§2 データリテラシー (2)　演習問題　解答

問題の難易度の目安【易】★☆☆　【基礎】★★☆　【標準】★★★

1 (★★☆)(統計的観点からの考察)

ある大学の学生の成績表には，S(秀)，A(優)，B(良)，C(可) の 4 種類と不合格 F のみが記載されている．S は 90 点から 100 点，A は 80 点から 89 点，B は 70 点から 79 点，C は 60 点から 69 点を意味する．この成績表から素点換算する必要のあった X さんは，S，A，B，C をそれぞれ 90 点，80 点，70 点，60 点とし，対応する成績科目の個数を掛け合わせて，科目個数の合計で割ったものをその学生の成績の換算点とした．統計的観点から，この計算の問題点を指摘し，適切な総合評価の方法を示せ．ここでは GPA における計算方法との比較などは考えなくて良い．

解　統計的観点から，この X さんが間違ってしまったのは以下の点である：

各階級における階級値という度数分布表の考え方から言えば，S は 90 点から 100 点であるので階級値は 95 点，A は 80 点から 89 点なので階級値は 84.5 点，B は 70 点から 79 点ゆえ階級値は 74.5 点，C は 60 点から 69 点であるから階級値は 64.5 点である．$n(\mathrm{Y})$ を成績 Y $(\mathrm{Y}=\mathrm{S},\,\mathrm{A},\,\mathrm{B},\,\mathrm{C})$ で取得した単位数を表すと，

$$\text{総合評価} = \frac{95\times n(\mathrm{S}) + 84.5\times n(\mathrm{A}) + 74.5\times n(\mathrm{B}) + 64.5\times n(\mathrm{C})}{\text{取得した単位数}}$$

を，この度数分布表における総合評価にするべきであった．∎

2 (★★☆)(重み付き分散)

データ $x_1,\ldots,x_n$ における重み付き平均を

$$\overline{x}_w := \sum_{i=1}^{n} w_i x_i, \qquad \forall w_i \geqq 0, \quad \sum_{i=1}^{n} w_i = 1$$

と定義するとき，通常の標本分散の定義において，通常の標本平均の代わりに重み付き平均を用いた分散を

$$s_w^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_w)^2$$

とする．このとき，通常の標本分散 s_x^2 と比べてどちらの方が大きいか．

解　重み付き分散 s_w^2 について，

$$s_w^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_w)^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - \overline{x}_w)]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\overline{x}_w) + \frac{1}{n}\sum_{i=1}^{n}(\bar{x}-\overline{x}_w)^2 .$$

ここで，右辺第 3 項は (0 以上だから) 落として，右辺第 2 項について

$$\frac{2}{n}\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\overline{x}_w) = \frac{2}{n}(\bar{x}-\overline{x}_w)\sum_{i=1}^{n}(x_i-\bar{x})$$

$$= 2(\bar{x}-\overline{x}_w)(\bar{x}-\bar{x}) = 0.$$

したがって，

$$s_w^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(\bar{x}-\overline{x}_w)^2 \geqq \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = s_x^2.$$

ゆえに，重み付き分散 s_w^2 が，通常の s_x^2 に比べて大きく，等号成立は $\bar{x}=\bar{x}_w$ ∎

3 (★★★)(重み付き相加相乗平均の不等式)

正値データ $x_1,\ldots,x_n$ に対する重み付き相加相乗平均の不等式

$$\overline{x}_w = \sum_{i=1}^{n} w_i x_i \geqq x_1^{w_1}\cdots x_n^{w_n}$$

を示せ．ここに，重み $(w_i)_{i=1}^n$ は $w_i \geqq 0$, $\sum_{i=1}^{n} w_i = 1$ を満たすとする．

解 $f(x) := -\log x$ $(x>0)$ とおくと，$f''(x) = 1/x^2 > 0$ であるから，(下に) 凸関数．従って，Jensen の不等式より

$$f\left(\sum_{i=1}^{n} w_i x_i\right) \quad \leqq \quad \sum_{i=1}^{n} w_i f(x_i)$$

$$\iff -\log\left(\sum_{i=1}^{n} w_i x_i\right) \quad \leqq \quad -\sum_{i=1}^{n} w_i \log x_i$$

$$= \quad -\sum_{i=1}^{n} \log x_i^{w_i} = -\log\left(x_1^{w_1}\cdots x_n^{w_n}\right)$$

$$\iff \sum_{i=1}^{n} w_i x_i \quad \geqq \quad x_1^{w_1}\cdots x_n^{w_n}$$

となって，結論に達する． ∎

4 (★★☆)(回帰直線)

以下のような 2 次元データ (x, y) を得た：

x	6	8	10	12	14	16
y	6	11	20	29	27	33

以下の問いに答えよ．

(1) x の平均 $\bar{x}$ と標本分散 s_x^2 を求めよ．

(2) y の標準偏差 s_y を求めよ．

(3) (x, y) の共分散 s_{xy} を求めよ．

(4) y の x への回帰直線の式を $y = ax + b$ の形で表せ．

解 (1)　与えられた 2 次元データについて，

$$x\text{の平均} = \frac{6+8+10+12+14+16}{6} = \mathbf{11}.$$

また，標本分散 s_x^2 は一般に $s_x^2 = \overline{x^2} - \bar{x}^2$ で与えられることに注意する．データ x^2 について表にまとめると

x^2	36	64	100	144	196	256

であるから，平均は $\overline{x^2} = \frac{36+64+100+144+196+256}{6} = 132.7$．したがって，$s_x^2 = 132.7 - 11^2 = \mathbf{11.7}$.

(2)　同様に，$\bar{y} = \frac{6+11+20+29+27+33}{6} = 21$ であり，データ y^2 について

y^2	36	121	400	841	729	1089

ゆえ，平均は $\overline{y^2} = \frac{36+121+400+841+729+1089}{6} = 536$．したがって，$s_y^2 = 536 - 21^2 = 95$ であるから，y の標準偏差は $s_y = \sqrt{95} \approx \mathbf{9.7}$.

(3)　(x, y) の共分散 s_{xy} は $s_{xy} = \overline{xy} - \bar{x}\bar{y}$ で与えられる．データ xy について

xy	36	88	200	348	378	528

だから，$\overline{xy} = \frac{36+88+200+348+378+528}{6} = 263$．したがって，$s_{xy} = 263 - 11 \cdot 21 = \mathbf{32}$.

(4)　(1)–(3) で求めた各データに対して，

$$a := \frac{s_{xy}}{s_x^2} = \frac{32}{11.7} \approx 2.7 \quad ; \quad b := \bar{y} - a\bar{x} = 21 - \frac{32}{11.7} \cdot 11 = -9.1$$

とおけば，y の x への回帰直線は $\boldsymbol{y = 2.7x - 9.1}$ で与えられる．∎

5 (★★☆)(スクリーニング検査 (PCR 検査))

スクリーニング検査 (PCR 検査) において，各数値の求め方は次である：

	病気に罹患している人	病気に罹患していない人	計
陽性	x	z	$x+z$
陰性	y	w	$y+w$
計	$x+y$	$z+w$	$x+y+z+w$

- 有病率 $=\frac{\text{病人}}{\text{母集団}}=\frac{x+y}{x+y+z+w}$
- 感度 $=\frac{\text{真に陽性}}{\text{病人}}=\frac{x}{x+y}$
- 特異度 $=\frac{\text{真に陰性}}{\text{非病人}}=\frac{w}{z+w}$
- 偽陰性率 $=1-$ 感度
- 偽陽性率 $=1-$ 特異度
- 陽性適中度 $=\frac{x}{x+z}=$ 検査陽性者中における病人の割合
- 陰性適中度 $=\frac{w}{y+w}=$ 検査陰性者中における非病人の割合

この定義から，有病率が低いほど陽性適中度も低く，逆に有病率が低いほど陽性適中度も低いことがわかる．以下，ある感染症 A, B に対して，感度 95%，特異度 95% のスクリーニング検査を実施したところ，以下の表を得た．ただし，感染症 A は有病率 10%，感染症 B は有病率 1% であるとする：

【表 1】

	感染症 A に罹患している人	感染症 A に罹患していない人	計
陽性	950	450	1400
陰性	50	8550	8600
計	1000	9000	10000

【表 2】

	感染症 B に罹患している人	感染症 B に罹患していない人	計
陽性	95	495	590
陰性	5	9405	9410
計	100	9900	10000

このとき，以下の問いに答えよ．

(1) 【表 1】【表 2】をもとに，感染症 A, B に対する偽陽性率，陰性適中度，陽性適中度を求めよ．

(2) (1) の結果から分かることを論ぜよ．

解 (1) 【表 1】より感染症 A に対して

$$\text{偽陽性率}=\frac{450}{9000}=\mathbf{5}\%,\quad \text{陰性適中度}=\frac{8550}{8600}\approx\mathbf{99.4}\%$$

$$\text{陰性適中度} = \frac{950}{1400} \approx \mathbf{67.9}\%$$

同様に，【表 2】より感染症 B に対して

$$\text{偽陽性率} = \frac{495}{9900} \approx \mathbf{5}\%, \quad \text{陰性適中度} = \frac{9405}{9900} \approx \mathbf{95}\%$$

$$\text{陰性適中度} = \frac{95}{590} \approx \mathbf{16.1}\%$$

(2)　(1) から分かることとして，たとえば

- 有病率の低い疾患の場合には，スクリーニング検査を実施しても，大量の偽陽性者が出る (偽陽性率は低くても，偽陽性者数は増える).
- 有病率の低い疾患には，検査対象をハイリスク群に絞って，スクリーニング検査を行わなければならない.

■