統計の手法レポート No.1 —解説と解答例

担当： 渕野 昌 (2005 年 4 月 28 日/6 月 25 日)

以下は，5 月 19 日に提出してもらったレポートの解説と解答例です．この文書は

`http://math.cs.kitami-it.ac.jp/fuchino/chubu/statistics-05s-report01.pdf`

から downloadable です．

`http://math.cs.kitami-it.ac.jp/fuchino/chubu/statistics-05s.html`

にはこのファイルをはじめ，講義に関連する資料のリンクがあります．

次の表は，あるクラスで行なったテスト（100 点満点）の得点のデータです：

57 71 62 65 62 88 95 66 100 71 73 79 45 32 78 75 86 67 50 56 100 73
79 35 75 50 56 88 75 90 63 82 63 82 63 72 87 49 48 69 45 90 73 82 78
72 83 55 80 54

1. このデータを度数分布表にまとめるとき，階級の個数は何個ぐらいが適当でしょうか？ スタージェスの公式，鈴木の公式，および教科書の p.5 にのっている目安からの結論を比較して考察してください．

このデータはサイズが 50 なので，$N = 50$ をスタージェスの式に代入すると，

$$1 + \frac{\log_{10} 50}{\log_{10} 2} = 6.64\cdots$$

となり，四捨五入して 7 程度と結論される．一方，鈴木の公式に $N = 50$ を代入すると，

$$1.7\sqrt[3]{50} = 6.26\cdots$$

となり，四捨五入して 6 程度と結論される．また，教科書 p.5 にある目安では，$5 \leq n \leq 7$ である．一方，このデータの値は最低が 32 点 最高が 100 点となっているが，100 点満点のテストの点は通常 30 点台， 40 点台，... と分類されることが多いので，ここでも点数を 10 点きざみに 30 点台， 40 点台，と分類して最後の 100 点は 90 点台に繰り込むことにすれば（これは例えば 29.35 から始めて階級の幅を 10.1 きざみにすると考えれば実現できる）階級の個数は 7 となり，上の目安とも一致するものとなる．

2. 1. の考察をふまえて，上のデータの度数分布表を作成し，ヒストグラムを作図してください．

1. の最後で与えた階級への分類による度数分布表は，次のようになる:

階級	人数
$29.35 \sim 39.45$ (30 点台)	2
$39.45 \sim 49.55$ (40 点台)	4
$49.55 \sim 59.65$ (50 点台)	7
$59.65 \sim 69.75$ (60 点台)	9
$69.75 \sim 79.85$ (70 点台)	14
$79.85 \sim 89.95$ (80 点台)	9
$89.95 \sim 100.05$ (90 点台)	5
計	50

ヒストグラムは省略する．

3. a) 上のデータの平均値とメディアンを求めてください．

データーの平均値は

$$\frac{57+71+62+65+62+88+\cdots}{50} = 69.78$$

メディアンについては，まずデータを値の小さい順に整列すると，

32 35 45 45 48 49 50 50 54 55 56 56 57 62 62 63 63 63 65 66 67 69 71 71 72 72 73 73 73 75 75 75 78 78 79 79 80 82 82 82 83 86 87 88 88 90 90 95 100 100

となる．この表の 25 番目と 26 番目の点数である 72 点がメディアンになる．

b) 2. で作成した度数分布表での平均値とモードを求めてください．

この度数分布表での各階級の階級値は，$29.35+\frac{10.1}{2}=34.4,\ 39.45+\frac{10.1}{2}=44.5,\ \ldots$ となるから，この度数分布表での平均値は，

$$\frac{34.4\times 2+44.5\times 4+54.6\times 7+\cdots}{50} = 69.952$$

である．また，この度数分布表でのモードは，最高頻度の 70 点台の階級値 74.8 である．

4. 上のデータの分散と標準偏差を求めなさい．

3. で求めたように，このデータの平均は，69.78 なので，分散は，

$$\frac{(57-69.78)^2+(71-69.78)^2+(62-69.78)^2+\cdots}{50} \approx 250.61$$

したがって，標準偏差は，これのルートだから，

$$\sqrt{250.6116} \approx 15.83$$

である．

5. このテストを実施した先生は，学生の素点が x 点のとき，$20+\dfrac{4}{5}x$ 点をこの学生の成績として成績表につけることにしました．このとき，成績表上の点数のデータの平均

値，分散，標準偏差は何になるでしょうか？ （ヒント： 教科書 p.26 の “ポイント” に書いてある事実を使うと 3. a) と 4. の結果から簡単に計算できます.）

$y = 20 + \frac{4}{5}x$ とすると，教科書 p.26 の “ポイント” から， $\bar{y} = 20 + \frac{4}{5}\bar{x}$ となるが，x の平均値 $\bar{x}$ は 3. から 69.78 だったから，成績表上の点数のデータの平均値は，$\bar{y} = 20 + \frac{4}{5} \times 69.78 \approx 75.82$ となる．一方，$\sigma^2(y) = (\frac{4}{5})^2\sigma^2(x)$ で，x の分散は $\sigma^2(x) = 250.61$ だから，成績表上の点数のデータの分散は，$\sigma^2(y) = (\frac{4}{5})^2 \times 250.61 \approx 160.39$. 標準偏差はこれのルートをとって $\sigma(y) \approx \sqrt{160.39} \approx 12.66$.

6. 平均点を中心にして ± 標準偏差の範囲の点をとっている学生の割合は全体の何 % でしょうか？ これをチビチェフの定理と比較して論じてください.

2. と 4. の結果から，平均点を中心にして ± 標準偏差の範囲は，69.78 ± 15.83 だから，54 点から 85 点までの間になるが，この範囲の点数をとっている学生は 50 人中 33 人いるから，この割合は 66% である．また，平均点を中心にして ± 2× 標準偏差の範囲は，39 点から 100 点までになるが，この範囲に入る点をとっている学生の数は 二人を除いて全員だから，その割合は，96% となり，この値は，チビチェフの定理が保証している全体の $\frac{3}{4}$ 以上という値をはるかに上まわるものとなっている．