情報科学演習 第 13 回
表計算ソフトを用いた統計処理

1　本日の目標

- 分散, 標準偏差など統計の基本用語の定義と意味を知る.
- テキストファイルで書かれた表計算のデータを LibreOffice で読む方法について知る.

前回に続き, LibreOffice.calc の使用法を学びます. 今回は成績処理に関するデータの扱いを例に, 統計の基本用語とその定義について解説します. 今回は, 無駄を承知で定義に基づいた計算と, ソフトウェアが装備している関数を利用した両方の計算を実行します.

2　本日の実習

まゆ, りの, ゆき, じゅりな, れな, さやかの 6 人がある試験でそれぞれ, 3 点, 4 点, 8 点, 10 点, 7 点, 5 点を取ったとします. これらのデータをもとに, 平均点と各人の偏差値を計算します.

	A	B	C	D	E	F
1	名前	得点	平均との差	その2乗	得点の2乗	偏差値
2	まゆ	3				
3	りの	4				
4	ゆき	8				
5	じゅりな	10				
6	れな	7				
7	さやか	5				
8	平均					
9	分散					
10	標準偏差					
11						

偏差値は, 素点を x とすると, 標準偏差 σ と平均 $\bar{x}$ を使って次の式で定義されます.

$$10 \times \frac{x - \bar{x}}{\sigma} + 50$$

標準偏差は, 偏差 (平均値からの偏り) の平均です. 正確には次のように, **分散**の平方根として定義されます:

n 人の人の点数が, $x_1, x_2, \ldots, x_n$ とし, 平均を $\bar{x}$, 分散を V, 標準偏差を σ とすると,

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$
$$V = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{x_1^2 + \cdots + x_n^2}{n} - \left(\frac{x_1 + \cdots + x_n}{n}\right)^2$$
$$\sigma = \sqrt{V}$$

となります. 分散の式の 2 番目の等式は簡単に証明できるので, 証明してみて下さい.

なお, 偏差値は受験用語で数学用語 (統計用語) ではありません (純粋に日本語で, 外国語, 例えば英語に対応する言葉はありません. deviation value と英語に直訳すると違う意味になります) が, 分散, 標準偏差は, 医学や工学など実験系はもちろん, 経済学, 社会学, 教育学などデータ分析をするときには必要とされる基本用語です. 定義も簡単なので, この機会に覚えて下さい.

次の指示に従い, 図 1 の表を完成させます.

1. 図 1 にあるデータおよび項目名を入力して下さい.
2. 次に, 関数 AVERAGE を使用して, セル B8 に 6 人の成績の平均値を計算する式を書きます.
   セルに`=AVERAGE(B2:B7)` と入力します.
3. C2 に「まゆ」の得点 (B2) と 6 人の平均点 (B8) の差を式で入力します.
   この時,「=B2-B8」と入力してしまうと, C3 にこの式をコピーした時に「=B3-B9」が入力され, 本来求める値とは異なる計算結果になります. (このようなセルの参照を「相対参照」といいます.)
   これに対して, 平均点の記述されたセル (B8) のようにどのセルからもそのセルの値を共通に利用したい場合, 「**絶対参照**」という方法を用います. 絶対参照では, セルの行番号と列のアルファベットの前に, `$`を入れます. 例えば, セル B8 を絶対参照するには, `$B$8` とします. したがって, C2 に入れる式は, `B2-$B$8` となります. 他に「**複合参照**」がありますが, こちらは自習して下さい.
   絶対参照とコピー&ペーストを利用して, 表を完成させます.

1. C2 をコピーしてから C3〜C7 にペーストします
2. D2〜D7 には C 列の 2 乗を式で入力します.
3. E 列には, B 列の 2 乗が入るように式を入力します.
4. E8 には「各人の得点の 2 乗」の平均を入力します. (E2〜E7 の平均を計算する式を入力.)
5. B9 に 6 人の成績の分散を入力します.(分散は「2 乗の平均-平均の 2 乗」ですから, E8 から B8 の 2 乗を引いた式を書くことになります.)
6. D8 に D2〜D7 の平均を計算する式を入力します. (この値は「各人の得点から平均点を引いたもの」の 2 乗ですから, 分散の定義式です. B9 の値と一致することを確認して下さい.)
7. B10 に標準偏差を入力します. 平方根を求めるには, SQRT という関数を利用します.)
8. F2〜F7 に各人の偏差値を計算する式を入力します.
9. C8 に C2〜C7 の平均を計算する式を書きます. (これは, 理論上 0 となりますが, $x.xxxxE-10$ のように表示されることがあります. これは, $x.xxxx \times 10^{-10}$ の意味で 0 に近い値です. 小数計算では, 計算機は無限小数や小さい数を途中で値を四捨五入するため, 理論値との誤差を生じることがあります.)

注意

1. 分散や標準偏差は標準的な統計関数なので, それを求める関数が備わっています. 但し, それを Help で正確に探すのは難しいです. 理由は, 抽出調査をして検定, 推定をする場合の不偏分散と言う概念と, 母集団の分散 (母分散) と言う概念があり, これらの正確な説明が Help に書かれていないことがあります (Excel は少しマシ). 今のバージョンの LibreOffice.calc では, VAR(), VARP(), VARPA() が分散を計算する関数としてあるようですが, 実際の計算式を書いていないので, どれが何を計算しているのかがわかりません (おそらく, VAR() が不偏分散で, VARP() が母分散). この辺が,「出が会計ソフト」という部分でもあります.
2. 統計データの使い方は, 3 年次以降の統計関連の授業で詳しく勉強して下さい. たとえば, データの散らばり具合の指標である標準偏差を上のように定義することが合理的である理由を, 私は初等的に説明することができません. ($\frac{1}{n}\sum|x_i-\bar{x}|$ もデータの散らばり具合を表現するはずですが, なぜこれを利用しないかを説明できますか?)
3. 日本で日常的に用いられている「偏差値」に関しては, 統計上の意味が全くありませんので, それを計算する関数が, 備わっていることはありません.
4. 上の成績を後述の 5 段階相対評価で評価すると, 2 が 2 人, 3 が 3 人, 5 が 1 人となります.

2.1 練習問題 1

例題を参考に次のような表を作成してみます. 図では省略されていますが, これは 45 人の成績からなる古いデータで, 元となるデータはこの講義のページにありますからファイルの取り寄せます.

	A	B	C	D	E	F	G
1		全体	09年度生	10年度生	11年度生		
2	平均	59.5556	65.66667	60.53333	52.46667		
3	分散	294.158	279.8222	199.5822	314.5156		
4	標準偏差	17.151	16.72789	14.12736	17.73459		
5							
6	No.	学籍番号	文理	得点	得点の2乗	偏差値	合否
7	1	093101	理	86	7396	65.419	○
8	2	093102	文	65	4225	53.174	○
9	3	093103	文	46	2116	42.096	×
10	4	093104	文	51	2601	45.012	×
11	5	093105	文	62	3844	51.425	○
12	6	093106	理	95	9025	70.666	○

1. firefox で `http://www.math.u-ryukyu.ac.jp/~suga/joho/sampledata.txt` を表示します
2.「ウィンドウを右クリックして → 名前を付けてページを保存」で保存します.

注意

このように, 統計処理されるもとのデータは, テキストファイルで保存するのが基本です. 例えば, `http://www.math.u-ryukyu.ac.jp/~suga/joho/sampledata.pdf` のように PDF(Portable Document Format) 形式にしますと, 閲覧や印刷はできますが, それを元にしたデータ処理をしようとすると, 改めてそのデータをコンピュータに入力しなければなりません (PDF は, 上手に作ってあれば, それを元にテキストデータを作るツールはありますが...). それには, 手間もかかりますし, ミスも起きますし, データ量が多ければ, 不可能になることもあります.

震災に伴う福島の原発事故では, 初期の頃, 行政や東京電力がこの間違いを犯しました. つまり, 放射線データを PDF で公開したのです. それに対して, データ処理ができないという苦情が多く寄せられたようで, その後は, 東京電力よりテキスト形式 (前回述べた, CSV, Comma Separated Values 形式) でデータ公開がされました. こういう重要データの処理は, 様々な場所で別々の方法で行うことで, その結果の分析が正確になりますので,「データは使いやすい形で提供する」というのは, 重要なことです.

2.1.1 テキストデータを LibreOffice Calc に取り込む

ファイルマネージャーを起動して, ダウンロードフォルダに取り寄せたファイルがある事を確認して下さい. それをダブルクリックして見て下さい. これは, 各項目がタブと改行で区切られたテキストデータです. このファイルを LibreOffice Calc で読み込みます.

1.「ファイル」メニューから「新規作成 → 表計算ドキュメント」を選んで新しい表を作る.
2.「ファイル」メニューから「開く」を選んで, 左側の「ダウンロード」フォルダをクリックし, 右側の欄に現れる `sampledata.txt` を選択します. もしこのファイルが右側の欄に現れない場合, 右下にあるファイル形式選択の部分をクリックして,「すべてのファイル」を選べば, 現れます.
3. 右下の「開く」をクリックします.
4. 現れるウィンドウの下部に開くファイルのプレヴュー画面が現れますが, 文字化けが起こっているはずです. この文字化けを次の操作で解消します.
   (a) 一番上の「文字エンコーディング」のところをクリックして, 「日本語 (EUC-JP)」を探して選びます. これで, 最下部のプレヴィュー画面の文字化けがなくなります.
   (b)「区切りオプション」の所で「タブ」にチェックが入っている事を確認する.
   (c) 1 番下のプレヴュー画面の, 学籍番号の上の標準と書いてある文字をクリックすると, 上に「列の種類」の欄の「標準」をクリックして,「テキスト」に変更する. 0 から始まる文字列は, 標準で解釈すると数値と解釈されて, 先頭の 0 が無くなります.
5. 右下の OK を押すと, 保存したデータが LibreOffice Calc に読み込まれます.

上の文字コードで EUC-JP とありますが, EUC は, Extended Unix Code の略です. 10 年位前では, 標準的な Linux の日本語の文字コードでした. テキストデータは, 新しい文字コードが標準となっても, ソフトウェアで変換できますので, 寿命の長いデータになることは知っておいて下さい.

2.1.2 成績表の作成

次にこれらのデータから, 学年別の平均点と全体のデータにおける各人の偏差値を計算し, もとの表に加えます. (平均, 分散, 標準偏差を求める式は, 前に書いてあるものを参考にして下さい.)

1. 1 行目の行番号をクリックし, 1 行目をハイライト表示にします.
2. メニューバーで「シート → 行の挿入 → 行の上」を選びます.(全体のデータが 1 行繰り下げられます.)
3. 上の操作をあと 4 回繰り返し, 1 行から 5 行まで空の行を作ります.
4. 図に従って, セル A2, A3, A4, B1, D1, D1, E1, E6 の項目をタイプします.
5. セル E6 に「得点の 2 乗」とタイプし, 改行キーを押します. (セルの幅は, 列の名前が表示されている行をクリックした後, セルの境界部分にマウスを持っていくと ↔ の形のポインタが現れますので, マウスでドラッグして調整します.)
6. セル E7 に式「=D7*D7」を入力します.
7. セル E7 をセル E8 からセル E51 にコピーします.
8. セル B2,B3,B4 に必要な値が得られるように計算式をタイプして下さい.
9. 同様にして, C2, C3, C4, D2, D3, D4, E2, E3, E4 にも式をタイプします.
10. セル F6 に「偏差値」とタイプします.
11. セル F7 から F51 に全体のデータにおける各人の偏差値が入るように式を入れて下さい.
12. 練習問題 2 に進む.

2.2 練習問題 2

練習問題 1 のデータを使い成績評価をします. まず得点の隣の列に合否 (○, ×) を書き込みます.

1. セル G6 に「合否」と入力します.
2. セル G7 に次の式タイプします.

   ```
   =IF(D7>=60;"○";"×")
   ```

   ×は, 機種依存文字でない乗算記号を使います. >= は数学でいう ≧ の意味です.
3. セル G7 をセル G8 からセル G51 にコピーします.

次に隣の列に成績 (優, 良, 可, 不可) を書き込みます.

1. セル H6 に「絶対評価」と書き込みます.
2. セル H7 に次の式をタイプします.

   ```
   =IF(D7>=80;"優"; IF(D7>=70; "良"; IF(D7>=60; "可"; "不可")))
   ```
3. セル H7 をセル H8 からセル H51 にコピーします.
4. 終わった人は発展問題に進んで下さい.

2.3 発展問題

相対評価と最後のページにあるような統計表を作成します.

- I 列に相対評価を記述します. 偏差値をもとに, I 列に A,B,C,D,E からなる相対評価を入れて下さい. 相対評価の基準は次のようにします. 偏差値 65 以上 A, 偏差値 55 以上 65 未満 B, 偏差値 45 以上 55 未満 C, 偏差値 35 以上 45 未満 D, 偏差値 35 未満 E.
- 図のような成績分布の表を作って下さい. 例えばセル H7 から H53 の中にある優の数を数えるには, `COUNTIF(H7:H53;"優")` と入力します (コロン : と セミコロン ; に注意).
- 成績分布の表をもとに, 成績分布のヒストグラム (柱状グラフ) を作って下さい. グラフの作成方法は, 前回やった事を思い出して下さい.

ここでは, 成績処理を取り上げましたが, 最近は実験装置もコンピュータにつながれており, 実験結果も, ここでやったようなテキストデータで得られるのが普通です. それを元に, 統計処理ソフトを用いて, 標準偏差や相関係数を計算します. ただし, 分散や標準偏差は, ここでやったものではなく, 不偏分散, 不偏標準偏差と呼ばれるものを計算します.

LibreOffice Calc のような表計算ソフトの基本は, ここで取りあげたデータ処理です. ただし, 本格的な統計処理 (特に多量のデータがある場合) をするには専用のソフトを用いるのが普通で, 無料ソフトだと R といわれるものがよく使われるようです. 表計算ソフトでは,「表」の作成部分が処理の邪魔をして, 実際の処理速度は遅くなりがちです. そもそも会計ソフトなので, 数値計算の誤差が, 統計処理専用のソフトより大きいとも言われています. LibreOffice Calc, Excel ともに「きれいな図表を作るため」のソフトではありません. また, データの作成や表示に対して, あまり「表形式」には拘らないで下さい.

		成績分布		
絶対評価	人数		相対評価	人数
優	5		A	2
良	7		B	11
可	14		C	19
不可	19		D	8
			E	5