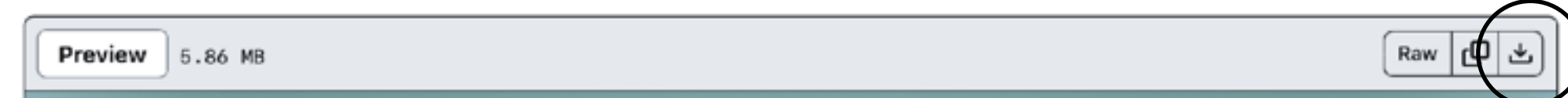
講義内容（予定）

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INNV1250>



ダウンロード可能



1. ガイダンス、データサイエンスとは何か
2. プログラミング基礎
3. 再現可能性
4. データ処理の手法
5. データの要約
6. データの可視化
7. データと確率
8. データからの推論

9. 複数のデータを比較する

10. 統計のウソ
11. 統計的モデリング
12. 統計的学習
13. さまざまなデータサイエンスの手法
14. 機械学習と人工知能（AI）
15. 期末試験
16. 振り返りと統括

今日の目標

意思決定を促進するための

関係の把握と比較の方法を学ぶ

【課題】 関係・比較についてのクイズ

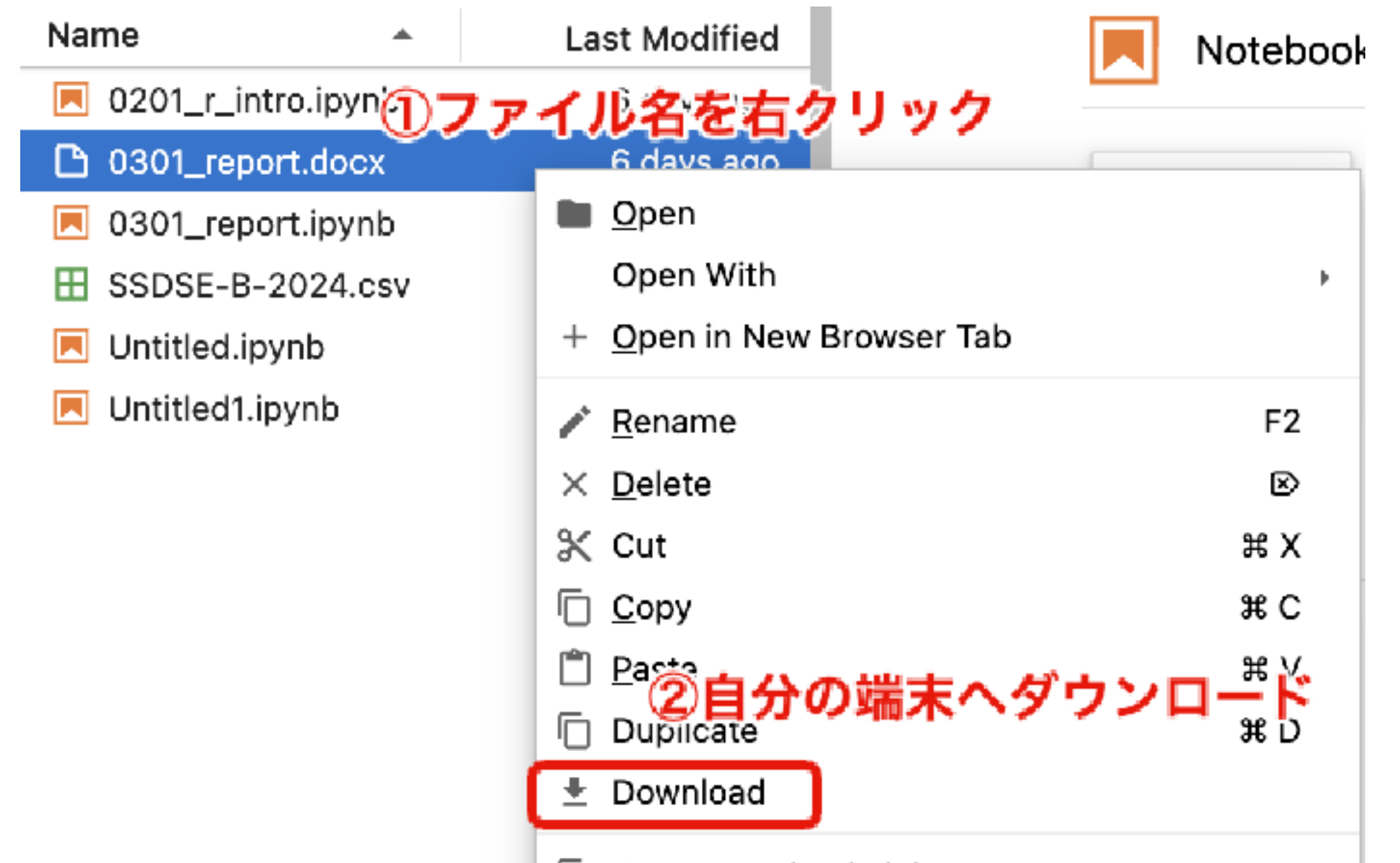
提出期限: 来週の講義開始前まで

手順

1. 添付ファイルをダウンロード
2. JupyterHubへアップロード
3. コードやコメントを記述、実行
4. 保存
5. ダウンロードしたファイルをmanabaへアップロード

manabaのレポートとして提出してください

メニュー上の「ファイル」から「ダウンロード」



注意: ファイル名は英数字のみにすること

日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない

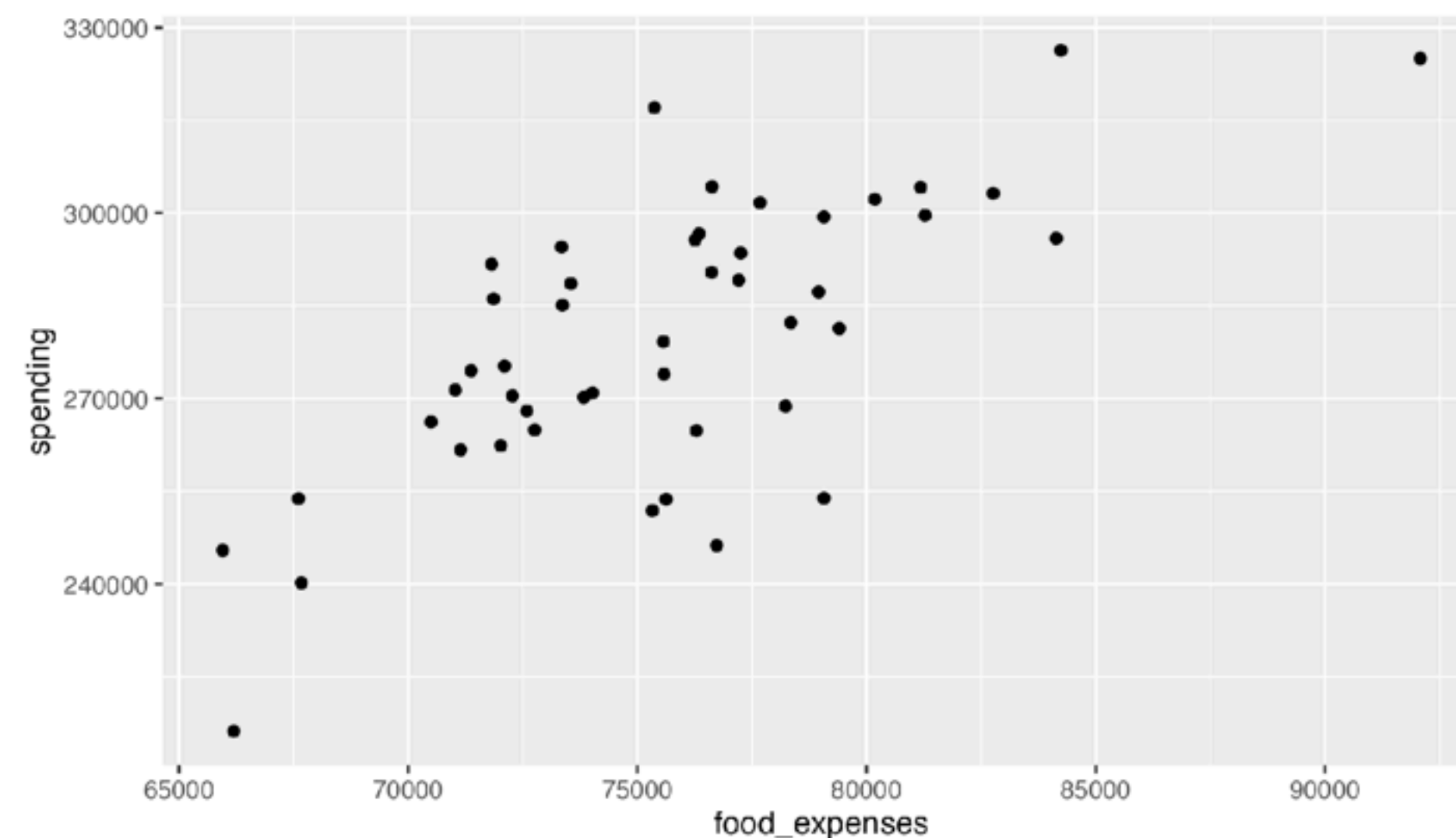
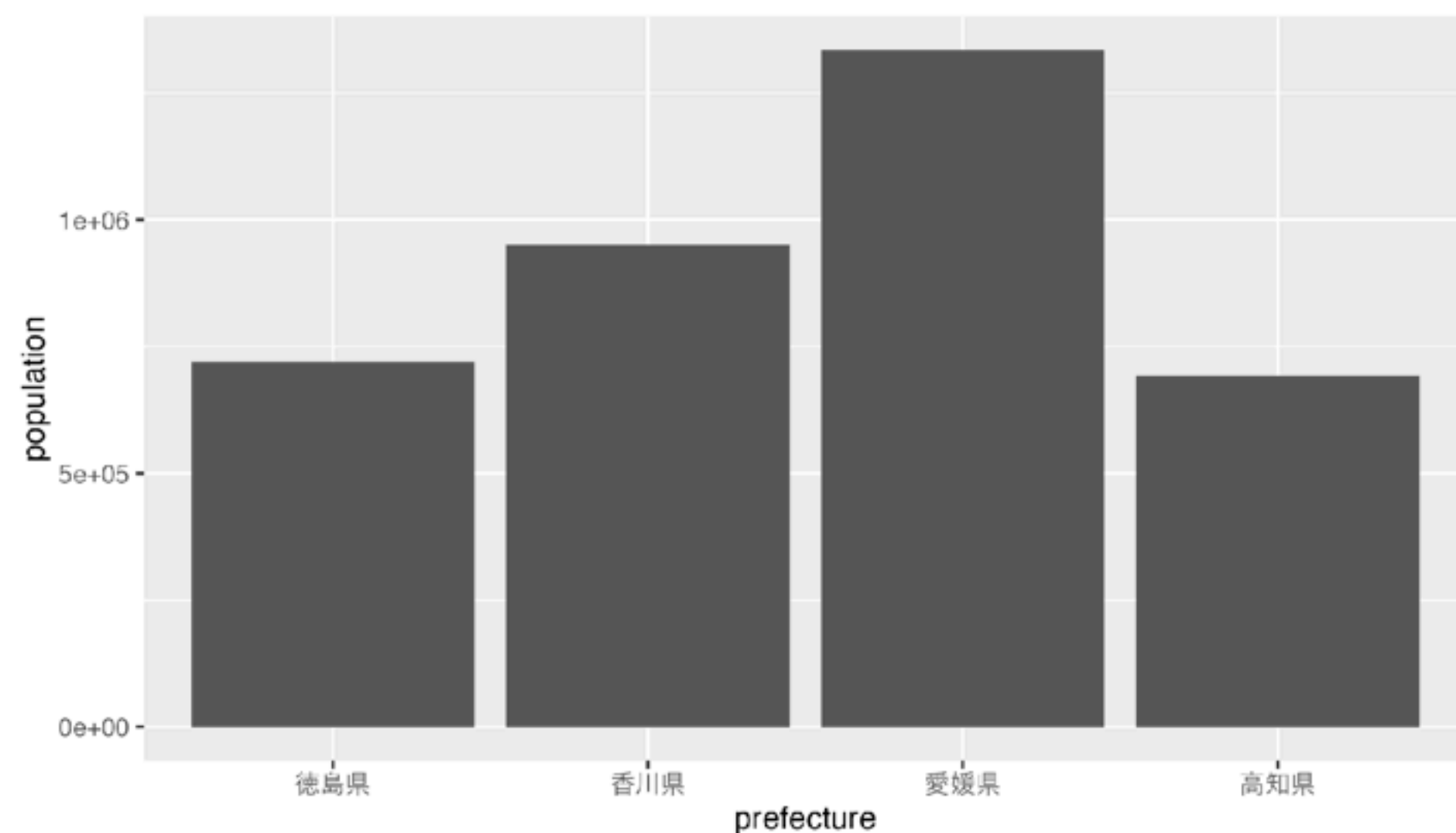
ファイルをダウンロードしても開けなくても問題ない（気にしない）

内容の確認、編集はJupyterHub上で行う

可視化によって関係・比較を可能にする

棒グラフ、散布図

パターンの発見、直感の検証、問題の特定



差がありそう、関連がありそう… 主観的な「ありそう」を客観的に評価するには？

相関分析

2つの変数の間の関係を調べるための手法、関連の程度の推定

統計的仮説検定

データを用いた統計的推論により仮説が真であるかを評価

関係を見る

データ分析における2つの関係

複数の変数がともに変化する状態

データ分析では**相関関係**と**因果関係**の2つの関係を扱う（似て非なるもの）

相関関係

ある出来事や物事と別の出来事や物事の間に関係があるもの

ペンギン個体の翼の長さ \longleftrightarrow ペンギン個体のくちばしの長さ

因果関係

ある出来事や物事が**原因**となって、別の出来事や物事（**結果**）が起こるもの

ある水道会社の利用を止める \longrightarrow 水道を利用していた地域のコレラ患者が減る

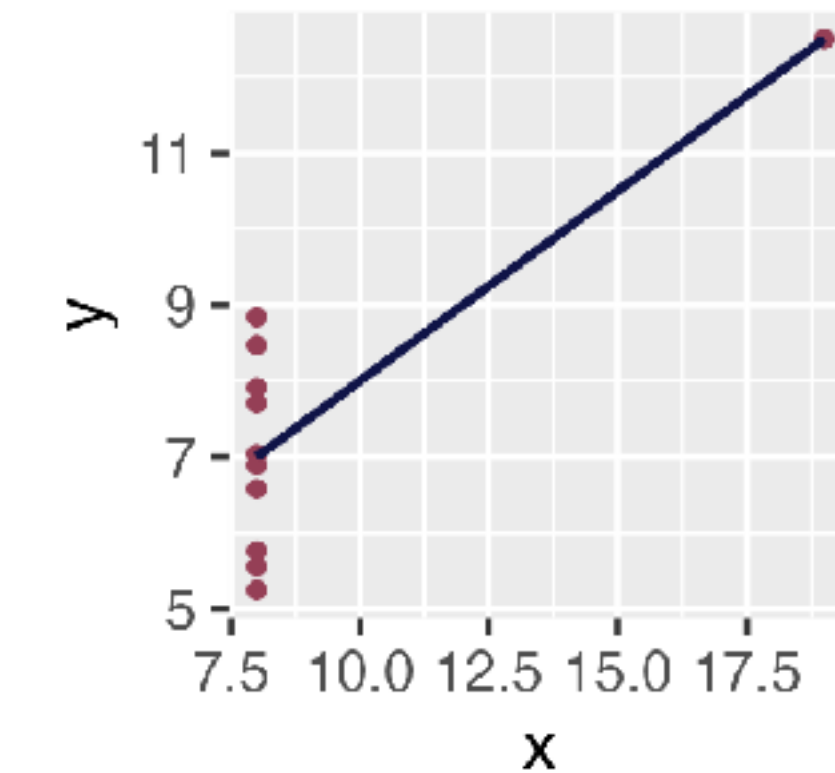
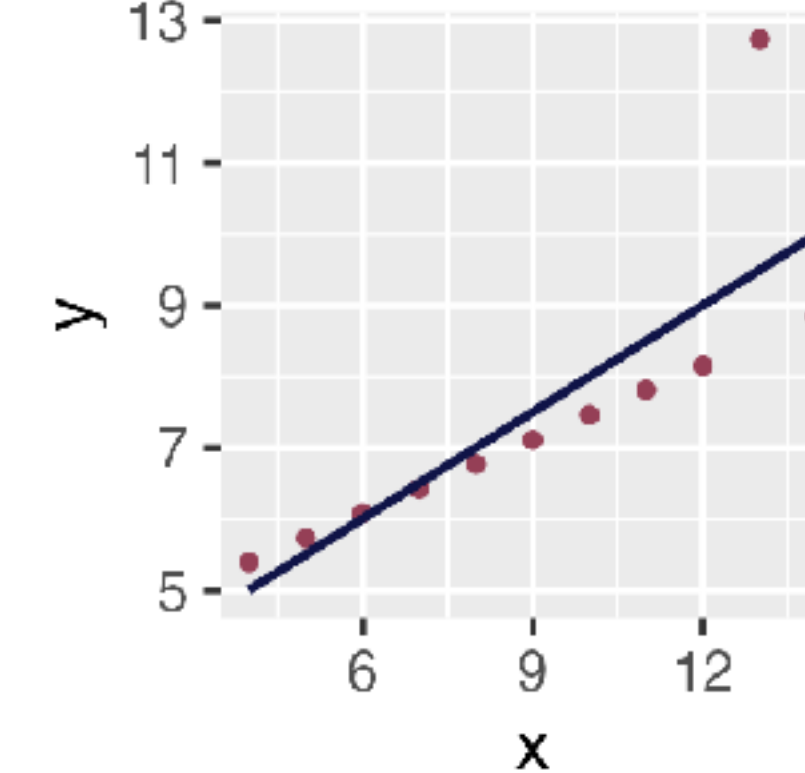
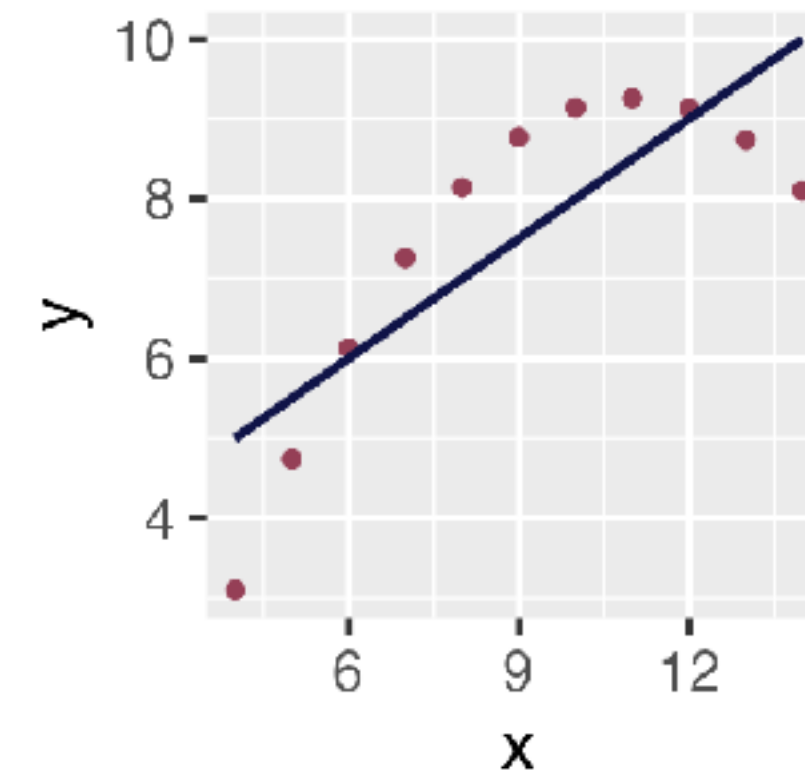
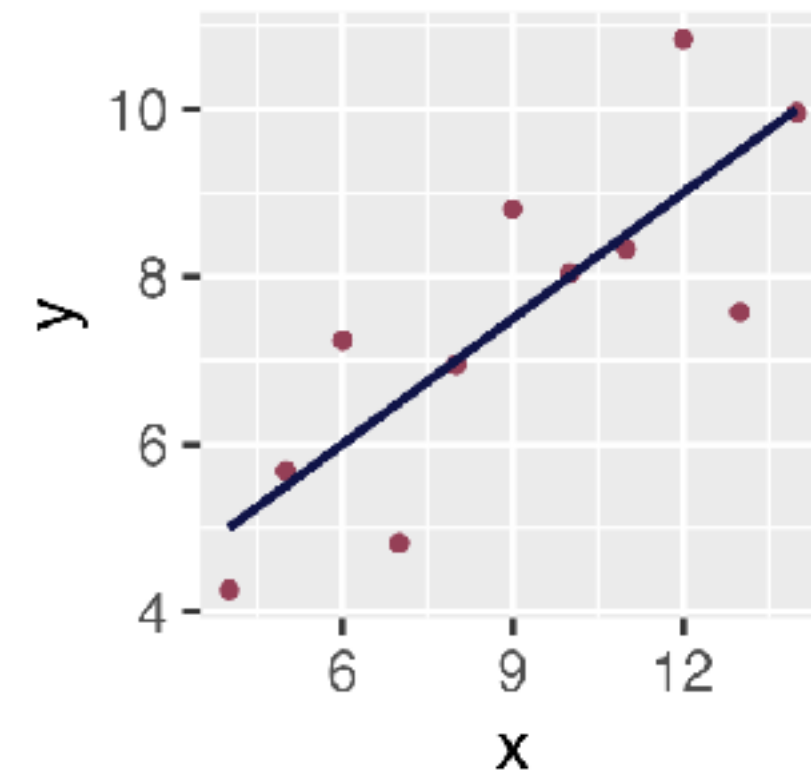
相関関係があるからと言って必ずしも因果関係があるわけではない
→ 因果関係を調べる「因果推論」

【再】アンスコム の例

参考) 第6回: データの可視化

統計量、相関係数がほぼ同じ値になる4種のデータセット
→ 散布図を描くとデータの傾向が大きく異なる

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89



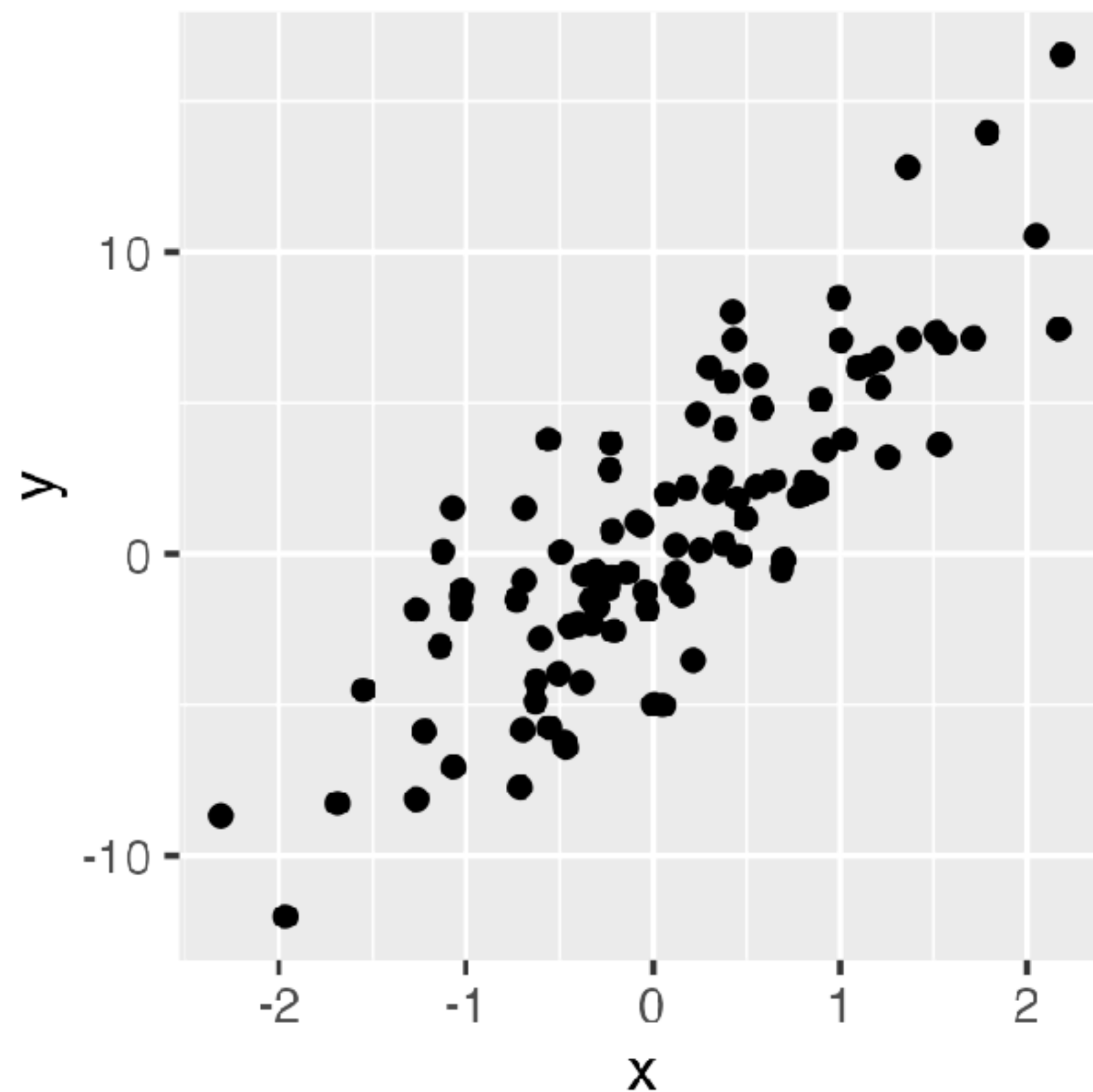
統計量（相関係数など）と可視化を
セットで評価することが重要

相関

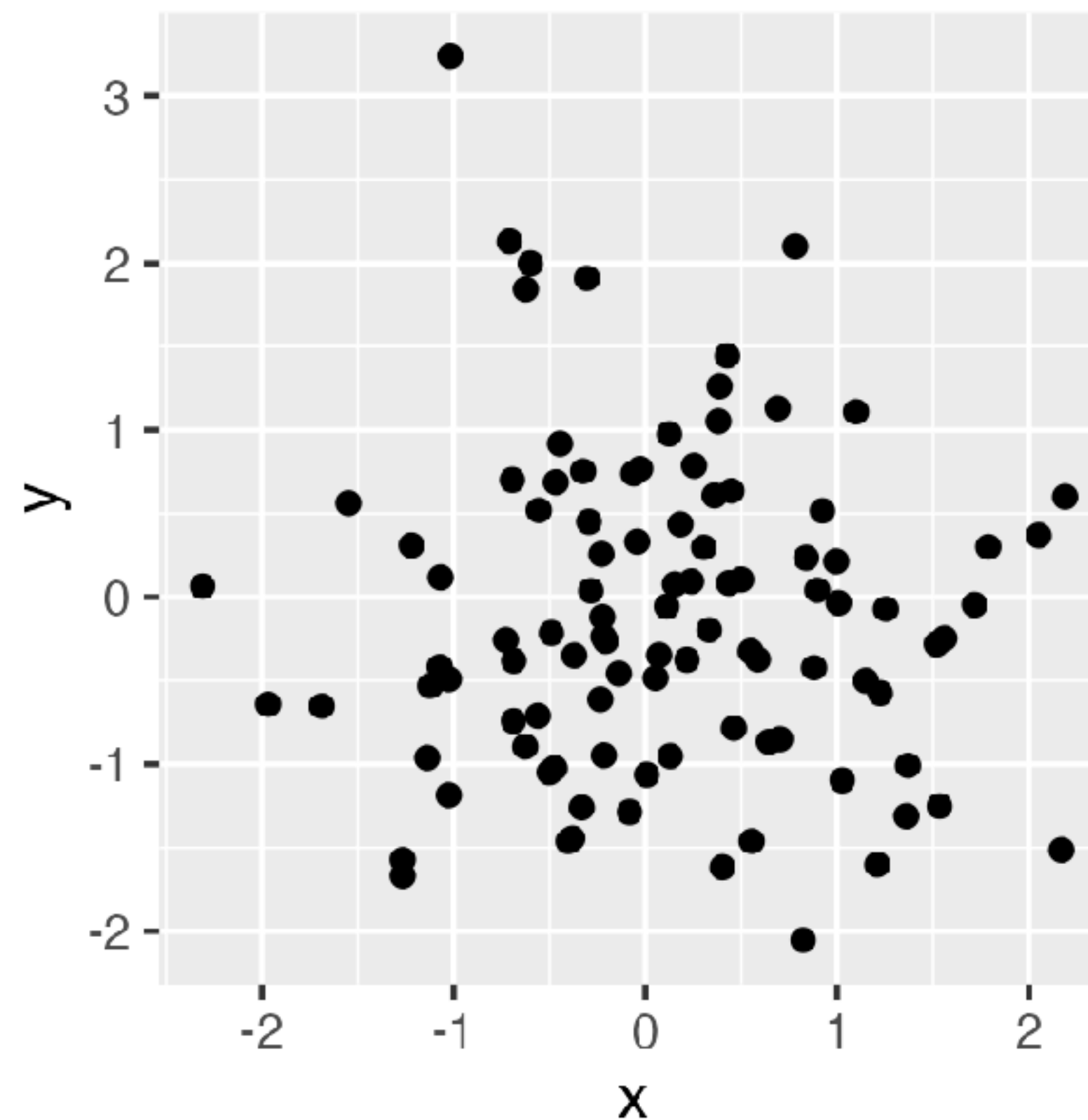
2つの変数間で起こる関係を表す

散布図（後述）としてグラフ上に可視化することで傾向を把握しやすくなる

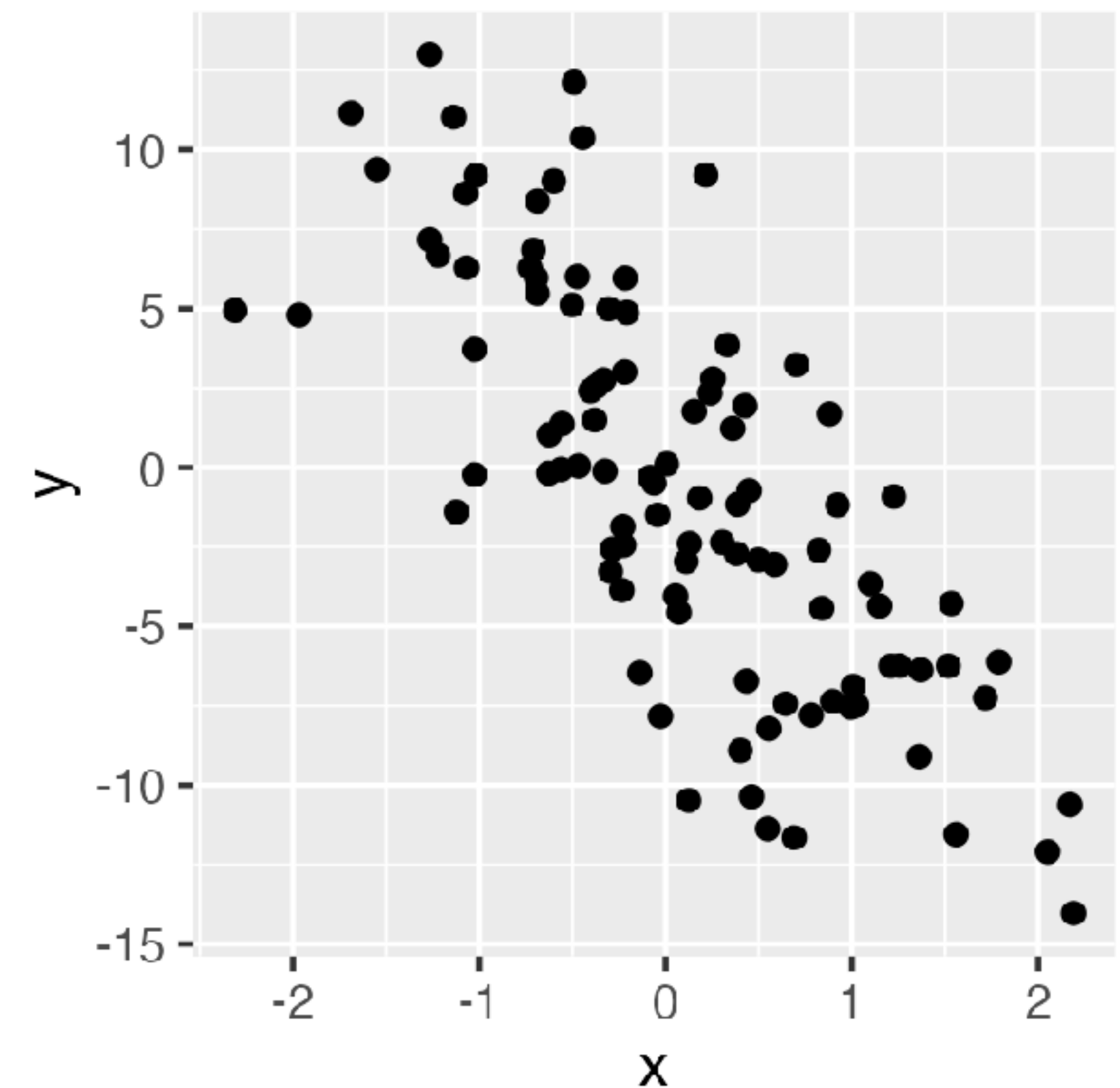
正の相関関係



無相関



負の相関関係



関係の数値化1: 共分散 covariance

2つの変数(x と y)についての共分散は次のように求められる

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

手順1/2

①、②変数 x (y) の値から変数 x (y) の平均値を引く → **偏差**

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

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{②}$$

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{③ 偏差の積を求める}$$

関係の数値化1: 共分散 covariance

手順2/2

④ n (すべてのデータ) まで右の処理を行い、それを足し合わせる

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

データの $i = 1$ 番目から

⑤ 変数 x と変数 y の各値に対して偏差を求め、それを掛け合わせたものを足す

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

n (データ数) で割る


```
# Rの標準関数で共分散を求めるとデータの数 - 1で割る不偏共分散になる
cov(penguins$bill_length_mm, penguins$bill_depth_mm, use = "complete.obs")
#> [1] -2.534234
```



共分散の特徴

値が大きいほど2変数の関係が強いことを示す
変数の単位に依存して値が変わる

```
cov(  
  # cm を m に  
  set_units(set_units(df_animal$body_length_cm, cm), m),  
  df_animal$weight_kg,  
  use = "complete.obs")  
#> [1] 66.19572  
cov(  
  # cm を m に  
  set_units(set_units(df_animal$body_length_cm, cm), m),  
  # kg を g に  
  set_units(set_units(df_animal$weight_kg, kg), g),  
  use = "complete.obs")  
#> [1] 66195.72
```



→標準化によって変数間のスケールを揃える

関係の数値化2: 相関係数

共分散の単位依存の問題を解消する指標

共分散を各変数の標準偏差の積で割ることで算出される

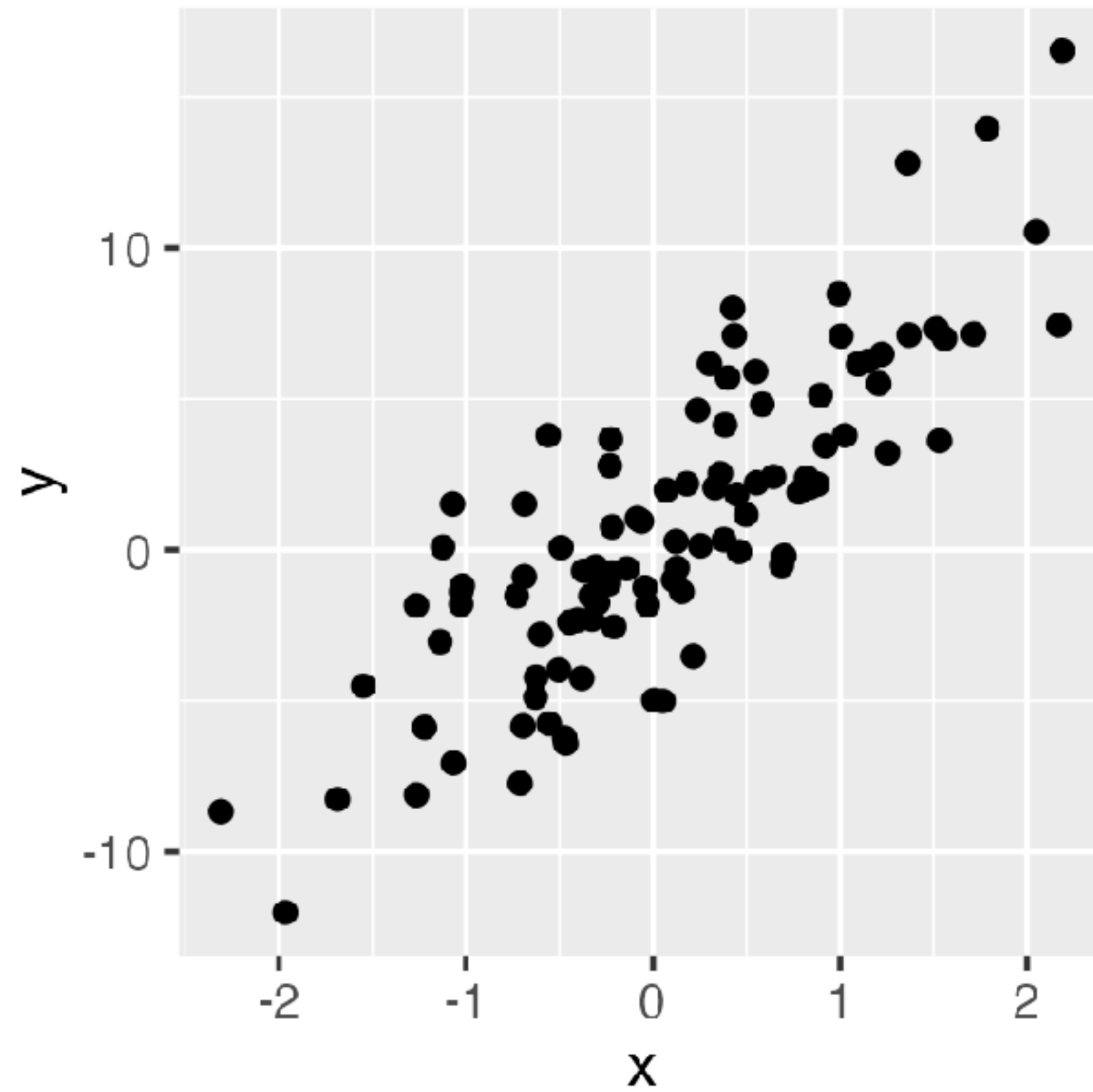
$$r = \frac{Cov_{xy}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

-1から1までの値をとる

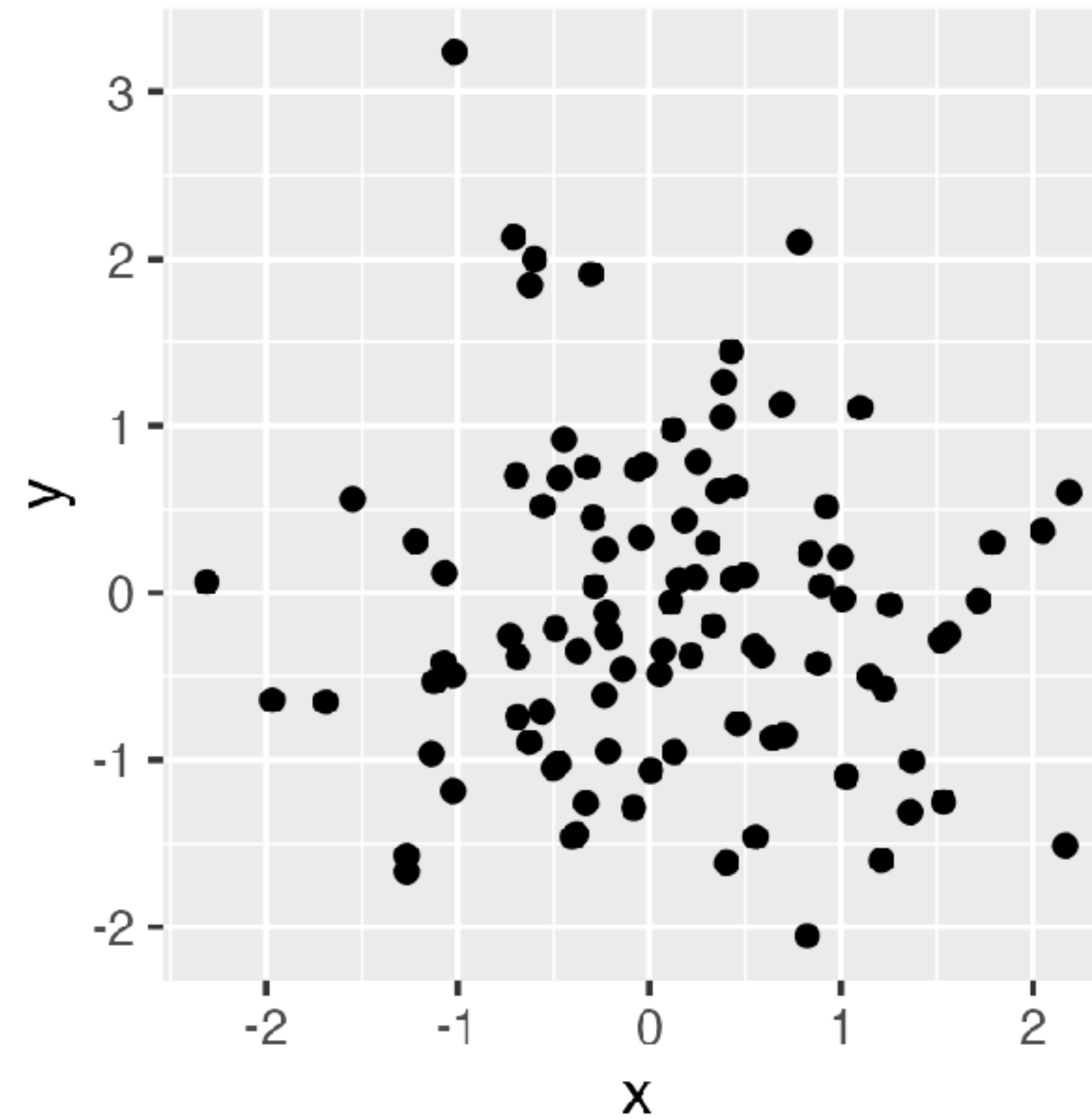
変数の関係が強いほど絶対値が1に近づく

```
cor(penguins$bill_length_mm, penguins$body_mass_g, use = "complete.obs")  
#> [1] 0.5951098
```

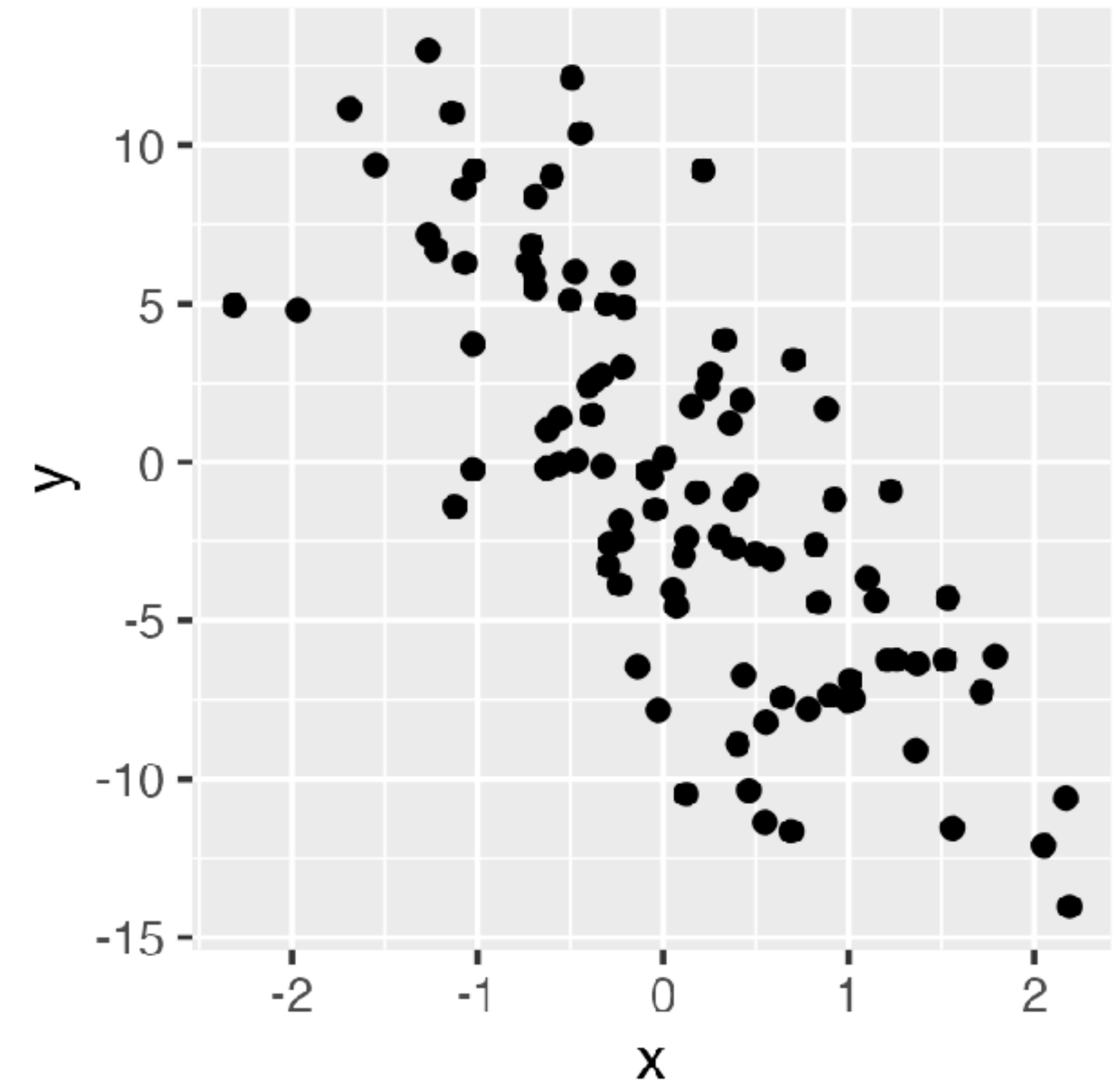
正の相関関係 $r = 0.829446$



無相関 $r = -0.04953215$



負の相関関係 $r = -0.7527922$



相関係数の大きさの目安

相関係数	相関の強さ
± 0.7 以上	とても強い
$\pm 0.4 \sim 0.7$	やや強い
$\pm 0.2 \sim 0.4$	弱い
± 0.2 以下	ほとんどなし

相関係数行列

変数のペアごとに計算した相関係数を行列形式で表現する

```
cor(penguins[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")],  
    use = "complete.obs")
```



	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
bill_length_mm	1.0000000	-0.2350529	0.6561813	0.5951098
bill_depth_mm	-0.2350529	1.0000000	-0.5838512	-0.4719156
flipper_length_mm	0.6561813	-0.5838512	1.0000000	0.8712018
body_mass_g	0.5951098	-0.4719156	0.8712018	1.0000000

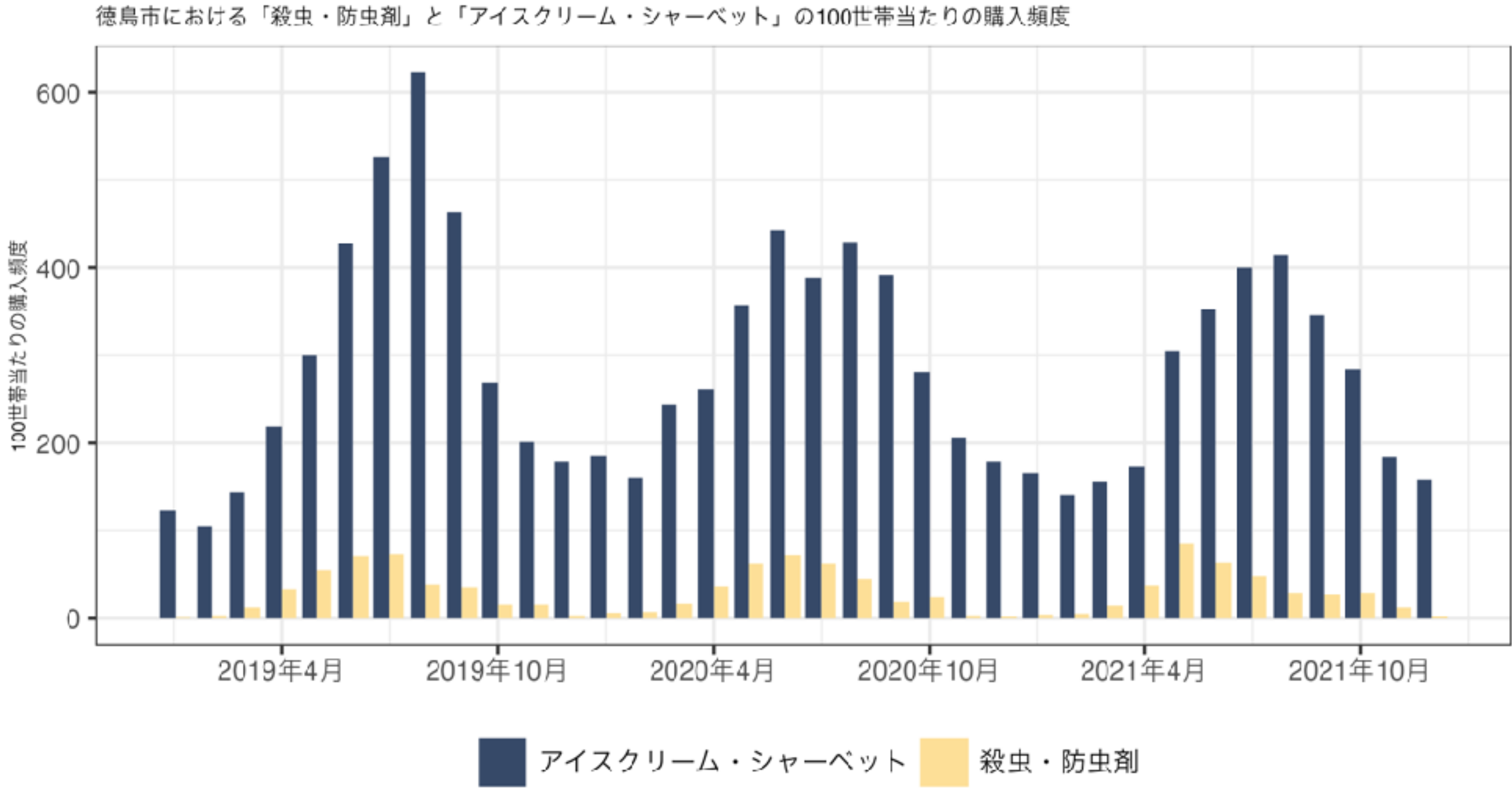
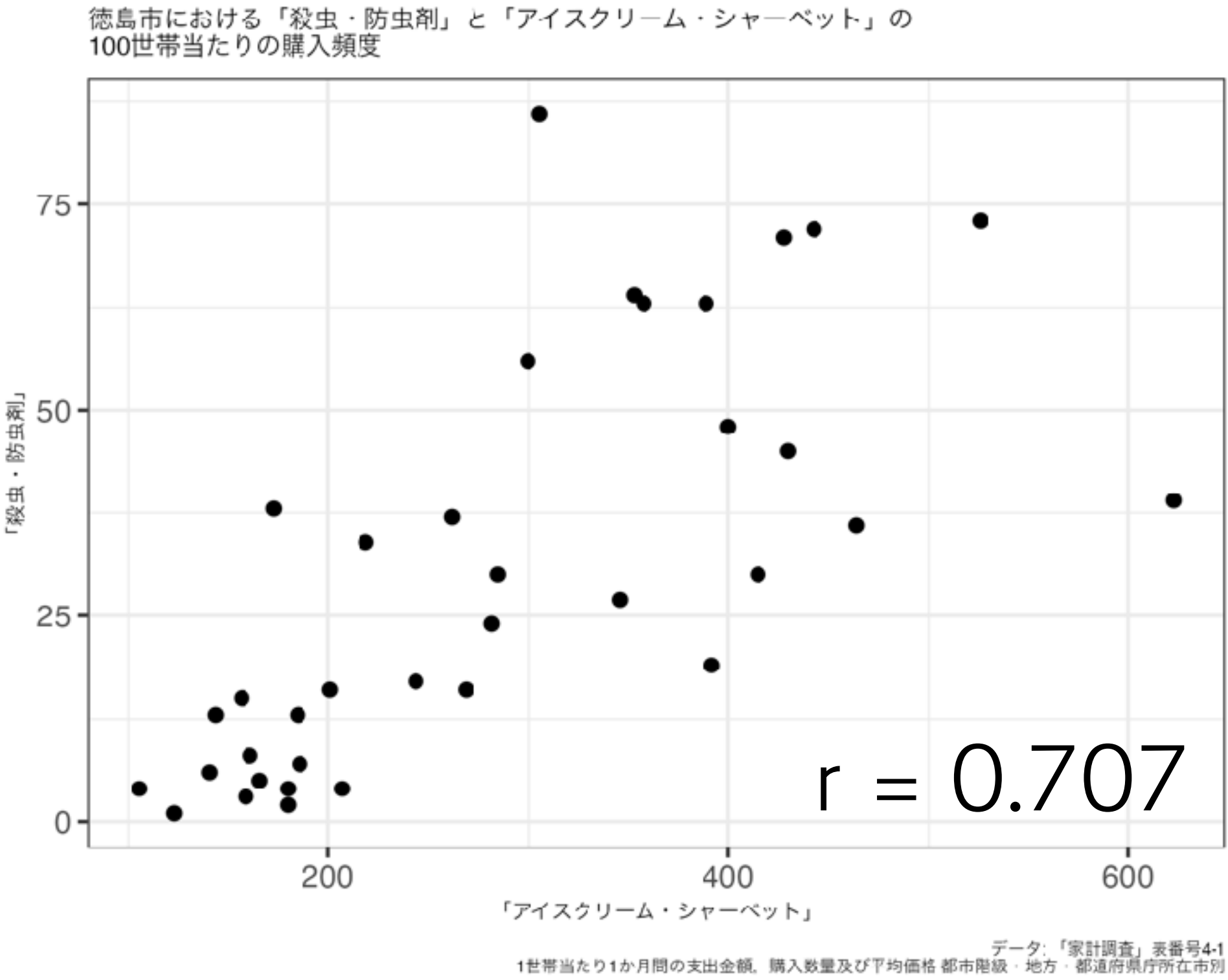
自分自身との相関係数は1

疑似相関（見せかけの相関）

次回、関連する話題を扱います

因果関係がありそうに見える二変数の関係が、
観測されていない第三の変数（潜在変数） の効果によってもたらされるもの

例）「殺虫・防虫剤」と「アイスクリーム・シャーベット」の購入頻度… **どちらも気温の影響を受ける**



データ:「家計調査」表番号4-1
1世帯当たり1か月間の支出金額、購入数量及び平均価格 都市階級・地方・都道府県庁所在市別

→相関から因果関係を導き出すのは難しい。データを得る際の設計・計画が重要

比較する

仮説検定の手順

参考) 第8回: データからの推論

あるクラスの生徒の身長について、男女で平均値に差があるか
→2つのデータ間の平均値の差の検定には**t検定**を利用

1. 帰無仮説と対立仮説を設定する

帰無仮説

H_0 →男女間で身長の平均値の差はない

対立仮説

H_1 →男女間で身長の平均値の差がある

2. 有意水準 α を設定する(5%、1%など)

3. 検定統計量を計算する

t検定ではt値を検定統計量とする

4. 検定統計量の分布を決定する

t検定ではt値がt分布に従う確率分布として扱う

5. 検定統計量の実現値を計算する

自由度(データ数-1)によって形状が変わる分布

6. p値を計算する

棄却域に含まれるかを評価

得られたデータから仮説が正しいか、稀なことか(偶然ではないこと)を評価する

【再】 仮説検定の手順

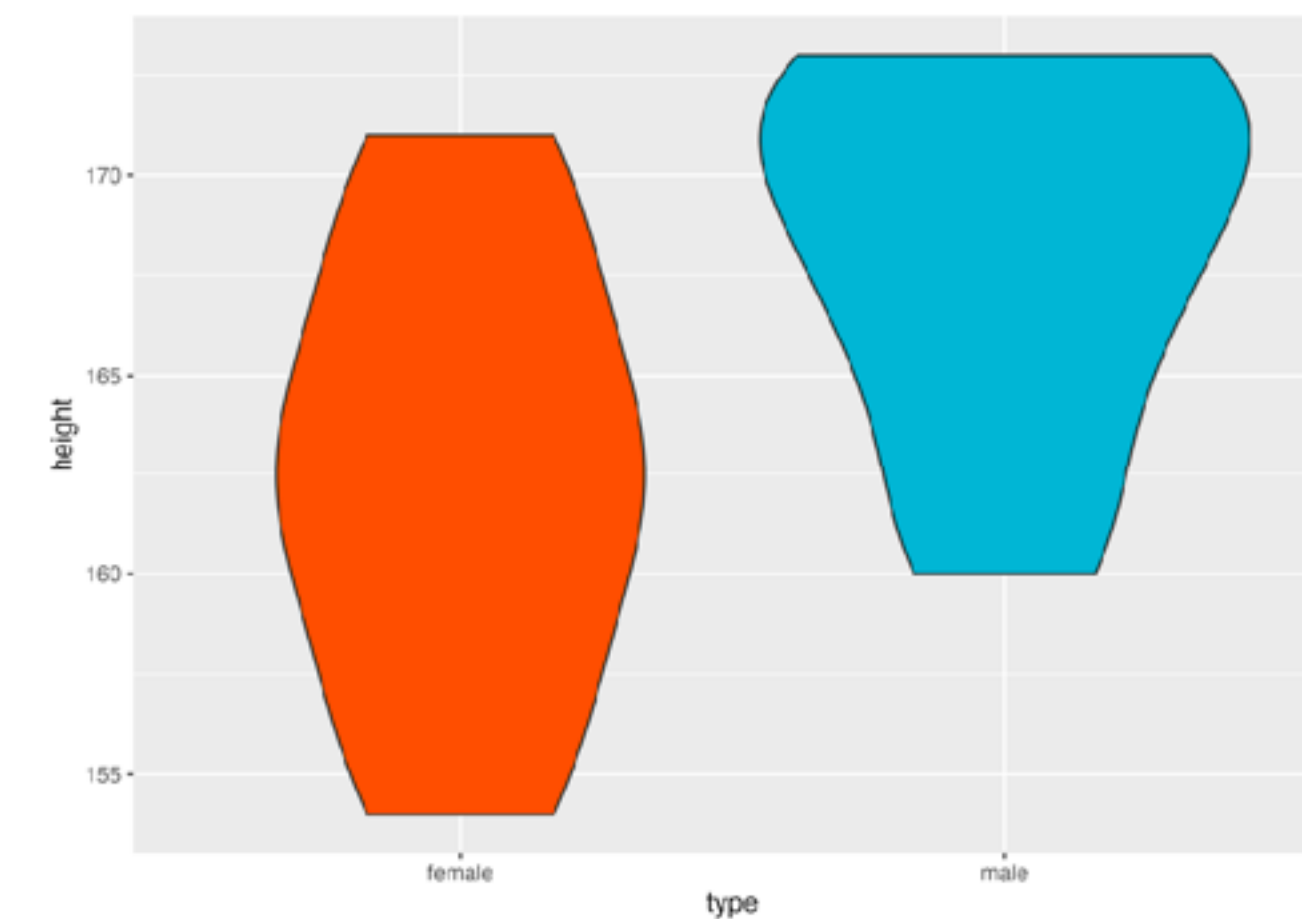
あるクラスの生徒の身長について、男女で平均値に差があるか

男女別でのクラスの生徒の身長

```
height_male <-  
  c(160, 162, 165, 167, 168, 170, 171, 172, 173, 173)  
height_female <-  
  c(156, 158, 161, 163, 164, 162, 167, 154, 169, 171)
```

```
t.test(height_male, height_female, var.equal = TRUE)  
Two Sample t-test
```

```
data: height_male and height_female  
t = 2.4679, df = 18, p-value = 0.02384  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.8327635 10.3672365  
sample estimates:  
mean of x mean of y  
 168.1    162.5
```



比較対象（変数）が1つの場合

1つの母集団の母数に関する検定

例) ある学校の生徒の身長の平均が全国平均と等しいかどうか

帰無仮説: 標本平均は全国平均と差がない $H_0 : \mu = \mu_0$

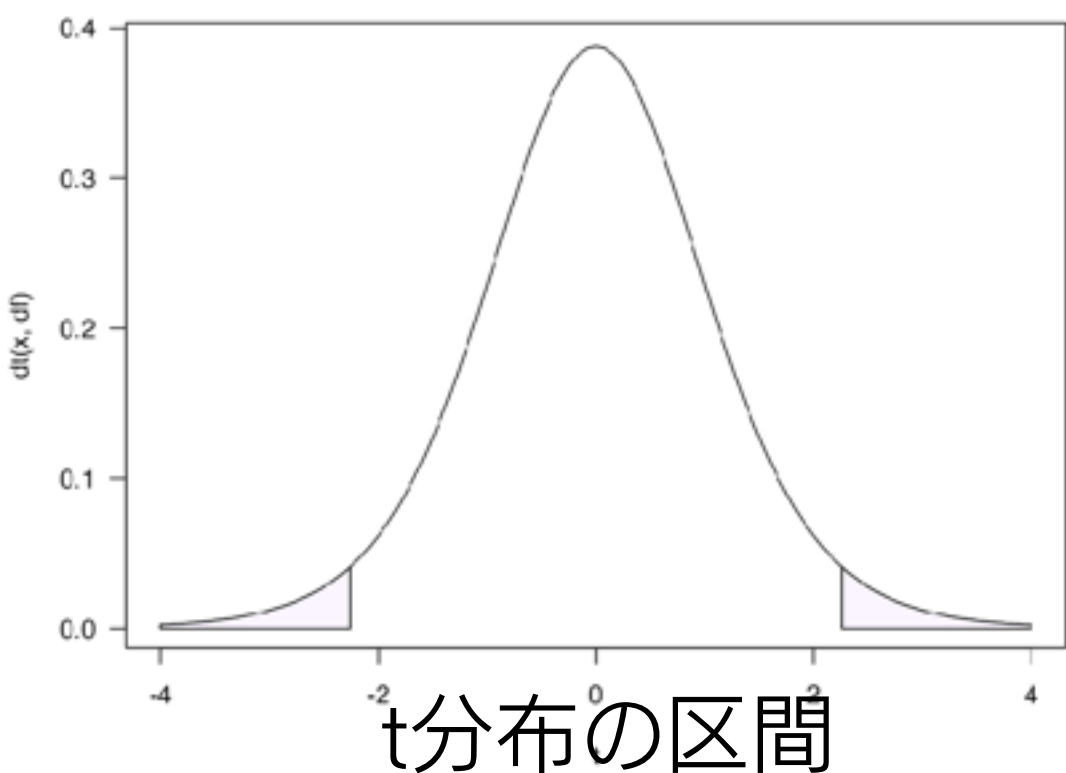
対立仮説: 身長の標本平均は全国平均に対して差がある $H_1 : \mu \neq \mu_0$



標本の件数、標本平均、帰無仮説が仮定する平均値から検定統計量（t値）を計算

有意水準5% t分布の両側5%の区間にt値が含まれるか→

YES。身長差がないとする帰無仮説を棄却。統計的に有意な差異があるとする



統計的な差がある

t値

統計的な差はない 統計的な差がある

t値

t値



t分布の区間

t値: t検定における検定統計量

帰無仮説（例えば、2つの標本間での平均に差がない）が真と仮定した際、その値がどれだけ異常であるかを示す

帰無仮説が真である場合の期待される分布（t分布）に基づき、**p値**を計算する

t値よりも極端な値が得られる確率を示す

帰無仮説を棄却するための判断材料

p値 < 有意水準となれば帰無仮説は棄却される

比較対象が複数（2つ）の場合

2つの標本間に対応関係があるかないか

（同一個体の）前後の比較

対応のあるt検定

$$t = \frac{\bar{x} - \bar{y}}{\text{var}(\bar{x} - \bar{y})} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

運動を行う前の血圧

```
x1 <- c(104, 120, 114, 122, 114, 134, 120,  
        100, 114, 130, 110, 114, 108)
```

運動を行った後の血圧

```
x2 <- c(144, 128, 134, 126, 134, 114, 128,  
        148, 134, 118, 138, 134, 140)
```

```
t.test(x1, x2, mu = 0, paired = TRUE, var.equal = TRUE)
```

差がないことを仮定し、mu = 0 とする




Paired t-test


```
data: x1 and x2  
t = -3.1246, df = 12, p-value = 0.008779  
alternative hypothesis: true mean difference is not equal to 0  
95 percent confidence interval:  
 -28.20129 -5.02948  
sample estimates:  
mean difference  
 -16.61538
```


p < 0.05 で帰無仮説は棄却される


運動


参考資料・URL

- 

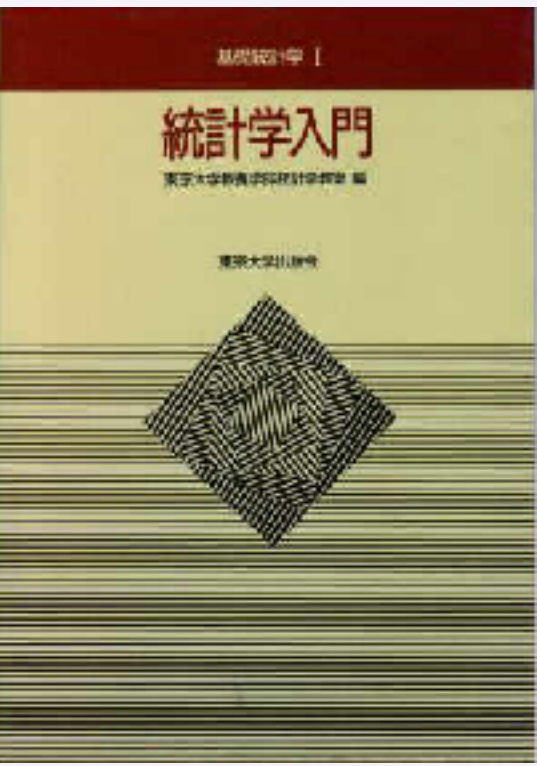
東京大学教養学部統計学教室（編）『基礎統計学I: 統計学入門』（1991）
東京大学出版会. ISBN: 4-13-042065-8
瓜生居室: あり、徳大図書館: あり、[市立図書館](#): なし、[県立図書館](#): あり
- 

滋賀大学データサイエンス学部, 長崎大学情報データ科学部（編）『データサイエンスの歩き方』（2022）学術図書出版社. ISBN: 978-4-7806-0936-3
瓜生居室: あり（電子版）、[徳大図書館](#): あり、[市立図書館](#): なし、[県立図書館](#): なし
- 


中西啓喜（編）耳塚寛明（監修）『教育を読み解くデータサイエンス：データ収集と分析の論理』（2021）ミネルヴァ書房. ISBN: 978-4-623-09172-0
瓜生居室: あり、[徳大図書館](#): なし、[市立図書館](#): なし、[県立図書館](#): なし
- 

Peter Bruce, Andrew Bruce, Peter Gedeck（著）, 黒川利明（訳）
『データサイエンスのための統計学入門（第二版）』（2020）オライリー・ジャパン. ISBN: 978-4-87311-926-7
瓜生居室: あり（電子版）、徳大図書館あり（第一版）、[市立図書館](#): なし、[県立図書館](#): あり
- 

阿部真人『統計学入門：データ分析に必須の知識・考え方』（2021）ソシム.
ISBN: 978-4-8026-1319-4
瓜生居室: あり（電子版）、徳大図書館: あり、[市立図書館](#): なし、[県立図書館](#): なし




参考資料・URL



日本統計学会（編）『データの分析：日本統計学会公式認定統計検定3級対応』（2020）東京図書.
ISBN: 978-4-489-02332-3

瓜生居室: あり、徳大図書館: あり（初版）、
、市立図書館: なし、県立図書館: なし



日本統計学会（編）『統計学基礎：日本統計学会公式認定統計検定2級対応』（2015）東京図書.
ISBN: 978-4-489-02227-2

瓜生居室: あり、徳大図書館: あり、
、市立図書館: なし、県立図書館: なし

