統計検定「データサイエンス発展」サンプル問題解説

2020 年 8 月

1. 出題の考え方

データサイエンスでは、数理的な基礎の上に情報学および統計学の知識を活用するとともに、分析結果の倫理的な解釈も求められます。このように複数の領域を組み合わせる力がデータサイエンス力の重要な要素です。この観点から「データサイエンス発展」では出題範囲表にある複数の項目に関連する「複合問題」も出題します。サンプル問題では、複合問題を例示しています。「データサイエンス発展」では問題の分類の便宜のため、次の 4 つの領域を設けています。

1)倫理・AI、2) 数理、3) 情報、4) 統計・可視化

範囲表の大項目とこれらの 4 つの領域の関係は以下のようになります。

範囲表の大項目	領域
社会におけるデータ・AI 利活用	倫理・AI、情報及び統計・可視化
データ・AI 利活用における留意事項	倫理・AI
数理基礎	数理
データ取得とオープンデータ	倫理・AI 及び統計・可視化
デジタル情報とコンピュータの仕組み、アルゴリズム基礎、データ構造とプログラミング基礎	情報
データリテラシー、確率と確率分布、統計的推測	統計・可視化
データハンドリング、種々のデータ解析、データ活用実践	情報及び統計・可視化

2. 解答形式

以下のいずれかの形式で解答します

- ラジオボタン方式：択一問題の解答の一つを、マウスでクリックして選択する形式
- 数値入力方式:計算結果等を半角数値で解答する形式(問題に指定された桁数で入力。数値が負となる場合には符号も入力)

3. サンプル問題の解答・解説

問 1

正解：5.

情報と統計・可視化の複合問題である。x_1からx_{50}までを加算し、その和を用いて標本平均を出力するプログラムでは、i をカウンターとして 1 から 50 まで変化させながら x_i を S へ加算することを繰り返して和を計算する。最後に i=51 になっていることに注意し、i-1 で割り算した値を出力する。

最初に、（ア）で変数を初期化するので、S=0, i=1 である。また（イ）において、i>50 が成立するかどうかを判断する必要がある。さらに（ウ）では、i=51 になっていることから、標本平均 S/(i-1)を出力する。よって、正解は 5. である。

問 2

正解：3.

倫理・AI と数理の複合問題である。日本の年齢階級別総人口 S_k は J 個存在する基本単位区について合計するため、$S_k = \sum_{j=1}^{J} A_{jk}$ となる。基本単位区平均人口は基本単位区ごとの人口の合計$\sum_{k=1}^{K} A_{jk}$から算出される 1 基本単位区当たりの人口である。これは、日本の総人口 $\sum_{j=1}^{J}\sum_{k=1}^{K} A_{jk}$を基本単位区数 J で割ったものだから、$M = \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K} A_{jk}$ となる。よって、正解は 3. である。

問 3

正解：4.

数理と統計の複合問題である。Vをcで微分すると、合成関数の微分により$-2\sum_{i=1}^{n}(x_i - c)$ となる。これを 0 とおいてcを求めると$c = \frac{1}{n}\sum_{i=1}^{n} x_i$となり平均値が得られる。よって、正解は 4. である。

問 4

正解：2.

数理と情報の複合問題である。関数 f(x) で定義された関数が 0 となるような方程式を解く python のプログラムとなっている。関数 f(x) の返り値が 0.5-math.log(x/2.0-1.0)/math.log(0.25) であるので、ネイピア数 e を底とする対数（自然対数）を含む方程式 $\log_e(x/2\text{-}1)/\log_e(1/4)=0.5$ の解を二分法で求めるプログラムとなっている。対数の中は常に非負である必要があるので、$x/2\text{-}1>0$ から $x>2$ の範囲で解を求めなければならない。

$\log_e(a)/\log_e(b)=\log_b(a)$ の関係を使うと $\log_{\frac{1}{4}}\left(\frac{x}{2}-1\right)=\frac{1}{2}\log_{\frac{1}{4}}\left(\frac{1}{4}\right)$ なので、$\log_{\frac{1}{4}}\left(\frac{x}{2}-1\right)=$

$\log_{\frac{1}{4}}\left(\frac{1}{4}\right)^{\frac{1}{2}}$より$\frac{x}{2}-1=\frac{1}{2}$となり、$x$=3 を得る。これは、選択肢の中で 2.，3.，5. のいずれかである。さらに、このプログラムが探索するべき値の範囲は x>2 でなければならないので、3. は誤りであるとわかる。また、2 分法では探索範囲を決める初期値は求めるべき解をその区間内に含んでいなければならない。よって 5. では求めるべき解である x=3 は得られないことが分かる。よって、正解は 2. である。

問 5

正解:（ア）3　　（イ）20

倫理と統計の複合問題である。性別と居住地の組合せのそれぞれで、10 歳刻みの年齢が同一階級となり、（ア）変換されたデータベースは 3-匿名性を持つ。各レコードの年齢を階級値（15 歳および 25 歳）で置き換えた平均値は（15×6＋25×6）÷12＝20 より、（イ）20 歳である。