

# 数理・データサイエンス・AI入門

## 第5回 データサイエンス知識(3)

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

0

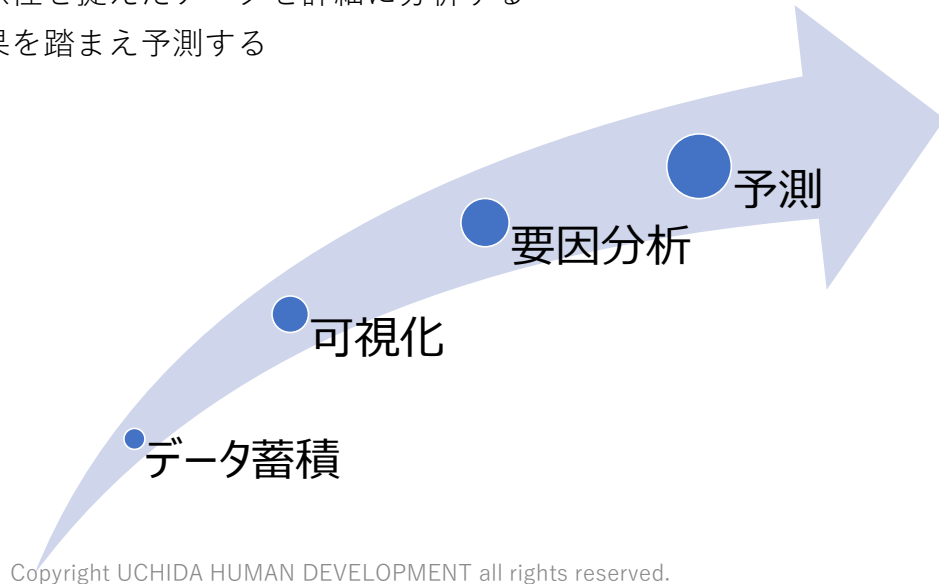
## データの分析と予測

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

1

# データサイエンスの一般的な進め方

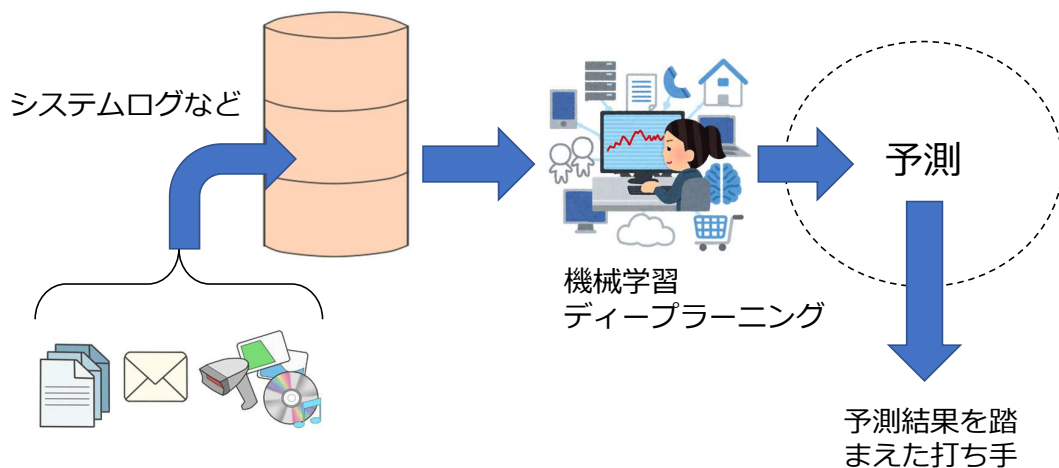
- データ蓄積・・・単年分の蓄積からまずは始めて複数年分蓄積
- 可視化・・・蓄積したデータを可視化して大きな傾向を捉える
- 要因分析・・・関係性を捉えたデータを詳細に分析する
- 予測・・・分析結果を踏まえ予測する



2

## 予測のしくみ

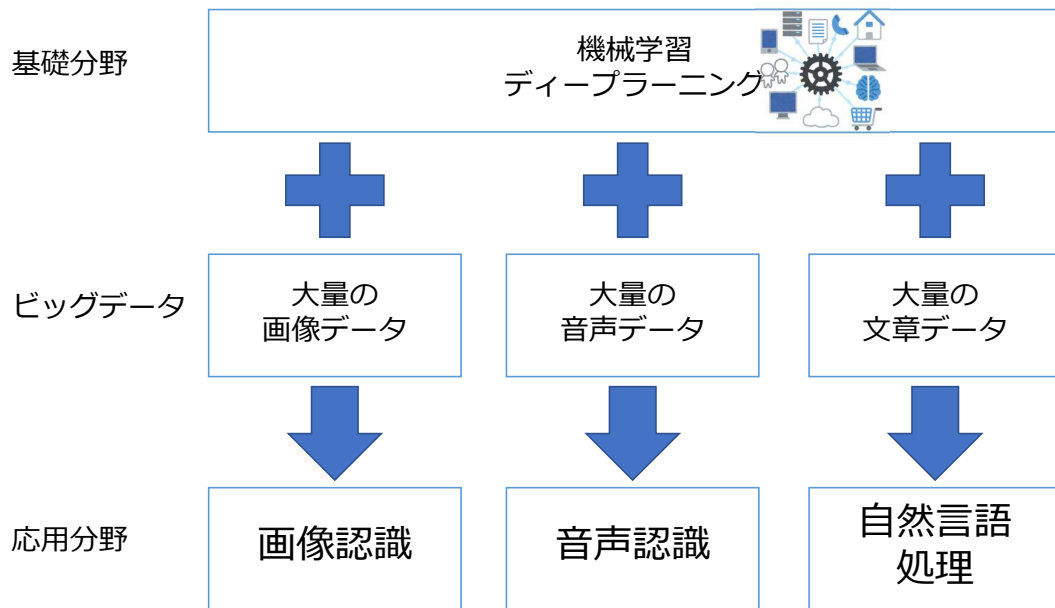
- 機械学習を使って予測するためのモデルを作る



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

3

# 機械学習

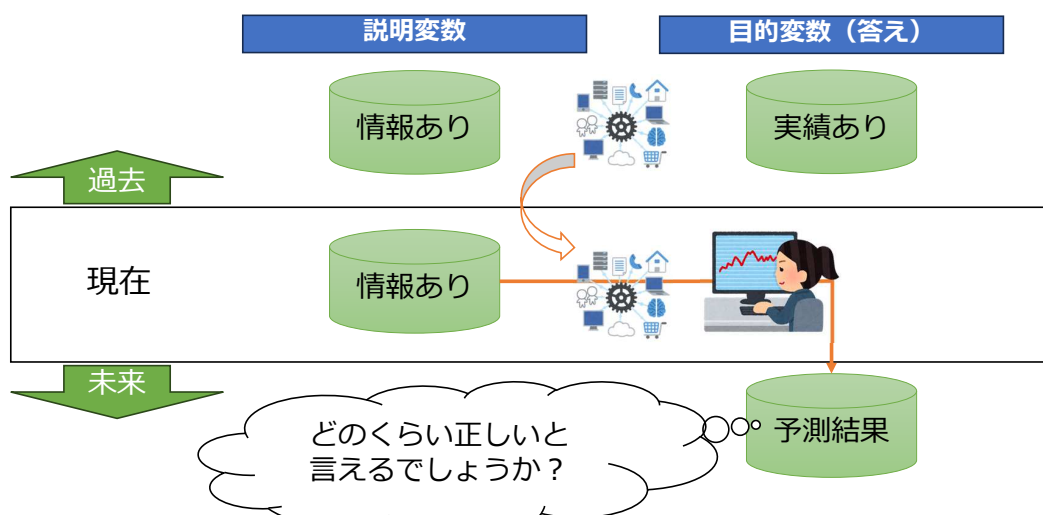


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

4

## 予測モデル作成とは

- 作成したモデルを現在の情報に適用し、未来の結果を予測

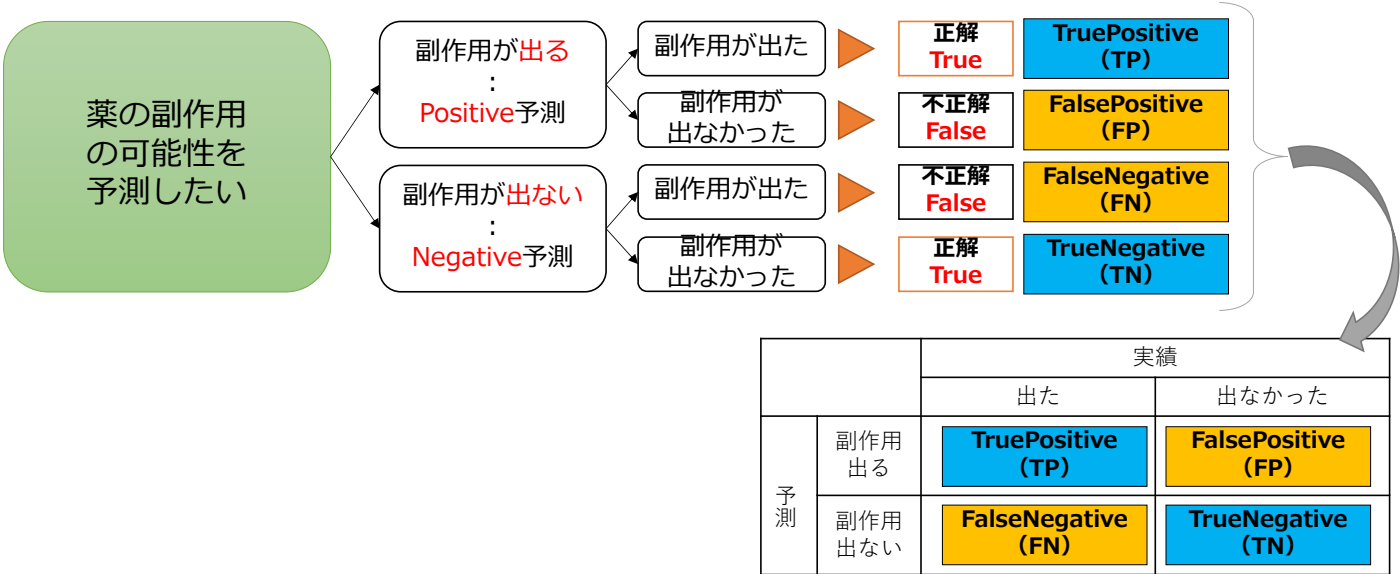


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

5

# 予測精度を検証する方法

## 4 パターンそれぞれの名称



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

6

## 【クイズ】

### 下記 2 つの結果、どちらのほうが精度が高いと考えますか？

- 新薬の副作用発生の予測

予測		実績		
		出た	出なかった	合計
副作用 出る	450	100	350	450
副作用 出ない	350	0	350	350
合計	800	100	700	800

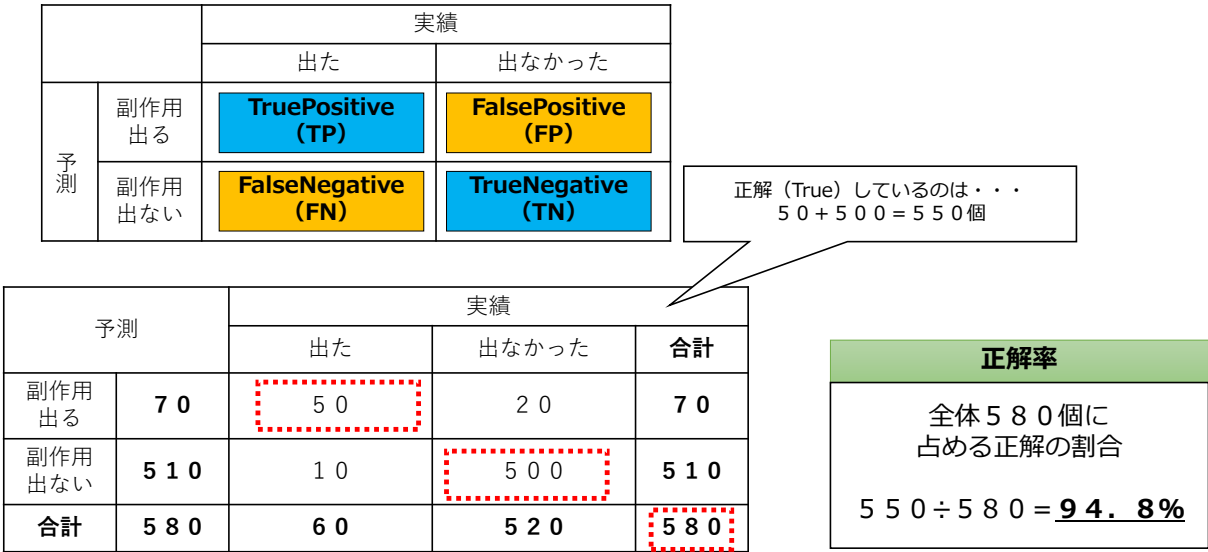
予測		実績		
		出た	出なかった	合計
副作用 出る	4	1	3	4
副作用 出ない	796	99	697	796
合計	800	100	700	800

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

7

# 精度を測る指標

精度を捉える一番シンプルな方法は、正解数の割合で見する方法



# 精度を測る指標

- 目的にあわせて精度の考え方は変わる
- 正解をどれだけ正しく予測できたかが重要なとき・・・再現率
  - 予測がどれだけ当たったかが重要なとき・・・適合率

# 精度検証のポイント

---

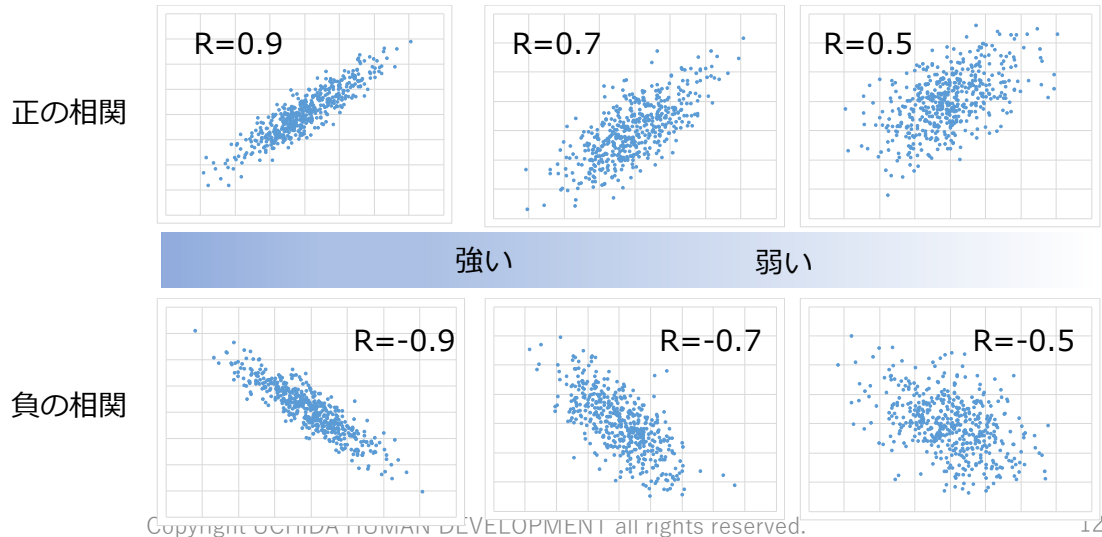
## □ 精度検証の結果を踏まえて次の分析を行う

- 精度を上げるためには、説明変数を変えながら何度も分析
- 再現率を上げると適合率が下がる、あるいは逆に適合率を上げると再現率が下がるということが発生
  - どちらかを100%とすることが求められるのか、あるいは双方がなるべく高い状態が良いのか、課題によって異なります

# 相関と回帰

# 相関

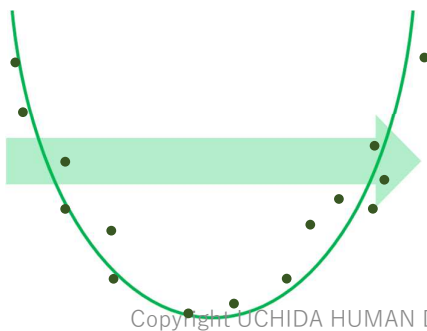
- 2種類のデータの関係性を調べる方法
- 相関係数で関係性の強さを見る
  - 相関係数「R」は、-1~1の間に収まる



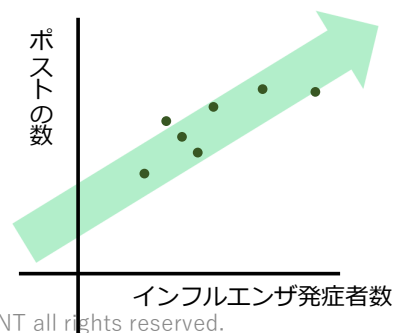
## 相関係数の注意点

- 直線の関係性は計測できるが、曲線の関係性は計測できない
- 分かるのは関係性の有無で、因果関係があるとは限らない

例) 明らかに関係性のある下図のような散布図も相関係数の値は0に近くなる。

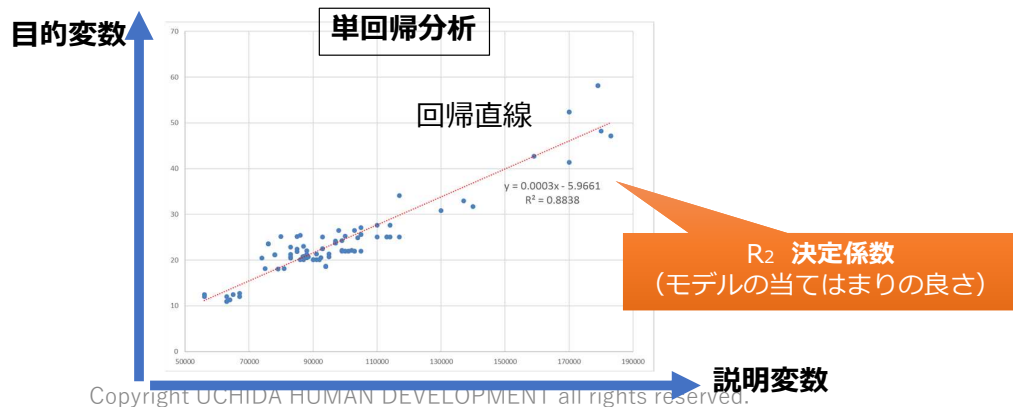


例) インフルエンザの発症数とポストの数に相関があった。ポストを減らせば発症を抑えられる？



# 回帰分析

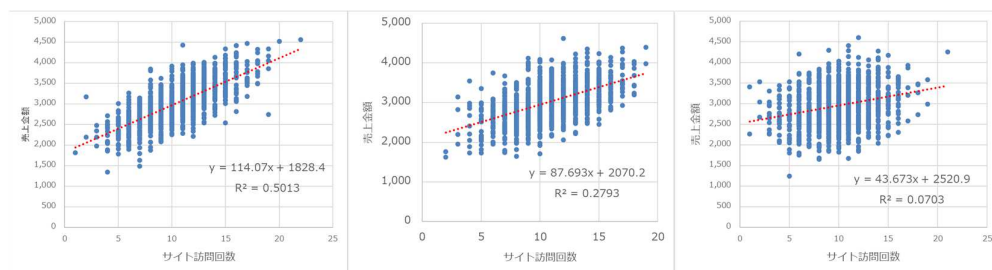
- 予測したい値（目的変数）をいくつかの要因となる項目（説明変数）の影響度で関係を説明する方法
  - 説明変数が1つの場合：単回帰分析
  - 説明変数が複数の場合：重回帰分析
- 単回帰分析は「散布図」から算出することができる
- 単回帰分析のモデル式は、「 $y = ax + b$ 」となる



14

## 回帰分析の結果の読み方

- 決定係数がモデルの当てはまりの良さを表す
  - 決定係数が0.5以上あれば当てはまりがよいと言える
  - 0.2未満であればモデルを見直す



決定係数

$R^2 = 0.5013$

$R^2 = 0.2793$

$R^2 = 0.0703$

相関係数

2乗したもの  
 $R = 0.708$

$R = 0.5285$

$R = 0.2652$



# 重回帰分析

□ 回帰分析で説明変数が複数ならば、重回帰分析

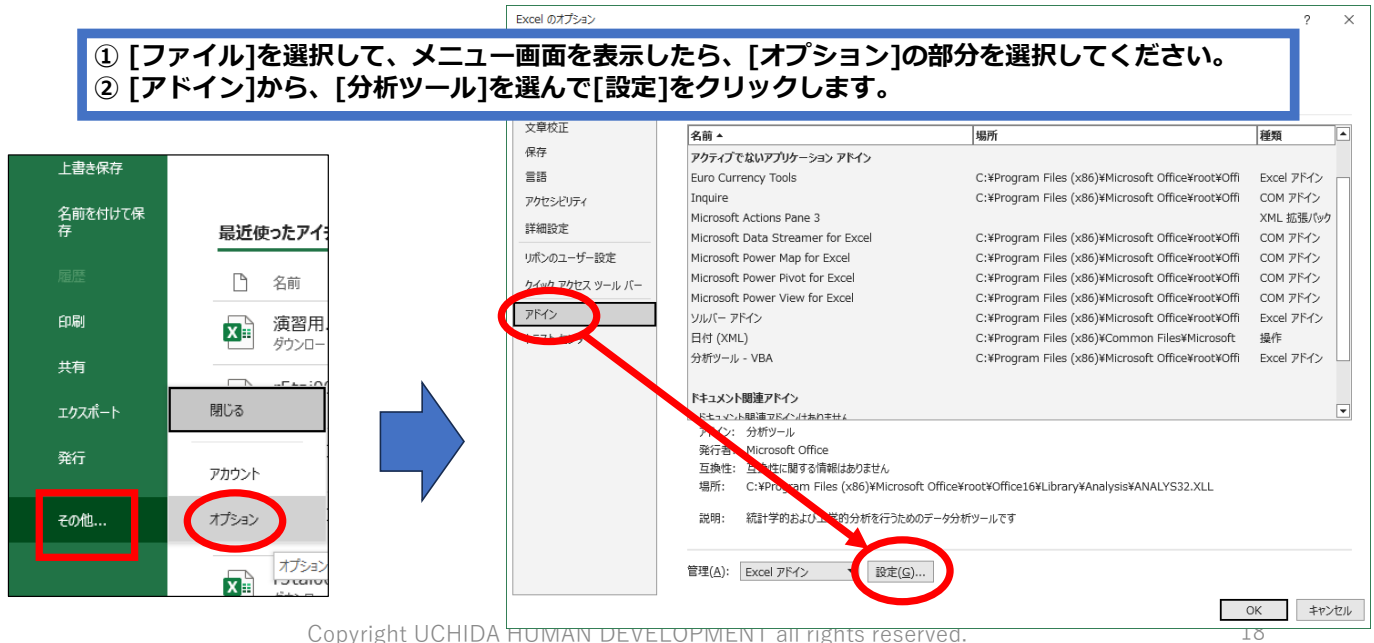


# Excelを使用した分析方法

# Excelで相関・回帰分析

## ■ Excelのアドインを設定するとデータ分析機能が使えます

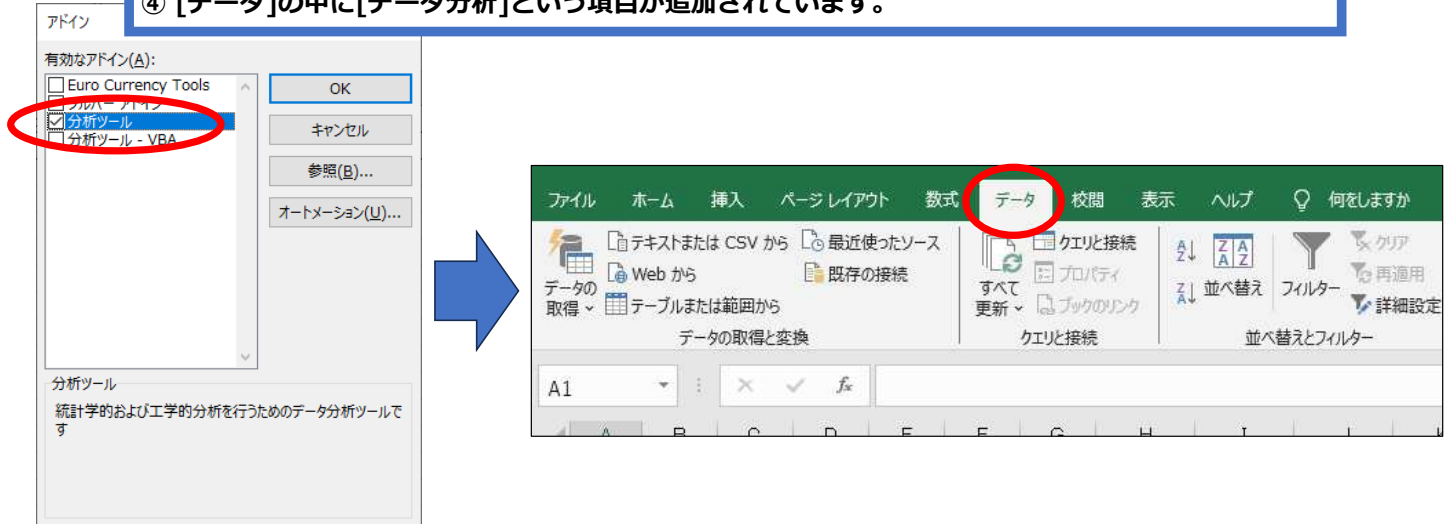
- ① [ファイル]を選択して、メニュー画面を表示したら、[オプション]の部分を選択してください。
- ② [アドイン]から、[分析ツール]を選んで[設定]をクリックします。



# Excelで相関・回帰分析

## ■ Excelのアドインを設定するとデータ分析機能が使えます

- ③ [分析ツール]のチェックボックスをクリックします。
- ④ [データ]の中に[データ分析]という項目が追加されています。



# 相関分析

## □ 相関分析をしてみましょう

## □ データの準備

- e-Stat (<https://www.e-stat.go.jp/>)



20

# データの準備

## □ e-Stat (<https://www.e-stat.go.jp/>)

教育・文化・スポーツ・生活

主な調査

- 社会生活基本調査
- 学校基本調査
- 学校教員統計調査
- 社会教育調査

すべて見る (28 調査)

00400205	外国人の子供の就学状況等調査	詳細
00400304	児童生徒の問題行動・不登校等生徒指導上の諸課題に関する調査	詳細
00400305	日本語指導が必要な児童生徒の受入状況等に関する調査	詳細
00400306	学校における教育の情報化の実態等に関する調査	詳細
00400401	大学・短期大学・高等専門学校における学生のキャリア形成支援活動実施状況等調査	詳細
00400802	学校給食実施状況等調査	詳細
00400803	学校給食栄養報告	詳細
00401101	宗教統計調査	詳細
00402101	体育・スポーツ施設現況調査	詳細
00402102	体力・運動能力調査	詳細
00700002	留学生受入実績	詳細

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

21

# データの準備

## 令和5年度（速報）のデータを使用します

体力・運動能力調査

一覧形式で表示

体力・運動能力調査は、国民の体力・運動能力の現状を明らかにするため、最初の東京オリンピックが開催された昭和39年以来、毎年実施されている調査です。これだけ継続的に幅広い年齢層（6～11歳、12歳～19歳、20歳～64歳、65歳～79歳）を対象として、国民の体力・運動能力を把握する調査は世界にも類を見ない貴重なものであり、本調査の結果は、国民の体力づくり、健康の保持・増進に資するとともに、体育・スポーツ活動の指導と行政上の基礎資料として広く活用されています。

体力・運動能力調査	データベース	件数 更新日	ファイル	件数 更新日	概要
体力・運動能力調査		36件 2020-01-08		2,822件 2024-03-29	



体力・運動能力調査

詳細

体力・運動能力調査は、国民の体力・運動能力の現状を明らかにするため、毎年実施されている調査です。これだけ継続的に幅広い年齢層（6～11歳、12歳～19歳、20歳～64歳、65歳～79歳）を対象として、国民の体力・運動能力を把握する調査は世界にも類を見ない貴重なものであり、本調査の結果は、国民の体力づくり、健康の保持・増進に資するとともに、体育・スポーツ活動の指導と行政上の基礎資料として広く活用されています。

体力・運動能力調査

- 令和5年度（速報） [13件]
- 令和4年度 [158件]
- 令和3年度 [158件]
- 令和2年度 [158件]
- 令和元年度 [158件]

# データの準備

## 今回データは3種類使用します

データセット一覧

戻る

URLをコピー

一覧形式で表示

政府統計名	体力・運動能力調査	詳細
提供統計名	体力・運動能力調査	
提供分類1	令和5年度（速報）	

表番号	統計表	調査年月	公開（更新）日	表示・ダウンロード
1	年齢別テストの結果			
	握力、上体起こし、長座体前屈	2023年度	2024-03-29	
	反復横とび、20mシャトルラン、持久走・急歩	2023年度	2024-03-29	
	50m走、立ち幅とび、ソフトボール投げ・ハンドボール投げ	2023年度	2024-03-29	
	開眼片足立ち、10m障害物歩行、6分間歩行	2023年度	2024-03-29	
	合計点	2023年度	2024-03-29	

# データの準備

## 分析に使用する列だけをコピー（3 ファイル分）

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S			
3		反復横とび					(点)	20mシャトルラン					(折り返し数)			持久走・急歩						(秒)
4	年齢	男 子			女 子			男 子			女 子			男 子			女 子					
5		標本数	平均値	標準偏差	標本数	平均値	標準偏差	標本数	平均値	標準偏差	標本数	平均値	標準偏差	標本数	平均値	標準偏差	標本数	平均値	標準偏差			
6	6	1099	27.65	5.24	1085	26.92	4.93	1095	18.97	9.81	108	15.70	6.55	-	-	-	-	-	-			
7	7	1100	31.12	6.36	1085	30.07	5.59	1108	28.01	13.67	109	20.93	9.44	-	-	-	-	-	-			
8	8	1105	34.51	7.69	1095	33.19	7.04	1121	35.44	17.10	111	27.43	12.21	-	-	-	-	-	-			
9	9	1105	38.47	7.41	1105	36.76	7.05	1125	43.91	19.11	112	33.12	14.83	-	-	-	-	-	-			
10	10	1115	42.13	7.33	1120	40.37	6.98	1126	50.40	21.65	112	40.24	16.39	-	-	-	-	-	-			
11	11	1112	45.55	7.18	1116	42.66	6.48	1118	59.46	22.76	111	44.43	17.93	-	-	-	-	-	-			
12	12	1345	49.80	7.05	1355	45.77	6.10	1104	68.91	24.19	110	48.37	19.20	289	426.39	62.15	282	304.56	39.23			
13	13	1364	53.02	7.53	1366	47.12	6.29	1120	82.63	23.51	112	53.74	20.36	297	388.55	52.91	271	292.56	41.54			
14	14	1375	55.88	7.38	1387	48.35	6.39	1126	90.43	24.14	112	53.82	19.37	286	379.35	53.43	272	289.14	38.95			
15	15	1418	55.38	7.52	1421	47.13	6.38	1067	77.69	26.12	106	44.83	18.17	355	388.55	54.32	357	309.01	42.31			
16	16	1434	56.93	7.80	1421	47.95	6.47	1101	85.50	27.86	106	46.80	-	-	-	-	-	-	-			
17	17	1431	57.47	7.81	1430	48.01	6.92	1074	83.78	28.57	105	46.11	19.88	367	371.23	54.09	369	296.23	43.52			
18	18	1091	57.18	7.75	1063	47.75	6.67	785	74.01	26.12	70	40.36	16.12	312	401.75	57.78	367	324.54	41.29			
19	19	776	56.29	8.02	752	46.72	6.77	430	73.48	25.67	45	39.42	15.03	220	409.33	60.64	231	323.59	41.83			
20	20-24	1001	55.03	7.50	815	46.51	6.78	819	69.23	24.06	64	37.73	15.65	132	690.10	110.23	138	520.93	55.06			
21	25-29	967	54.50	7.27	760	44.96	6.95	794	66.70	23.45	57	34.28	14.04	134	685.09	104.48	144	534.61	52.59			
22	30-34	1036	52.51	7.26	803	43.55	6.52	813	57.93	23.52	58	30.21	12.78	157	686.16	97.99	150	532.28	46.91			
23	35-39	1227	50.55	6.99	1107	41.70	6.26	924	56.89	23.65	77	26.10	11.62	180	699.53	85.01	177	536.62	45.86			
24	40-44	1213	48.21	6.93	1225	40.62	6.35	886	50.85	22.50	82	24.51	11.44	202	724.38	83.79	222	541.50	55.10			
25	45-49	1242	46.86	6.62	1205	40.38	6.01	933	46.16	21.30	82	24.24	11.17	216	733.78	97.23	236	544.02	56.67			
26	50-54	1121	45.36	6.71	1084	39.30	6.09	801	40.30	18.78	73	22.16	10.89	229	746.35	99.89	250	542.22	60.83			
27	55-59	855	44.08	6.21	985	38.26	5.71	622	35.12	17.51	68	20.25	10.34	191	741.63	106.81	216	547.32	61.26			
28	60-64	922	42.07	6.47	1023	37.12	5.98	694	29.89	15.84	70	17.11	8.97	184	773.10	91.59	246	555.65	62.90			

新しいシート

	A	B	C
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

# データの準備

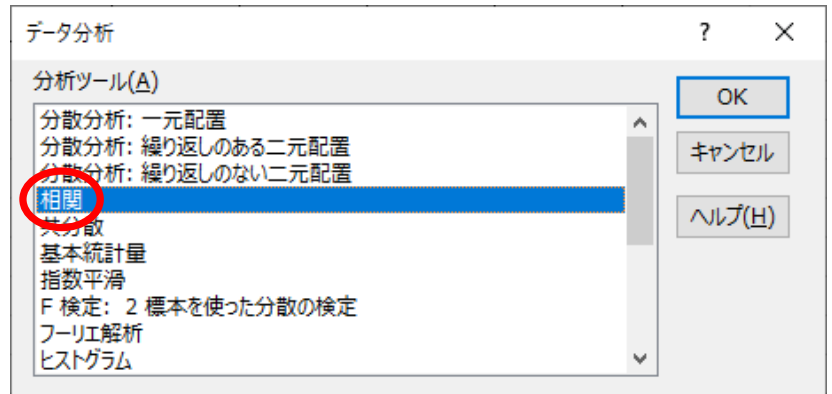
## 完成データ

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		男 子	男 子	男 子	男 子	男 子	男 子	女 子	女 子	女 子	女 子	女 子	女 子
2		握力	上体起こし	長座体前屈	反復横とび	20mシャトルラン	立ち幅とび	握力	上体起こし	長座体前屈	反復横とび	20mシャトルラン	立ち幅とび
3	6	9.17	11.47	26.70	27.65	18.97	117.36	8.56	11.07	28.87	26.92	15.70	108.49
4	7	10.83	14.40	28.14	31.12	28.01	127.40	10.16	13.69	30.78	30.07	20.93	117.82
5	8	12.51	15.82	29.36	34.51	35.44	135.68	11.89	15.24	32.51	33.19	27.43	127.78
6	9	14.43	17.91	31.55	38.47	43.91	144.48	14.02	17.32	35.23	36.76	33.12	137.53
7	10	16.72	19.95	33.63	42.13	50.40	154.62	16.56	18.89	38.40	40.37	40.24	147.39
8	11	19.88	22.07	36.20	45.55	59.46	166.70	19.36	19.89	40.92	42.66	44.43	155.26
9	12	24.30	23.48	41.68	49.80	68.91	184.99	21.93	20.76	44.04	45.77	48.37	167.52
10	13	30.24	26.42	45.16	53.02	82.63	203.14	23.90	22.18	46.51	47.12	53.74	172.45
11	14	34.77	29.00	49.40	55.88	90.43	216.86	25.20	23.46	48.63	48.35	53.82	174.37
12	15	36.83	27.80	47.04	55.38	77.69	217.84	25.05	21.80	47.43	47.13	44.83	169.73
13	16	39.32	29.27	49.75	56.93	85.50	223.57	26.08	23.01	48.78	47.95	46.80	171.43
14	17	40.52	30.15	51.50	57.47	83.78	227.13	26.53	23.62	49.58	48.01	46.11	172.06
15	18	39.26	29.18	48.82	57.18	74.01	226.36	25.87	22.68	47.59	47.75	40.36	167.98
16	19	39.96	29.65	49.97	56.29	73.48	225.19	26.13	22.92	47.84	46.72	39.42	167.34
17	20-24	44.11	28.37	44.57	55.03	69.23	224.57	26.84	21.18	45.21	46.51	37.73	167.28
18	25-29	45.60	28.05	43.97	54.50	66.70	221.83	27.66	20.14	44.05	44.96	34.28	163.16
19	30-34	45.67	26.87	42.93	52.51	57.93	215.43	27.78	18.39	43.32	43.55	30.21	158.16
20	35-39	46.28	25.75	41.73	50.55	56.89	211.59	28.13	17.02	42.93	41.70	26.10	153.80
21	40-44	45.78	23.77	40.12	48.21	50.85	202.66	28.16	16.08	40.88	40.62	24.51	149.02
22	45-49	45.30	22.46	38.54	46.86	46.16	196.02	27.84	15.33	41.55	40.38	24.24	146.62
23	50-54	44.31	21.85	38.06	45.36	40.30	190.63	27.05	14.04	40.73	39.30	22.16	139.68
24	55-59	43.41	20.54	37.12	44.08	35.12	183.46	26.78	13.64	41.47	38.26	20.25	135.71
25	60-64	41.94	19.12	36.75	42.07	29.89	174.85	26.08	12.59	41.25	37.12	17.11	129.95

Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

# Excelで相関分析

- アドイン機能を使って相関分析が行えます

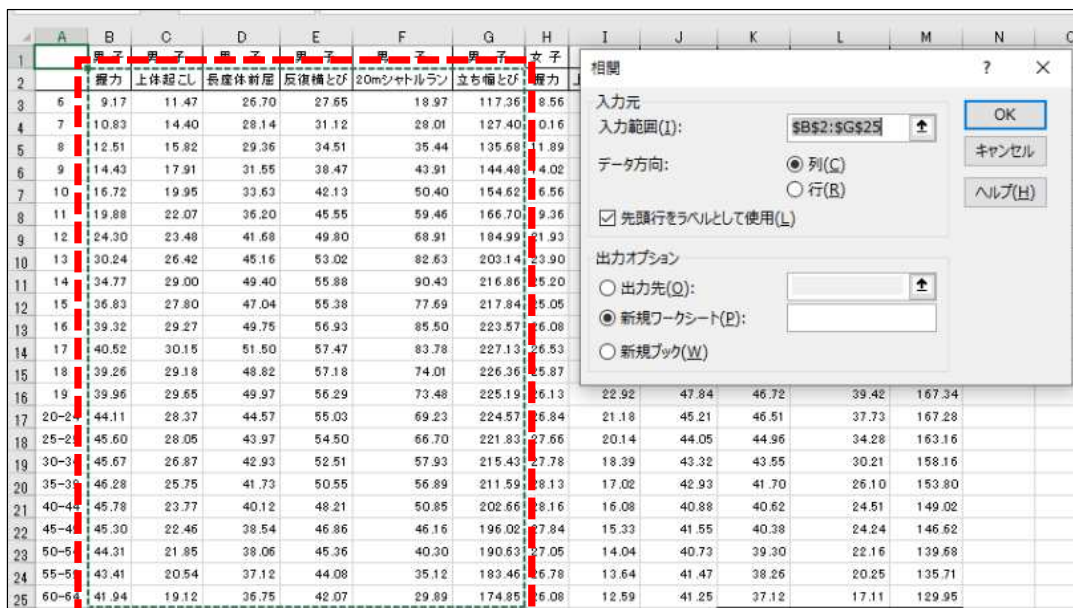


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

26

## 相関分析

- 関係性を見たい2列を選びます（3列以上も可能）



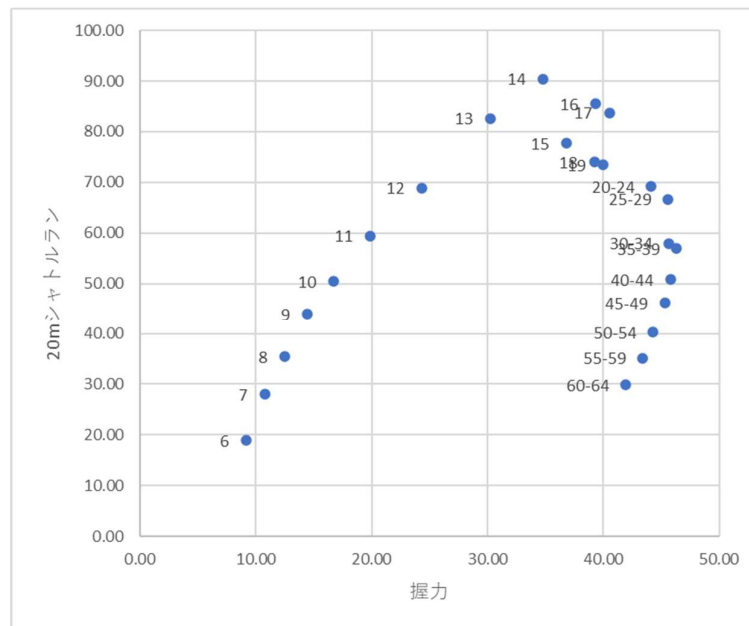
Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

27



# 相関分析

## □ 散布図を描いて相関係数と見比べましょう

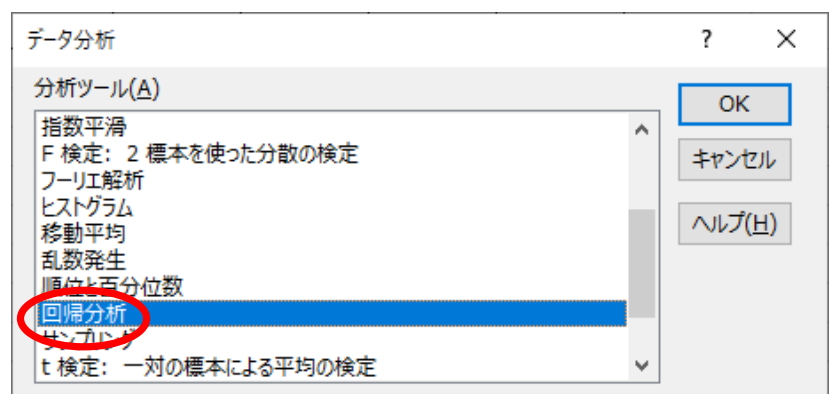


Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

28

# Excelで回帰分析

## □ アドイン機能で回帰分析も行えます



Copyright UCHIDA HUMAN DEVELOPMENT all rights reserved.

29