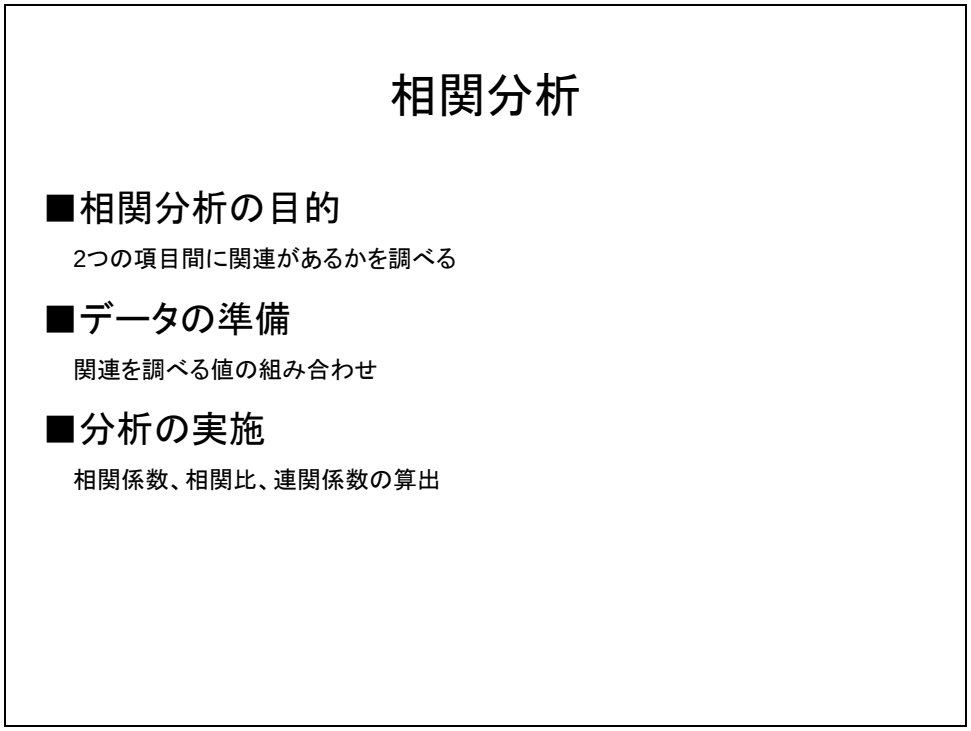
4．1 相関分析



相関分析とは、2 つの項目間の関連の強さを調べる分析です。項目間の関連の強さを調べるために、相関係数などの関連の強さを表現する値を計算します。

項目間に関連があるかどうかは、散布図や多次元分析により推測することは可能ですが、あくまで憶測です。相関分析は、客観的に項目間の関連の強さを調べることができます。

【相関分析を実施するケース】

理想とするデータの特徴	分析結果の活用方法
値が高い(低い)ほど良い	・管理する値に影響の強い項目を探し出し、その項目に働きかけ、値を向上(低下させる)
特定の値が望ましい	・管理する値に影響の項目を探し出し、その項目に働きかけ、値を安定させる

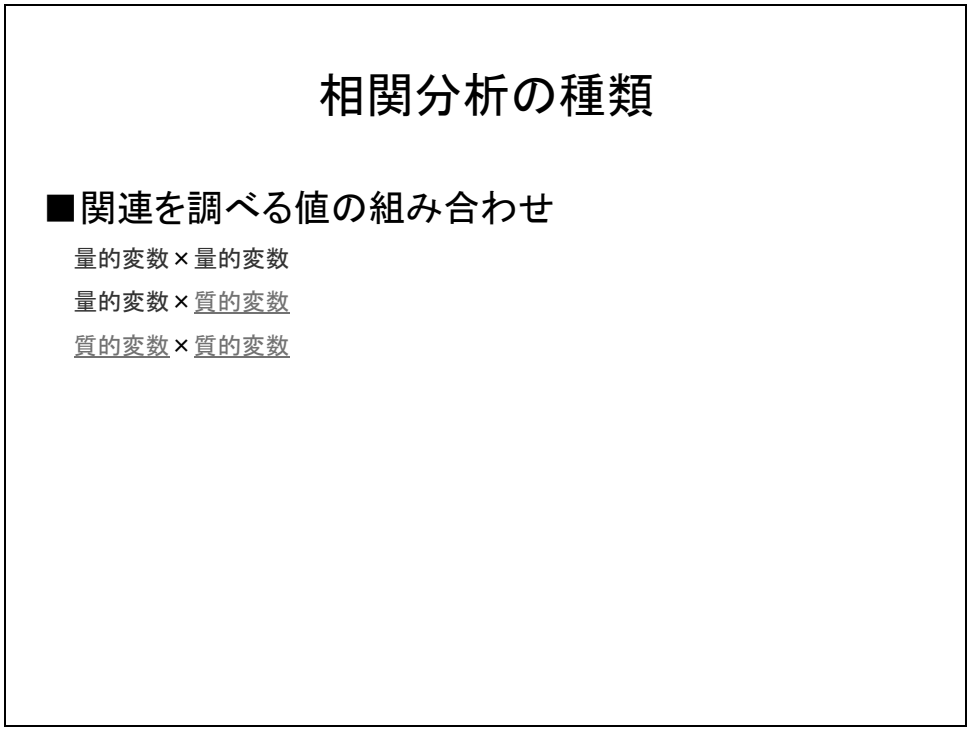
【活用事例】

・つぶやきのキーワード数と株価との関係を調べ、株価予測に活用する。

・Web サイトのデザインと購買実績との関係を調べ、それをもとに Web サイトを改善し、売上を向上する。

・ある商品とある商品の購買に関連があるかを調べ、関連の強い商品の組み合わせをリコメンドする。

4．2 相関分析の種類



相関分析には、関連があるかを調べたい2つの項目の組み合わせにより、
「量的変数×量的変数」「量的変数×質的変数」「質的変数×質的変数」の 3 つのパターンが考えられます。

■ 量的変数×量的変数

連続した数値をもつ2つの項目の組み合わせを用意します。例えば、「アクセス数×売上数」、「部署人数×消費電力」などの組み合わせです。

アクセス数が増えるほど売上数が増える関係なのか、アクセス数が増えるほど売上数が減る関係なのか、ほとんど関係がないのかを調べることができます。

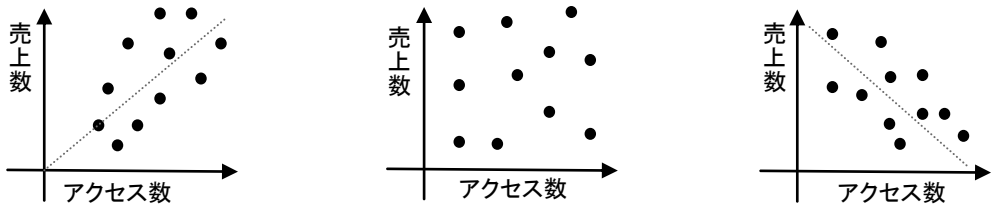
【サンプル】Web サイトアクセス数と売上数

アクセス数	売上数
9	4
30	5
11	11
…	…

量的変数×量的変数の関係を調べるためには、**相関係数**を算出します。相関係数の値の解釈は、以下のようになります。

0.3 ＜【相関係数】≦ 1 :正の相関あり
-0.3 ≦【相関係数】≦ 0.3 :相関なし
-1 ≦【相関係数】＜ -0.3 :負の相関あり

2 つの項目で散布図を描くと、以下のような特徴になります。



正の相関あり　　相関なし　　負の相関あり

相関係数は、2 つの項目間に直線的な関係があるかを調べる方法です。2 つの項目間に関係があったとしても、その関係が直線的ではないと相関係数は、0 付近の値になってしまいます。そのため、散布図で項目間にどのような関係があるかを把握し、相関係数を適切に算出する必要があります。

■ 量的変数×質的変数

連続した値をもつ項目と連続しない値をもつ項目の組み合わせを用意します。例えば、「都道府県×売上数」「部門名×消費電力」などの組み合わせです。

質的変数の値が異なると量的変数が変化する関係なのか、ほとんど関係ないのかを調べることができます。

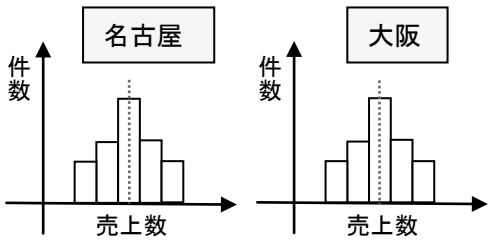
【サンプル】都道府県ごとのマンション販売数

都道府県	売上数
名古屋	4
名古屋	5
名古屋	11
大阪	6
大阪	2
大阪	8

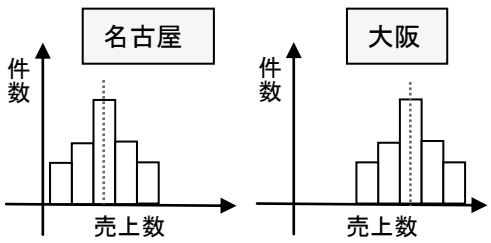
量的変数と質的変数の関係を調べるためには、**相関比**を算出します。相関比の値の解釈は、以下のようになります。

0 ≦【相関比】≦ 0.1 : 相関なし
0.1 ＜【相関比】≦ 0.25 : 相関あり
0.25 ＜【相関比】≦ 1 : 強い相関あり

質的変数の値ごとにヒストグラムを描くと、以下のようになります。



相関なし



相関あり

■ 質的変数×質的変数

連続しない値をもつ2つの項目の組み合わせと各組み合わせの件数を用意します。例えば、「都道府県×商品」「商品 A×商品 B」などの組み合わせです。

一方の質的変数が決まると対応するもう一方の質的変数が決まる傾向が強いのか、ほとんど関係無いのかを調べることができます。

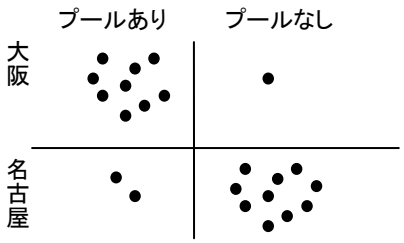
【サンプル】都道府県、プール有無ごとのマンション販売数

	プールあり	プールなし	合計
大阪	60	24	**84**
名古屋	37	76	**113**
合計	**97**	**100**	**197**

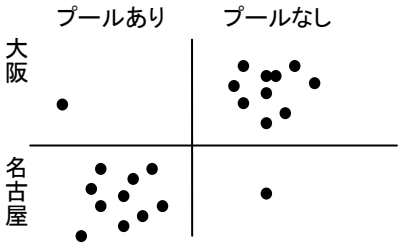
質的変数と質的変数の関係を調べるためには、値の組み合わせとその件数により、**連関係数**を算出します。連関係数の値の解釈は、以下のようになります。

0 ≦【連関係数】≦ 0.1 :相関なし
0.1 <【連関係数】≦ 0.25 :相関あり
0.25 <【連関係数】≦ 1 :強い相関あり

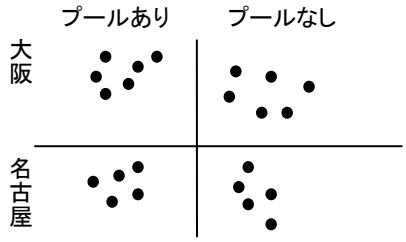
2 つの項目の組み合わせデータの数と連関係数の関係は、以下のようになります。



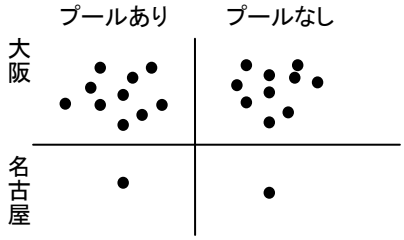
強い相関あり



強い相関あり



相関なし



相関なし

4．3 相関係数の算出

■ 相関係数の算出

「量的変数×量的変数」の項目間の関連の強さを調べるためには、相関係数を算出します。
相関係数の算出方法を以下に紹介します。

$$\text{相関係数} = \frac{\text{「項目1の偏差」と「項目2の偏差」の積の合計}}{\sqrt{\text{「項目1の偏差」の 2 乗の合計} \times \text{「項目2の偏差」の 2 乗の合計}}}$$

[計算例]

	アクセス数	販売数	② アクセス数の偏差	売上数の偏差	③ アクセス数の偏差の 2 乗	売上数の偏差の 2 乗	④ 偏差の積
	9	4	-13	-2.9	169	8.3	37.6
	30	5	8	-1.9	64	3.6	-15.1
	11	11	-11	4.1	121	16.9	-45.2
	35	6	13	-0.9	169	0.8	-11.6
	9	2	-13	-4.9	169	23.9	63.6
	7	8	-15	1.1	225	1.2	-16.7
	4	4	-18	-2.9	324	8.3	52.0
	48	7	26	0.1	676	0.0	2.9
①	45	15	23	8.1	529	65.8	186.6
平均	22	6.9	⑤	合計	2446	128.9	254

① 各項目の平均値を算出します
② 各項目の偏差を求めます。偏差は、「個別値-平均値」です。
③ 各項目の偏差の値を 2 乗します。
④ 各項目の偏差の積を算出します。
⑤ ③と④の合計値を算出します。
⑥ 以下の式で相関係数を算出します。

$$\text{「アクセス数」と「販売数」の相関係数} = \frac{254}{\sqrt{2446 \times 128.9}} = 0.45$$

4. 4 相関比の算出

■ 相関比の算出

「量的変数×質的変数」の項目間の関連の強さを調べるためには、相関比を算出します。
相関比の算出方法を以下に紹介します。

$$\text{相関比} = \frac{\text{グループ間平方和}}{\text{全体の偏差平方和}}$$

[計算例]

地区	販売数
大阪	6
大阪	2
大阪	1
大阪	4
大阪	5
名古屋	15
名古屋	13
名古屋	11
名古屋	15
名古屋	11

地区	件数	平均値	平均値-全体平均値	「②の 2 乗」×件数
大阪	5	3.60	-4.70	110.45
名古屋	5	13.00	4.70	110.45
全体	**10**	**8.30**		

①　②　③

グループ間平方和	220.9
全体の偏差平方和	254.1

④　⑤

① 質的変数の値ごとの件数、平均値を算出します。

② 「質的変数の値ごとの平均値」と「全体平均値」の差を求めます。

③ 「②の 2 乗」とデータ件数の積を求めます。

④ グループ間平方和を求めます。質的変数の値ごとの③の値の合計です。

≪グループ間平方和≫ = 110.45 + 110.45 = 220.9

⑤ 全体の偏差平方和を求めます。全体の偏差平方和は、以下の式で求めます。

【全体の偏差平方和】= (個別値-全体の平均値)^2 の合計

≪全体の偏差平方和≫ = $(6-8.3)^2 + (2-8.3)^2 + \cdots\cdots + (15-8.3)^2 + (11-8.3)^2$

= 254.1

⑥ 以下の式で相関比を算出します。

$$\text{地区と販売数の相関比} = \frac{220.9}{254.1} \fallingdotseq 0.87$$

4．5 連関係数の算出

■ 連関係数の算出

「質的変数×質的変数」の項目間の関連の強さを調べるためには、連関係数を算出します。
連関係数の算出方法を以下に紹介します。

$$\text{連関係数} = \sqrt{\frac{\text{カイ2乗値}}{\text{全体データ件数} \times (\text{2項目のうちの種類が少ないほうの種類数}-1)}}$$

[計算例]

【個別値】

	プールあり	プールなし	合計
大阪	60	24	**84**
名古屋	37	76	**113**
合計	**97**	**100**	**197**

【期待度数】 ①

	プールあり	プールなし	合計
大阪	41.36	42.64	**84**
名古屋	55.64	57.36	**113**
合計	**97**	**100**	**197**

【(個別値-期待度数)^2/期待度数】 ②

	プールあり	プールなし
大阪	8.400	8.148
名古屋	6.244	6.057

① 各組み合わせの期待度数を求めます。期待度数は、以下の式で求めます。

【列項目の値の数】×【行項目の値の数】/ 全体の値の数

≪大阪-プールあり≫ 84 × 97 / 197 ≒ 41.36

② 各組み合わせの【(個別値-期待度数)^2/ 期待度数】を求めます。

≪大阪-プールあり≫ (60 - 41.36)^2 / 41.36 ≒ 8.4

③【(個別値-期待度数)^2/期待度数】の合計(カイ 2 乗値)を求めます。

【カイ 2 乗値】= 8.4 + 8.148 + 6.244 +6.057 ≒ 28.850

④ 連関係数を以下の式で求めます。

「地区」と「プール」の連環係数 $= \sqrt{28.850 / (197 \times (2-1))}$

$= 0.383$

【メモ】