

相関と回帰

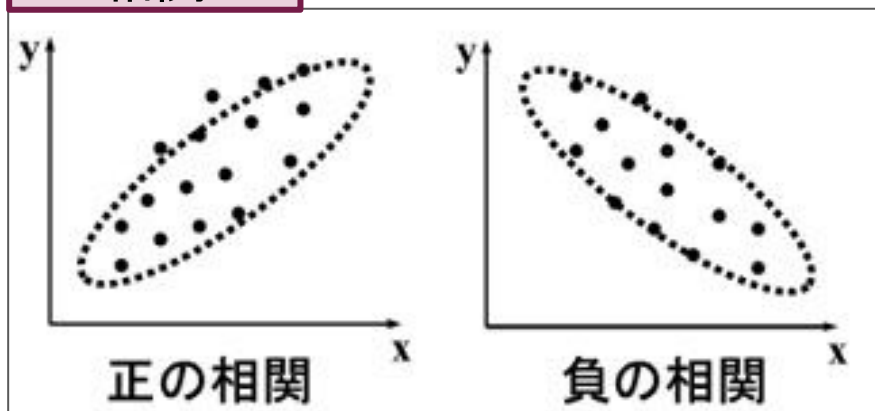
道具としての統計学
著: 奥田千恵子

60171033 築山文香

相関と回帰の違い

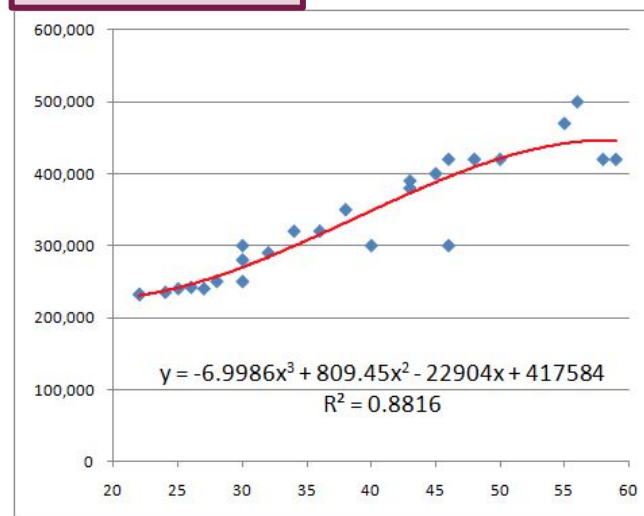
- ❖ どちらも**2つの変数の関係**を示す
- ❖ 回帰の関係にある場合
 - XとYは入れ替え不可
 - 回帰は直線関係とは限らない
- ❖ **相関は直線関係**を指す
 - XとYは入れ替え可

相関



<http://www.agri.tohoku.ac.jp/iden/toukei3.html>

回帰



https://www.rosei.jp/jinjour/article.php?entry_no=55309

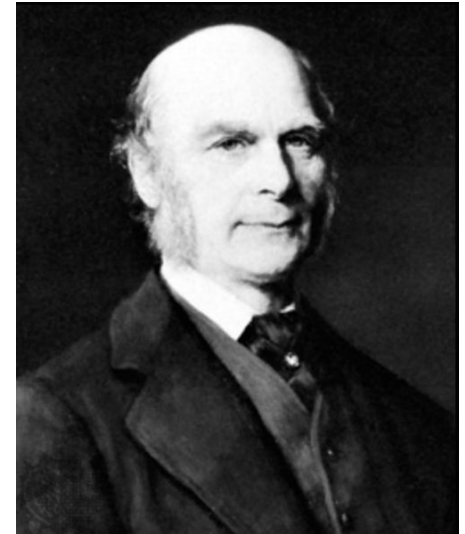
フランシス・ゴールトン(1822-1911)

❖ 優生学と現代統計学の父

❖ スイートピーの種子を栽培・比較(1877)

- 種子の重量・平均直径: **子の特徴は親よりも平均に近づいていく**
- 「平均回帰」現象を発見
- 先祖帰りと表現
- しかし、これは生物学的な特徴ではなく、**データの扱いの問題**であることに気づいて、**回帰係数**と変更

❖ **回帰**の語源



<http://www.stat.go.jp/dss/course/701.htm>

確率的に極端な値が出たあとは平均的な値が出るよね...

続、フランシス・ゴールトン(1822-1911)

- ❖ 以後、研究を重ね
- ❖ **相関**を表す数値を「**相関係数 r** 」とする
- ❖ なお、相関の概念を一般化したのは後継者のピアソン
 - 一般的に相関係数は**ピアソンの積率相関係数**という

相関

相関

❖ 相関係数

- 2つの変数の間の直線関係の強さをはかる尺度
- **ピアソンの積率相関係数**とも呼ばれる

相関係数を用いる条件

1. 標本は母集団から無作為抽出されている
2. 標本が互いに独立している
3. 各標本が対応する2つの変数をもつ
4. 2つの変数が独立して測定されている
5. **2つの変数がともに正規分布している**
 - これが満たされない場合、**ノンパラメトリックな相関係数**を使用
6. 2つの変数が直線関係にある

相関係数の計算手順

1. n個の対になっている変数から**ピアソンの積率相関係数r**を求める

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

2. 相関係数rの値が正規分布するように変換

➤ フィッシャーのz変換

$$Z = \frac{0.5 \times \ln\left(\frac{1+r}{1-r}\right)}{\sqrt{\frac{1}{n-3}}}$$

3. 有意水準、 α を求める (両側検定)
4. 標準正規分布を用いて、P(z変換されたrに対応する確率)が α より小さい
➤ 「2変数が相関でない($\rho=0$)」という帰無仮説を棄却

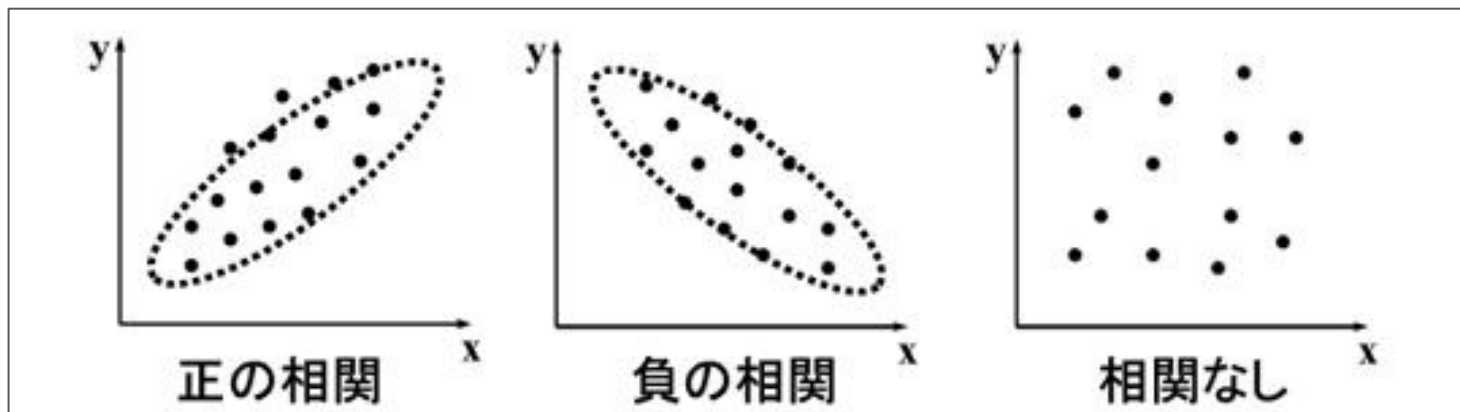
相関係数 r

❖ 相関係数 r : $-1 < r < 1$

- $r > 0$ であれば正の相関
- $r < 0$ であれば負の相関

❖ 散布図を描くようにする

- 相関係数は外れ値に大きな影響を受けやすい



<http://www.agri.tohoku.ac.jp/iden/toukei3.html>

相関係数の有意差の意味

❖ 有意差5%の場合

➤ P値はデータ数に依存

サンプル数50のとき、
 $r=0.3$ 以下でも
有意差があることになる...

データ数	有意差が認められる $ r $ の最小値
10	0.63
20	0.44
30	0.36
40	0.31
50	0.28

その相関が本当に意味があるのかは
吟味する必要性がある

ノンパラメトリックな相関係数

❖ スピアマンの順位相関係数

- ピアソンの積率相関係数を求める式に、**データの大きさの順位**を入れる
 - ピアソンの積率相関係数と同様、 -1 から 1 の間の値
 - $r < 0$: 正の相関
 - $r > 0$: 負の相関

❖ ケンドールの順位相関係数

回歸

回帰

❖ 単回帰

- **2つの変数が相関関係にある場合、最も適合度の高い関数である回帰直線の方程式を求めること**

単回帰を行う条件

1. 標本は母集団から無作為抽出されている
2. 標本が互いに独立している
3. 各標本が対応する2つの変数を持つ
4. 2つの変数がともに正規分布している
5. 2つの変数が直線関係にある

❖ 独立変数と従属変数は入れ替え不可

- 入れ替えると結果が変わる

一般的に、
独立変数: x
従属変数: y

回帰直線の計算手順

1. 最小2乗法(least square methods)による計算式

- y切片:a、傾き:b
- $\bar{X}$:xの平均、 $\bar{Y}$:yの平均

$$b = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - \bar{X})^2}, \quad a = \bar{Y} - b\bar{X}$$

2. 回帰直線の適合度の指標

- 相関係数 r
- 決定係数 r^2 (相関係数の2乗)

回帰直線の信頼区間

- ❖ 回帰直線はデータから次のYを予測する式
 - データ(標本)から母集団の回帰直線を求めたい
- ❖ 統計ソフトを使って、母回帰直線の信頼区間を求めることができる

その他の回帰直線

❖ 重回帰分析

- 独立変数が2つ以上
- 適合度の指標: R (重相関係数)

❖ 非線形回帰

- 多項式回帰

❖ ロジスティック回帰

- 従属変数: ある出来事が**起きる/起きないの2値変数**
- 独立変数: 連続量、離散量どちらも可
- ある事象 y が起こる確率 p をオッズ比の形で予測

まとめ

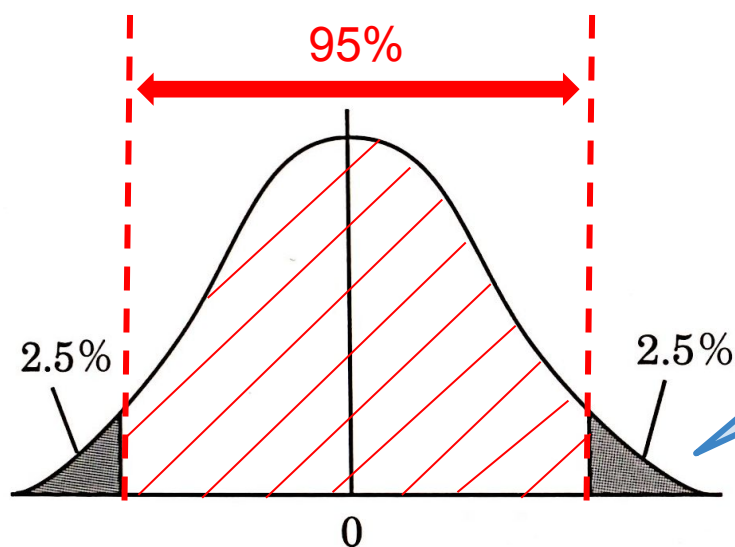
- ❖ 相関はXとYの直線関係を示したもの
 - XとYは正or負の相関がある
- ❖ 回帰は相関関係にあるXとYから、データ(標本)をとって標本にないYを予測するためのもの
 - $Y = aX_1 + bX_2 + c$
- ❖ 相関は標本数が少ない場合、相関係数が小さくても相関関係の有意差があると判断される場合もある
 - 相関関係にあるのか吟味の必要性
- ❖ 詳しい計算方法は使うときにお願いします...

おわり

両側検定：8章統計的仮説検定

帰無仮説：AとBの間に差がない： $A=B$

対立仮説：AとBの間に差がある： $A<B$, または, $A>B$
(A、Bは2つの母平均値、2つの母比率など)



有意差5%の場合、
両端2.5%ずつ

両側検定
(有意水準 5%)

有意水準

- ❖ 5%か1%が多い
- ❖ 有意水準5%
 - 5%は違う結果が出るということ

オッズ比

- ❖ ○倍というわかりやすい指標で表現できる
 - 喫煙する人は喫煙しない人に比べて病気になるリスクが2倍
- ❖ 単純にオッズ比に倍数をかけてはいけない