

第 3 章の答え

1. 正となる場合は、身長と体重、年齢と給料、都道府県別の人口と車保有台数、家賃と敷地面積など。負となる場合は、遊び時間と成績、家賃と築年数、家賃と駅からの距離、喫煙量と寿命など。
2. 定義式から、 x と y を入れ替えても標本相関係数は同じ。
3. 別の変数 z が、 x と y に影響を与えている「見せかけの相関」の可能性もあります。
4. x と y の両方とも平均 3 です。よって、標本共分散は、

$$s_{xy} = \frac{(1-3)(3-3) + (2-3)(2-3) + (3-3)(5-3) + (4-3)(1-3) + (5-3)(4-3)}{5-1} = 1/4$$

となり、標本標準偏差は

$$s_x = \sqrt{\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5-1}} = \frac{\sqrt{10}}{2}$$

$$s_y = \sqrt{\frac{(3-3)^2 + (2-3)^2 + (5-3)^2 + (1-3)^2 + (4-3)^2}{5-1}} = \frac{\sqrt{10}}{2}$$

となります。以上から、標本相関係数は

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1/4}{10/4} = 1/10$$

5. 標本共分散は 0 ですから標本相関係数も 0 です。式を書くと、 $\bar{x} = 0$ 、 $\bar{y} = 11.6$ から、標本共分散 S_{xy} は

$$\frac{(-5-0)(25-11.6) + (-2-0)(4-11.6) + (0-0)(0-11.6) + (2-0)(4-11.6) + (5-0)(25-11.6)}{5-1}$$

となり、これを計算すると 0 となります。標本共分散は 0 ですから、標本相関係数も 0 となります。

6. 強いチームなら人気も上がり収入も増えますから、多額の年俵を支払うことができるといえます。また、多額の年俵を支払えば良い選手を高額で引き抜くことができ、強いチームになるともいえます。よって、①チームの強さ→年俵、②年俵（選手のやる気がでたり、補強を行える）→チームの強さという双方からの因果関係がありそうです。相関関係を見ても、①と②の効果が混在しているため、②の効果が本当にあるのか、また、どの程度あるのかは分かりません。

7. 日本は年功序列が一般的です。よって、年齢とともに所得は上昇する傾向があります。年齢とともに血圧は上昇する傾向があります。以上から、両者の関係は「見せかけの相関」と考えられます。

8. 間違っています。アイスクリームの消費も夏に増える傾向があり、「見せかけの相関」の可能性がります。

9. 1)学科別の合格率をみると、男女によって合格率の差はほとんどないか、もしくは女性の方が高くなっています。全体の合格率に差が出た理由は、男性が難易度の低い学科を志望し、逆に、女性が難易度の高い学科を志望しているからです。たとえば、男性志願者は難易度の低い学科 A、B に多く、女性志願者は難易度の高い学科 C、E に多くなっています。この例のように、全体の傾向と層別の傾向が異なってしまう現象は、**シンプソンのパラドックス (Simpson's paradox)** といわれます。

2)男性の合格率は、男性だけの志願者数を加重としています。これは以下のよう確認できます。

$$\begin{aligned} \frac{512 + 353 + 120 + 138 + 53 + 22}{2691} &= \frac{825}{2691} \frac{512}{825} + \frac{560}{2691} \frac{353}{560} + \frac{325}{2691} \frac{120}{325} + \frac{417}{2691} \frac{138}{417} + \frac{191}{2691} \frac{53}{191} + \frac{373}{2691} \frac{22}{373} \\ &= \frac{825}{2691} \cdot 0.621 + \frac{560}{2691} \cdot 0.630 + \frac{325}{2691} \cdot 0.369 + \frac{417}{2691} \cdot 0.331 + \frac{191}{2691} \cdot 0.277 + \frac{373}{2691} \cdot 0.059 = 0.445 \end{aligned}$$

同様に、女性の合格率は女性の志願者数を加重としています。

$$\frac{108}{1835}0.824 + \frac{25}{1835}0.680 + \frac{593}{1835}0.341 + \frac{375}{1835}0.349 + \frac{393}{1835}0.239 + \frac{341}{1835}0.070 = 0.304$$

3)男女全体の受験者数で加重を付けると

$$\frac{933}{4526}0.621 + \frac{585}{4526}0.630 + \frac{918}{4526}0.369 + \frac{792}{4526}0.331 + \frac{584}{4526}0.277 + \frac{714}{4526}0.059 = 0.387$$

$$\frac{933}{4526}0.824 + \frac{585}{4526}0.680 + \frac{918}{4526}0.341 + \frac{792}{4526}0.349 + \frac{584}{4526}0.239 + \frac{714}{4526}0.070 = 0.430$$

であり、男女別の合格率にほとんど差はありません。

10. 2001年9月に発症したBSEによって、2002年に牛肉の輸入量が急減しました。BSE問題は2003年には沈静化し、牛肉輸入量はもとの水準に回復しました。BSEが原因で2003年の輸入量の増加率は一時的に高くなっており、セーフガードの発動は不適切だったと考えられます。

11. (1) 偏差の積和は以下として展開できます。

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i + \bar{x} \bar{y} - \bar{x} y_i - \bar{y} x_i) = \sum_{i=1}^n x_i y_i + n \bar{x} \bar{y} - \bar{x} \sum_{i=1}^n y_i - \bar{y} \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_i y_i + n \bar{x} \bar{y} - \bar{x} n \bar{y} - \bar{y} n \bar{x} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \end{aligned}$$

(2) 偏差の積和は以下として表現できます。

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i - \sum_{i=1}^n (x_i - \bar{x}) \bar{y}$$

偏差の和は0ですから、上式2項目は0となります。つまり、

$$\sum_{i=1}^n (x_i - \bar{x}) \bar{y} = \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \bar{y} \times 0 = 0$$

です。よって、 $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$ となります。同様に、

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (y_i - \bar{y}) x_i - \sum_{i=1}^n (y_i - \bar{y}) \bar{x}$$

と書けます。偏差の和は0から、上式2項目は0となります。よって、

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (y_i - \bar{y}) x_i$$