



相関分析・偏相関分析

教育学研究科修士課程1回生

田中友香理

MENU

- 相関とは
- 相関分析とは(パラメトリックな手法)
 - Pearsonの相関係数について
 - SPSSによる相関係数
 - 偏相関係数
 - SPSSによる偏相関係数
- 順位相関係数とは(ノンパラメトリックな手法)
 - SPSSによる順位相関係数
- おまけ(時間があれば)
 - 回帰分析で2変数間の関係を出す
 - 曲線回帰分析を行う



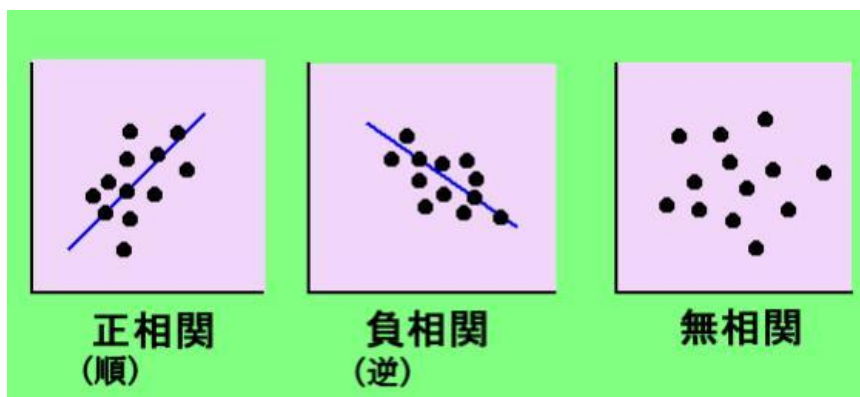
相関とは

- 2変数間の直線関係の強さを見ることを相関という。

Ex. 教師が経験を積むにつれ、教育熱心になる

テレビを見る時間が長いほど、宿題をする時間が短くなる

- 変数(x)の値が大きいほど他方の変数(y)の値も大きい傾向にある時⇒正の相関関係がある。
- 変数(x)の値が大きいほど他方の変数(y)の値が小さい傾向にある時⇒負の相関関係がある。



相関分析とは(パラメトリックな手法)

○ Pearsonの相関係数について

- 一般に、パラメトリックな手法では、Pearson の積率相関係数(Pearson's product-moment correlation coefficient: r)を用いる。
- 計算式(x,y の共分散を x と y の各分散幾何平均値で割った値)

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

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



相関分析とは(パラメトリックな手法)

- 相関係数の値の範囲

- 相関係数の値の範囲は、 $-1 \leq r \leq 1$ であり、相関係数の大きさの評価について、以下が基準となる。

$.0 \leq |r| \leq .2$ ほとんど相関なし

$.2 < |r| \leq .4$ 弱い相関あり

$.4 < |r| \leq .7$ 比較的強い相関あり

$.7 < |r| \leq 1.0$ 強い相関あり



相関分析とは(パラメトリックな手法)

○ 相関分析のポイント

- 複数の変数間の関係を探る最初的手段として有効
- 相関関係と因果関係は異なる。
- 相関係数が強くても、本質的な因果関係が成立しているとは限らない。
- Cf.回帰分析(一方の変数から他方の変数を予測する)
- 結果だけではなく、散布図も必ず確認する。
- 直線以外の関係性が見えることもある。



相関分析とは(パラメトリックな手法)

○ SPSSで相関分析

- excelデータをダウンロードして、保存する。
- SPSSを立ち上げ、ファイルからexcelデータを読み込む。
- 【分析(A)】⇒【相関(C)】⇒【2変量(B)】を選ぶ。
- ダイアログボックスで、対象とする変数を選ぶ。ここでは、【出生体重】と【在胎週数】を選び、▶をクリックして右のボックスへ移す。
- 【相関係数】の【Pearson(N)】にチェックする。
- 【有意な相関係数に星印をつける(F)】をチェックする。
- 【OK】をクリックする。
- 2変数間の相関を複数見たいときには、3つ以上の変数を選ぶと、一気にそれぞれの相関を出してくれる。



相関分析とは(パラメトリックな手法)

○ 結果

- .585が胎盤重量と出生体重の相関係数である。
- 出生体重と胎盤重量は1%未満で有意な正の相関があった。
 - 同じ変数同士の相関は1と表記され、対角線として右上と左下部分に同じ値が出力されるので、いずれか一方を見ればよい。
- 「検定の結果 $p < 0.01$ で有意となり、相関係数 $r = 0.585$ で比較的強い相関があった。」

相関係数

		出生体重	胎盤重量
出生体重	Pearson の相関係数	1	.585**
	有意確率(両側)		.000
	N	1423	1423
胎盤重量	Pearson の相関係数	.585**	1
	有意確率(両側)	.000	
	N	1423	1423

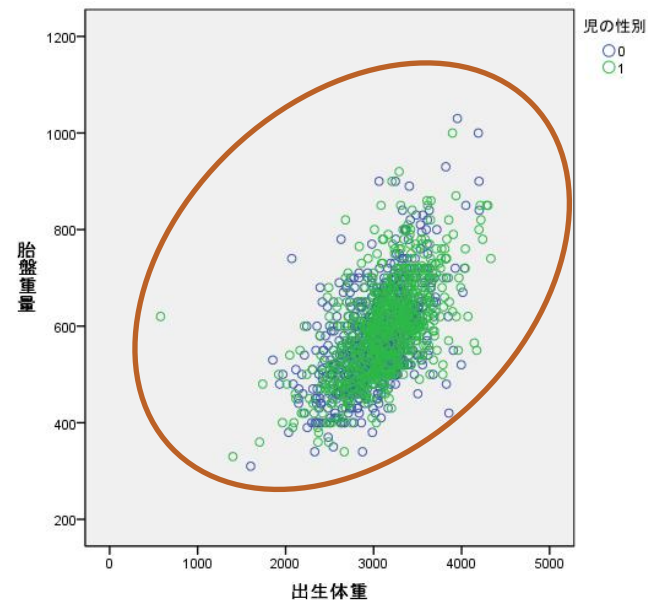
**．相関係数は 1% 水準で有意(両側)です。



相関分析とは(パラメトリックな手法)

○ 散布図を描いてみる

- 【グラフ(G)】⇒【レガシーダイアログ(L)】を選ぶ。
- サブメニューから【散布図/ドット(S)】を選択する。
- 【単純】をクリックし、【定義】をクリックする。
- X軸に「出生体重」Y軸に「胎盤重量」をドラック&ドロップ
- 【OK】をクリックする。



相関分析とは(パラメトリックな手法)

○ 相関分析をするときの注意

- 疑似相関の危険性
- 2変数間に相関関係が認められた場合には、必ず他の変数の影響をうたがう。Ex. 知能指数と身長が高い相関関係にある。⇒年齢の影響が背後に存在する。
- 年齢を制御変数とした偏相関係数を算出する。
- 通常は、3変数以上の相関係数を求め、相関の強い組合せが多いときに、偏相関係数を求めて疑似相関を発見する作業を行う。



相関分析とは(パラメトリックな手法)

○ 相関分析をするときの注意

- 混合標本に注意する。
- **Ex.**全体としてみれば、正の相関だが、学年ごとに区切ってみると、負の相関になる。

- 外れ値の扱い
- 散布図中に飛び離れた値があると、外れ値(観察上、他とは明らかに大きな・小さな値を取るデータ)の可能性がある。
- 相関係数と順位相関係数を求めて比較する。
- 明らかに外れ値とわかる場合には、その値を除外する。



相関分析とは(パラメトリックな手法)

○ 相関分析をするときの注意

- 曲線的な相関が存在しないか?
- 散布図を出し、点の並びが直線的な関係を示さない場合、順位相関係数の適用を考える。
- 回帰分析であれば、曲線回帰分析を適用する。

→おまけ

- 時系列データには使用できない。
- 同じ被験者で時系列的にとったデータをまとめ、時間がたつごとに増える・減るという関係を出すのはよくない(相関係数は出るが、個人内の変化を見ているわけではない)。
- こういう場合、回帰分析や、分散分析を用いる。



相関分析とは(パラメトリックな手法)

- 偏相関係数 (partial correlation coefficient)
 - 見かけ上の相関(疑似相関 spurious correlation)を危惧してひとつの変数の影響を除いた他の二つの変数の相関係数を求める手法。
 - x, y, z の変数があり、 x, y の相関係数を r_{xy} 、 x と z の相関係数を r_{xz} 、 y と z の相関係数を r_{yz} としたとき、 z を制御変数とした偏相関係数 $r_{xy.z}$ は以下の式になる。

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2} \sqrt{1 - r_{yz}^2}}$$

- 偏相関係数も、 $-1 \leq r \leq 1$ の範囲をとり、絶対値が1に近いほど関係は強いことを示す。



相関分析とは(パラメトリックな手法)

○ SPSSによる偏相関係数

- 先ほどと同じデータを使用する。
- **【分析(A)】⇒【相関(C)】⇒【偏相関(R)】**をクリックする。
- ダイアルボックス中の、**【出生体重】**と**【胎盤重量】**を**【変数(V)】**に移動する。
- 影響を取り除きたい変数**【母親の年齢】**を**【制御変数(C)】**に移動する。
- **【OK】**をクリックする。

- ついでに、出生体重・胎盤重量・母親の年齢の相関係数を求めてみる。(手順は先ほどと同じ)



相関分析とは(パラメトリックな手法)

- .585が偏相関係数である。有意確率は.001未満

相関係数

制御変数			出生体重	胎盤重量
母親の年齢	出生体重	相関	1.000	<u>.585</u>
		有意確率(両側)	.	<u>.000</u>
		df	0	1419
胎盤重量	胎盤重量	相関	.585	1.000
		有意確率(両側)	.000	.
		df	1419	0

- 相関係数の表:年齢と出生体重・胎盤重量の間の相関が低かったため、年齢を制御変数としても結果はあまり変わらなかった。

相関係数

		胎盤重量	出生体重	母親の年齢
胎盤重量	Pearson の相関係数	1	.585**	<u>.017</u>
	有意確率(両側)		.000	.531
	N	1423	1423	1422
出生体重	Pearson の相関係数	.585**	1	<u>-.004</u>
	有意確率(両側)	.000		.871
	N	1423	1423	1422
母親の年齢	Pearson の相関係数	.017	-.004	1
	有意確率(両側)	.531	.871	
	N	1422	1422	1422

** 相関係数は 1% 水準で有意(両側)です。



順位相関係数とは(ノンパラメトリックな手法)

- データの分布が、正規分布以外となるときには、順位相関係数を用いる。

- Spearmanの順位相関係数 (spearman's rank correlation coefficient) ρ (ロー)
- Kendallの順位相関係数 (Kendall's rank correlation coefficient) τ -b(タウ)
- とともに-1から1の値を取り、1だと同順、-1だと逆順になる。
- 計算式

- Spearmanの順位相関係数

- Kendallの順位相関係数

$$r = 1 - \frac{\sum_{i=1}^n (x_i - y_i)^2}{\frac{n(n^2 - 1)}{6}} = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n(n^2 - 1)}$$

$$r_k = \frac{\sum_{i=1}^n P_i - \sum_{i=1}^n Q_i}{\frac{n(n-1)}{2}}$$



順位相関係数とは(ノンパラメトリックな手法)

- SPSSによる順位相関係数
 - 先ほどと同じデータを使う。
 - **【分析(A)】⇒【相関(C)】⇒【2変量(B)】**を選ぶ。
 - ダイアログボックスで、出生体重と、在胎週数をクリックする。▶
で変数を右のボックスに移す。
 - **【相関係数】**の、**【Kendallのタウb(K)】**、**【Spearman(S)】**にチェック
 - **【有意な相関係数に星印をつける(F)】**をチェック
 - **【OK】**をクリック



順位相関係数とは(ノンパラメトリックな手法)

○ 結果

- 表の見方は、相関係数と同じ。上段がKendallのタウbの結果であり、下段がSpearmanの ρ の結果。

相関係数

			出生体重	在胎週数
Kendallのタウb	出生体重	相関係数	1.000	<u>.275**</u>
		有意確率(両側)	.	<u>.000</u>
		N	1423	1423
	在胎週数	相関係数	.275**	1.000
		有意確率(両側)	.000	.
		N	1423	1423
Spearmanの ρ	出生体重	相関係数	1.000	<u>.362**</u>
		有意確率(両側)	.	<u>.000</u>
		N	1423	1423
	在胎週数	相関係数	.362**	1.000
		有意確率(両側)	.000	.
		N	1423	1423

** 相関は、1%水準で有意となります(両側)。



おまけ(時間があれば)

- 2変数の関係を、回帰分析を用いて調べてみる。
 - 単回帰分析の詳細は省略(復習)
 - 先ほどと同じデータを使用。
 - 【分析(A)】→【回帰(R)】→【線形(L)】を選択する。
 - 【従属変数(D)】に「胎盤重量」を入れる。
 - 【独立変数(I)】に「出生体重」を入れる。
 - 【統計量(S)】をクリックし、【推定値(E)】【信頼区間(N)】【モデルの適合度(M)】をチェックする。
 - 【続行】をクリックし、【OK】をクリック



おまけ (時間があれば)

○ 結果

- 分散分析の表の有意確率を確認 ($p < .05$ でないと回帰式は役に立たない)

分散分析^b

モデル	平方和 (分散成分)	自由度	平均平方	F 値	有意確率
1 回帰	4758465.290	1	4758465.290	739.841	.000 ^a
残差 (分散分析)	9139503.605	1421	6431.741		
合計 (ピボットテーブル)	13897968.89	1422			

a. 予測値: (定数)、出生体重。

b. 従属変数 胎盤重量

- 係数の表の有意確率と非標準化係数を見る。

○ 胎盤重量 = $131.783 + 0.147 \times$ 出生体重

係数^a

モデル	標準化されていない係数		標準化係数	t 値	有意確率	B の 95.0% 信頼区間		共線性の統計量	
	B	標準誤差	ベータ			下限	上限	許容度	VIF
1 (定数)	131.783	16.733		7.875	.000	98.958	164.608		
出生体重	.147	.005	.585	27.200	.000	.136	.157	1.000	1.000

a. 従属変数 胎盤重量



おまけ(時間があれば)

- モデル集計のR・R²を見る(1に近いほど予測精度が高い)
 - R² = .342なので、予測精度はあまりよくない。

モデル集計

モデル	R	R2乗	調整済み R2乗	推定値の標準誤差
1	.585 ^a	.342	.342	80.198

a. 予測値: (定数)、出生体重。

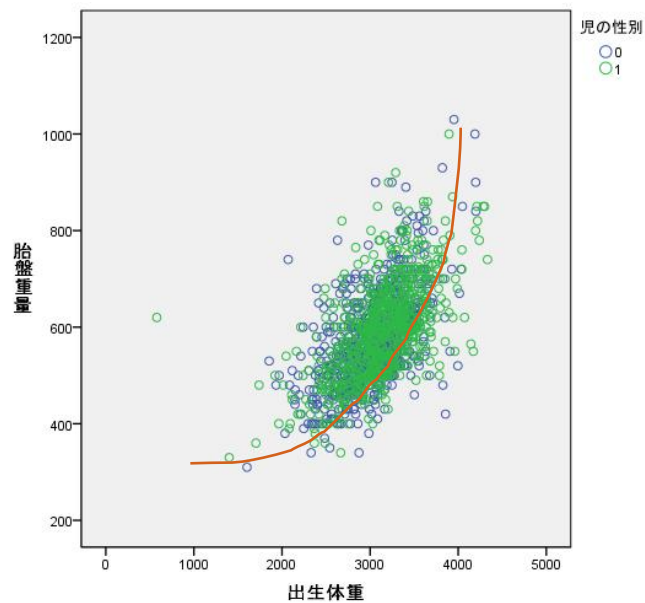
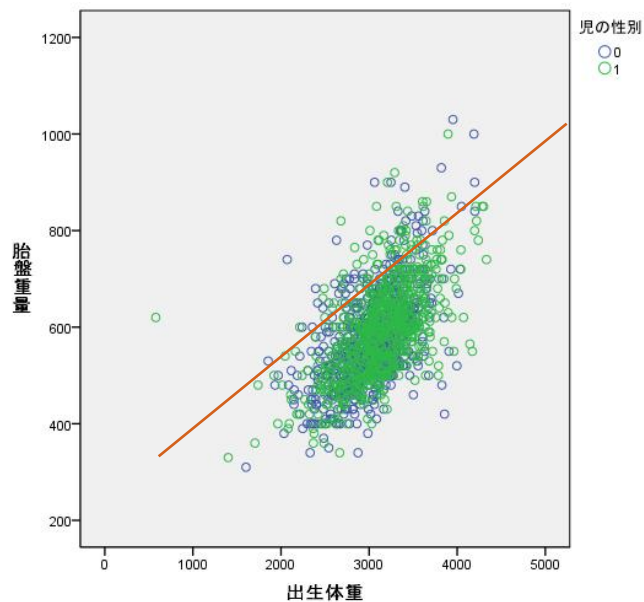
- この回帰式は分散分析表より、 $p < 0.01$ で有意であり、回帰係数も、 $p < 0.01$ で有意であった。しかし、決定係数はR² = .342 と小さく、予測精度は高くないことがわかった。



おまけ (時間があれば)

- 散布図を見る

もしかして、曲線の関係で捉えたほうがいい・・・？



おまけ(時間があれば)

- 曲線回帰分析をやってみる。
 - 2変数の散布図を書き、曲線的な関係であると考えられる場合には、曲線回帰を行う。
 - 【分析(A)】→【回帰(R)】→【曲線推定(C)】
 - 「出生体重」を【独立変数】の【変数(V)】へ、「胎盤重量」を【従属変数(D)】へ移動する
 - 当てはめたい曲線式にチェックする。(散布図をみて、どの回帰式がふさわしいか考える)
 - 今回は、【線形(L)】、【二次(Q)】にチェックする。
 - 【OK】をクリックする。



おまけ(時間があれば)

○ 結果

- モデル集計とパラメータ測定値の表を見る。
 - モデル集計のR²乗の値が各回帰式の精度である。
 - 2次がR²=.353で線型よりも適合度がいいようだが、極端に違う場合、線形の方が無難？
- パラメータ測定値の定数が回帰式の値。
 - 2次: $447.475 - 0.066x + 0.00003532x^2$

モデル集計とパラメータ推定値

従属変数:胎盤重量

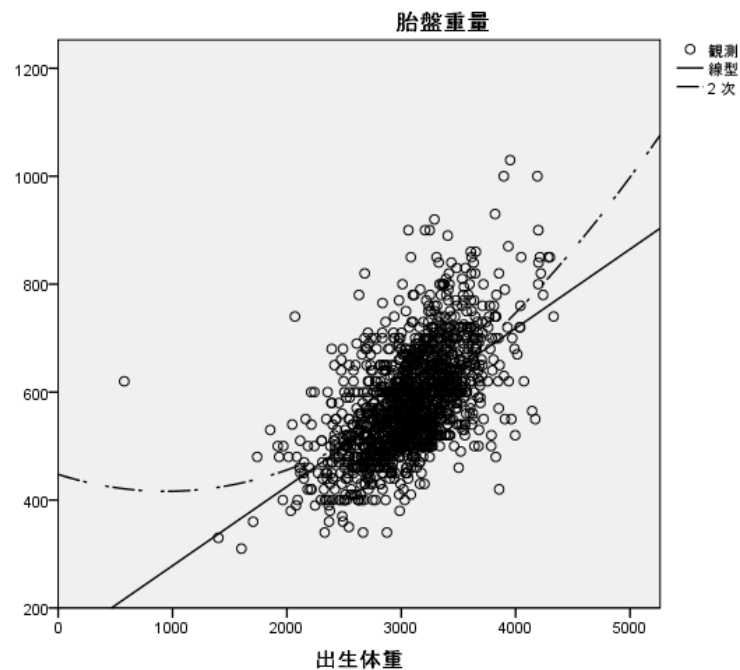
方程式(等式)	モデル集計					パラメータ推定値		
	R ² 乗(決定係数)	F	df1	df2	有意確率	定数	b1	b2
線型(1次)	.342	739.841	1	1421	.000	131.783	.147	
2次	.353	387.922	2	1420	.000	447.475	-.066	3.532E-5

独立変数は出生体重です。



おまけ (時間があれば)

- 散布図には、実際の曲線が描かれている。



まとめ

- 相関分析・・・2つの変数の比例関係を数値的に表す。
例「体重が重くなるほど身長が高くなる」
 - 2変数間の関係を調べるのには、便利な方法
 - 擬似相関、混合標本などに注意。偏相関分析や、群ごとに分けて相関分析をする。
 - とったデータに偏りがある場合や、曲線関係になるときには、順位相関分析をする。
- 回帰分析・・・一つの変数が他方の変数を予測する。
例「体重○kgの人が平均的にどれくらいの身長となるか」
 - 説明力(決定係数)が出せる。
 - 曲線回帰分析ができる。



参考文献

- 森敏昭・吉田寿夫(1990). 心理学のためのデータ解析テクニカルブック 北大路書房
- 対馬栄輝(2007). SPSSで学ぶ医療系データ解析 東京図書
- 青木繁伸(2009). Rによる統計解析 オーム社
- 石村貞夫・石村光資郎(2007). SPSSでやさしく学ぶ統計解析第3版 東京図書
- 小田利勝(2007). SPSSによる統計解析入門 プレアデス出版

