統計処理NO10

相関と回帰分析 (1/2)

2012年12月11日

後　保範

0. 分析ツールの使用

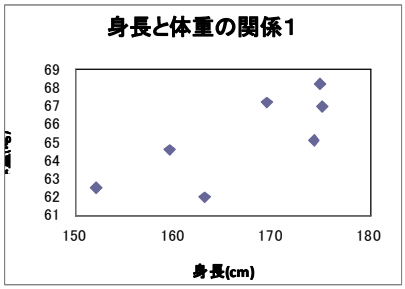
- Excelの分析ツールの使用準備
  (1) Excel左上の「ファイル」をクリック
  (2) 「Excelオプション」をクリック
  (3) 「アドイン」をクリック
  (4) 管理に「Excelアドイン」を選定
  (5) 「設定」をクリック
  (6) 「分析ツール」をオンにし「OK」をクリック

1. 共分散

- 2組の対応するデータ間での、平均からの差の積の平均値
- 共分散が正なら一方が大きくなれば他方も大きく、負なら一方が大きくなれば他方は小さくなる関係
- 共分散は「関係の向き」を見るもので、「関係の大きさ」は見れない → 相関係数
- 2変数x,yの共分散S(x,y)は下記で計算

$$S(x,y)=\sum_i (x_i-\bar{x})(y_i-\bar{y}), \quad \bar{x},\bar{y}は各平均$$

1.1 散布図で示す共分散



1.2 共分散の計算

- 共分散の計算(F51.xls シート1)
  データ → データ分析 → 共分散 → 入力範囲「2組のデータ」 → 先頭の行をラベルに使用 → 出力先 → 同一画面へ分析結果出力
- 計算結果の例

	身長	体重
身長	68.628	
体重	14.224	4.8763

2. 相関係数

- 共分散では「関係の向き」だけに対し、相関係数では「関係の強さ」も示す
- データを標準化(平均0、標準偏差1)にして、共分散を計算したものと同じ
- 相関係数は-1～1の範囲の値となる
- 2変数x,yの相関係数rは下記で計算

$$r=\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2 \sum_i (y_i-\bar{y})^2}}, \quad \bar{x},\bar{y}は各平均$$

2.1 相関係数の計算

- 相関係数の計算(F51.xlsx シート2)
  データ → データ分析 → 相関 → 入力範囲
  「2組のデータ」→ 先頭の行をラベルに使用
  → 出力先 → 同一画面へ分析結果出力
- 計算結果の例

(a) 弱い相関

	身長	体重
身長	1	
体重	0.778	1

(b) 強い相関

	身長	体重
身長	1	
体重	0.922	1

2.2 相関関係の検定

- 相関係数の値の大きさから、統計的に相関関係があるか検定する
- 相関係数rとr_0より大(r>r_0)なら相関ありと判断

$$r_0 = \sqrt{4/(n+2)}, \quad n\text{はデータ数}$$

- 前記の例はn=7なのでr_0を計算すると

$$r_0 = \sqrt{4/(7+2)} = 0.667$$

となり、共にr_0より大きいので身長と体重は相関ありと判定でき、(a)より(b)の方が強い相関である

3. 回帰分析

- 回帰分析とは2組の変数x,yの関係を直線y=ax+bで近似し、係数a,bを求めることである
- n個のデータの組(x_i,y_i)に対しf(x)=ax+bと置き、最小2乗法を使用し下記が最小となるa,bを求める

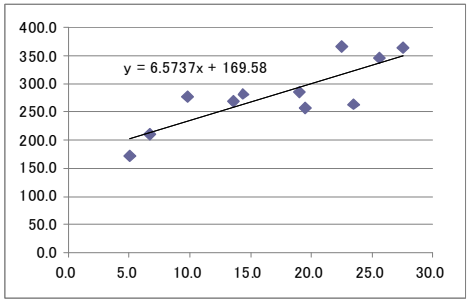
$$\sum_i (f(x_i) - y_i)^2 \Rightarrow \text{最小にする} a,b\text{を求める}$$

- 回帰分析はExcelの「散布図」及び、データ分析の「回帰分析」を使用して計算できる

3.1 回帰分析(散布図)

- 「散布図」の作成(F51.xlsx シート3)
  x,yの両データ範囲 → 挿入 → 散布図 →
  出力する図を選ぶ → 散布図が出力 →
  目的に合わせて整形
- 「回帰曲線」と「回帰式」の追加
  「散布図」の任意のプロット点上で「右クリック」→
  表示メニュー「近似曲線の追加」→「線形近似」、
  「グラフに数式を表示」にチェック →「閉じる」

3.1 回帰分析(散布図例)



3.2 回帰分析(データ分析)

- データ分析を利用した回帰分析
  データ → データ分析 → 回帰分析 →
  入力Y範囲「yの値として入力のデータ」→
  入力X範囲「xの値として入力のデータ」→ ラベル
  →出力先 → 同一画面へ分析結果出力

- F51.xlsxのシート4の図5.4 月平均気温とビールの集荷量では入力Y範囲は「国内出荷量」ので、入力X範囲は「月平均気温」の1～11月のデータ

3.2 回帰分析(分析結果)

概要

回帰統計	
重相関 R	0.829
重決定 R2	0.688
補正 R2	0.653
標準誤差	35.697
観測数	11

→ 相関係数

分散分析表

	自由度	変動	分散	された分	有意 F
回帰	1	25271.6	25271.6	19.8	0.0016
残差	9	11468.4	1274.3		
合計	10	36740.0			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	169.58	27.33	6.21	0.00016	107.76	231.40
平均気温	6.57	1.48	4.45	0.00159	3.23	9.91

3.2 回帰分析の評価

- 相関関係の評価
  相関係数(r) = 0.829 (重相関Rの値)
  r0 = sqrt(4/(n+2)) = sqrt(4/13) = 0.55
  r > r0のため、平均気温とビールの出荷量は
  相関が有ると言える。
- 近似曲線(直線)
  出荷量(Ml) = 169.58 + 16.57x平均気温

4. 演習問題

- F5-1A.xlsxのシート1の図1. 年齢と最高血圧を使用して、下記を作成せよ

(1) 散布図の作成

(2) 散布図に「回帰曲線」と「回帰式」を追記

(3) 年齢と最高血圧の相関係数を算出

(4) 統計的に相関関係があるかの判定