6　基本統計量の計算

表計算ソフトである Excel 2007 の基本的な機能を紹介しながら，平均や分散といった基本的統計量について復習する。この章では主に次の書物を参考にした。

- 菅　民男，『Excel で学ぶ統計解析入門』，オーム社，東京，1999

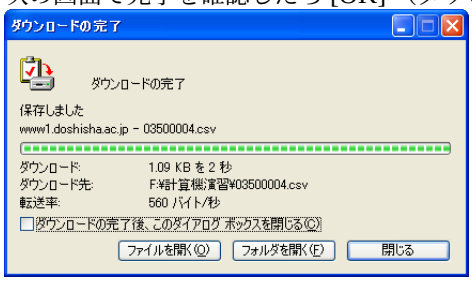
6.1　Web ページからのデータのダウンロード

6.1.1　ダウンロード

まずはじめに，CSV 形式のファイルを Web ページからダウンロードする。CSV ファイル (Comma Separated Value) は，値をコンマで区切って列挙したデータ形式で，どの表計算ソフトでも読み込める

Internet Explorer

1. Internet Explorer を起動する。
2.「機能分子工学科計算機演習」の Web ページを開く。
   URL (Uniform Resource Locator) http://www1.doshisha.ac.jp/~kibuki/computer/index.html
   (kibuki の前の ~（印刷物によっては ˜ ）は「チルダ」という。キーボードでは，数字が並んだ一番上の列で，0 の二つ右にある。Shift+^)
3. [授業資料・データ] → 第 6 回 [データ 06]。
   (i)「説明用データ」右クリック →「対象をファイルに保存」。
   (ii) ファイル名は一旦そのままで，保存先はリムーバブルディスクのの「計算機演習」フォルダとする。
   (iii) 次の画面で完了を確認したら [OK] （ダウンロード先等は各自が指定したものになる）。

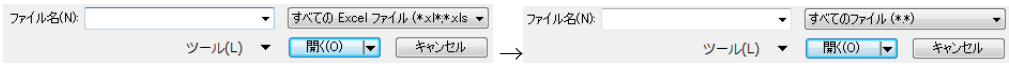


6.1.2　保存したファイルを Excel 2007 で開く

CSV ファイルを Excel 2007 で開く。ただし，CSV 形式はただの数値の並びなので，作業後に上書き保存しても，Excel の様々な機能を保ったまま保存できない。よって，作業後のファイルは必ず Excel 形式で保存しておく。通常，Excel 2007 のファイル形式である xlsx ファイルを作ればよいが，Excel 2003 でも作業したい場合には Excel 97-2003 の形式である xls ファイルを作る。

Excel 2007

1. Office ボタン → [開く]。
2.「ファイル名」の右横にファイルの種類の種類を指定するボックスがあるので，「すべてのファイル (*.*)」にする。



3. 「ファイルの場所」,「ファイル名」に適当なものを選ぶ (この場合，H ドライブの「計算機演習」フォルダ) → [開く]。
4. Excel 形式 (xlsx ファイル，あるいは必要なら xls ファイル) で保存しておく。
   (i) Office ボタン → [名前を付けて保存] → 「Excel ブック」(Excel 2003 でもファイルを使いたい場合は「Excel 97-2003 ブック」)。
   (ii)「ファイルの場所」に適当なものを選び,「ファイル名」は課題提出用にする → [保存]。
5. こまめに上書き保存すること。

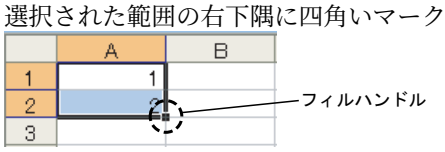
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	第5章	番号	元データ	ソート	累計	対数	偏差	偏差の平方	基準値	偏差値		定義	Excel		区切り	範囲	度数分布
2	基本統計量の計算		13								データ数						
3			11								最大値						
4			5								最小値						
5			13								和						
6			14								算術平均						
7			6								対数の和						
8			8								幾何平均						
9			8								最頻値						
10			7								中央値						
11			7								範囲						
12			17								偏差平方和						
13			12								偏差の和						
14			13								標本分散						
15			15								不偏分散						
16			14								標本標準偏差						
17			13								標準偏差						
18			5														

6.1.3 番号付け

説明用のファイルには，整数のデータが C 列の C2:C101 の範囲に 100 個並んでいる。あとの作業をわかりやすくするために B 列にデータ番号をつける。二つの方法を紹介する。

Excel 2007

- 「フィル」の機能を利用する。
  1. 新しいワークシートの A2 に「1」，A3 に「2」を入力。
  2. A2 と A3 の 2 つのセルを選択する。
  3. 選択された範囲の右下隅に四角いマーク「フィルハンドル」がある。



  そこにマウスポインタをあわせるとポインタの形が「+」になる。その状態でマウスを下の方へドラッグしてボタンをはなすと順々に数が入る。
  4. ひとつだけのセルの選択から始めれば同じ数値が入る。
  5. 1 おき，2 おきといった等差数列もできる。
  6. フィルには便利な使い方があるので，いろいろ試してみること。
- 数式を利用する。
  1. B2 に数字「 1 」を入力。
  2. B3 に数字「 =B2+1 」を入力。
  3. B3 をコピーし，B4:B101 に貼り付け。

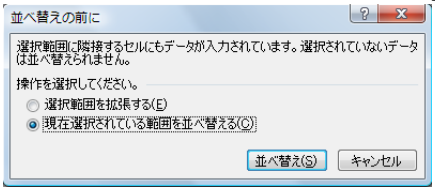
6.2 ソート

6.2.1 ソート（並べ替え）

データを小さいものからおおきいものへ並べたり（昇順），大きいものから小さいものを並べたり（降順）する事を，「ソート」という。

Excel 2007

1. 元のデータを残しておくため，C2:C101 を D2:D101 にコピー & ペーストする。
2. D 列のみを並べ替える。
   (i) D 列を選択 →「データ」タブ →「並べ替えとフィルタ」グループ →「昇順」 クリック。
   (ii)「並べ替えの前に」で「現在選択されている範囲を並べ替え」を ON →[並べ替え]。



3. Excel には，一列のデータをソートする機能だけでなく，表全体のデータをある列のデータにしたがってソートする機能もある。「並べ替えとフィルタ」グループの「昇順」 や「降順」 ではなく，「並べ替え」 

を使うと，より高度な並べ替えができる。失敗してもクイックアクセスツールバーの「元に戻す」  で元に戻れるので，操作してみること。

6.2.2 データ数

Excel 2007

- M2 に「=COUNT(C2:C101)」
- L2 に「=B101」（データ数は最後の番号）

[Excel 関数の説明]「=COUNT」は，指定した範囲の中で数値の表示されているセルの数を求める。空欄は数えない。

6.2.3 最大値と最小値

Excel 2007

1. 最大値
   - M3 に「=MAX(C2:C101)」
   - L3 に「=D101」（最大値は昇順にソートしたときの最後の数である）
2. 最小値
   - M4 に「=MIN(C2:C101)」
   - L4 に「=D2」（最小値は昇順にソートしたときの最初の数である）

[Excel 関数の説明]「MAX」は指定した範囲の中で最大値を求める関数。「MIN」は指定した範囲の中で最小値を求める関数。

6.2.4 総　和

Excel 2007

1. M5 に「=SUM(C2:C101)」
   [Excel 関数の説明]「SUM」指定した範囲の総和を求める。
2. 順に i 番目までの累計を求める
   (i) E2 に「=C2」。
   (ii) E3 に「=E2+C3」。
   (iii) E3 をコピーし E4:E101 にペースト。
   (iv) L5 に「=E101」（最後までの累計が総和）。

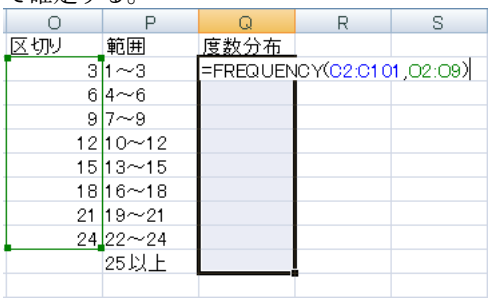
6.3 分　布

6.3.1 度数分布

データの範囲をある値ごとに区切り，それぞれの領域に含まれるデータ数を数えたものを「度数分布」という。

Excel 2007

1. O2 に「3」，O3 に「6」，…，O9 に「24」と入力する。ここでは，例として 3 ずつに区切ってみることにする。数値の入力にフィルや数式を用いて入力してもかまわない。
2. P2 に「1〜3」，P3 に「4〜6」，…，P9 に「22〜24」，P10 に「25 **以上**」と入力する。
3. Q2:Q10 をすべて選択した状態で「=FREQUENCY(C2:C101,O2:O9)」と入力し，Ctrl+Shift+Enter で確定する。



4. これで Q2 には $x \leq 3$，Q3 には $3 < x \leq 6$，…，Q9 には $21 < x \leq 24$ のデータ数がはいり，Q10 には $24 < x$ のデータ数がはいる。
5. 何かエラーがおきて抜け出せなくなった場合，Esc キーを押してみること。

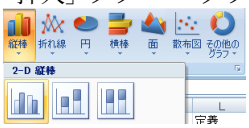
[Excel 関数の説明]「FREQUENCY」では，引数（ひきすう，カッコの中の値）に範囲を 2 つ指定する。はじめの範囲はデータの範囲，次の範囲は区切値の数列が書いてある範囲である。この関数では，Q2:Q10 の範囲すべてを用いて処理の結果を表示する。このような関数を「配列関数」といい，確定するときには必ず Ctrl+Shift+Enter を用いる。この例では，区切値の数列は O2:O9 なのに，関数の表示範囲は Q2:Q10 と 1 セル多くしているのがポイントである。これで範囲を超えたデータがないかどうか確認できる

6.3.2 ヒストグラム

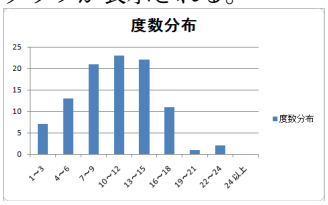
度数分布を棒グラフにしたものを「ヒストグラム」という。折れ線グラフにした場合は「度数多角形」という。範囲を隙間無く区切っているので，グラフの棒も隙間無くうめるのがヒストグラムの通常の書き方である。

Excel 2007

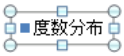
1. P1:Q10 の四角い範囲を選択。
2. 「挿入」タブ → 「グラフ」グループ → 「縦棒」で現れるメニューで，「2-D 縦棒」の一番左を選ぶ。



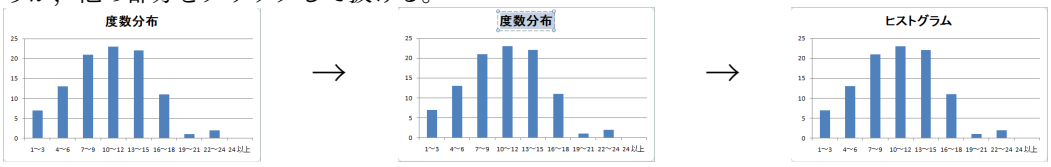
3. グラフが表示される。



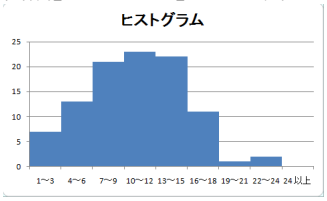
4. 右横の凡例「■度数分布」を選択し Delete。



5. 上のグラフタイトル「度数分布」の部分をダブルクリックし，テキストを編集できる状態にして，「ヒストグラム」に書き換える。Enter ではテキストボックス内で改行されるので，Esc をうまく使うか，他の部分をクリックして抜ける。



6. どれかの棒の上で右クリックして「データ系列の書式設定」 →「系列のオプション」 → 「要素の間隔」→「なし」 → 「閉じる」

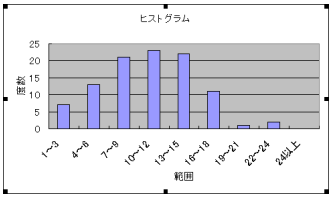


7. 必要に応じて色や線種を変更する。

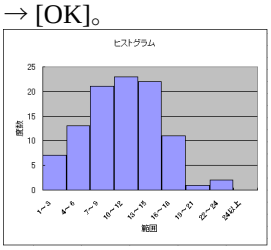
Excel 2003

1. P1:Q10 の四角い範囲を選択。
2. メニュー [挿入] → [グラフ] で「グラフの種類」は「縦棒」，「形式」は左上 → [次へ]。

3. 「データ範囲」はそのまま [次へ]。
4. 「タイトルとラベル」の「グラフタイトル」は「ヒストグラム」,「X/項目軸」は「範囲」,「Y/数値軸」は「度数」。
5. 「凡例」で「凡例を表示する」のチェックをはずす →「次へ」。
6. 「グラフの場所」は「オブジェクト」で「Sheet 1」→ [完了]。



7. 「グラフエリア」（外枠の中）をクリックして選択し，辺の中央や四隅のボタンをドラッグして大きさ調整する。
8. どれかの棒の上をダブルクリックして [データ系列の書式設定] → [オプション],「棒の間隔」を「0」→ [OK]。



6.4 代表値

$x_1, x_2, \cdots, x_N$ と N 個のデータがあるとする。ソートしたあとのデータは $X_1, X_2, \cdots, X_N$ と書く。これを用いて一般的な説明をする。Excel 2007 の説明は説明用のデータを例にして書く。

6.4.1 **算術平均（相加平均）**arithmetic mean

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{6.1}$$

単に「平均」といえば算術平均のことである。

Excel 2007

- ひとつの関数で書けば「=AVERAGE(C2:C101)」（M6 に入力）。
- 定義式に沿った書き方なら「=SUM(C2:C101)/COUNT(C2:C101)」（L6 に入力）。あるいは「=L5/L2」でもよい。

6.4.2 **幾何平均（相乗平均）**geometric mean

$$M_{\mathrm{G}} = \left(\prod_{i=1}^{N} x_i \right)^{1/N} \tag{6.2}$$

次のようにも書ける。

$$\ln M_{\mathrm{G}} = \frac{1}{N} \sum_{i=1}^{N} \ln x_i \tag{6.3}$$

データの中に 1 つでもゼロまたは負があってはならない。

Excel 2007

- ひとつの関数で書けば「=GEOMEAN(C2:C101)」（M8 に入力）。
- 定義式に沿った書き方なら，
  1. F 列にまず元データの自然対数を表示（F2 を「=LN(C2)」として，これを F3:F101 にコピー & ペースト）。
  2. F 列の和を L7 に計算する。
  3. その結果をデータ数で割ったものの exponential を L8 に表示する（「=EXP(L7/L2)」）。

[Excel 関数の説明] exponential を計算する Excel 関数は「EXP」である。

どの関数でも，引数の中にさらに数式を使ってもよい。

6.4.3 **最頻値（モード）**mode

最も頻繁に現れるデータの値。

Excel 2007

- ひとつの関数で表すと「=MODE(C2:C101)」（M9 に入力）。

最頻値が複数ある場合にはデータ列ではじめに現れたものが表示される。時間があれば，この関数を用いずに最頻値を求める手順を考えてみること。

6.4.4 **中央値（メディアン）**median

ソートしたあとのデータ列のちょうど真ん中にある値。

N が奇数のとき。

$$M_{\mathrm{D}} = X_m, \quad m = \frac{N+1}{2} \tag{6.4}$$

N が偶数のとき。

$$M_{\mathrm{D}} = \frac{X_m + X_{m+1}}{2}, \quad m = \frac{N}{2} \tag{6.5}$$

Excel 2007

- ひとつの関数で書けば「=MEDIAN(C2:C101)」（M10 に表示）。
- 定義式に沿った書なら，

1. まず元データをソートする（D 列）。
2. ソートした数列の適当なセルを参照する（今の例では D51 と D52 の平均を L10 に表示する）。

6.5 散布度

6.5.1 範囲（レンジ）range

データのうち最大のものを $x_{\max}$，最小のものを x_{mim} とする範囲 R は次のように定義できる。

$$R = x_{\max} - x_{\mathrm{mim}} = X_N - X_1 \tag{6.6}$$

Excel 2007

- Excel 2007 では，範囲をひとつの関数で求める方法はない。
- 定義式に沿った書き方なら，
  1.「=MAX(C2:C101)−MIN(C2:C101)」（M11 に入力）。
  2. ソートしたデータの終わりの値からはじめの値を引いてもよい（D 列の先頭と末尾を参照して O11 に計算）。

6.5.2 偏差二乗和 sum of square deviations

偏差 deviation とは，あるデータと平均との差 $(x_i - \langle x \rangle)$ である。よって，偏差二乗和は次のように定義される。

$$S = \sum_{i=1}^{N} (x_i - \langle x \rangle)^2 \tag{6.7}$$

偏差の和は，定義により 0 になる。ただし Excel で実際に偏差の和を求めた場合，計算誤差のため，厳密に 0 にはならないのが普通である（コンピュータは有限桁の計算しかできない）。偏差の 2 乗の和は，必ず正の値をもつ。

Excel 2007

- ひとつの関数で求めるには「=DEVSQ(C2:C101)」（M12 に入力）。
- 定義式に沿った書き方をすれば，
  1. 先に求めた算術平均（L6）を絶対参照して，G 列に偏差を計算する。
  2. H 列に偏差の二乗を計算する（「=G2^2」で G2 の 2 乗）。
  3.「SUM」を用いて偏差の二乗の和を求める（L12 に表示）。
- 偏差の和も計算してみること（G 列の和を L13 に表示）。計算誤差のため，正確にゼロにはならない。

6.5.3 分散 variance

データのばらつきを表すのによく分散 V が用いられる。分散が大きいということは分布が拡がっている事を意味する。単に「分散」といったとき，2 つの定義が考えられる。

$$V = \frac{S}{n} \tag{6.8}$$

$$V = \frac{S}{n-1} \tag{6.9}$$

これらを区別するために，式 (6.8) の方を「標本分散」sample variance あるいは「母分散」population variance といい，式 (6.9) の方を「不偏分散」unbiased variance という事もある。本によって呼び名が違うことがあるので注意すること。この授業では混乱を避けたい場合，前者を「標本分散」，後者を「不偏分散」と呼ぶことにする。2 つの量の意味合いの違いについては，あとの授業で取り扱う。

Excel 2007

- 標本分散をひとつの関数で求めるには「=VARP(C2:C101)」（M14 に入力）。
- 不偏分散をひとつの関数で求めるには「=VAR(C2:C101)」（M15 に入力）。
- 定義式に沿った書き方をすれば，
  1. 標本分散の場合，L12 に計算した偏差二乗和 S をデータ数で割る（L14 に表示）。
  2. 不偏分散の場合，L12 に計算した偏差二乗和 S をデータ数から 1 引いた数で割る（L15 に表示）。

6.5.4 標準偏差 standard deviation

分散はデータを 2 乗しているので，その平方根をとることが多い。これを標準偏差 σ という。

$$\sigma = \sqrt{V} \tag{6.10}$$

ここで V は，目的に応じて標本分散であったり不偏分散であったりする。標本分散から求めたものを標本標準偏差，不偏分散から求めたものを単に標準偏差と呼ぶことにする。

Excel 2007

- 標本標準偏差をひとつの関数で求めるには =STDEVP(C2:C101)」（M16 に入力）。
- 不偏分散から求めた標準偏差をひとつの関数で求めるには =STDEV(C2:C101)」（M17 に入力）。
- 定義式に沿った書き方なら，
  1. いずれの場合も，先に求めた分散の平方根をとればよい（L16 には L14 の平方根を，L17 には L15 の平方根を表示）。

[Excel 関数の説明] 平方根 square root を計算する Excel 関数は「SQRT」である。

6.6 基準値と偏差値

6.6.1 基準値 normalize score と偏差値 deviation score

平均 0，分散 1 になるように個々のデータを変換したものを基準値 z という。偏差を標準偏差で割ったものである。

$$z_i = \frac{x_i - \langle x \rangle}{\sigma} \tag{6.11}$$

基準値を次のように変換したものを偏差値という。

$$h_i = 10 z_i + 50 \tag{6.12}$$

Excel 2007

- 先に求めた偏差（G 列）と標準偏差（不偏分散から求めたもので L 列のもの）を利用して I 列に基準値を計算する。
- 基準値をもとに J 列に偏差値を計算する。