§1 データリテラシー (1)　演習問題　解答

問題の難易度の目安【易】★☆☆　【基礎】★★☆　【標準】★★★

1 (★☆☆)(定量的データと定性的データ)

(1) 次の語群を，定量的データと定性的データに分類せよ：

5 段階アンケート調査，　不快度指数，　新生児指数，　気温，　国籍，　距離，　職種，　価格，　視聴率，　摂取カロリー，　血液型，　好きな食べもの．

(2) (1) で分類した各データはさらに細かく分類することができる．実際に，定量的データを**間隔尺度データ** (観測値間の差に意味があるもの)，**比率尺度データ** (観測値間の比率に意味があるもの) に分類せよ．また，定性的データを**名義尺度データ** (他と区別し分するもの)，**順序尺度データ** (順序には意味があるが間隔には意味がないもの) に分類せよ．

解 (1)　与えられた語群を定量的データと定性的データに分類すると，次の表のようになる：

定量的データ	定性的データ
不快度指数	5 段階アンケート調査
気温	新生児指数
距離	国籍
価格	職種
視聴率	血液型
摂取カロリー	好きな食べもの

表 1: 定量的データと定性的データの分類

(2)　(1) の表 1 をさらに細かく分類すると，次の表のようになる：

間隔尺度データ	比率尺度データ	名義尺度データ	順序尺度データ
気温	不快度指数	国籍	5 段階アンケート調査
距離	価格	職種	新生児指数
摂取カロリー	視聴率	血液型	好きな食べもの

表 2: 定量的・定性的データの細かな分類

■

2 (★☆☆)(定性的データに対する度数分布表)

次の表はある町に住む成人の母集団から，無作為に選んだ 500 人の血液型のデータである：

血液型	人数	相対度数 [%]
A 型	163	
B 型	146	
O 型	137	
AB 型	54	
計	500	100

このとき各血液型の相対度数を求め，データの中で最も現れる値 (**最頻値**) を求めよ．

解 たとえば，B 型の人の相対度数は $\frac{146}{500}\cdot 100 = 29.2\%$．以下同様に，他の血液型に対する相対度数を求めると，下の表のようになる：

血液型	人数	相対度数 [%]
A 型	163	**32.6**
B 型	146	**29.2**
O 型	137	**27.4**
AB 型	54	**10.8**
計	500	100

またこれより，最頻値は A 型である． ∎

3 (★☆☆)(定量的データに対する度数分布表)

次の表は，高速道路のある地点における，走行中の自動車 500 台の時速 km/h を計測したものである．

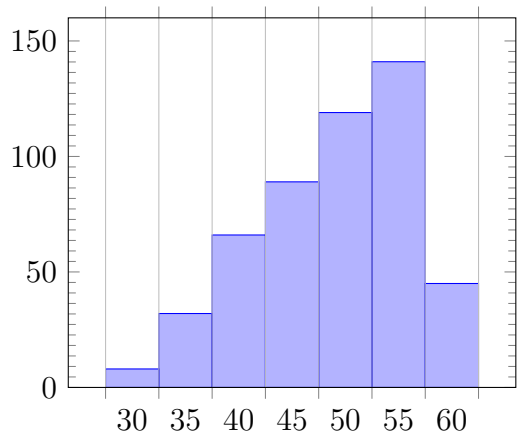
時速	27.5 ～ 32.5	32.5 ～ 37.5	37.5 ～ 42.5	42.5 ～ 47.5	47.5 ～ 52.5	52.5 ～ 57.5	57.5 ～ 62.5
台数	8	32	66	89	119	141	45

このとき，度数分布表とヒストグラムを作成し，それからわかることを述べよ．

解 たとえば，時速 27.5 ～ 32.5 km/h の自動車の相対度数は $\frac{8}{500}\cdot 100 = 1.6\%$．以下同様に，他の項目に対する相対度数を求めると，下の表のようになる：

時速	27.5 ～ 32.5	32.5 ～ 37.5	37.5 ～ 42.5	42.5 ～ 47.5	47.5 ～ 52.5	52.5 ～ 57.5	57.5 ～ 62.5
台数	8	32	66	89	119	141	45
度数	**1.6**%	**6.4**%	**13.2**%	**17.8**%	**23.8**%	**28.2**%	**9.0**%

また，ヒストグラムを図示すると次のようになる．ただし，縦軸は台数 (度数) を表し，横軸は時速 km/h を表す．



また，このヒストグラムより 45 ～ 55km/h 前後の自動車が多いことがわかる． ∎

4 (★☆☆)(無作為抽出)

ある母集団から無作為にサンプルが抽出された標本のことを**無作為標本**という．この無作為標本によって帰納的にその母集団を推察することができる．この無作為抽出について，次の問いに答えよ：

ある地域に住んでいる 100,000 人の中から 5,000 人を無作為に抽出して，ある感染症に罹患しているか否か検査を行うこととなった．この無作為に抽出する手順について，簡単に述べよ．

解 100,000 人の一人一人に 00000 から 99999 までの 5 桁の数字を割り当てる．次に，5 個の乱数の中からなる 5 桁の数字を無作為に縦横に並べた表 (=乱数表) を用いて，次のステップに基づいて 5,000 個の数字を拾えあげれば，5,000 人を無作為に抽出することができる．

Step 1：スタートする行および列を選択．

Step 2：縦，横，斜め好きな方向に数字を選ぶ． ∎

5 (★☆☆)(平均値と中央値の違いの例)

プロ野球のある球団における選手の年棒データに対して，平均値は球団側にとって有益であるが，選手側にとってあまり有益ではなく，反対に中央値は球団側にとってはあまり有益ではないが，選手側にとっては有益である．この理由を述べよ．

解 理由：球団側は支払っている念棒総額を得るために，平均値を使うのに対して，選手側は自分の年棒の位置付け (すなわち，ほぼ真ん中かそれ以上かそれ以下) を得るために，中央値を用いるから． ∎

6 (★☆☆)(分散に関する等式)

N 個の実数値のデータ $x_1, x_2, \ldots, x_n$ に対して，$\bar{x}$ をその平均とする．このとき分散 s^2 について次の等式が成り立つことを示せ：

$$s^2 = \overline{x^2} - \bar{x}^2.$$

ただし，$\overline{x^2} := \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i^2$.

解 直接計算による．

$$\begin{aligned} s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i^2}_{=\overline{x^2}} - 2\bar{x}\underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{=\bar{x}} + \bar{x}^2\underbrace{\frac{1}{N}\sum_{i=1}^{N} 1}_{=1} = \overline{x^2} - \bar{x}^2. \end{aligned}$$

∎

7 (★☆☆)(Cauchy-Schwarz の不等式)

a_i, b_i $(i = 1, \ldots, n)$ を実数とする．任意の実数 t に対して $\displaystyle\sum_{i=1}^{n}(a_i t + b_i)^2 \geqq 0$ が成り立つことから，次の Cauchy-Schwarz の不等式を導け：

$$\sum_{i=1}^{n} a_i b_i \leqq \left(\sum_{i=1}^{n} a_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} b_i^2\right)^{\frac{1}{2}}$$

解 ・$a_1 = \cdots = a_n = 0$ のとき：所望の Cauchy-Schwarz の不等式の成立は自明である．

・$a_1, \cdots, a_n$ のうち少なくとも 1 つが 0 でないとき：実数全体を $\mathbb{R}$ で表すとき，

$$\begin{aligned} &\sum_{i=1}^{n}(a_i t + b_i)^2 \geqq 0, \quad \forall t \in \mathbb{R} \\ \iff &\left(\sum_{i=1}^{n} a_i^2\right) t^2 - 2\left(\sum_{i=1}^{n} a_i b_i\right) t + \left(\sum_{i=1}^{n} b_i^2\right) \geqq 0, \quad \forall t \in \mathbb{R} \quad \cdots ①. \end{aligned}$$

$a_1, \cdots, a_n$ のうち少なくとも 1 つが 0 でないから，$\displaystyle\sum_{i=1}^{n} a_i^2 > 0$ に注意する．①がすべての $t \in \mathbb{R}$ に対して成り立つための必要十分条件は，判別式/4 $\leqq 0$，すなわち

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 - \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \leqq 0$$

となり，所望の式に逢着する．∎

8 (★☆☆)(相関係数の範囲)

s_x, s_y をそれぞれ変数 x，変数 y のデータに関する標準偏差とする．このとき，相関係数 $r_{xy} := \dfrac{s_{xy}}{s_x s_y}$ は $-1 \leqq r_{xy} \leqq 1$ を満たすことを，Cauchy-Schwarz の不等式を用いることで示せ．

解 共分散 s_{xy} について Cauchy-Schwarz の不等式より

$$
\begin{aligned}
|s_{xy}| = \left| \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \right| &\leqq \sum_{i=1}^{N} \frac{|x_i - \bar{x}|}{\sqrt{N}} \cdot \frac{|y_i - \bar{y}|}{\sqrt{N}} \\
&\leqq \left(\sum_{i=1}^{N} \frac{|x_i - \bar{x}|^2}{N} \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} \frac{|y_i - \bar{y}|^2}{N} \right)^{\frac{1}{2}} \\
&= (s_x^2)^{\frac{1}{2}} (s_y^2)^{\frac{1}{2}} = |s_x||s_y|.
\end{aligned}
$$

■

9 (★☆☆)(相関の調査)

圧延加工したアルミニウム板の高度と抗張力を測定して以下の表のデータを得た．硬度と抗張力の相関を調べよ．ただし，計算にあたり電卓，スマホ，PC を用いてもよいとする．

硬度 x ($\mathrm{H_g}$)	25	26	25	25	26	27	25	25	28	27
抗張力 y ($\mathrm{kg/m^2}$)	14.2	15.9	13.7	14.3	15.1	17.6	15.9	16.0	18.0	17.8

解 $x' = x - 25$, $y' = (y - 15)/0.1$ と変換することにより，右の表を得る：ゆえに，

x'	y'	$x'y'$	$(x')^2$	$(y')^2$
0	−8	0	0	64
1	9	9	1	81
0	−13	0	0	169
0	−7	0	0	49
1	1	1	1	1
2	26	52	4	676
0	0	0	0	0
0	10	0	0	100
0	10	0	0	100
3	30	90	9	900
2	28	56	4	784

$$
\begin{aligned}
s_{x'y'} &= 208 - \frac{9 \times 76}{10} = 139.6 \\
s_{x'} &= 19 - \frac{9^2}{10} = 10.9 \\
s_{y'} &= 2824 - \frac{76^2}{10} = 2246.4 \\
r &= \frac{139.6}{\sqrt{10.9 \times 2246.4}} = 0.892
\end{aligned}
$$

となるから，硬度と抗張力には**かなり強い相関がある**．

■