

度数分布表と相関表について

中 馬 正 博

序

全数調査などから得られる膨大な数のデータを対象とする場合、分析者がいわゆる「木を見て森を見ず」という状況に陥り易い。記述統計学はこれを回避するための方法として、①度数分布表の作成、②特性値の計算という2つの統計処理を提示している。

度数分布表は、最小値から最大値に到る範囲を実数区間であることを前提に、階級と呼ばれる隙間のない複数の小区間に分割し、各区間に属するデータの度数を示したものである。階級数は少なすぎても多すぎても、度数分布の姿を適切に捉えることを難しくする。最も基本的なことでありながら、原データから度数分布表という集計データを作成する方法について丁寧に触れている教科書は意外に少ない。本稿では、数少ない教科書の例として、宮川 [10] を参照する。

度数分布表や相関表は、岡田 [3] によると、かつては原データの数が膨大で直接計算することが困難な場合に、平均値、標準偏差、メジアン、相関係数などの特性値を、比較的簡単に計算する仲介手段として大きな役割を果たしていたが、計算環境が大幅に改善した今日では、その重要性は分布全体の姿を見通すという役割に移っているという。

そのため、今日のどのような入門的教科書においても、度数分布表を取り上げないで済ませることはない。しかし、2次元データの度数分布表と考えられる相関表を取り上げて、度数分布表の場合のように特性値の計算を示しているものは少ない。特に近年出版されている教科書では取り上げられていない場合が多いようである。

東大教養学部統計学教室 [8]、松原 [9] では、イギリスの人類遺伝学者、統計学者であるフランシス・ゴルトン (1822-1911) が、初めて相関という考え方を表す上で用いたという身長と上腕の長さに関する相関表が示されている。相関表を観察すると、厳密ではないが明らかに正の相関を示す直線的な傾向が認められると説明している。しかし、相関表に基づく具体的な計算方法までは示されていない。

相関表のデータに基づく相関係数の求め方を示している教科書として、例えば児玉 [4] は相関表をデータ数が多い場合の相関係数の求め方として取り上げている。森田ほか [12] では勤労者世帯の実収入と実支出、高校生17歳男子の身長と体重を取り上げているものの、観察数がそれほど多くなければ、「分類されない変数」の組の形で処理できると説明している。つまり、相関表として集計化したデータとするまでもなく、2次元の原データを用いた計算で済ませることができるとしているのである。したがって、多くの教科書において、具体的な計算方法は、原データによる場合のみで事足りると考えられているものと思われる。

また、石井 [1]、小寺 [5]、小針 [6] 等の教科書においても、周辺度数分布を経て、平均、分散、共分散、相関係数を求めるまでの説明に止められており、相関表を構成する2変量間の単回帰の計算方法について説明するものは、管見する限り武藤 [13] のみである。これは、恐らく相関表データについても、原データに基づく計算方法によれば良いので、改めて取り上げるまでもないという考えに基づくものであろう。

本論文の目的は、度数分布表と相関表について再検討すると共に、相関表に基づく特性値、特に単回帰の計算方法を再確認することにある。第1章では、まず度数分布表の階級数に関するスタージェスの公式の意味を、二項分布から同公式が導出されているとするヒンドマン [16] の指摘に基づいて再吟味した上で、第2節以降では階級値の選び方の観点から、平均、分散に関する原データと度数分布表データの計算結果を比較検討する。第2章では、2次元の度数分布表である相関表について、複数の教科書で取り上げられている平均、分散、共分散、相関係数の計算方法を再確認する。また、相関表が単に分布の特性値の計算を容易化する目的だけではなく、度数分布の状態を視覚的に確認する上

で意味があることを確認するために、2つの比較的大きめの相関表を具体例として取り上げ、2変量ヒストグラムを描いた。第3章では、第1節で相関表における単回帰の計算方法を再確認した。第2節では、その計算方法によって、前章で取り上げた総度数がかなり大きな相関表のデータについて、実際に回帰係数、決定係数の計算結果を求めた。

第1章 度数分布表

度数分布表の作成は、データがどのように分布しているかを捉える最も基本的な手段である。これは原データの持つ不要な情報を減らす情報の縮約化の作業でもあり、適切に行うことが肝要である。原データの持つ情報を減らし過ぎれば、全く意味のない作業にもなりかねず、作業結果にも分析者ごとに若干の差が現れる余地を残している。

しかし、教科書の多くはあまり悩む余地の少ない簡単な例を示すだけで済ませており、その作成方法について、丁寧な説明を加えている教科書も少ない。きちんと取り上げている数少ない教科書の一つ宮川 [10] は、第2章2節において、度数分布に大きな歪みがないことを前提とする場合の基本的な注意事項として、①適切な階級数、②均一な階級幅、③明瞭な級限界（階級間の境界）、④階級内に度数の集中点がある場合の取り扱い、⑤片方の開いた階級に関する取扱いの5項目を挙げている。階級に重複が生じてはいけないこと、片方の開いた階級について、度数分布表の作成者には合計あるいは平均、最小値あるいは最大値を注記する配慮、読者にはこれらの表記がないかについて注意が必要であるとまとめている。

第1章では、度数分布表の作成に伴う原データ集計化の作業が、データの持つ情報にどのような影響を与えるか、特に平均や分散の大きさにどのような影響を与えるか検討する。第1節では階級数についてスタージェスの公式の導出過程を検討し、第2節では平均について、階級下限と階級上限の真ん中の値を階級値とする通常の度数分布表、階級内平均を階級値とする度数分布表を比較する。第3節では分散について、階級値の求め方によりどのような違いが生じ

るか検討する。第4節では階級内のデータの散らばりを無視する通常の度数分布表の処理の妥当性を評価するため、データが階級内に均一に散らばっていると想定した場合について、級内平均が通常の階級下限と階級上限の真ん中の値に一致することを確認し、分散がどのように表せるか検討する。第5節では原データの具体例を挙げ、本章の検討内容を計算結果から評価する。

1.1 階級数

東大教養学部統計学教室 [8] を始めとする教科書の多くが、度数分布表における適切な階級数の目安として、スタージェス (Herbert Arthur Sturges 1882-1958) の公式を紹介している。宮川 [10] は一応の目安として公式の機械的な適用を戒めながらも、公式の示す適切な階級数 m は総度数の増加に比例して大きくなるものではなく、極めて大きな総度数に対しても案外大きな値とならないと説明している。また歪みの大きい分布について分布形を捉えるには、範囲を等間隔の階級幅で構成するのではなく、度数の集中する領域の階級幅を細かくする必要性も指摘している。

スタージェスの公式は、適切な階級数 m を総度数 n により次のように定めるものである。スタージェス [15] は、その適切な階級数を前提にして範囲を R とするとき、適切な階級幅 C は、 R を m で割ったものと推定されると述べている。

$$m = 1 + \log_2 n$$

$$C = \frac{R}{m} \doteq \frac{R}{1+3.322 \log n}$$

スタージェスの公式の導出方法については、ヒンドマン [16] がスタージェスは理想的な度数分布として、 m 階級の二項係数を度数とする二項分布を想定していると述べており、それを理由にこの公式は誤りであると結論づけている。同論文に基づいてスタージェスの公式の具体的な導出方法については、斎木 [7] が同論文を引用した上で説明を加えている。本論文では、もう少し公式の導出方法を見直して、その妥当性に検討を加えることにする。

まず二項分布 $B_i(m-1, \frac{1}{2})$ に従う確率変数 X を取り上げる。この二項分布は、確率関数が二項係数のみで決まるので、パスカルの三角形として示される二項係数の性質から左右対称の歪みのない分布となる。確率変数 X は $0, 1, 2, 3 \dots, m-1$ という m 通りの値をとる離散型変量であるが、適切な階級数を求めるという目的に鑑みて、二項分布の正規近似に際して利用される連続補正の考え方に従うものとする。確率変数 X のとる値 x は、 $x-0.5$ 以上 $x+0.5$ 未満の基準化された階級幅 1 の階級に対応するものとする。

$$X \sim B_i(m-1, \frac{1}{2}) \quad \therefore \quad P(X=x) = {}_{m-1}C_x \cdot \left(\frac{1}{2}\right)^{m-1}$$

この二項分布の確率が増加から減少に転じる境界値は $\frac{m}{2}$ であり、これを軸として左右対称の歪みのない分布となっている。従って、その最小確率は $P(X=0)$, $P(X=m-1)$ であり、 m 階級の度数分布表を構成するには、その確率に総度数 n をかけた値は作成しようとしている度数分布表の度数 1 以上の値に対応づけられる必要がある。

$$n \cdot P(X=0) = n \cdot {}_{m-1}C_0 \cdot \left(\frac{1}{2}\right)^{m-1} = n \cdot \left(\frac{1}{2}\right)^{m-1} \geq 1$$

$$\therefore \quad n \geq 2^{m-1} \quad \therefore \quad m \leq 1 + \log_2 n$$

従って、スタージェスの公式の示す階級数は、歪みのない分布を捉える上で適切な階級数の最大値を示していることになる。

それでは、同様の方法で歪みのある分布について、何か示唆は得られるだろうか。確率変数 X が二項分布 $B_i(m-1, p)$ に従うものとするとき、母数 p が $\frac{1}{2}$ より小さい場合は正の歪み、 $\frac{1}{2}$ より大きい場合は負の歪みを持つ分布となる。

$$X \sim B_i(m-1, p) \quad \therefore \quad P(X=x) = {}_{m-1}C_x \cdot p^x (1-p)^{m-1-x}$$

$$\therefore \quad P(X=0) = {}_{m-1}C_0 \cdot p^0 (1-p)^{m-1} = (1-p)^{m-1}$$

$$P(X=m-1) = {}_{m-1}C_{m-1} \cdot p^{m-1} (1-p)^0 = p^{m-1}$$

母数 p が $\frac{1}{2}$ より小さいとき、二項分布の最小確率は $P(X=m-1)$ となり、適切な階級数を m とするには次の関係が必要となる。

$$n \cdot p^{m-1} \geq 1 \quad \therefore \quad p^{m-1} \geq \frac{1}{n}$$

$$\therefore m \leq 1 - \log_p n < 1 + \log_2 n \quad (\because p < \frac{1}{2})$$

その一方、母数 p が $\frac{1}{2}$ より大きいとき、二項分布の最小確率は $P(X=0)$ となり、階級数を m とするには次の関係が必要となる。

$$n \cdot (1-p)^{m-1} \geq 1 \quad \therefore \quad (1-p)^{m-1} \geq \frac{1}{n}$$

$$\therefore m \leq 1 - \log_{1-p} n < 1 + \log_2 n \quad (\because p > \frac{1}{2})$$

この結果は、歪みのある場合の適切な階級数 m が、歪みのない場合より小さくなることを示している。二項分布 $B_i(m-1, p)$ の分散は $(m-1) \cdot p(1-p)$ であり、 p が $\frac{1}{2}$ に等しいときに最大となることから、分散がより小さい分布については階級数が少なくても良いとする妥当な結果を示しているものと思われる。しかし、歪みの大きい分布では、階級幅を統一した度数分布表を作成すると度数が集中する階級が生じる可能性が高まる。度数が集中する階級は、さらに細かい階級幅で度数を調べる必要があることを考えると、階級幅を統一した度数分布表を作成して済ませることは妥当ではない。ヒンドマン [16] は否定的であるが、ステージスの公式の利用は歪みのない分布、或いは歪みの小さい分布に限定した一応の目安に止めるとする従来どおりの取り扱いに従う限り、特に問題は見当たらず妥当と言えよう。

1.2 平均

原データに基づく平均 $\bar{x}$ は、原データ $x_1, x_2, x_3, \dots, x_n$ を用いて次のように表される。

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

これに対して度数分布表に基づく平均 $\bar{X}$ は、階級数 m の度数分布表のデータが、階級値と度数の組み合わせ $(X_1, n_1), (X_2, n_2), (X_3, n_3), \dots, (X_m, n_m)$ とまとめられている場合、次のように定義される。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^m X_i n_i$$

ここで求められる平均 $\bar{X}$ は、度数分布表が適切に作成されていたとしても、原データから求められる平均 $\bar{x}$ とは若干異なる値となる。しかし、データの集計方法によっては、両者を一致させることもできる。以下では、まず原データを予め昇順に並べた上で、統一された階級幅で集計作業を行い、階級値 X_i を第 i 階級の代表値と考え、第 i 階級を構成するデータの平均として定義する。第 i 階級の度数を n_i 、第 i 階級を構成する第 j 番目の原データを x_{ij} と表す場合、階級値 X_i は次のように記述できる。

$$X_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

各階級を構成するデータの平均を階級値と定義する場合、度数分布表の平均 $\bar{X}$ は原データの平均 $\bar{x}$ に一致することになる。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^m X_i n_i = \frac{1}{n} \sum_{i=1}^m \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \cdot n_i = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

但し、この度数分布表のデータについては、階級下限から階級上限に到る階級幅は統一されており、度数分布表作成の基準は一応満たされているものの、階級値の増分は一般的には一定とはならない。

手計算で簡便に度数分布表から平均や分散を求めるには、一次変換を利用した方法が有効である。階級値 X_i を階級下限と階級上限の真ん中の値とする場

合、階級幅を一定とすれば、階級値の間隔も一定となる。階級幅を d とすれば、階級値の間隔も一定値 d となることを前提に以下のような一次変換を行い、新たな階級値を U_i と表す。

$$U_i = \frac{X_i - X_0}{d} \quad \therefore X_i = X_0 + d \cdot U_i$$

このとき、 X_0 を第 1 階級の階級値 X_1 とすれば、 U_i は $0, 1, 2, \dots, m-1$ となるため、その平均と分散は以下のとおりである。 X_0 としては、小針 [6]、宮川 [10] は算術平均を含むと思われる階級の階級値、岡田 [3]、小寺 [5] は、度数最大の階級の階級値、つまり最頻値を用いている。いずれにしても、 $U_i n_i$ と $U_i^2 n_i$ の計算値はその階級で 0 となり、 $\bar{U}$ は下記の計算値よりさらに小さい値となることが期待される。

$$\bar{U} = \frac{1}{n} \sum_{i=1}^m U_i n_i = \frac{1}{n} \sum_{i=1}^m (i-1) n_i = \frac{1}{n} \sum_{i=1}^m i \cdot n_i - 1$$

$$s_{\bar{U}}^2 = \frac{1}{n} \sum_{i=1}^m U_i^2 n_i - \bar{U}^2 = \frac{1}{n} \sum_{i=1}^m (i-1)^2 n_i - \bar{U}^2 = \frac{1}{n} \sum_{i=1}^m i^2 n_i - \frac{1}{n^2} (\sum_{i=1}^m i n_i)^2$$

さらに以下の関係を利用すれば、度数分布表データから比較的容易に平均と分散を求めることができることが示されている。

$$\bar{X} = X_0 + d \cdot \bar{U}, \quad s_{\bar{X}}^2 = d^2 \cdot s_{\bar{U}}^2$$

級内平均を度数分布表の階級値とする場合には、以上のような手計算を前提に一次変換の性質を利用して考案された計算の簡素化を行うことはできない。しかし、少なくとも今日の関数電卓を利用できる環境であれば、多くの場合について一次変換の性質を用いて計算値を小さくする工夫を行わなくても、計算を進めることに大きな困難はないように思われる。

1.3 分散

本節では原データ $x_1, x_2, x_3, \dots, x_n$ に基づく分散 s_x^2 と度数分布表データの基づく分散 s_X^2 の関係をどのように理解したら良いか検討する。まず原データに基づく分散 s_x^2 の定義は、次のとおりである。

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

これに対して度数分布表として集計されたデータに基づく分散は、推測統計学を意識して不偏分散を分散と定義する立場を取らなければ、多くの教科書で特に詳しい説明もなく、分散は平均からの偏差の2乗の平均として、次のように定義されている。

$$s_X^2 = \frac{1}{n} \sum_{i=1}^m (X_i - \bar{X})^2 n_i$$

2つの分散 s_x^2 と s_X^2 の間の関係は、どのように説明できるのだろうか。管見する限り、両者は全く独立した形で定義され、適切な度数分布表による限り、両者には分析に支障の出ない範囲でわずかな差が生じる。しかし、両者の値が一致しないのはむしろ当然という説明に止められている。

本論文では、先に用いた度数分布表のデータにより、両者の関係を確認するため、原データに基づく分散を、度数分布表の階級値を用いて分解する。

$$\begin{aligned} s_x^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} \{(x_{ij} - X_i) + (X_i - \bar{x})\}^2 \\ &= \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - X_i)^2 + \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (X_i - \bar{x})^2 + \frac{2}{n} \sum_{i=1}^m (X_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - X_i) \dots\dots\dots \textcircled{1} \end{aligned}$$

平均からの偏差の総和は0となるので、級内平均を階級値 X_i とする場合、上式の第3項は0となり、原データの分散 s_x^2 はいわば各階級の級内分散からなる部分と級間分散に分解される。

$$s_x^2 = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - X_i)^2 + \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (X_i - \bar{x})^2 \quad \dots\dots\dots ②$$

第1項の級内分散に相当する部分 s_A^2 は、次のように変形すれば、相対度数を重みとして級内分散を加重算術平均したものであることがわかる。

$$s_A^2 = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - X_i)^2 = \sum_{i=1}^m \frac{n_i}{n} \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - X_i)^2$$

級内分散の部分 s_A^2 は非負なので、①の第1項からこれを除くと、原データの分散は級間分散に相当する部分よりも大きい、または等しいということになる。

$$\begin{aligned} s_x^2 &\geq \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (X_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^m (X_i - \bar{x})^2 n_i = \frac{1}{n} \sum_{i=1}^m (X_i - \bar{X})^2 n_i = s_X^2 \end{aligned}$$

よって、級内平均を階級値とする場合は、原データによる分散 s_x^2 は度数分布表による分散 s_X^2 よりも大きい、または等しいことになる。これは度数分布表による分散 s_X^2 が級内分散を無視していること、つまり階級を構成するデータが全て階級値に等しいものと単純化して計算することによる結果である。

但し、以上のように単純に説明できるのは、あくまで階級値を級内平均とする特殊な度数分布表データに基づくものであり、階級値を階級下限と階級上限の真ん中の値とする一般的な度数分布表データから計算される分散は、①式の第1項と第3項の相殺の程度により、原データによる分散より大きくなることも小さくなることもあり得る。

1.4 階級値

階級値としては、階級下限と階級上限の真ん中の値が用いられることが通常である。原データが示されず、度数分布表のデータから計算を進める必要がある場合、なぜこのように定めることになるのだろうか。階級内のデータの散らばりを見捨て、階級内のデータ分布が階級の中央で潰れていると想定して計算を進める通常の計算方法とは対極にあり、矛盾するようにも思われるが、散らばりが不明な場合のもう一つの単純な想定としては、逆にデータが階級内に均一に散らばっている状態を考えることもできる。このように想定する時、第 i 階級の下限を X_L 、上限を X_U 、度数を n_i とすれば、第 i 階級を構成する j 番目のデータ x_{ij} は次のように表せる。

$$x_{ij} = X_L + j \cdot \frac{X_U - X_L}{n_i + 1} \quad \text{但し, } j = 1, 2, 3, \dots, n_i \quad \dots\dots ②$$

このとき第 i 階級を代表する階級値を、階級を構成するデータの平均とすれば、階級下限と階級上限の和の半分、つまり通常の階級値 X_i の定義式が得られることがわかる。

$$\begin{aligned} X_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(X_L + j \cdot \frac{X_U - X_L}{n_i + 1} \right) = \frac{1}{n_i} \cdot n_i X_L + \frac{X_U - X_L}{n_i(n_i + 1)} \sum_{j=1}^{n_i} j \\ &= \frac{1}{n_i} \cdot n_i X_L + \frac{X_U - X_L}{n_i(n_i + 1)} \cdot \frac{n_i(n_i + 1)}{2} = \frac{X_L + X_U}{2} \end{aligned}$$

さらに第 i 階級のデータについて、階級内平均からの偏差を求めると、次のとおりである。

$$\begin{aligned} x_{ij} - X_i &= X_L + j \cdot \frac{X_U - X_L}{n_i + 1} - X_i = X_L + j \cdot \frac{X_U - X_L}{n_i + 1} - \frac{X_L + X_U}{2} = \frac{X_L - X_U}{2} + j \cdot \frac{X_U - X_L}{n_i + 1} \\ &= (X_U - X_L) \cdot \left(j \cdot \frac{1}{n_i + 1} - \frac{1}{2} \right) \end{aligned}$$

$$\begin{aligned}
\therefore s_{X_i}^2 &= \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - X_i)^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_U - X_L)^2 \left(j \cdot \frac{1}{n_i+1} - \frac{1}{2} \right)^2 \\
&= \frac{(X_U - X_L)^2}{n_i} \sum_{j=1}^{n_i} \left\{ j^2 \frac{1}{(n_i+1)^2} - j \frac{1}{n_i+1} + \frac{1}{4} \right\} \\
&= \frac{(X_U - X_L)^2}{n_i} \left\{ \frac{1}{(n_i+1)^2} \sum_{j=1}^{n_i} j^2 - \frac{1}{n_i+1} \sum_{j=1}^{n_i} j + \sum_{j=1}^{n_i} \frac{1}{4} \right\} \\
&= \frac{(X_U - X_L)^2}{n_i} \left\{ \frac{1}{(n_i+1)^2} \cdot \frac{n_i(n_i+1)(2n_i+1)}{6} - \frac{1}{n_i+1} \cdot \frac{n_i(n_i+1)}{2} + \frac{n_i}{4} \right\} \\
&= \frac{(X_U - X_L)^2}{n_i} \left\{ \frac{n_i(2n_i+1)}{6(n_i+1)} - \frac{n_i}{4} \right\} = \frac{(X_U - X_L)^2}{n_i} \cdot \frac{2n_i(2n_i+1) - 3n_i(n_i+1)}{12(n_i+1)} = \frac{(X_U - X_L)^2}{n_i} \cdot \frac{n_i^2 - n_i}{12(n_i+1)} \\
&= \frac{(X_U - X_L)^2(n_i-1)}{12(n_i+1)} \\
\therefore s_A^2 &= \sum_{i=1}^m \frac{n_i}{n} \cdot \frac{(X_U - X_L)^2(n_i-1)}{12(n_i+1)}
\end{aligned}$$

1.5 計算値の具体例

自ら調査しない限り、原データをそのまま利用できる機会は少ない。教室で入手できる試験の成績データについても、きちんと試験対策して臨んだグループと、ほとんど準備なく受験しているグループの2つから構成されていることを反映してか、双峰の分布となりがちである。そこで、本節では教科書で説明のために示されているデータを取り上げて、原データから得られる特性値と、度数分布表から得られる特性値を比較検討する。

表1-1は、蓑谷 [11] p.12に示されている2008年の東京都内S区の小学校6年生男子の身長を母集団として抽出された標本の大きさ100の身長データから作成した度数分布表であり、これに対応するヒストグラムを図1-1に示した。また同書 p.13には9階級の度数分布表が示されているが、スタージェスの公式に従い $1 + \log_2 100$ の値を求めると、約7.6となるので、本節では8階級を選び、階級幅5.5cmの度数分布表とした。

表 1-1 小学校6年生男子100人の身長（8階級）

階級	階級下限	階級上限	階級値(1)	階級値(2)	度数
1	122.5	128.0	125.25	125.65	2
2	128.0	133.5	130.75	131.54	7
3	133.5	139.0	136.25	136.58	15
4	139.0	144.5	141.75	141.78	25
5	144.5	150.0	147.25	147.08	29
6	150.0	155.5	152.75	152.22	16
7	155.5	161.0	158.25	157.73	4
8	161.0	166.5	163.75	163.85	2
総計					100

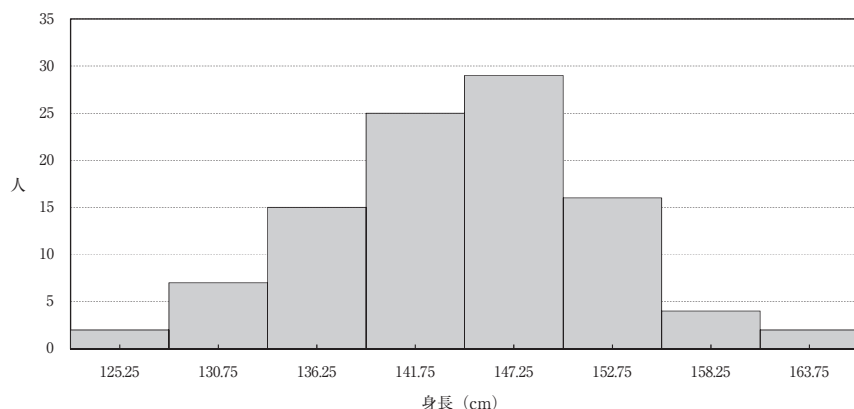


図 1-1 小学6年生男子100人の身長分布（ヒストグラム）

原データから直接計算した平均は144.25、分散は58.4255である。一方、階級下限と階級上限の真ん中を階級値とする階級値(1)の場合の平均は144.28となり、原データのそれよりわずかに大きい。また、分散は級内分散を無視して、級間分散のみを求めているにもかかわらず、60.7541と原データによる分散より大きくなり、級内分散2.4286を加えると63.1827とさらに大きくなる。

これに対して、各階級内の平均を階級値とする階級値(2)によって求めた平均は144.25と、原データから直接計算した平均に等しく、分散は級間分散で

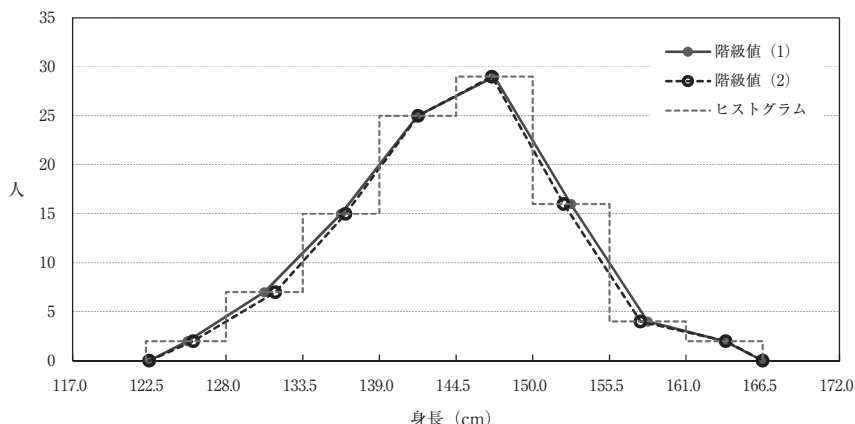


図 1-2 小学6年生男子100人の身長分布（階級値）

55.9947と級内分散 S_A^2 で2.4308, これらを加えると原データによる分散に一致する。第2節で示した結果が数値例でも確認された。

図 1-2 は図 1-1 に示したヒストグラムに度数折れ線を重ねて、階級値 (1) と階級値 (2) の違いを明示的に示そうとしたものである。階級値 (1) は階級の中央にあるが、階級値 (2) のずれ方から、度数折れ線からも散らばりがやや小さくなっていることを窺うことができる。

図 1-3 は階級下限と階級上限の真ん中の値で、階級内の分布が階級値で潰れているとした場合の想定データ (1) と原データの対応関係を示している。原データは階級内で変化するが、想定データ (1) は階級値に固定されているため、階段状に点が並ぶ図となっている。

これに対して図 1-4 では、度数分布表である表 1-1 から出発し、第 1 章 4 節の②式に従って、階級内に均一にデータが散らばることを想定した場合を表示している。点が45度線より上にあれば、想定したデータが原データを上回っていることを、45度線より下にあれば、想定したデータが原データを下回っていることを示している。各点の推移を見る限りほぼ45度線上にあり、想定データが原データにほぼ対応した系列となっている。

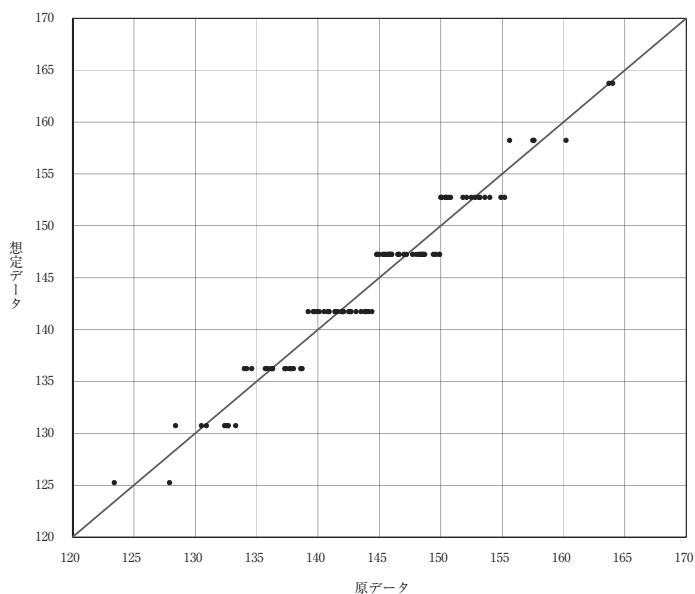


図 1-3 原データと度数分布表から想定されるデータ (1) の対応関係

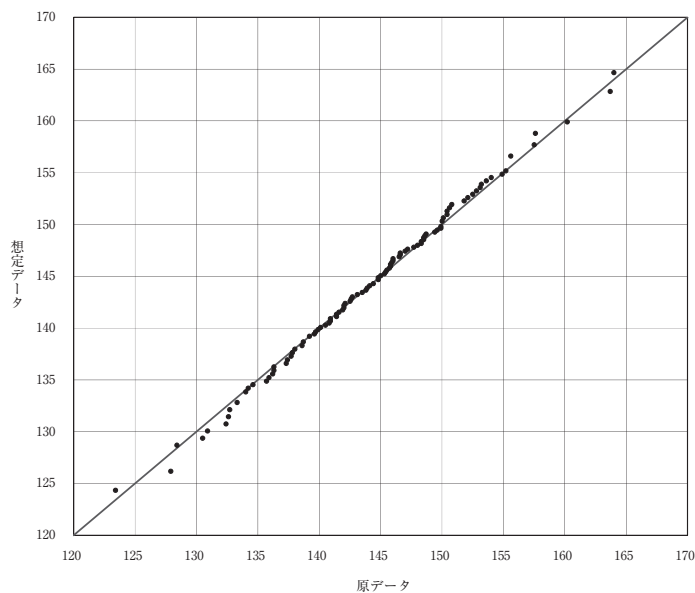


図 1-4 原データと度数分布表から想定されるデータ (2) の対応関係

図 1 - 4 に示したとおりデータが階級内に均一に散らばっていることを想定した場合については、級間分散に相当する s_x^2 が 60.7541、級内分散に相当する s_A^2 が 2.1772、両者を合計した分散は 62.9313 となり、原データから直接計算される分散 s_x^2 の 58.4255 を約 4.5 上回ることになる。

この結果に基づく限り、階級下限と階級上限の真ん中の値を階級値として分散を求める場合、級内分散の相当する部分を無視しても、度数分布表による分散は原データから計算される分散を超える可能性があり、データが階級内に均一に散らばることを合わせて想定すれば、計算される分散はさらに過大になることがわかる。

第 2 章 相関表

相関表 correlation table は、分割表、クロス集計表における 2 つの標識が特に量的変量である場合と説明されることもあるが、要するに 2 次元データに関する度数分布表のことである。相関表は 2 変量の同時度数分布を示している。従って、相関表の縦の合計値、横の合計値をまとめたものは、各変量の周辺分布を表す 1 次元の度数分布表となる。

本章の目的は、相関表から相関係数を求める一連の過程を要約することである。近年、パソコンの性能が大幅に向上し、手頃な価格で入手できるようになった。統計ソフトや統計パッケージを用いて容易に相関係数を求めることができることに加え、相関表がかつては大量データの計算に対処する一つ的手段と考えられていたこともあってか、相関表を 2 次元の度数分布表として紹介し、相関表に基づく相関係数の計算法を説明する教科書も減っているように思われる。

度数分布表を図示したヒストグラムに対して、相関表を図示すると立体的なものとなる。この立体図を Kendall [14] は 2 変量ヒストグラム bivariate histogram と呼んでいる。猪間 [2] はステレオグラム stereogram と呼び、手書きする場合の要領まで説明している。しかし、この作図はそれほど容易なものではない。2 変量ヒストグラムでは分布の山の向こうは隠れることになり、全体像を静止

画一枚に表示するのは困難で、試行錯誤的にできるだけ分布の概観を把握し易い向きから描いたものとせざるを得ない。本論文では、比較的大きな相関表2枚を例示的に取り上げ、エクセルの作図機能を用いて2変量ヒストグラムを描いた。

本章では、第1節で相関表における相関係数までの一連の計算を確認し、石井 [1]、児玉 [4]、小寺 [5]、小針 [6] が取り上げている一次変換による相関係数に到る計算法について整理する。第2節では、相関表の具体例を取り上げ、その計算結果を確認する。

2.1 相関表における特性値の計算

本節ではまず相関表から平均、分散、共分散、相関係数の計算式を確認する。基本的な相関表として2変量の同時分布を示す表2-1が与えられ、変量 X について M 階級の階級値 X_i 、変量について N 階級の階級値 Y_j の同時分布が示されているものとする。

表2-1 相関表

	Y_1	Y_2	Y_3	...	...	Y_j	...	...	Y_N	周辺分布
X_1	n_{11}	n_{12}	n_{13}	...	...	n_{1j}	...	...	n_{1N}	$n_{1\cdot}$
X_2	n_{21}	n_{22}	n_{23}	...	...	n_{2j}	...	...	n_{2N}	$n_{2\cdot}$
X_3	n_{31}	n_{32}	n_{33}	...	...	n_{3j}	...	...	n_{3N}	$n_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$			$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$			$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	n_{i3}	...	...	n_{ij}	...	...	n_{iN}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$			$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$			$\vdots$	$\vdots$
X_M	n_{M1}	n_{M2}	n_{M3}	...	...	n_{Mj}	...	...	n_{MN}	$n_{M\cdot}$
周辺分布	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	...	...	$n_{\cdot j}$	...	...	$n_{\cdot N}$	n

表 2 - 1 において、変量 X の度数分布は相関表の度数 n_{ij} を横に集計した度数が示す周辺分布 $n_{i\cdot}$ であり、変量 Y の度数分布は相関表の度数 n_{ij} を縦に集計した度数が示す周辺分布 $n_{\cdot j}$ である。また表右下には周辺分布の度数を合計した総度数が n として示されている。

$$n_{i\cdot} = \sum_{j=1}^N n_{ij} \quad (i = 1, 2, 3, \dots, M)$$

$$n_{\cdot j} = \sum_{i=1}^M n_{ij} \quad (j = 1, 2, 3, \dots, N)$$

$$\sum_{i=1}^M n_{i\cdot} = \sum_{j=1}^N n_{\cdot j} = \sum_{i=1}^M \sum_{j=1}^N n_{ij} = n$$

表 2 - 1 の 2 次元データについて、2 変量 X , Y の平均 $\bar{X}$, $\bar{Y}$, 分散 s_X^2 , s_Y^2 , 共分散 s_{XY} , 相関係数 r_{XY} は、度数 n_{ij} を用いて次のように定義することができる。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^M X_i n_{i\cdot} = \frac{1}{n} \sum_{i=1}^M X_i \cdot \sum_{j=1}^N n_{ij}$$

$$s_X^2 = \frac{1}{n} \sum_{i=1}^M (X_i - \bar{X})^2 n_{i\cdot} = \frac{1}{n} \sum_{i=1}^M X_i^2 n_{i\cdot} - \bar{X}^2$$

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^N Y_j n_{\cdot j} = \frac{1}{n} \sum_{j=1}^N Y_j \cdot \sum_{i=1}^M n_{ij}$$

$$s_Y^2 = \frac{1}{n} \sum_{j=1}^N (Y_j - \bar{Y})^2 n_{\cdot j} = \frac{1}{n} \sum_{j=1}^N Y_j^2 n_{\cdot j} - \bar{Y}^2$$

$$s_{XY} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N (X_i - \bar{X})(Y_j - \bar{Y}) n_{ij} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} - \bar{X} \cdot \bar{Y}$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

表計算では、 $\sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij}$ を計算するための表を 1 枚追加すれば、容易に以上の計算式に従って計算を進めることができる。関数電卓による場合でも、

やや煩雑ではあるものの、無理なく相関係数を求めることができると思われるが、石井 [1]、児玉 [4]、小寺 [5]、小針 [6] には度数分布表に基づく場合のように、一次変換に伴う性質を利用して計算の負担を減らす方法が示されている。変量 X , Y について、各階級の階級幅がそれぞれ d_X , d_Y に統一され、階級値が階級下限と階級上限の真ん中とされていることを前提として、最頻値 $\tilde{X}$, $\tilde{Y}$ により U_i , V_j を以下のように定義する。

$$U_i = \frac{x_i - \tilde{X}}{d_X}, \quad V_j = \frac{y_j - \tilde{Y}}{d_Y}$$

$$\therefore X_i = \tilde{X} + d_X \cdot U_i, \quad Y_j = \tilde{Y} + d_Y \cdot V_j$$

よって、変量 X , Y の平均は、変量 U , V の平均を用いて次のように表せる。

$$\bar{X} = \tilde{X} + d_X \cdot \bar{U}$$

$$\bar{Y} = \tilde{Y} + d_Y \cdot \bar{V}$$

これにより、変量 X , Y の分散 s_X^2 , s_Y^2 , 共分散 s_{XY} , 相関係数 r_{XY} は、変量 U , V のそれら s_U^2 , s_V^2 , s_{UV} , r_{UV} と以下のような関係にあることがわかる。特に相関係数については、変換前と変換後の値が同じになり、計算の負担がかなり軽減されるものと思われる。

$$s_X^2 = \frac{1}{n} \sum_{i=1}^M (X_i - \bar{X})^2 n_i = \frac{1}{n} \sum_{i=1}^M \{d_X (U_i - \bar{U})\}^2 n_i = d_X^2 \cdot \frac{1}{n} \sum_{i=1}^M (U_i - \bar{U})^2 = d_X^2 \cdot s_U^2$$

$$s_Y^2 = \frac{1}{n} \sum_{j=1}^N (Y_j - \bar{Y})^2 n_j = \frac{1}{n} \sum_{j=1}^N \{d_Y (V_j - \bar{V})\}^2 n_j = d_Y^2 \cdot \frac{1}{n} \sum_{j=1}^N (V_j - \bar{V})^2 = d_Y^2 \cdot s_V^2$$

$$s_{XY} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N (X_i - \bar{X})(Y_j - \bar{Y}) n_{ij} = d_X d_Y \cdot \frac{1}{n} \sum_{i=1}^M (U_i - \bar{U})(V_j - \bar{V}) = d_X d_Y \cdot s_{UV}$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{d_X d_Y \cdot s_{UV}}{d_X d_Y \cdot s_U s_V} = \frac{s_{UV}}{s_U s_V} = r_{UV}$$

表 2 - 2 は石井 [1], 児玉 [4], 小寺 [5], 小針 [6] に示されている計算表を要約したものである。本表の集計欄を完成させた上で以下の計算を行えば, 変量 U , V の平均, 分散, 共分散, 相関係数を求めることができる。

$$\bar{U} = \frac{1}{n} \sum_{i=1}^M U_i n_{i\cdot} = \frac{1}{n} \sum_{i=1}^M U_i \sum_{j=1}^N n_{ij} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N U_i n_{ij}$$

$$s_U^2 = \frac{1}{n} \sum_{i=1}^M U_i^2 n_{i\cdot} - \bar{U}^2$$

$$\bar{V} = \frac{1}{n} \sum_{j=1}^N V_j n_{\cdot j} = \frac{1}{n} \sum_{j=1}^N V_j \sum_{i=1}^M n_{ij} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N V_j n_{ij}$$

$$s_V^2 = \frac{1}{n} \sum_{j=1}^N V_j^2 n_{\cdot j} - \bar{V}^2$$

$$s_{UV} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N U_i V_j n_{ij} - \bar{U} \bar{V} = \frac{1}{n} \sum_{i=1}^M U_i \sum_{j=1}^N V_j n_{ij} - \bar{U} \bar{V} = \frac{1}{n} \sum_{j=1}^N V_j \sum_{i=1}^M U_i n_{ij} - \bar{U} \bar{V}$$

$$r_{UV} = \frac{s_{UV}}{s_U s_V}$$

表 2 - 2 は一見わかりにくく, かなり煩雑な印象を受けるが, 表 2 - 2 を完成させると, 右下の L 字形の領域, 9 個のセルから構成される集計欄の計算値から, 上記の関係に基づいて計算することにより, 平均, 分散, 共分散, 相関係数を求めることができる。集計欄の要素を確認すると, 右に倒れた L 字型の領域の角の欄は総度数 n , その右に変量 U の総和 $\sum_{i=1}^M U_i n_{i\cdot}$, 変量 U^2 の総計 $\sum_{i=1}^M U_i^2 n_{i\cdot}$, その右には $\sum_{i=1}^M \sum_{j=1}^N V_j n_{ij}$ と続くが, これは変量 V の総和 $\sum_{j=1}^N V_j n_{\cdot j}$ に等しい。総度数 n の欄の下には変量 V の総和 $\sum_{j=1}^N V_j n_{\cdot j}$, 変量 V^2 の総和 $\sum_{j=1}^N V_j^2 n_{\cdot j}$, その下には $\sum_{j=1}^N \sum_{i=1}^M U_i n_{ij}$ と続くが, これは変量 U のとる値の総和 $\sum_{i=1}^M U_i n_{i\cdot}$ に等しい。また, 集計欄右端の欄 $\sum_{i=1}^M U_i \sum_{j=1}^N V_j n_{ij}$ と最下段の欄 $\sum_{j=1}^N V_j \sum_{i=1}^M U_i n_{ij}$ は, いずれも 2 変量 U , V の積の総和 $\sum_{i=1}^M \sum_{j=1}^N U_i V_j n_{ij}$ に等しく, 同じ値を取ることになる。

表2-2 一次変換を用いた相関係数の計算表

	V_1	V_2	V_3	...	V_j	...	V_M	周辺分布				
U_1	m_{11}	m_{12}	m_{13}	...	m_{1j}	...	m_{1M}	$n_{1\cdot}$	$U_1 m_{1\cdot}$	$U_1^2 m_{1\cdot}$	$\sum_{j=1}^N V_j m_{1j}$	$U_1 \sum_{j=1}^N V_j m_{1j}$
U_2	m_{21}	m_{22}	m_{23}	...	m_{2j}	...	m_{2M}	$n_{2\cdot}$	$U_2 m_{2\cdot}$	$U_2^2 m_{2\cdot}$	$\sum_{j=1}^N V_j m_{2j}$	$U_2 \sum_{j=1}^N V_j m_{2j}$
U_3	m_{31}	m_{32}	m_{33}	...	m_{3j}	...	m_{3M}	$n_{3\cdot}$	$U_3 m_{3\cdot}$	$U_3^2 m_{3\cdot}$	$\sum_{j=1}^N V_j m_{3j}$	$U_3 \sum_{j=1}^N V_j m_{3j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
U_i	m_{i1}	m_{i2}	m_{i3}	...	m_{ij}	...	m_{iM}	$n_{i\cdot}$	$U_i m_{i\cdot}$	$U_i^2 m_{i\cdot}$	$\sum_{j=1}^N V_j m_{ij}$	$U_i \sum_{j=1}^N V_j m_{ij}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
U_M	m_{M1}	m_{M2}	m_{M3}	...	m_{Mj}	...	m_{MN}	$n_{M\cdot}$	$U_M m_{M\cdot}$	$U_M^2 m_{M\cdot}$	$\sum_{j=1}^N V_j m_{Mj}$	$U_M \sum_{j=1}^N V_j m_{Mj}$
周辺分布	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	...	$n_{\cdot j}$	...	$n_{\cdot N}$	n	$\sum_{i=1}^M U_i m_{i\cdot}$	$\sum_{i=1}^M U_i^2 m_{i\cdot}$	$\sum_{i=1}^M \sum_{j=1}^N V_j m_{ij}$	$\sum_{i=1}^M U_i \sum_{j=1}^N V_j m_{ij}$
	$V_1 n_{\cdot 1}$	$V_2 n_{\cdot 2}$	$V_3 n_{\cdot 3}$	...	$V_j n_{\cdot j}$	...	$V_M n_{\cdot N}$	$\sum_{j=1}^N V_j n_{\cdot j}$				
	$V_1^2 n_{\cdot 1}$	$V_2^2 n_{\cdot 2}$	$V_3^2 n_{\cdot 3}$	...	$V_j^2 n_{\cdot j}$	...	$V_M^2 n_{\cdot N}$	$\sum_{j=1}^N V_j^2 n_{\cdot j}$				
	$\sum_{i=1}^M U_i m_{i1}$	$\sum_{i=1}^M U_i m_{i2}$	$\sum_{i=1}^M U_i m_{i3}$	...	$\sum_{i=1}^M U_i m_{ij}$	...	$\sum_{i=1}^M U_i m_{iN}$	$\sum_{j=1}^N \sum_{i=1}^M U_i m_{ij}$				
	$V_1 \sum_{i=1}^M U_i m_{i1}$	$V_2 \sum_{i=1}^M U_i m_{i2}$	$V_3 \sum_{i=1}^M U_i m_{i3}$	...	$V_j \sum_{i=1}^M U_i m_{ij}$	...	$V_M \sum_{i=1}^M U_i m_{iN}$	$\sum_{j=1}^N V_j \sum_{i=1}^M U_i m_{ij}$				

2.2 相関表の具体例における計算結果

表 2-3 はイギリスの遺伝学者ガルトン Sir Francis Galton (1822-1911) に
よってまとめられた相関表で、松原 [9] によると相関関係という考え方を初
めて示した歴史的に有名な表であるとのことである。

表 2-3 のデータについて平均、分散、共分散、相関係数を求めると次のと
おりである。上腕長の平均は18.12インチで、1 インチは約2.54cm なので、
46.02cm ということになり、分散は0.6708となる。身長は平均は67.63インチ
で171.78cm となり、分散は6.7496である。また、上腕長と身長は共分散は
1.5974、これより相関係数は0.7507となり、両変量の間には正の相関が認めら
れる。

表 2-3 からわかるように変量 X_i の最頻値は18.25、階級幅 d_x は0.5、変量
 Y_j の最頻値は67.5、階級幅 d_y は1である。変量 X_i, Y_j を次のように変量 U_i, V_j
に一次変換した上で、表の 2-3 の数値を示したものが表 2-4 である。

$$U_i = \frac{X_i - 18.25}{0.5}$$

$$V_j = \frac{Y_j - 67.5}{1}$$

$$\therefore X_i = 18.25 + 0.5 U_i, Y_j = 67.5 + V_j$$

$$\therefore \bar{X} = 18.25 + 0.5 \bar{U}, \bar{Y} = 67.5 + \bar{V}$$

$$\therefore s_X^2 = 0.25 s_U^2, s_Y^2 = s_V^2, s_{XY} = 0.5 s_{UV}, r_{XY} = r_{UV}$$

表 2-4 を構成する数値は、この一次変換によって整数化されている。計算に
際してどの程度の負担軽減が図られるのかを、共分散を求める上で不可欠な変量
 X, Y の積の総和 $\sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij}$ と、変量 U, V の積の総和 $\sum_{i=1}^M \sum_{j=1}^N U_i V_j n_{ij}$
の数値で比較すると、前者は420,858.88であり、後者は1,084.00となる。前者
は関数電卓でないと計算が難しいのに対して、後者は算盤的な電卓でも対処で

表 2-3 ゴルフトンによる身長と上腕の長さの調査データ (インチ)

身長 : Y 上腕長 : X		身長 : Y																			周辺分布	
		階級下限	59.00	60.00	61.00	62.00	63.00	64.00	65.00	66.00	67.00	68.00	69.00	70.00	71.00	72.00	73.00	74.00				
階級下限	階級上限	階級値	60.00	61.00	62.00	63.00	64.00	65.00	66.00	67.00	68.00	69.00	70.00	71.00	72.00	73.00	74.00	75.00				
15.50	16.00	15.75		1														74.50				
16.00	16.50	16.25	1	1	1	3	2											8				
16.50	17.00	16.75			1		6	5	4	1	1	1	1					20				
17.00	17.50	17.25			1	3	6	11	10	7	7	3	1					49				
17.50	18.00	17.75				2	1	2	12	18	15	7	2	1	1			61				
18.00	18.50	18.25						3	8	15	28	14	25	5	2	2		102				
18.50	19.00	18.75					1		2	6	8	7	15	13	2	1		55				
19.00	19.50	19.25									2	4	6	11	8	4	2	1	38			
19.50	20.00	19.75										2			1	2		5				
20.00	20.50	20.25														1		1				
20.50	21.00	20.75													1	1		2				
21.00	21.50	21.25																1	1			
		周辺分布	1	2	3	8	16	21	36	47	61	38	50	30	15	9	4	2	343			

表2-4 一次変換した変量 U と V の相関表

	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	n_i	$U_i n_i$	$U_i^2 n_i$	$\sum_{j=1}^N V_j n_{ij}$	$U_i \sum_{j=1}^N V_j n_{ij}$
-5		1															1	-5	25	-7	35
-4	1	1	1	3	2												8	-32	128	-44	176
-3			1		6	5	4	1	1	1	1						20	-60	180	-51	153
-2			1	3	6	11	10	7	7	3	1						49	-98	196	-100	200
-1				2	1	2	12	18	15	7	2	1	1				61	-61	61	-44	44
0						3	8	15	28	14	25	5	2	2			102	0	0	57	0
1					1		2	6	8	7	15	13	2	1			55	55	55	75	75
2									2	4	6	11	8	4	2	1	38	76	152	120	240
3										2			1		2		5	15	45	18	54
4														1			1	4	16	5	20
5													1	1			2	10	50	9	45
6																1	1	6	36	7	42
n_j	1	2	3	8	16	21	36	47	61	38	50	30	15	9	4	2	343	-90	944	45	1084
$V_j n_j$	-8	-14	-18	-40	-64	-63	-72	-47	0	38	100	90	60	45	24	14	45				
$V_j^2 n_j$	64	98	108	200	256	189	144	47	0	38	200	270	240	225	144	98	2321				
$\sum_{j=1}^M U_i n_{ij}$	-4	-9	-9	-20	-38	-39	-42	-29	-20	5	20	34	25	18	10	8	-90				
$V_j \sum_{i=1}^M U_i n_{ij}$	32	63	54	100	152	117	84	29	0	5	40	102	100	90	60	56	1084				

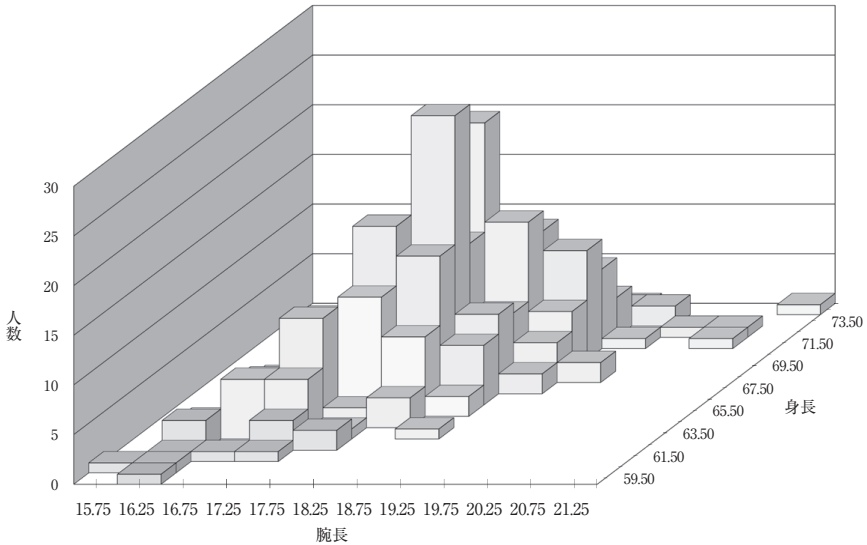


図2-1 ゴルトンの身長と上腕の長さ（インチ）

きそうである。因みに共分散 s_{XY} は1.5974, 共分散 s_{UV} の計算値3.1948の0.5倍となっている。

表2-5は人口動態調査人口動態統計から2022年の男女初婚年齢に関する相関表を引用したものである。夫については17歳から79歳まで1歳刻みで、最後の階級は80歳以上と表示されているので、17歳以上18歳未満、18歳以上19歳未満、…、78歳以上79歳未満の階級、最後の階級はとりあえず80歳以上90歳未満と表示した。一方、妻については15歳から79歳まで1歳刻みで、最後の階級は80歳以上と表示されているので、夫の場合と同様に15歳以上16歳未満、16歳以上17歳未満、…、78歳以上79歳未満、最後の階級は80歳以上90歳未満と表示した。

表2-5には紙幅の都合で表示できなかった42歳以上43歳未満、43歳以上44歳未満、…、79歳以上80歳未満の階級を含む全データについて、平均、分散、共分散、相関係数を求めると、次のとおりである。妻の年齢の平均は31.70歳、分散は71.9709、夫の年齢の平均は33.75歳、分散は94.3253、共分散は68.9705、

相関係数は0.8371で、両者に当然予想される高い相関関係があることを示している。

なお、図2－2は表2－5に基づく2変量ヒストグラムであり、縦軸は夫の初婚年齢、横軸は妻の初婚年齢を示している。

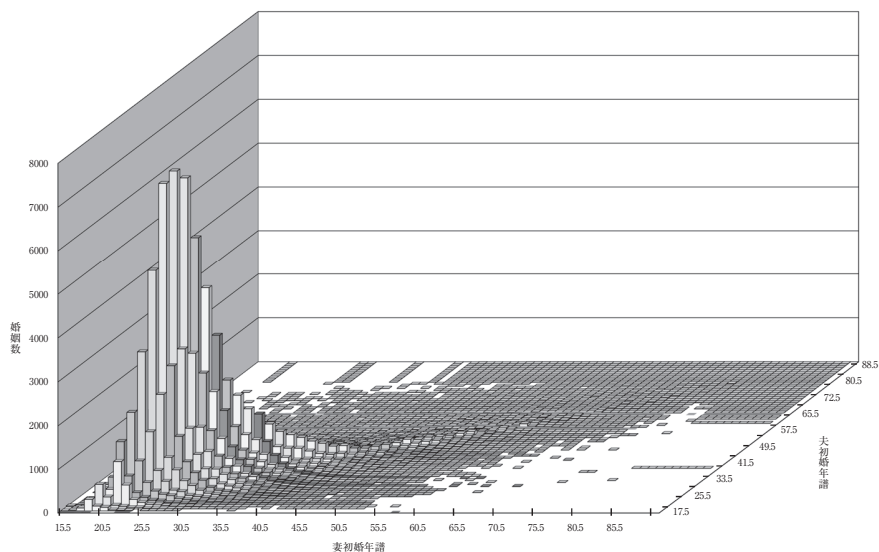


図2－2 初婚夫と初婚妻の年齢分布

e-Stat 統計でみる日本 <https://www.e-stat.go.jp/index.php/dbview?sid=0003411856>

[illegible]

第3章 相関表データに基づく単回帰の計算

相関表として与えられる2次元データに対して、線形回帰を行う場合、どのような計算を行うべきか、武藤 [13] は「同一の階級に属するデータが全て階級値に等しい観測値を持つとして2次元データ系列に変換し、度数分布表化されていないデータと同様の処理を行えば良い」と説明している。総度数が小さければ、データの変換も容易で、さらに統計処理ソフトを前提にすれば、容易に回帰計算を行えることから、このような説明になったものと思われる。しかし、総度数が大きい場合には、このデータ形式の変換もかなり煩雑な作業となるのではなかろうか。相関表から直接計算できる方が良いように思われる。本章では、まず第1節で度数分布表のデータのまま、単回帰の計算がどのように表せるか確認する。第2節では、本論文で取り上げた2つの相関表について、回帰係数を求めることにする。

3.1 回帰モデル

相関表においては、説明変数 X_i と被説明変数 Y_j の階級値に対して、度数 n_{ij} ないし相対度数 f_{ij} が与えられている。相対度数は度数 n_{ij} を総度数 n で割った値であり、計算処理においては総度数で割る処理が記述されないだけなので、とりあえず本節では表2-1に示した相関表が与えられているものとする。

相関表のデータについて線形回帰を行う場合、相関表の柁目ごとに残差を考える必要がある。残差を e_{ij} と表すと、被説明変数 Y_j は説明変数 X_i の線形回帰式 $a + bX_i$ と残差 e_{ij} を用いて、次のように表せる。

$$Y_j = a + bX_i + e_{ij} \quad \text{但し, } i = 1, 2, 3, \dots, M, \quad j = 1, 2, 3, \dots, N$$

相関表データの場合では、線形回帰の定式化において、説明変数の添え字と被説明変数の添え字が異なり、残差の添え字は説明変数の添え字と被説明変数の添え字を反映するものとなっている。最小二乗法を適用して線形回帰を行うとき、残差平方和は相関表の柁目ごとの残差 e_{ij} の2乗に度数 n_{ij} をかけ、全て

の添え字について集計して次のように表せる。

$$\sum_{i=1}^M \sum_{j=1}^N e_{ij}^2 n_{ij}$$

残差 e_{ij} が実績値 Y_j と推定値 $a + bX_i$ との差であることを利用して変形すると、この残差平方和は原データの場合と同様、 a 、 b に関する 2 次関数となり、残差平方和を最小にするような回帰直線の定数項 a と傾き b の値を求めることができる。残差平方和は説明変数、被説明変数で書き改めると、 a 、 b の関数とみなすことができるので $S(a, b)$ と表し、これを a 、 b それぞれで偏微分して 0 と置いた関係式は以下のとおりである。

$$S(a, b) = \sum_{i=1}^M \sum_{j=1}^N e_{ij}^2 n_{ij} = \sum_{i=1}^M \sum_{j=1}^N (Y_j - a - bX_i)^2 n_{ij}$$

$$\therefore \frac{\partial S}{\partial a} = -2 \sum_{i=1}^M \sum_{j=1}^N (Y_j - a - bX_i) n_{ij} = 0$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^M \sum_{j=1}^N X_i (Y_j - a - bX_i) n_{ij} = 0$$

$$\therefore a \sum_{i=1}^M \sum_{j=1}^N n_{ij} + b \sum_{i=1}^M X_i n_{i\cdot} = \sum_{j=1}^N Y_j n_{\cdot j}$$

$$a \sum_{i=1}^M X_i n_{i\cdot} + b \sum_{i=1}^M X_i^2 n_{i\cdot} = \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij}$$

以上の両式を総度数 n で割ると次のように表せる。

$$a + b\bar{X} = \bar{Y}$$

$$a\bar{X} + b\frac{1}{n} \sum_{i=1}^M X_i^2 n_{i\cdot} = \frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij}$$

第 1 式の両辺に $\bar{X}$ をかけた上で第 2 式から辺々差し引くことにより b が求められ、この結果と第 1 式より a が求められる。得られた結果は原データによ

る単回帰の場合と全く同様である。

$$b = \frac{\frac{1}{n} \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} - \bar{X} \bar{Y}}{\frac{1}{n} \sum_{i=1}^M X_i^2 n_{i\cdot} - \bar{X}^2} = \frac{s_{XY}}{s_X^2}$$

$$a = \bar{Y} - b\bar{X}$$

平方完成によっても同様の結果を求めることができる。まず残差平方和 $\sum_{i=1}^M \sum_{j=1}^N e_{ij}^2 n_{ij}$ を a について整理すると以下のとおりである。

$$\begin{aligned} \sum_{i=1}^M \sum_{j=1}^N e_{ij}^2 n_{ij} &= \sum_{i=1}^M \sum_{j=1}^N (Y_j - a - bX_i)^2 n_{ij} \\ &= \sum_{i=1}^M \sum_{j=1}^N (Y_j^2 + a^2 + b^2 X_i^2 - 2aY_j + 2abX_i - 2bX_i Y_j) n_{ij} \\ &= a^2 \sum_{i=1}^M \sum_{j=1}^N n_{ij} - 2a \sum_{i=1}^M \sum_{j=1}^N (Y_j - bX_i) n_{ij} \\ &\quad + \sum_{i=1}^M \sum_{j=1}^N Y_j^2 n_{ij} + b^2 \sum_{i=1}^M \sum_{j=1}^N X_i^2 n_{ij} - 2b \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} \\ &= na^2 - 2a \left(\sum_{j=1}^N Y_j \sum_{i=1}^M n_{ij} - b \sum_{i=1}^M X_i \sum_{j=1}^N n_{ij} \right) \\ &\quad + \sum_{j=1}^N Y_j^2 \sum_{i=1}^M n_{ij} + b^2 \sum_{i=1}^M X_i^2 \sum_{j=1}^N n_{ij} - 2b \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} \\ &= na^2 - 2a \left(\sum_{j=1}^N Y_j n_{\cdot j} - b \sum_{i=1}^M X_i n_{i\cdot} \right) + \sum_{j=1}^N Y_j^2 n_{\cdot j} + b^2 \sum_{i=1}^M X_i^2 n_{i\cdot} - 2b \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} \\ &= na^2 - 2na(\bar{Y} - b\bar{X}) + \sum_{j=1}^N Y_j^2 n_{\cdot j} + b^2 \sum_{i=1}^M X_i^2 n_{i\cdot} - 2b \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} \end{aligned}$$

a について平方完成すれば次のとおり。

$$\begin{aligned}
 \sum_{i=1}^M \sum_{j=1}^N e_{ij}^2 n_{ij} &= n\{a^2 - 2(\bar{Y} - b\bar{X})a\} + \sum_{j=1}^N Y_j^2 n_{.j} + b^2 \sum_{i=1}^M X_i^2 n_{i.} - 2b \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} \\
 &= n\{a - (\bar{Y} - b\bar{X})\}^2 - n(\bar{Y} - b\bar{X})^2 + \sum_{j=1}^N Y_j^2 n_{.j} + b^2 \sum_{i=1}^M X_i^2 n_{i.} - 2b \sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} \\
 &= n\{a - (\bar{Y} - b\bar{X})\}^2 + (\sum_{i=1}^M X_i^2 n_{i.} - n\bar{X}^2)b^2 - 2(\sum_{i=1}^M \sum_{j=1}^N X_i Y_j n_{ij} - n\bar{X}\bar{Y})b \\
 &\quad + \sum_{j=1}^N Y_j^2 n_{.j} - n\bar{Y}^2 \\
 &= n\{a - (\bar{Y} - b\bar{X})\}^2 + n\{s_X^2 b^2 - 2s_{XY}b + s_Y^2\}
 \end{aligned}$$

従って、残差平方和は a が $\bar{Y} - b\bar{X}$ に等しいとき最小となり、その最小値は $n\{s_X^2 b^2 - 2s_{XY}b + s_Y^2\}$ と表せる。相関表のデータを前提にする場合も、線形回帰で得られる直線が説明変数と被説明変数の平均に対応する点 $(\bar{X}, \bar{Y})$ を通るという性質が成立している。

$$a = \bar{Y} - b\bar{X} \quad \therefore \bar{Y} = a + b\bar{X}$$

また、以上の関係が成立する時の残差平方和は、回帰直線の傾き b の 2 次関数となっており、 b について平方完成することにより、この残差平方和を最小化する b の値が求められる。

$$\begin{aligned}
 \sum_{i=1}^M \sum_{j=1}^N e_{ij}^2 n_{ij} &= n\{s_X^2 b^2 - 2s_{XY}b + s_Y^2\} \\
 &= n\left\{s_X^2 \left(b^2 - 2\frac{s_{XY}}{s_X^2}b\right) + s_Y^2\right\} = ns_X^2 \left(b - \frac{s_{XY}}{s_X^2}\right)^2 + ns_Y^2 - n\frac{s_{XY}^2}{s_X^2}
 \end{aligned}$$

こうして二次元データへの変換を行わず相関表のデータから直接、最小二乗法による回帰直線が平均を座標とする点 $(\bar{X}, \bar{Y})$ を通り、傾きは共分散 s_{XY} を説明変数の分散 s_X^2 で割ったものに等しい事を示せる。被説明変数 Y_j を説明変数

X_i に回帰することにより、残差平方和の最小値は平均 $\bar{Y}$ を推定値とした場合の残差平方和 ns_Y^2 と比べて $n \frac{s_{XY}^2}{s_X^2}$ だけ小さくなっており、これが説明変数 X_i による説明力向上の程度を示している。従って、線形回帰の当てはまりの良さを示す決定係数 R^2 は、原データの場合と全く同様に相関係数 r_{XY} の2乗に等しい。

$$R^2 = \frac{n \frac{s_{XY}^2}{s_X^2}}{ns_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = r_{XY}^2$$

単回帰における決定係数は、以下のように変形することもできる。

$$R^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2} = \frac{s_{XY}}{s_X^2} \cdot \frac{s_{XY}}{s_Y^2}$$

右辺第1項は変数 Y を変数 X で説明した回帰直線の傾き b であり、右辺第2項は変数 X を変数 Y で説明した回帰直線の傾き b' であることに注目すると、決定係数は次のように表せる。

$$R^2 = b \cdot b'$$

相関表のデータに基づく単回帰から得られる結果は、定式化に際して残差の添え字が2次元になり、やや煩雑に見えるものの、原データから得られるものと全く同様に説明できる。

3.2 数値例

本節では、第1節で確認した計算方法を、2つの具体的な相関表に適用して、単回帰の計算結果がどのように求められるかを確認することにする。

3.2.1 ゴルトンの相関表

第2章で松原 [9] から引用したゴルトンの相関表を再び用いて、相関表データに基づく回帰計算の結果を示すと次のとおりである。この相関表における総度数は343であり、相関表からのデータの変換は単純な処理とは言え、特

注のプログラムなしでは難しいように思われる。第2章において既に表示した平均、分散、共分散、相関係数の計算結果は以下のとおりである。

$$\bar{X} \cong 18.10, \quad s_X^2 \cong 0.6877, \quad \bar{Y} \cong 67.56, \quad s_Y^2 \cong 7.0254, \quad s_{XY} \cong 1.6739, \quad r_{XY} = 0.7615$$

この計算結果に基づいて、身長 Y_j を上腕の長さ X_i で説明する回帰式は次のとおりである。

$$Y_j = 23.505028 \cdots + 2.433953 \cdots X_i + e_{ij}$$

また、逆に上腕の長さ X_i を身長 Y_j で説明する回帰式は次のとおりである。

$$X_i = 2.002697 \cdots + 0.238263 \cdots Y_j + e'_{ij}$$

相関係数の2乗は $0.5799921 \cdots$ となるが、これは以上の2つの回帰式の決定係数ということになる。実際に両式の説明変数の係数をかけ合わせた結果 $2.4340953 \cdots \times 0.23823 \cdots$ もこれに一致している。

3.2.2 男女の初婚年齢

第2章で示した人口動態統計から得られる初婚年齢の組み合わせについて、妻の初婚年齢 X と夫の初婚年齢 Y について、第2章において既に表示した平均、分散、共分散、相関係数の計算結果を再掲すると、次のとおりである。

$$\bar{X} \cong 31.70, \quad s_X^2 \cong 71.9709, \quad \bar{Y} \cong 33.75, \quad s_Y^2 \cong 94.3253, \quad s_{XY} \cong 68.9705, \quad r_{XY} = 0.8371$$

夫の年齢 Y_j を妻の年齢 X_i で説明する回帰式は次のとおりである。

$$Y_j = 3.371546 \cdots + 0.958312 \cdots X_i + e_{ij}$$

また、逆に妻の年齢 X_i を夫の年齢 Y_j で説明する回帰式は次のとおりである。

$$X_i = 7.022449 \cdots + 0.731979 \cdots Y_j + e'_{ij}$$

相関係数の2乗は0.700715…となるが、これが回帰式の決定係数ということになる。両回帰式の説明変数の係数を実際にかけ合わせた結果 $0.958312 \cdots \times 0.731979 \cdots$ もこれに一致している。

結 語

本論文は、統計学を学び始めて最初に学ぶ一変量の度数分布表、二変量の度数分布表である相関表について、その取扱いを再確認することを目的としている。最も基本的な話題でありながら、その作成方法と作成結果、原データと度数分布表の関係などについては、じっくりしない点も若干残されがちである。

度数分布表の作成に際しては、同じ原データを対象にしながら、各分析者の作成結果にある程度の差が生じ得るからである。度数分布表の作成は原データの持つ情報を減らすことによって、木を見て森を見ずという状況を回避する1つの手段として説明されるものの、原データから直接読み取れる平均や分散に関する情報と、あらかじめ作成された度数分布表から読み取れるそれらの近似的情報にどのような関係があるのかなど、気にし始めれば切りがないようにも思われる。

本論文ではこのような問題意識を念頭に、度数分布表と相関表の作成、それらを図示したヒストグラム、度数分布表からの平均、分散、相関表からの相関係数、回帰係数の計算方法について検討した。

第1章では、まず第1節でデータの総度数と適切な階級数の目安を示すスタージェスの公式について検討した。その導出には歪みのない二項分布が前提とされているというヒンドマン [16] の指摘を手掛かりに、同公式が階級幅の統一された度数分布表の作成に際して必要とされる階級数の上限の値を示していることを確認した。本公式を厳密に捉えて否定する見解もあるが、従来どお

り歪みのない、あるいは歪みの小さい分布に関する目安として利用する限り、十分に利用価値があるものとする。第2節においては、平均の計算について、度数分布表に基づく平均が原データの平均に一致するのは、各階級を構成するデータの平均が階級値とされる場合であることを確認した上で、第3節ではそのような度数分布表については、分散が階級内の分散の相対度数を重みとする加重平均と、度数分布表に基づく分散の総和と表せることを確めた。階級内データの平均を階級値とする場合に限定したものではあるが、度数分布表に基づく通常の分散の定義式は前者を無視したものとなり、このことから通常の定義式の分散は、原データに基づく分散以下の値となることを明らかに示すことができた。第4節では階級値が階級下限と階級上限の値の真ん中の値とする取り扱いについて、一般的には階級内のデータに散らばりはないと想定したものとされるが、逆に階級内に均一に散らばっているものとする場合について、平均と分散がどのように表せるか検討した。第5節では教科書に示されている原データを例として取り上げて度数分布表を作成し、前節までに取り上げた指標を用いて、度数分布表に基づく分散が、原データに基づく分散とどの程度の差があるか確認した。

第2章では相関表を取り上げ、まず第1節で教科書のいくつかで取り上げられている平均、分散、共分散、相関係数の計算方法、特にデータの一次変換を利用した計算方法についてまとめた。第2節では、具体的な数値例として、ゴルトンによる腕の長さと言長の間の相関関係を示す相関表、人口動態統計から男女の初婚年齢に関する相関表について、平均、分散、共分散、相関係数の計算結果を示すと共に、2変量ヒストグラムがどのように表せるか示した。階級幅が統一されている相関表の場合、一次変換による計算の簡素化を行えるが、電卓による手計算を前提にして、ゴルトンの相関表に基づく表2-4を見る限り、その効果はかなり大きいことがわかる。

第3章では、まず第1節で相関表のデータから単回帰の計算を直接行えることを再確認した。そこで得られる最小二乗推定量は、相関表のデータを2次元のデータ系列に置き換えて計算する場合と全く同様になる。管見する限り単回帰については、相関表から普通の2次元データの系列を求めた上で、通常の回

帰計算を行えば良いという武藤 [13] の説明しか見当たらないが、例として取り上げたゴルトンの相関表の総度数は343、初婚男女の年齢の相関表の総度数は34万4,823もあり、この方法では事前に2次元データの系列を作成するためのプログラムが必要になりそうだ。第2節では、数値例として前章で取り上げた2枚の相関表について、前節で求めた推定量に基づく計算法により、単回帰の回帰係数、決定係数を求め、その計算結果を示した。

参考文献

- [1] 石井満『経済・経営系の統計学』森北出版, 1991年5月
- [2] 猪間驥一『統計図表の見方画き方使い方』東洋経済新報社, 1971年6月
- [3] 岡田泰栄『統計学概論』共立出版, 2021年3月
- [4] 児玉正憲『基本数理統計学』牧野書店, 1992年10月
- [5] 小寺平治『新統計入門』裳華房, 1996年10月
- [6] 小針現宏『確率・統計入門』岩波書店, 1974年12月
- [7] 斎木清治「ヒストグラムは怖い」『雑感/高校数学の問題を作る－工夫・コツとデータ－』, No.108, 2013年10月, <http://www10.plala.or.jp/mondai/column/hist.pdf>
- [8] 東京大学教養学部統計学教室編『統計学入門』東京大学出版会, 2023年2月
- [9] 松原望『統計学100のキーワード』弘文社, 2005年4月
- [10] 宮川公男『基本統計学』第5版, 有斐閣, 2022年4月
- [11] 蓑谷千風彦『これからはじめる統計学』, 東京図書, 2009年10月
- [12] 森田優三, 久次智雄『新統計概論』改訂版, 日本評論社, 2004年4月
- [13] 武藤真介『統計解析ハンドブック』, 1995年6月
- [14] Alan Stuart & J.Keith Ord, “Kendall’s Advanced Theory of Statistics”, 5th ed. Vol.1, 1987
- [15] Herbert A. Sturges, ‘The Choice of a Class Interval’, “Journal of the American Statistical Association”, Mar. 1926, Vol.21, No.153, pp.65-66
- [16] Rob J Hyndman, ‘The problem with Sturges’ rule for constructing histograms’, July 1995, <https://robjhyndman.com/papers/sturges.pdf>