第6回 4月28日の授業内容

- §2.4 データ分布の量的記述 (続)
  - (Note) 分散の定義について
  - §2.4.3 その他の代表値
- §2.5 二変量データの記述
  - 散布図
  - 同時度数分布、二変数ヒストグラム
  - 共分散
  - 相関係数

§2.4.3 その他の代表値

- 変動係数 coefficient of variation
  - 相対的な散らばりの大きさを測る指標
    変動係数=標準偏差/平均
  - 単位はない (無名数)
    - 単位が異なるもの同士の相対的な散らばり具合を比較可能
  - データの値がすべて正の値をとるときに用いる

§2.4.3 その他の代表値(2)

- 歪度 skewness
  - 分布の歪みを測る指標

$$\omega \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{\sigma^3}$$

  - 左右対称の形に近ければ、0近辺の値をとる。
    - 歪度が正値のとき、右側に歪んだ分布
    - 歪度が負値のとき、左側に歪んだ分布

§2.4.3 その他の代表値(3)

- 尖度 kurtosis
  - 分布の尖り、裾の厚さを測る指標

$$\xi \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{\sigma^4}$$

  - 正規分布 (後述) の尖度は3であるので、正規分布と比較して議論することが普通
    ⇒ 3を引いて定義することも

$$\xi^* \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{\sigma^4}-3$$

§2.4.3 その他の代表値(4)

- Note: 偏差値
  - 平均が50、標準偏差が10になるように変換したもの

$$H_i = \frac{(X_i-\bar{X})}{\sigma}\times 10+50$$

- Note: 標準化 (基準化)
  - 平均が0、標準偏差が1になるように変換すること

$$Y_i = \frac{(X_i-\bar{X})}{\sigma}$$

§2.5 二変量データの記述

- ペア(x, y)になったデータの記述

⇓

x, y それぞれ (一変量) のデータ分布の記述に加えて、xとyの関係についての記述も行われる
  - 共分散
  - 相関

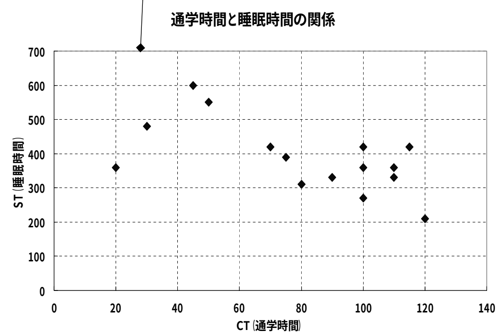
という概念が導入。

二変量データ分布の記述手法

- 散布図： 視覚からの把握
- 同時度数分布表、二変数ヒストグラム
- 共分散： xとyが共に動く大きさを示す代表値（共変動）
- 相関係数： xとyの関係が線形に近いかを示す代表値

散布図

データセットの大きさが大きくなる（＝個体数が大である）時には重なり合う点が多くなってしまう

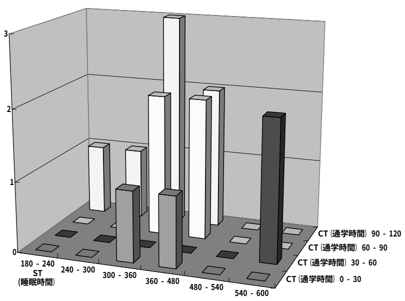


同時度数分布表

		CT（通学時間） 0 - 30	30 - 60	60 - 90	90 - 120	計
ST（睡眠時間）	180 - 240	0	0	0	1	1
	240 - 300	0	0	0	1	1
	300 - 360	1	0	2	3	6
	360 - 480	1	0	2	2	5
	480 - 540	0	0	0	0	0
	540 - 600	0	2	0	0	2
計		2	2	4	7	15

- 縦ないし横の「計」のところだけ見れば、一変量の度数分布
  - 周辺（度数）分布表
- 「相関表」と呼ぶことも

二変数ヒストグラム



共分散 covariance

- 二変量（ペアになった変数）の共変動の大きさを表す代表値
- 定義

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

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{n}$$

相関係数 correlation coefficient

- 二変量間の線形関係の強弱を表す代表値
  - -1≦r≦1
- 定義

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

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