レポート課題その2（4月23日分）要点解説

問1

以下の尺度について，外れ値に対して頑健でないと思われるものを選び（複数可），理由とともに示せ.

- 第2四分位数
- 第3四分位数
- 範囲
- 分散

要点解説

この中では範囲と分散が外れ値に対して頑健でない.

他のデータより極端に大きかったり小さかったりするデータ（つまり，外れ値）があると，最大値や最小値が大きく変化してしまう．範囲とは $\text{最大値} - \text{最小値}$ であるから，範囲も外れ値に対して大きな影響を受けやすい．対して，四分位数は少々の外れ値に対しては頑健であり，そのため四分位範囲も外れ値に対しては頑健である.

分散についても，外れ値をx_iとすると，$\left(x_i - \bar{x}\right)^2$が極端に大きな値になってしまうため，分散も外れ値に対して頑健であるとはいえない.

問2

データ $x_1, x_2, \ldots, x_n$ の平均，分散，標準偏差をそれぞれ $\bar{x}, S^2, S$ とする． いま，個々のデータ x_i に対して，$x_i^{'} = ax_i + b$ という変換を行った．変換後のデータ $x'_1, x'_2, \ldots, x'_n$ の平均，分散，標準偏差を $\bar{x}, S, a, b$ を用いてそれぞれ表せ.

要点解説

単に結果を覚えるだけでなく，次の問3も含めて結果を「導出」できるようになってください．データの変換や標準化の話は講義後半にも出てきます．また，統計学では，和記号

（$\sum$）や積分記号（$\int$）が多く出現しますので，その扱いに慣れるためにも，この種の問題は解けるようにしてください．

平均

変換後のデータの平均を $\bar{x'}$ とおくと，

$$\bar{x'} = \frac{1}{n}\sum_{i=1}^{n} x'_i$$

上式に $x'_i = ax_i + b$ を代入すると（以降，簡略化のため，$\sum_{i=1}^{n}$ は単に $\sum_i$ とだけ記載）

$$\begin{aligned}\bar{x'} &= \frac{1}{n}\sum_i (ax_i + b) \\ &= \frac{1}{n}\sum_i ax_i + \frac{1}{n}\sum_i b \\ &= a \cdot \frac{1}{n}\sum_i x_i + \frac{1}{n}\sum_i b\end{aligned}$$

いま，$\frac{1}{n}\sum_i x_i = \bar{x}$ なので，

$$\bar{x'} = a\bar{x} + \frac{1}{n}\sum_i b$$

また，$\sum_{i=1}^{n} b = nb$ なので（b を n 回足しているだけ），

$$\bar{x'} = a\bar{x} + b$$

よって，$\bar{x'} = a\bar{x} + b$.

分散

変換後のデータの分散を $S^2_{x'}$ とおくと，

$$S^2_{x'} = \frac{1}{n}\sum_{i=1}^{n} (x'_i - \bar{x'})^2$$

上式に，$x'_i = ax_i + b$，および，さきほど求めた $\bar{x'} = a\bar{x} + b$ を代入すると，

$$
\begin{aligned}
S_{x'}^2 &= \frac{1}{n}\sum_i (ax_i + b - (a\bar{x} + b))^2 \\
&= \frac{1}{n}\sum_i (a(x_i - \bar{x}))^2 \\
&= \frac{1}{n}\sum_i (a^2(x_i - \bar{x})^2) \\
&= a^2 \cdot \frac{1}{n}\sum_i (x_i - \bar{x})^2
\end{aligned}
$$

いま，$\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = S^2$ より，

$$
S_{x'}^2 = a^2 S^2
$$

よって，$S_{x'}^2 = a^2 S^2$.

この結果から分かるように，各データを a 倍すると，分散は a^2 倍されます．また，各データに定数（この例だと b に相当）を加えても分散は変化しません．

標準偏差

$S_{x'} = \sqrt{S_{x'}^2}$ に先ほどの分散の結果（$S_{x'}^2 = a^2 S^2$)を代入すると，

$$
S_{x'} = \sqrt{a^2 S^2}
$$

標準偏差は正であるから（すみません，ここ講義中に厳密に説明していなかったと思います），

$$
S_{x'} = |a| S
$$

この結果から分かるように，各データを a 倍すると，標準偏差は $|a|$ 倍されます．また，分散と同様に，定数を加えても標準偏差は変化しません．

問3

データ $x_1, x_2, \ldots, x_n$ を標準化したデータ $z_1, z_2, \ldots, z_n$ の平均は0，標準偏差は1であることを示せ．

要点解説

問2が解ければ，同じようなやり方でこの問題も解けます．$x_1, x_2, \ldots, x_n$の平均を $\bar{x}$, 標準偏差を S_x とすると，標準化されたデータは以下のように表されます（標準化の式はよく出てきますので覚えてください）．

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

これを使って，$z_1, z_2, \ldots, z_n$の平均と標準偏差を定義に従って求めればよいだけです.

平均

$z_1, z_2, \ldots, z_n$の平均を $\bar{z}$ とおくと,

$$\begin{aligned}
\bar{z} &= \frac{1}{n}\sum_{i=1}^{n} z_i \\
&= \frac{1}{n}\sum_{i} \frac{x_i - \bar{x}}{S_x} \\
&= \frac{1}{S_x} \cdot \frac{1}{n}\sum_{i} (x_i - \bar{x}) \\
&= \frac{1}{S_x}(\frac{1}{n}\sum_{i} x_i - \frac{1}{n}\sum_{i} \bar{x})
\end{aligned}$$

いま，$\frac{1}{n}\sum_i x_i = \bar{x}$ なので,

$$\begin{aligned}
\bar{z} &= \frac{1}{S_x}(\bar{x} - \frac{1}{n} \cdot n\bar{x}) \\
&= \frac{1}{S_x}(\bar{x} - \bar{x}) \\
&= 0
\end{aligned}$$

標準偏差

$z_1, z_2, \ldots, z_n$ の標準偏差を S_z とおくと,

$$S_z^2 = \frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})^2$$

先ほど求めた $\bar{z} = 0$ を代入すると,

$$S_z^2 = \frac{1}{n}\sum_{i} z_i^2$$

上式に $z_i = \frac{x_i - \bar{x}}{S_x}$ を代入すると,

$$\begin{aligned}
S_z^2 &= \frac{1}{n}\sum_{i} (\frac{x_i - \bar{x}}{S_x})^2 \\
&= \frac{1}{S_x^2} \cdot \frac{1}{n}\sum_{i} (x_i - \bar{x})^2
\end{aligned}$$

いま，$\frac{1}{n}\sum_i (x_i - \bar{x})^2 = S_x^2$ なので，

$$
\begin{aligned}
S_z^2 &= \frac{1}{S_x^2} \cdot S_x^2 \\
&= 1
\end{aligned}
$$

標準偏差は正なので，

$$
S_z = \sqrt{S_z^2} = \sqrt{1^2} = 1
$$

問4

それぞれの変数間の相関係数を表にまとめると下記の様になりました．

	通勤	課金額	発信回数	フォロワ	検索	運動時間	LINE
通勤	–	0.12	0.01	-0.20	0.15	-0.03	-0.09
課金額	–	–	-0.04	-0.04	-0.05	-0.07	-0.01
発信回数	–	–	–	-0.01	0.21	0.42	0.14
フォロワ	–	–	–	–	-0.05	0.10	0.12
検索	–	–	–	–	–	0.05	-0.01
運動時間	–	–	–	–	–	–	-0.03
LINE	–	–	–	–	–	–	–

ちなみに，この中で相関係数が0とはいえないと思われるものは「ソーシャルメディア上での発信回数」と「週の運動時間」のみでした（相関係数$r = 0.42$）．運動を良くしている人とソーシャルメディア上での発信回数にはなにか因果関係があるのかもしれませんね．部活や運動のことをソーシャルメディア上でつぶやく，とかでしょうか．

また，データを10個だけ抽出して計算したひとについては，相関係数が上記と大きく異なる結果を得たひともいると思います．それは計算結果が間違っているのではなく，少数のデータだけから相関係数を計算すると，本来の集団（そのような集団を母集団といいます）における値と異なった値がでてしまうということがよくあります．