統計基礎　第９回　(2017.12.4)

テーマ：正規分布（復習)、相関

1　はじめに

ビデオ　NHK　Biz プラス「ビッグデータを分析」(2013.3、5 分)

（用語）ビッグデータ（big data)（出典：IT 用語辞典 e-Words)

・　従来のデータベース管理システムなどでは記録や保管、解析が難しいような巨大なデータ群。明確な定義があるわけではない。多くの場合、ビッグデータとは単に量が多いだけでなく、様々な種類・形式が含まれるデータであり、さらに、日々膨大に生成・記録される時系列性・リアルタイム性のあるようなものを指すことが多い。今までは管理しきれないため見過ごされてきたそのようなデータ群を記録･保管して即座に解析することで､ビジネスや社会に有用な知見を得たり、これまでにないような新たな仕組みやシステムを産み出す可能性が高まるとされている。

2　正規分布（復習）

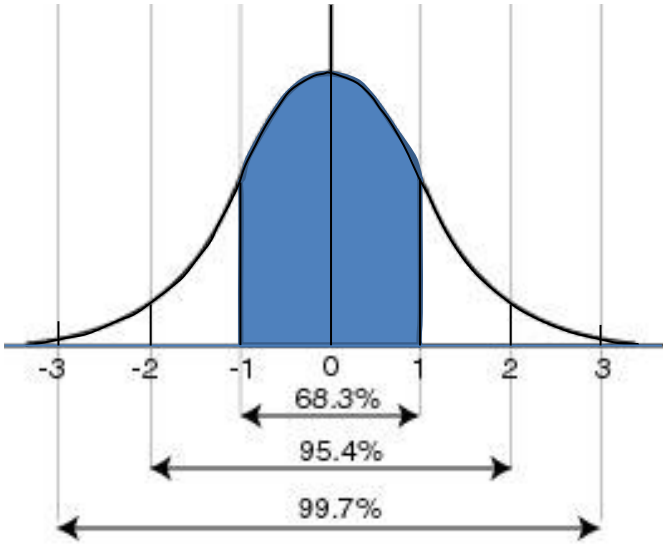
・ 正規分布では、「平均」と「標準偏差」がわかると、以下がわかります。

- ①度数分布図の形
- ②ある範囲にあるデータが全体に占める割合

・ 標準化係数とは？

データ値が平均値から標準偏差の何個分離れているのか？

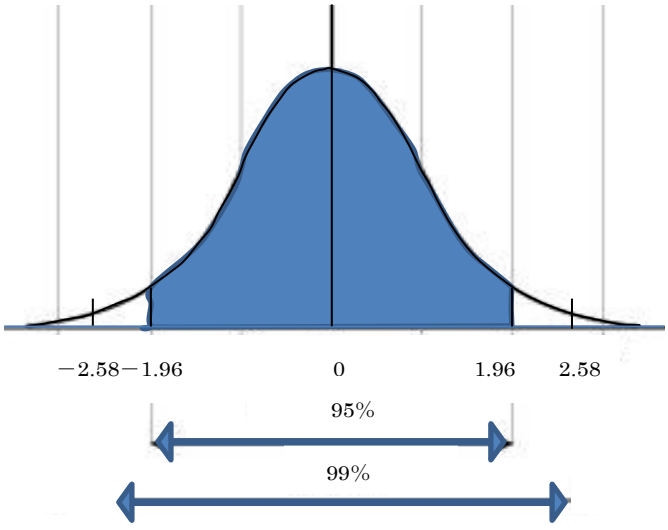
$$\text{標準化係数} = \frac{\text{データ値} - \text{平均値}}{\text{標準偏差}}$$



正規分布

←標準化係数

←ある範囲にあるデータが全体に占める割合



正規分布

←標準化係数

←ある範囲にあるデータが全体に占める割合

演習１　17 歳女子の平均身長は 158.0cm、標準偏差は 5.4cm です。17 歳女子学生の身長が正規分布に従っているものとすると、95%の 17 歳女子の身長は、何 cm から何 cm の間にあると推測できますか？

（ヒント）全体の 95%が含まれる範囲とは、

「平均値－標準偏差×1.96」～「平均値＋標準偏差×1.96」　の範囲　です。

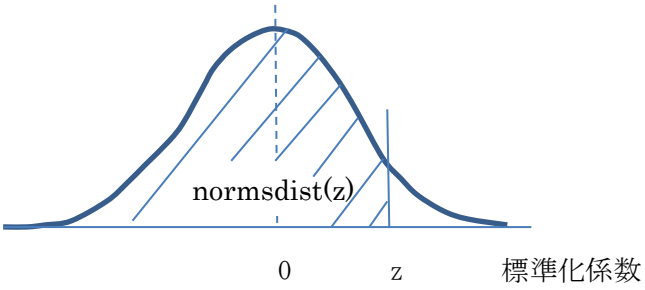
正規分布に関する EXCEL 関数

・　データが正規分布に従う場合に、

「標準化係数 z 以下のデータの個数が全体に占める割合」を求める EXCEL 関数：

=normsdist（標準化係数）

（注）％ではなく全体を 1 とした割合を返す



演習２　=normsdist(1.96)-normsdist(-1.96)を求めてください

（説明）

標準化係数が-1.96～1.96 のデータの個数が全体を占める割合を求めています。

それは、

（標準化係数が 1.96 以下のデータの個数が全体に占める割合）

－（標準化係数が-1.96 以下のデータの個数が全体に占める割合）

で求まります。

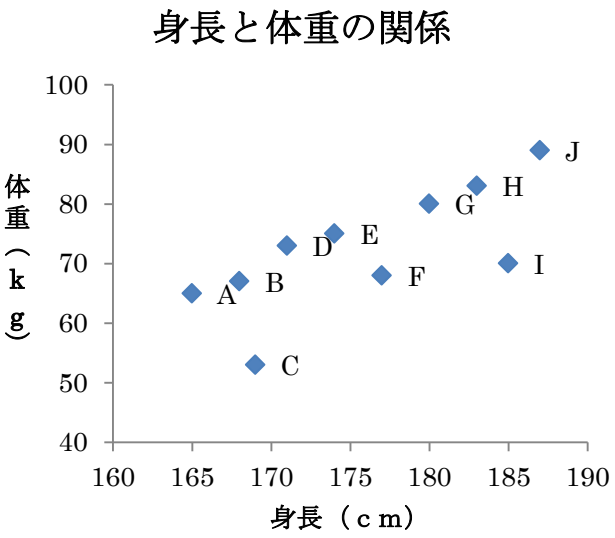
3　相関

散布図とは？

- ２種類の項目を縦軸と横軸にとり、プロット（打点）により作成される図のことです。散布図を作成することで、２種類の項目の間にどのような関係がある調べることが可能となります。

演習３　以下の表は身長、体重の組み合わせを示しています。散布図を描いて両者の関係を説明して下さい。

名前	身長 (cm)	体重 (kg)
A	165	65
B	168	67
C	169	53
D	171	73
E	174	75
F	177	68
G	180	80
H	183	83
I	185	70
J	187	89



（説明）＿＿＿と＿＿＿は＿＿＿＿＿＿＿＿＿＿＿＿。

相関とは？

- ２つのものの間に関連があることです。
  - **正の相関**　散布図において、プロットが右上がり。
    （＝一方の変量の値の増加にともない他方の変量の値も増加している）
    例：身長と体重
  - **負の相関**　散布図において、プロットが右下がり。
    （＝一方の変量の値の増加にともない他方の変量の値は減少している）
    例：マンションの価格と東京駅からの距離
- **無相関**　正の相関でも、負の相関でもない場合。「相関がない」ともいう。

演習４ 以下の表は、毎月の平均気温とカゼ薬への支出額を示しています。散布図を描いて、平均気温とカゼ薬への支出額の関係を述べてください。

月	平均気温(℃)	カゼ薬への支出　額（円）
1	4.4	241
2	4.8	263
3	7.6	235
4	14.3	160
5	17.3	155
6	22.7	118
7	25.0	111
8	27.3	100
9	23.8	113
10	18.0	204
11	11.0	268
12	4.4	314

【参考】Excel で散布図を描く手順

1　「平均気温の列」と「カゼ薬への支出額の列」を、列名を含め選択

- 最初に選択した列が横軸、次に選択した列が縦軸になる。
  - （参考）「離れた列を選択する方法」＝「ctrl キーを押して選択する」
  - グラフの横軸は「原因」と考えられる列、縦軸は「結果」と考えられる列にする。

2　「挿入」タブ→「グラフ」→「散布図」→「散布図（マーカーのみ）」

3　タイトル、軸ラベル、凡例

- グラフの余白をクリックしてグラフを選択すると、
  画面上方に「グラフツール」という緑色の帯と
  「デザイン」、「レイアウト」、「書式」のタブが現れるので、
  「レイアウト」を選択

- グラフタイトル→「平均気温とカゼ薬への支出額の関係」
- 軸ラベル→主軸横ラベル→（適宜）→「平均気温（° C）」
- 軸ラベル→主軸縦ラベル→（適宜）→「カゼ薬への支出額（円）」
- 凡例→なし
  - （ヒント）凡例があったら delete キーで削除、
    又は、「グラフツール」→「レイアウト」→「凡例」→（なし）

4（できれば）

- 横軸の目盛線を消す、（又は、縦軸にも目盛線を入れる）
  - 「グラフツール」→「レイアウト」→「目盛線」→「主軸横目盛線」→（なし）
  - （又は、「グラフツール」→「レイアウト」→「目盛線」→「主軸縦目盛線」→「目盛線」）
- 散布図を正方形にする
  - （グラフが選択された状態で）右下隅にポインターを合わせ、
    両矢印に変わったらドラッグし、グラフが正方形になるように調整する
  - （注）正方形の方が、２つの量の関係が見やすい

5　（時間が余ったら）

- 傾向線を入れる
  - 「グラフツール」→「レイアウト」→「近似曲線」→「線形近似曲線」

6　両者の関係の説明を書く

7　ヘッダーに学籍番号、名前を入れる

- グラフの選択の解除
  - グラフエリア以外の場所を一旦クリック
- 挿入　→ヘッダーとフッター　→（記入）