表計算 (Excel) とデータ処理 (7)

1 資料用ファイルのダウンロード

http://isl.sss.fukushima-u.ac.jp/ から本日の資料をダウンロードする。

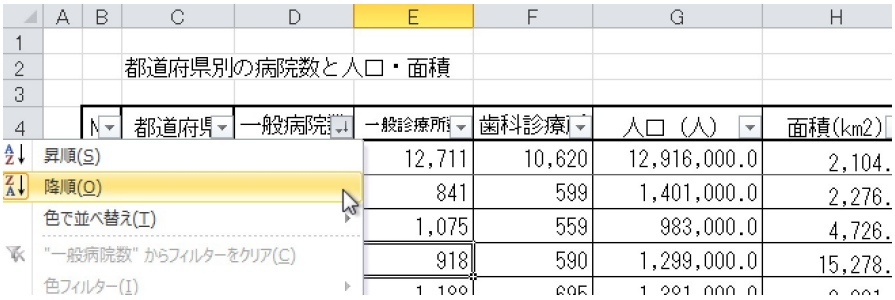
2.1 1つの変数の代表値

「都道府県別の病院数と人口・面積」ワークシートを使って，各尺度における代表値，散布度を求めてみよう。

1）最頻値（mode モード）（名義尺度以上）

「一般病院数が最も多い（最頻値） 自治体はどこか？」

病院総数が最大の自治体を見つければ良いが，ここでは「並べ替え」機能を使って見つけてみよう。



タイトルセル「一般病院数」の△印をクリックし「降順」に設定する。

すると，タイトル行のすぐ下から，病院総数の多い順に並べかえられる。

（答え）病院総数の最も多い自治体は「東京都」

2a）中央値（median メディアン）（順序尺度以上）

「病院総数で並べ替えたとき，ちょうど真中の順位（中央値）の病院数はいくつか，またその自治体はどこか？」

スクロールして24番めを探すと病院数127で，自治体は「新潟県」であることがわかる。

（答え）病院総数で並べ替えたとき，ちょうど真中の順位（24番目）（中央値）は127，自治体は「新潟県」

2b）間隔尺度のヒストグラムの中央値（median メディアン）（順序尺度以上）

ウェストサイズ10cmごとに区切って，99本のズボンの売り上げ数のヒストグラムを作ったとすると，中央値は50本目の値であるが，階級値で評価する場合は，50本目を含む階級から計算する。（階級値または，階級内で重みをつけた計算を行う）

3）算術平均（average アベレージ）（間隔尺度以上）

「病院総数の自治体平均はいくつか？」

算術平均とは，総数をデータ数で割った数である。

エクセルでは average() という関数で，自動的に平均を計算させることができる。練習のために，「一般病院数」についても平均を求めてみよう。

「平均」列の一般病院の平均として，D54 に以下の式を入れる。

=AVERAGE(D5:D51)

合計は関数 SUM() で計算する。

最後に，平均を計算したセルを横方向に「面積」までコピーする。

（答え）病院総数の自治体平均は 176.6。

2.2 1つの変数の散布度

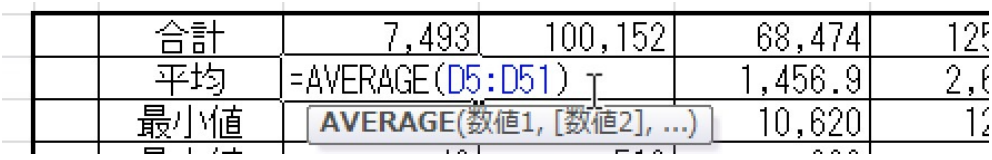
1）最大値／最小値，第1四分位／第3四分位（順序尺度以上）

「一般病院数の最大値，最小値はいくつか，またその自治体名は？」

「一般病院数の第1四分位／第3四分位はいくつか，またその自治体名は？」

先の「並べ替え」で総数の大きい順に並べ替え済みなので，最大値は1位，最小値は47位の自治体名と総数を答えれば良い。また，並べ替えなくても，max(), min() で，指定範囲の最大，最小値を計算できる。

（答え）最大値は東京都の592，最小値は鳥取県の43。



第1四分位とは，下から1/4の順位なので，36, 37位，第3四分位は下から3/4の12, 13位の値と都道府県名を答えれば良い。

（答え）第1四分位は岩手91，沖縄91，第3四分位は熊本211，静岡175。

エクセルでは「四分位数」を計算するための関数 QUARTILE.INC(**データ範囲, 分位指定**) がある。分位指定は 0:最小値，1:第１四分位（下から25%），2:中央値，3:第３四分位（下から75%），4:最大値，である。

※定義から，第一４分位と第三４分位の間には総数の半分のデータが存在する。

［練習］表の各変数について最大，最小値，第１，第３四分位数を求めなさい。

２）分散／標準偏差（間隔尺度以上）

分散(variance)とは，「各値と平均値の差の自乗」の平均で，以下の式で定義されている。

$$\mathrm{var}\,iance = \frac{\sum\left(x-\bar{x}\right)^2}{n} = \frac{n\sum x^2 - \left(\sum x\right)^2}{n^2}$$

エクセルでは関数 varp() で分散を計算することができる。

標準偏差(standard deviation)は，分散の平方根で，以下の式で定義されている。

$$s\tan dard\_deviation = \sqrt{\mathrm{var}\,iance} = \sqrt{\frac{n\sum x^2 - \left(\sum x\right)^2}{n^2}}$$

エクセルでは関数 stdevp() で標準偏差を計算することができる。

［練習］各変数の分散と標準偏差を求めなさい。また，エクセルには平方根を求める関数 sqrt() が準備されている。 sqrt() を使って標準偏差が分散の平方根になっていることを確かめなさい。

３）標準偏差の意味

正規分布（ガウス分布）という分布の変数では，平均と標準偏差の間には以下の関係がある。多くの変数は分布が正規分布に近いと考えられている。

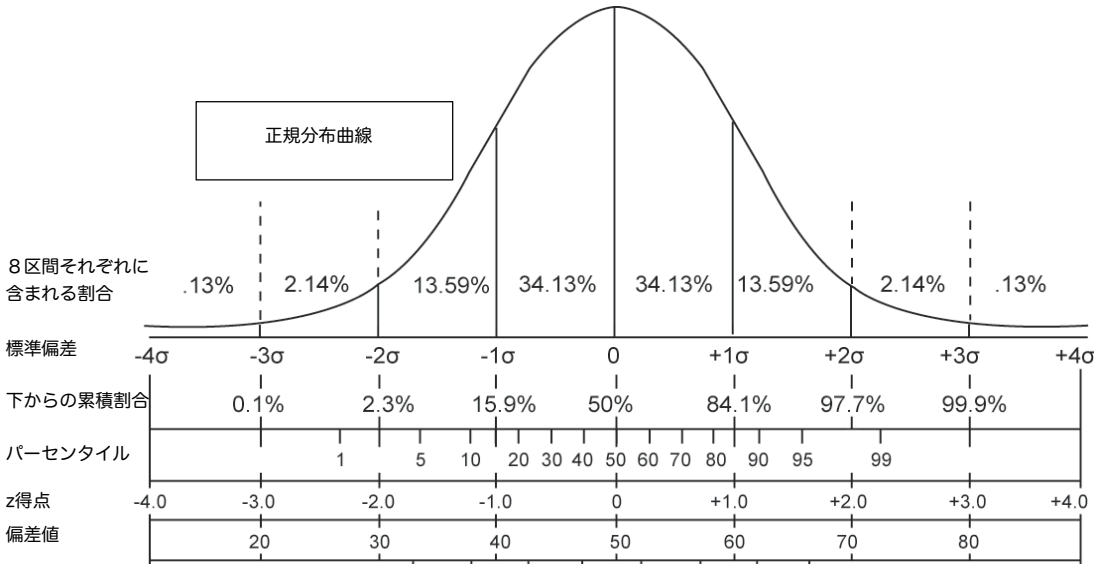
正規分布は左右対称の形をしていて，平均μ，標準偏差σの正規分布は以下の特徴を持つ。

総数（全体）を1とすると。

平均+/-標準偏差の範囲内には，全体の 0.683 のデータが入る

平均+/-標準偏差×２の範囲内には，全体の 0.954 のデータが入る

平均+/-標準偏差×３の範囲内には，全体の 0.997 のデータが入る



出典 WikiPedia http://ja.wikipedia.org/wiki/偏差値 を一部改変

「偏差値」とは平均を50, 標準偏差を10とした値である。

問題）10000人の集団の成績が正規分布になっているとき，偏差値50以上の人は何人か，また，70以上の人は何人か。