SSDSEを用いた分析事例1 Excelで行う相関分析



データ分析を行う場合、データの関係性を知ることは基本的かつ重要なことです。

データの関係性を明らかにする分析手法に相関分析があります。

相関分析では、大量のデータの中からデータ間の関係性を見つけることができ、関係性の有 無を客観的に判断することができます。

統計データ分析コンペティションの応募論文(高校生の部)中でも多く用いられる分析手法です。

相関分析とは

相関分析では、複数のデータ(変数)間にどのような関係性、すなわち「相関」があるか を見ます。変数間の相関を見る方法にはいくつかありますが、ここでは最も一般的な次の方 法を行います。

- ・散布図を描く
- ・相関係数を算出する

相関分析は様々なツールで行うことができますが、ここでは最も一般的で手軽な Excel を 使った方法を紹介します。

統計データ分析コンペティション 2022 の受賞論文(高校生の部:統計活用奨励賞)「**都市** 部と地方の教育格差の要因と課題~日本の教育現場において~」

(<u>https://www.nstac.go.jp/sys/files/static/statcompe/files/2022/2022H4-katsuyo.pdf</u>) を例に見てみましょう。

受賞論文では、都道府県間で教育格差の生じる要因を所得や教育費など様々なデータで相 関分析を行うことで明らかにしています。

論文中の図4は、都道府県別の「一人当たり県民所得」と「教育費」(一世帯当たりの月間 消費支出額)の関係性について相関分析を行った結果です。 これを実際に Excel で行ってみましょう。

分析で用いるデータは、「一人当たり県民所得」
 が SSDSE-基本素材 (SSDSE-E)の項目コード
 C121101 を、「教育費」が SSDSE-県別推移
 (SSDSE-B)の項目コード L322108 です。
 データは、SSDSE のウェブサイトからダウンロードできます。

(https://www.nstac.go.jp/use/literacy/ssdse/)



Excel で行う相関分析

散布図を描く

散布図は2変数を縦軸と横軸に設定し、データを点でプロットしたグラフのことです。 散布図を描くことで、データにどのような関係性があるか視覚的に確認することができます。 データを入力した Excel シート内の関係性を調べたい2変数を選択し、「挿入」タブ内の「グ ラフ」メニューから「散布図(X,Y)またはバブルチャートの挿入」⇒「散布図」を選択しま す。

ł	3 .				HEM分析 (回) _xitor - ExoH	
77	116 1	一厶 詞入	x-5 L-1791 Bost	データ 咬間 表示	■発 Q 前行したい作業を入力して(220 発共	H
100	子 おす か おす ブル ビボット テー	? すめ テーブル ブル	10. 10. 10 10. 10 10. 10 10. 10 10. 10 10. 10 10 10 10 10 10 10 10 10 10 10 10 10 1	2157 10 10 10 10 10 10 10 10 10 10 10 10 10		*
A.	L	1 8 1	× 1.	and the second se		14
4	.A.	В	C D	2530%BA	N O P Q	-
2		所得 教	電費	おすすめグラフ すべてのグラフ		1
2	北海道	2682	8848	 ・ ・ ・	Dellet Liber Li la 1a	-1
3	青炭県	2490	5095	# FSJU-N	A VO VA VA 08 08 08	1
4	治学業	2772	10722	1218		-1
0	白水色	2944	10218	必 新和線	散布闼	-
7	0.208	2033	9193	(B) (F)	19225110 BRB	-1
0	10.00	2023	8375	E 48		-
0	1010	3306	13836	No D	1 3 3 3 7 3 7 3 7 3 7 4 4 5 5 5 7 5 7 5 7 5 7 5 7 5 7 5 7 5	-1
10	N-+-R	2413	10706		the second that the second sec	-1
11	R.m.r.	3325	6493	100 0010 100 00010	at and a set of the set of the set of the	-1
12	15 E.B.	3067	20381	de la-d-		-11
13	千葉県	3193	14700	11 11-707		
14	-	5427	19935	(例) サンバースト		
15	神奈川虞	3227	14446	1 LZ1094		
16	新潟県	2873	11309	社会報びの国		
17	富山県	3319	8174	回 ウォーターフォール		
18	石川県	2962	17612	論、語み合わせ		
19	道井県	3265	11039			
20	山和風	2973	6606			
21	具野泉	2940	6496			
22	被車溝	2849	14598		OK: \$7721	
23	發出意	3388	8919	-		
24	爱知道	28.85	11680			•
-		一門相と取得費	<u>24</u> (+)			056

シート内にグラフ(散布図)が挿入されます。



挿入されたグラフはデフォルトの状態ですので、「グラフ要素を追加」メニューから、「グラ フタイトル」、「データラベル」、「軸ラベル」、「近似曲線」(回帰直線)等を挿入するなどして グラフを見やすく整えます。



相関係数を算出する

散布図を見ることでデータの関係性を視覚的に確認することができますが、関係性の解釈は 見る人によって異なるかもしれません。そこで、データの関係性を客観的に判断するために「相 関係数」という数値で表します。

「データ分析」ツールを用いる方法

Excel で相関係数を算出する方法はいくつかありますが、ここでは「データ分析」ツールを 用いる方法を紹介します。

「データ」タブ内の「データ分析」メニューを選択します。 分析ツールの中から「相関」を選択します。

	B 9.								相關分析(医									
71	イルホー	-ム 挿入	パージレイア	아 政	式 データ	校開	表示	開発 🖓	実行したい情	業を入力してください							8	共有
外部取出	Di データの 新 Di込み・ エ	 クェリの クェリの テーブル リ・ 最近使 取得と空き 	表示 ゆら すう ったソース 更	「日田」の日本	機 ロバティ い方の順生	<u>また また 並べ替え</u> 並べ替え	フィルター	こかで の両適用 と詳細設定 -	区切り位置	19 7500-2 7700 日重複の削除 19 データの入力用 データ	IP 読み R 0 UL- R 1 - 日 デー リソール		What-If	日本 日本 日本 日本 日本 日本 日本 日本 日本 日本	一部 グループ化 の の の の の の の の の の の の の の の の の の の	· (1) 除・13 ン 「	<u>デ</u> ータ分析 分析]
-	1	P	C	D	E	F	0	1		1.1	K	1 1	м	A.	0	P	0	1 1.
1	A	所得	教育費	5-11	0.15	L F	0	n,		2 X	n	L	IVI	19	0	r	X	
2	北海道	2682	8848	63.86														
3	香森県	2490	5095	57.01	ワール(日) 分析・ 編り返	しのない二元日	Pitte			OK								
4	岩手県	2772	10722	相比						キャンセル								
5	宫城県	2944	10218	長方	統計量					A # 7(H)								
6	秋田県	2699	9193	指数 F核	(平滑) 定:2標本を	使った分散の	検定			1002(0)								
7	山形県	2923	12377	7-1	」工解析													
8	福島県	2971	8375	移動	平均													
9	茨城県	3306	13836	「おし」の	饶生				~									
10	栃木県	3413	10706															
11	群馬県	3325	6493															
12	埼玉県	3067	20381															
13	千葉県	3193	14700															
14	東京都	5427	19935															
15	神奈川県	3227	14446															
16	新潟県	2873	11309															
17	富山県	3319	8174															
18	石川県	2962	17612															
19	福井県	3265	11039															
20	山梨県	2973	6606															
21	長野県	2940	6496															
22	岐阜県	2849	14598															
23	静岡県	3388	8919															
21	受利量	2685	11680	(+)		-	-				1	κ.					1	
準(第二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十二十	L mars lata		9								C.L		Ħ		1	+	100%

入力範囲のデータ列、出力先のセルを選択し OK ボタンを押します。

1	1.461				_			相關	分析 (图) 3039	- Expel									
外部取り	木 データの III 込み -	-ム 神入 -ム 神入 	К-21472) 16 16 Тл 19 Тл		F-7		ステ フィルター で はえとフィルター	19社 🛛 東日 クリア 村田田 2 1年16日23年	ELECTION DE LA	わらてください オシュ フィル 初の可以待 のの入力の データ	Р+ща чёц. лац. у-л	i -Sachal 9 Efinologi	What-If	200 201 201 201 201 201	7/1-7 17/1-7 17/1-7 17/1-7 17/1-7 17/1-7 17/1-7 17/1-7		२ २२ २ - १२३ क	,共称	
83		 • 	v Je																
4	A	В	c	D	Ē	F	G	н	1.	J	ĸ	L	М	N	0	P	Q	1 5	1
1		所得	2 音費	-															
2	北海道	2682	8848	1	11.00					×									
3	骨疫道	2490	5095		入力元		-		-										
4	沒手賃	2772	10722		入力範囲()):	\$8\$1:5	sCs48	UK										
5	宫城鼻	2944	10218		〒		● 利位	2	キャンセル	1									
6	秋田県	2699	9193				○行(8	0	ヘルプ(日)										
7	山形溝	2923	12377		回先翻行	もうべいとして信	8用(上)												
8	摄展展	2971	8375		出力オプショ	>													
9	茨城昌	3306	13836		④ 出力先	(0):	\$E\$3	15											
10	栃木県	3413	10706	_	〇新規7-	15-1(P):													
11	群馬湯	3325	6493		〇新規プロ	2(W)													
12	埼玉県	3067	20381																
13	千葉県	3193	14700																
14	東京都	5427	19935																
15	神奈川虞	3227	14446																
16	针测温	2873	11309																
17	富山県	3319	8174																
18	石川溝	2962	17612																
19	福井県	3265	11039																
20	山梨園	2973	6606																
21	長野県	2940	6496																
22	使尿道	2849	14598																
23	转回温	3388	8919																
24	受知道	2625	11680															1	ł
	-	所得上教育教	1 图4	Ð								4						2	
部計	a 🐮															-	+		

相関係数が算出された相関表がシート内に作成されます。

1	⊟ 5	er a				-		TENE	9町 (図) .xi	sx - Excel							- 0	×
2	r414	木-ム 挿入	A-3617	카 8	Et 7-2	2282	起示 開発	Q 88	-	A DUTKERN							8	共有
外音取得	D込み*	「クエリ 「「フーコ 「リー」」。 「日 単立 」「日 単立 」「日 単立	の表示 かから す 使ったソース 更	同時	ette Touros-A Is-Konnagos, B	21 (1) (1) (1) (1) (1) (1) (1) (1	▼ 100 トルター 下雨 ストンイルター	anii iz Anii iz	日本	フラッシュ フィル 重複の削除 テータの入力規 データ	1-2統合 中国リル・ 同一・ 同・ デー・ ガール	す - つらつつけば タモデルの管理	What-If	9년 予測 9년 주測 9년	間 グループ 間 グループ 間 小叶 アウトラ	£-≤ ∰-: Ф с	□ 〒-今分析 分析	f A
A	1	· •	1. Je															٣
4	A	В	c	D	E	F	G	H	1	1	ĸ	L	м	N	0	P	Q	
1		所得	教育費														-	
2	北房運	268	2 8848															1.1
3	青白星	249	0 5095		1	前得	教育費											
4	岩手県	277	2 10722		所得	1												
5	宫河鼻	294	4 10218		教育費	0.463216	1											121
6	秋田県	269	9 9193		-	1												
7	山形県	292	3 12377															
8	福島県	297	1 8375															
9	茨坡震	330	6 13836															
10	栃木県	341	3 10706															
11	群馬県	332	5 6493															
12	场宝课	306	7 20381															
13	千葉県	319	3 14700															
14	東京都	542	7 19935															
15	複索川	L 322	7 14446															
16	新潟県	287	3 11309															
17	富山県	331	9 8174															
18	石川県	296	2 17612															
19	播井県	326	5 11039															
20	山梨県	297	3 6606															
21	長野県	294	0 6496															
22	续阜溝	284	9 14598															
23	转回果	338	8 8919															
21	stat #	900 2010	6 17680 新聞 184	۲								40					1	
2	御売了	17		~										E	(C) (C)		+	100%

相関係数は-1~1の値をとり、-1に近いほど負の相関があり、1に近いほど正の相関があります。

算出した相関係数は、0.4632 でやや強い正の相関がみられましたが、散布図で描画した回帰直線の当てはまりの良さを示す決定係数(R²:相関係数の二乗で0~1の値をとり、値が高いほど回帰直線の説明力があることを意味する。)は0.2146 とそれほど高くないという結果となりました。

SSDSE を用いた相関分析の一例を紹介しました。

統計データを使った分析では、データ間に相関関係があるかどうか調べる機会は多く存在 します。

SSDSE をダウンロード(<u>https://www.nstac.go.jp/use/literacy/ssdse/</u>)して、身近なデ ータの中に存在する関係性を明らかにしてみてはいかがでしょうか。

 Excel で行う相関分析 (SSDSE を用いた分析事例)

 公開:令和5年3月22日

 作成:独立行政法人 統計センター https://www.nstac.go.jp/

 お問い合わせ先: 統計データ分析コンペティション担当

 〒162-8668 東京都新宿区若松町19-1

 統計センター 統計技術・提供部 技術研究開発課

 statcompe_atmark_nstac.go.jp (「_atmark_」を「@」に置き換えてください)