市民のための〈基礎から学ぶ〉統計学

平均・分散から検定力分析まで、
社会科学に最低限必要な統計学を解説してみた本
（疫学もあるよ）

後藤 和智（後藤和智事務所OffLine）

STATISTICS

市民のための〈基礎から学ぶ〉統計学

後藤和智

平成 22 年 10 月 24 日（サンシャインクリエイション 49）

0.1 まえがき

15 冊目の同人誌にして、初の総集編となる後藤和智です。本書は、これまで当サークルで出してきた同人誌、「市民のための統計学」シリーズの 3 冊…「度数分布・相関・回帰分析編」（サンシャインクリエイション 45)、「確率分布・検定編」（サンシャインクリエイション 46)、そして、偏相関係数と適合度の検定を取り扱った「Extend1」（サンシャインクリエイション 47）をまとめて再構成し、さらにいくつかの章を増補した総集編です。

そもそもこのシリーズは、『市民のための統計解析 [基礎編 Extend+ 多変量解析編]』（コミックマーケット 75。現在は『新・市民のための統計学』としてコミックマーケット 78 にて再版）に対する反響を見て、基礎から「社会科学のための統計学」を学ぶことのできる本が必要だ、と思い、始めたものです。そのため、最初の巻では若干高度な数式が出てくる統計的検定よりも、基礎的な統計量や回帰分析（そもそもこの分析法の概念自体は中学数学で理解可能だと思います）のを先に解説しました。そこから今度は「検定編はないんですか？」という声に応えて続刊では統計的検定と確率分布を採用。その後は補足的な内容の「Extend1」（本来であればサンクリ 48 で検定力分析を扱った「Extend2」を出す予定でしたが、結局のところ出せませんでした…）を刊行しております。

このシリーズの内、「度数分布・相関・回帰分析編」は、サンクリ 45 で刊行したオフセット版（初版 100 冊）は「第十回文学フリマ」にて完売宣言し、コミケ 78 及び「こみっく★トレジャー 16」ではコピー誌として再版しております。そもそも統計学シリーズをテーブルの上に並べているときに、もっともよく聞かれるのは「一番基礎的な内容はどれですか？」というものです。そこで「度数分布・相関・回帰分析編」を紹介すると、買っていってくれる、というパターンが大半でした。「確率分布・検定編」が本書刊行時点で初版 100 冊中 40 冊くらい売れ残っているのとは対照的ですね。

とはいえ、社会科学で使われている統計学は基礎的な平均や分散だけでなく、回帰分析や多変量解析（本書では取り扱いませんが…こちらについては『新・市民のための統計学』をご参照ください）に至るまで様々な分析手法が用いられます。それらの全てを網羅すべき、とは言いませんが（というか本書自体それらのすべてを網羅しておりません)、少なくとも基礎的なところくらいは本書で抑えられるようにはしたいと思っております。さらに言うと、特定の分析手法やパラメータに固執しないことも重要です。例えば、平均に固執すると分散や差の有意性に無頓着になってしまいますし、また差の有意性ばかり重視しすぎるとその差が本当に有効なものかどうかに対する感覚が麻痺してしまいます。そこで役に立つのが検定力分析だったりするのですが。

何はともあれ、統計学は社会科学の研究成果を読むためには必須のスキルです。とある研究のテクニカルな背景を理解しないまま、その主張を鵜呑みにしてしまうのは危険です。だからこそ、私は統計学の同人誌という形で、そのリスクを低下させる役割を果たしたいと思っております。（苦笑）

なお、総集編の追加要素として、社会調査法及び医療統計学に関する読み物を追加しました。よろしければ、こちらも興味をもってくださると幸いです。

注意

本書は、下記の同人誌をまとめて加筆・修正・再構成を行った総集編となっております。

1. 『市民のための統計解析 [基礎編 Extend+ 多変量解析編]』（コミックマーケット 75）の第 1 章・第 9 章の一部
2. 『市民のための統計学 [確率分布・検定編]』（サンシャインクリエイション 46）
3. 『市民のための統計学 Extend1』（サンシャインクリエイション 47）
4. 『市民のための統計学 [度数分布・相関・回帰分析編] ver1.1』（コミックマーケット 78 ／初版はサンシャインクリエイション 45）

目次

第 1 章

平均と分散

1.1 はじめに

特定のパラメータ群（人口、経済指標、学校や病院などの施設数、降水量、睡眠時間、読書量など…）の特徴は、ただパラメータのセットだけを見ていてもわかりません。例えば、新宿駅を午前 6 時台に利用する人は x_6 人、7 時台に利用する人は x_7 人…というふうに、データだけを並べてみても、そのデータがどのような特徴を持っており、またその特徴が何を意味するのか、ということは全くわからないでしょう。

データを集める際にまず必要なのは、特定の規則性に従って集計を行うことです。このような過程を経ないかぎり、持っている情報は何の役にも立たないただのゴミの山になってしまいます。

データの群れの意味を検討する指標として、「平均」や「分散」があります。少なくとも「平均」については多くの人が名前だけならよく聞くと思いますが、改めてここで平均や分散の意味、そして平均や分散だけを見るだけではわからないものを再確認してみましょう。

1.2 平均

平均とはデータの群れをならしたときにどのくらいになるかを示した値です。ならすということは、すなわち全てのデータを足して、その個数で割るということです。本章では、仮に x_1〜x_n という、n 個のデータがあると仮定してみます。また、一般に x の平均値は $E(X)$、μ_x などと表記されることもあります（本書では特に断りがない限り $E(X)$ と表記します）。

$$E(X) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

ちなみに $\sum$ は、高校の数学 A で学んだ、総和を表す記号で、

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{1.2}$$

を表します。

ここで、仮に埼京線の池袋〜大宮間の駅間距離の平均値を求めてみましょう。

駅間の距離を x と置くと、

区間	距離 [km]	区間	距離 [km]
池袋～板橋	1.8	戸田～北戸田	1.4
板橋～十条	1.7	北戸田～武蔵浦和	2.4
十条～赤羽	2.0	武蔵浦和～中浦和	1.2
赤羽～北赤羽	1.5	中浦和～南与野	1.7
北赤羽～浮間舟渡	1.6	南与野～与野本町	1.6
浮間舟渡～戸田公園	2.4	与野本町～北与野	1.1
戸田公園～戸田	1.3	北与野～大宮	1.8

表 1.1　埼京線・池袋～大宮間の駅間距離（出典：Wikipedia）

$$\begin{aligned} E(X) &= \frac{x_{\text{池袋～板橋}} + x_{\text{板橋～十条}} + \cdots + x_{\text{北与野～大宮}}}{14} \\ &= \frac{1.8+1.7+2.0+1.5+1.6+2.4+1.3+1.4+2.4+1.2+1.7+1.6+1.1+1.8}{14} \\ &= 1.6786 \end{aligned} \tag{1.3}$$

となり、埼京線の池袋～大宮間における駅間距離は 1.68km であることがわかります。このことから、この区間においては、だいたい 1.68km おきに駅が設置されていることがわかるのです。

しかし、それでいいのでしょうか？

1.3　分散

「平均値」というものは、データ群の特徴を示す指標として重用されがちです。確かに、平均値とは、データ群の値を均等にならしたものですから、確かにそのデータ群がどのような特徴を持っているかを知るための一つの指標にはなりうるでしょう。しかし、それだけの話です。

そこで、次の 2 つのデータ群を見てみましょう。

$$x \cdots 1, 1, 1, 2, 2, 3, 4, 4, 5, 5, 5 \tag{1.4}$$

$$y \cdots 1, 1, 1, 1, 1, 1, 2, 3, 5, 7, 10 \tag{1.5}$$

この二つのデータ群は、どちらも平均が 3 です。しかし、明らかにデータの分布が違っています。このように、平均だけに着目してしまうと、大事なものを見失ってしまう可能性が結構あるのです。

そこで、「分散」という概念を導入します。「分散」とは、まあ字義の如く、そのデータ群がどれだけ「散らばっている」かを測るための基準なのですが、「散らばっている」というからには、どこかしらを基準点として据える必要があります。

この場合、基準点となるのは平均値です。ただし、データから平均を引いた値の総和をとると、必ず 0 になってしまいます。

$$\mu = E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \text{とする。}$$

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \mu) &= \sum_{i=1}^{n} x_i - \mu\sum_{i=1}^{n} \\ &= n \times \frac{1}{n}\sum_{i=1}^{n} x_i - n\mu = n\mu - n\mu = 0 \end{aligned} \tag{1.6}$$

$x_i - \mu$ はプラスにもマイナスにもなり得ますので、データ群の平均値からの「離れ具合」の総和を求めるためには、なんとかしてプラスにする必要があります（この「なんとかしてプラスにする」という発想は、統計量の評価においては何度か出てくるので、ここで知っておくといいでしょう）。

プラスにする方法は、$x_i - \mu$ は実数ですから、絶対値をとる、あるいは複数乗をとるという方法が考えられます。しかし、高校の数学 A で体験した人も多いかと思いますが絶対値の扱いは難しいですし、複数乗するにしても指数が多いと x_i が平均値より大きく離れている場合、莫迦にならない値が出ることもあります（逆に、その「莫迦にならなさ」を逆に利用する指標もあるのですが）。

というわけで（？）$x_i - \mu$ の 2 乗を、x_i の平均からの「離れ具合」の指標として用います。この「離れ具合」の平均値が、分散となるのです（分散を x の二次モーメントという言い方をすることがあります。また、$(\mu - x_i)^n$ の平均ことを n 次モーメントと呼ぶこともあります）。

$$V(X) = E((X-\mu)^2) = \frac{1}{n}\sum_{i=1}^{n}(\mu - x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{1.7}$$

また、$V(X)$ は、以下のように変換することもできます。

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}({x_i}^2 - 2\mu x_i + \mu^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - 2\mu \times \frac{1}{n}\sum_{i=1}^{n}x_i + \mu^2\frac{1}{n}\sum_{i=1}^{n} \\ &= E(X^2) - 2\mu \times \mu + \mu^2 \times \frac{1}{n} \times n \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 = E(X^2) - E(X)^2 \end{aligned} \tag{1.8}$$

ここでいったん注意を喚起しておきたいのですが、$E(X^2)$、すなわち ${x_i}^2$ の平均を求めるとして、それは x_i が全て同じ値である場合を除いて、決して $E(X)^2$、すなわち x_i の平均の二乗にはならない、ということです。例えば先の埼京線の例を挙げると、駅間距離の二乗の平均はおよそ 2.961 ですが、平均の二乗はおよそ 2.818 です。

閑話休題、分散は、「二乗の平均」から「平均の二乗」を引いた値としても求めることができます。先ほどの埼京線の例ならば、二乗の平均が 2.961、平均の二乗が 2.818 ですから、分散は 0.413 となります。

また、分散の（正の）平方根は、データ群の「偏り具合」を示す指標として用いられ、これを「標準偏差」といい、σ と表記されることが多いです。また、データ群が正規分布と呼ばれる散らばり具合をしている場合、$\mu - \sigma$ から $\mu + \sigma$ の間にはデータのおよそ 68.3%、$\mu - 2\sigma$ から $\mu + 2\sigma$ の間にはおよそ 95.5%、$\mu - 3\sigma$ から $\mu + 3\sigma$ の間にはおよそ 99.7% が集まることが知られています（別に正規分布でなくとも、それなりの量のデータがあれば大数の法則と呼ばれる法則によりだいたいこうなります。なお、マーケティングなどでよく用いられる「シックスシグマ」という表現は、この現象から来ています）。埼京線の例なら、駅間距離の分散が 0.413 ですから、標準偏差は 0.642 となります。また、$\mu - \sigma$ と $\mu + \sigma$ を求めると、前者が $1.679 - 0.642 = 1.037$、後者が $1.679 + 0.642 = 2.321$ で、この範囲から外れるものはありません（まあデータが 14 個しかないから仕方がないのかもしれませんが…）。

1.3.1　歪度と尖度

ここでは余談として、平均、分散以外にデータの特徴を測る値をいくつか紹介しておきます。

まず、x の 3 次モーメント（$(x-\mu)^3$ の平均）を x の標準偏差の 3 乗で割ったものを歪度といいます。歪度はプラスにもマイナスにもなり得、x が平均よりプラスまたはマイナスのどちらに偏っているか（歪んでいるか）

を表す指標となります（歪度がプラスなら平均より大きい側、マイナスなら平均より小さい側に偏っている）。また、歪度の絶対値が小さいほど、x は対称性の高い歪みねぇ分布をしていることがわかります。

さらに、x の 4 次モーメント（$(x-\mu)^4$ の平均）を x の標準偏差の 4 乗（分散の 2 乗）で割ったものを尖度といい、x がどれだけ平均値の周りに集中しているか（次章で取り扱う度数分布表を書いたときどれだけ平均値の周りにデータが集まるか）、要するに分布が尖っているかを示す指標として使われることがあります。なお、一般的に x の尖度は、x の 4 次モーメントを標準偏差の 4 乗で割った値よりも、さらにそこから 3 を引いた値が用いられます（これは、標準正規分布と呼ばれる分布の尖度が 3 であることに由来しています）。どちらを使うにせよ、値が小さいほど、x の分布は尖っていると言えます。

1.3.2　標準化と偏差値

ここまで説明した通り、データ群は様々な平均や分散を持ちます。従って、例えば、クラス平均が 70 点の数学のテストにおける 85 点と、クラス平均が 60 点の数学のテストにおける 85 点では、クラス内での位置づけはそれなりに異なりますし、また平均が 60 点であっても 60 点の周辺に点数が固まっている場合と、90 点周辺と 30 点周辺にばかり固まっている場合とでは、やはり 85 点の位置づけは異なります。

その問題を解決する方法の一つとして、標準化と呼ばれる方法があります。これは、データ群 $\{x_i\}(=x_1,\ldots,x_n)$ の平均を μ、標準偏差を σ としたとき、以下の操作を行うことを示します。標準化を行った後の x_i を y_i とすると、

$$y_i = \frac{x_i - \mu}{\sigma} \tag{1.9}$$

このように作られたデータ群 $\{y_i\}$ は、必ず平均が 0、標準偏差が 1 になります。これで、x_i のデータ群の中における立ち位置を正確に把握できます。

似た方法として、偏差値を求めるというやり方があります。x_i の偏差値を z_i とした場合、z_i は以下の式で求めます。

$$z_i = 50 + 10 \times \frac{x_i - \mu}{\sigma} \tag{1.10}$$

この場合、$\{z_i\}$ は必ず平均 50、標準偏差 10 の分布に修正されます。このように偏差値とは極めて合理的な手法なのですけれども、一時期はこの「偏差値」が競争を煽るとして変に叩かれた時期があったわけなのですが…（まあ、その主張も、偏差値の性格というものを考えればそれなりに理にかなっていると言えなくもありません。あくまでもそれなりに、ですけど）。

なお、先ほど説明した歪度と尖度は、それぞれ、標準化した値の 3 乗平均、4 乗平均と言い換えることが可能です。

1.4　知っておくと便利な定理

データ群 $\{x_i\}$ 及び $\{y_i\}$、及び定数 a,b,c に対して、次の定理が成り立ちます。証明については「市民のための統計解析 [基礎編 Extend + 多変量解析編]」で行っているので、ここでは定理のみを示します。

$$E(aX+b) = a \times E(X) + b \tag{1.11}$$

$$E(aX+bY+c) = a \times E(X) + b \times E(Y) + c \tag{1.12}$$

$$V(aX+b) = a^2 \times V(X) \tag{1.13}$$

年齢	人口	比率 [%]
20	698	8.49
21	721	8.77
22	749	9.11
23	761	9.26
24	795	9.67
25	869	10.57
26	874	10.63
27	873	10.62
28	922	10.22
29	957	10.64
20 代合計	8219	100

表 1.2　宮城県多賀城市の 20〜29 歳の各年齢の人口分布

年齢	人口	比率 [%]
20〜24	3724	8.83
25〜29	4495	10.66
30〜34	4958	11.76
35〜39	5123	12.15
40〜44	4198	9.95
45〜49	3794	9.00
50〜54	3899	9.25
55〜59	4481	10.63
60〜64	4175	9.90
65〜69	3323	7.88
20〜60 代合計	42170	100

表 1.3　宮城県多賀城市の 20〜69 歳の 5 歳ごとの人口分布

1.5　度数分布表

1.5.1　度数分布表の基礎

さて、社会学などの研究においては、特定のデータ群に対して、一つ一つの値が示されることはないといっても過言ではありません。何せ社会学や教育学の量的研究は数百、場合によっては数千のデータを取り扱うので、例えば家計 1 における 1 年間の消費支出は x_1、家計 2 は x_2…などというデータをいちいち提示していては、本や報告書がものすごく分厚くなってしまいます。そこで用いられるのが度数分布表です。

度数分布表は、大別して二つに分かれますが、ここでは、実物を見たほうがいいでしょう。一つは宮城県多賀城市の 20〜29 歳の人口分布、もう一つは 20〜69 歳の 5 歳ごとの人口分布です（平成 21 年 8 月 31 日現在)[*1]。このように、特定の数値（ここでは年齢）に属するデータの数がどれだけであるかというものと、特定の数値の範囲に属するデータがどれだけであるかを表示するものに分かれるのです。

ただし、一般的に度数分布表は、それぞれの値、範囲に属するデータ数が与えられることは少なく、むしろ比率（%）で与えられることが多いです。もちろん、そこに属する実際のデータ数がどれだけになるかは、比率にデータの総数をかければいいので、それぞれに属するデータの数はわかるのですが、少なくとも私には、それぞれのデータの総数を開示しないのは、少々サーヴィス精神に欠けているな、と思ってしまいます。

1.5.2　度数分布表による平均・分散の求め方

度数分布表を使って、データ群における平均と分散を求めることも可能です。まずは、表 1-2 のように、データ群 $\{x_i\}(= x_1, x_2, \ldots, x_n)$ の内、値 $y_1, y_2, \ldots, y_k$ （一般にこの値は小さい順に並べます）に属するものが $n_1, n_2, \ldots, n_k$ 個、となっている度数分布表から求めて見ましょう。なお、データ群 $\{x_i\}$ の内、特定の数値 y_j に属する値を n_j 個とする場合（ただし、j は 1 から k までの整数とする)、n_j を数値 y_j に対する度数と言います。また、$y_1, y_2, \ldots, y_k$ が $\{x_i\}$ に存在する値を全て記録している場合、以下の関係が成り立つはずです。

$$n = n_1 + n_2 + \cdots + n_k \tag{1.14}$$

このとき、n_j を n で割った値 $p_j = \frac{n_j}{n}$ を、y_j に対する（度数）分布と言います。

[*1] 出典…http://www.city.tagajo.miyagi.jp/sisei/toukei/si-to-zinkou_danzyo.pdf より筆者作成。比率は加筆

値	度数	分布	値 × 分布	値 − 平均	(値 − 平均)^2	分布 × (値 − 平均)^2
y_1	n_1	p_1	y_1p_1	$y_1-\mu$	$(y_1-\mu)^2$	$p_1(y_1-\mu)^2$
y_2	n_2	p_2	y_2p_2	$y_2-\mu$	$(y_2-\mu)^2$	$p_2(y_2-\mu)^2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_k	n_k	p_k	y_kp_k	$y_k-\mu$	$(y_k-\mu)^2$	$p_k(y_k-\mu)^2$
合計	n	1	$\sum_{j=1}^{k} y_jp_j$	$\sum_{j=1}^{k}(y_j-\mu)$	$\sum_{j=1}^{k}(y_j-\mu)^2$	$\sum_{j=1}^{k} p_j(y_j-\mu)^2$

表 1.4　表 1-2 形式の度数分布表における平均と分散の求め方

上の考え方に従って度数分布表を作成すると、表 1-4 のようになります。次に、平均と分散の定義を、もう一度整理してみましょう。

$$E(X)=\mu=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1.15}$$

$$V(X)=\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{1.16}$$

さて、表 1-4 の度数分布表は、データ群 $\{x_i\}$ の内、値 y_1 であるものが n_1 個、y_2 であるものが y_2 個、$\cdots y_k$ であるものが n_k 個であるということを示しています。従って、表 1-4 の度数分布表で表されたデータ群 $\{x_i\}$ の総和は、次のように表されるはずです。

$$\sum_{i=1}^{n}=n_1y_1+n_2y_2+\cdots+n_ky_k=\sum_{j=1}^{k}n_jy_j \tag{1.17}$$

さて、今一度、平均の定義を振り返ると、それはデータの総和をデータの個数で割った値です。従って、平均 $\mu=E(X)$ は、以下の通りに表されます。

$$\mu=E(X)=\frac{1}{n}\sum_{j=1}^{k}n_jy_j \tag{1.18}$$

ところで、今一度、n_j に関して、以下の関係を振り返ってみます。

$$n=n_1+n_2+\cdots+n_k=\sum_{j=1}^{k}n_j \tag{1.19}$$

この等式の各辺を n で割ってみると、次の通りになります。

$$\frac{n_1+n_2+\cdots+n_k}{n}=\frac{n_1}{n}+\frac{n_2}{n}+\cdots+\frac{n_k}{n}=p_1+p_2+\cdots+p_k=\sum_{j=1}^{k}p_j=1 \tag{1.20}$$

というわけで、度数分布 p_j を全部足すと 1 となることがわかるわけですが（ただし、実際の度数分布表では、パーセンテージ表記時の端数処理の関係により、必ずしも 1 になるとは限らない）、これを踏まえると、平均を求める式を以下の通りに置き換えることが可能です。

$$\mu = E(X) = \frac{1}{n}\sum_{j=1}^{k} n_j y_j = \sum_{j=1}^{k} p_j y_j \tag{1.21}$$

従って、度数分布表を使えば、「値 × 分布」の総和が $\{x_i\}$ の平均値に等しくなることが示されるのです。

同じように分散も求めてみましょう。分散とは、それぞれの値の平均値 μ からの差の二乗、すなわち $\{(x_i-\mu)^2\}$ の平均値です。また、$\{x_i\}$ の内、y_j に属するものが n_j 個あるとすると、分散は次の通りに表されます。ここでは p_j を用いた表し方も一緒に示しましょう。

$$\sigma^2 = V(X) = \frac{n_1(y_1-\mu)^2 + \cdots + n_j(y_j-\mu)^2}{n} = \frac{1}{n}\sum_{j=1}^{k} n_j(y_j-\mu)^2 \tag{1.22}$$

$$= \sum_{j=1}^{k} p_j(y_j-\mu)^2 \tag{1.23}$$

従って、度数分布 × (値 − 平均)^2 の総和が、$\{x_i\}$ の分散となるわけです。

さて、表 1-2 のような度数分布表ならば、平均や分散の求め方は簡単に説明できますし、また結果も正確なものです。しかし、表 1-3 のような度数分布表ではどうでしょうか。

表 1-3 では年齢層の分布（離散量）を用いているので、比較的簡単に、それぞれの値を示すデータ数の分布を作成することができます。しかし、表 1-3 のような度数分布表は、一般には人口や、所得や消費支出などの経済統計、あるいは農地面積や単位人口あたりの医師数などといった連続量（あるいは連続量と見なせるもの）の分布を用いるため、それぞれの値は非常に多様になり、当然のことながら小数だって平気で出てきます。そうすると、例えばとある自治体において、世帯収入が 400 万円の世帯は $n_{4,000,000}$ 世帯、400 万 1 円のは $n_{4,000,001}$ 世帯…などといった度数分布表を作成すると、度数分布表それ自体がものすごく大きくなり、報告書に掲載することなど不可能になります。そこで、例えば世帯収入が 300 万円以上 400 万円未満の世帯は n_3 世帯、400 万円以上 500 万円未満は n_4 世帯…というように、$\{x_i\}$ 特定の範囲 j（一般的に、y_j 以上 y_{j+1} 未満という形式がとられることが多いです）に属するデータの数を n_j とする度数分布表を作成するとわかりやすくなります。それが表 1-3 の形式の度数分布表です。

そのような形式の度数分布表を一般的な形で表したものが、表 1-5 の左から 1〜3 列目です。

ただし、このような表からでは、データ群の正確な平均と分散（＋標準偏差）を求めることはできません。なぜならデータ群の分布が、特定の値ではなく範囲で示されているからです。そこで、それぞれの値の範囲を代表する（と思われる）値を領域ごとに求め、それを用いて平均及び分散を見る方法があります。このとき、領域ごとに定められた、その領域を代表する値を、代表値と言います。

一般に代表値には、領域の上限と下限の中央の値が用いられます。例えばとある領域が y_j 以上 y_{j+1} 未満の場合、代表値（仮に z_j と置きます）は、

$$z_j = \frac{y_j + y_{j+1}}{2} \tag{1.24}$$

とされることが多いです。

ただし、表 1-5 に示した通り、一般にこの形式の度数分布表は、表の一番下に位置される領域は、上限がない、青天井である場合が多いです（特に経済統計などは一般にこうなっています）。この場合、上限値は無限大ですから、下限値と上限値の中間をとって代表値としても、その値は無限大になり、結果として平均値まで無限大になってしまいます。こういった領域の代表値の定め方は、基本的に分析者の判断にゆだねられます（ちなみに私はその領域の下限値の 1.5 倍〜2 倍にすることが多いです）。

さて、代表値 z_1〜z_k を設定した上で、特定の領域 j（代表値 z_j）に属するデータの数を n_j とした場合、表 1-2 の形式で平均及び分散を求めた方法と同様にして、以下のように求めることができます。

値 以上		未満	代表値	度数	分布	分布 × 代表値
y_1	$\sim$	y_2	z_1	n_1	p_1	$p_1 z_1$
y_2	$\sim$	y_3	z_2	n_2	p_2	$p_2 z_2$
$\cdots$		$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_{k-1}	$\sim$	y_k	z_{k-1}	n_{k-1}	p_{k-1}	$p_{k-1} z_{k-1}$
y_k	$\sim$		z_k	n_k	p_k	$p_k z_k$
			$-$	n	1	$\sum_{j=1}^{k} p_j z_j$

表 1.5　表 1-3 形式の度数分布表での平均値の算出

値 以上		未満	代表値	度数	分布	代表値 − 平均	(代表値 − 平均)^2	分布 × (代表値 − 平均)^2
y_1	$\sim$	y_2	z_1	n_1	p_1	$z_1 - \mu$	$(z_1 - \mu)^2$	$p_1(z_1 - \mu)^2$
y_2	$\sim$	y_3	z_2	n_2	p_2	$z_2 - \mu$	$(z_2 - \mu)^2$	$p_2(z_2 - \mu)^2$
$\cdots$		$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_{k-1}	$\sim$	y_k	z_{k-1}	n_{k-1}	p_{k-1}	$z_{k-1} - \mu$	$(z_{k-1} - \mu)^2$	$p_1(z_{k-1} - \mu)^2$
y_k	$\sim$		z_k	n_k	p_k	$z_k - \mu$	$(z_k - \mu)^2$	$p_k(z_k - \mu)^2$
			$-$	n	1	$\sum_{j=1}^{k}(z_j - \mu)$	$\sum_{j=1}^{k}(z_j - \mu)^2$	$\sum_{j=1}^{k} p_j(z_j - \mu)^2$

表 1.6　表 1-3 形式の度数分布表での分散の算出（ただし、平均は概算平均）

$$\mu = E(X) = \frac{1}{n}\sum_{j=1}^{k} n_j z_j = \sum_{j=1}^{k} p_j z_j \tag{1.25}$$

$$\sigma^2 = V(X) = \frac{1}{n}\sum_{j=1}^{k} n_j (z_j - \mu)^2 = \sum_{j=1}^{k} p_j (z_j - \mu)^2 \tag{1.26}$$

ただし、ここで注意しておかなければならいのは、z_i とはあくまでも特定の領域を代表すると思われる値であり、その領域における x の値が全て z_i であると仮定した上での値であるので、ここで求めた平均及び分散はあくまでも概算値であることに注意しなければなりません。

1.5.3　度数分布表から平均・分散を求める

ここでは、実際の度数分布表を用いて、データ群の平均と分散を求めてみましょう。今回用いるデータは、平成 21 年 4 月 21 日に行われた、平成 21 年度全国学力・学習状況調査（通称：全国学力テスト）の内、中学 3 年を対象とした、岩手県における「国語 B：主として活用」（以下、「国語 B」）全 11 問の正答数の分布です。こちらは正答数ごとの度数が開示されているため、本項の教材としては結構都合がいいと思います。

正答数及びそれに対応する正答者数は表 1-6 の通りです[*2]。

まずは、正答数 $x_i (= i, 0 \leq i \leq 11)$ と、それに対応する正答者数 n_i を表に示し、$x_i n_i$ の値を求め、その総和を求めます。表にある通り、$x_i n_i$ の総和は 92406 で、また n_i の総和は 11326 ですから、平均正答数 μ は、

[*2] http://www.nier.go.jp/09chousakekka/09todoufuken_data/03_iwate/ 06_chuu_kyouka_chousakekkagaikyou_03iwate.pdf

正答数 x_i[問]	正答者数 n_i[人]	$x_i n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$n_i(x_i - \mu)^2$
0	130	0	−8.159	66.569	8654.0
1	210	210	−7.159	51.251	10762.7
2	222	444	−6.159	37.933	8421.1
3	303	909	−5.159	26.615	8064.3
4	440	1760	−4.159	17.297	7610.7
5	562	2810	−3.159	9.979	5608.2
6	770	4620	−2.159	4.661	3588.9
7	967	6769	−1.159	1.343	1298.7
8	1385	11080	−0.159	0.025	34.6
9	1779	16011	0.841	0.707	1257.8
10	2345	23450	1.841	3.389	7947.2
11	2213	24343	2.841	8.071	17861.1
合計	11326	92406	3.841	14.753	81109.3

表 1.7　岩手県・中学 3 年における「国語 B」の正答数（出典：国立教育政策研究所ウェブサイト）

$$\mu = \frac{\sum_{i=0}^{11} i n_i}{n} = \frac{92406}{11326} = 8.159 \tag{1.27}$$

となります。次に、分散を求めるために、この値と x_i の差を求め、さらに二乗したあと、それに n_i をかけます。これにより求めた $n_i(x_i - \mu)^2$ の和は 81109.3 ですから、分散 σ^2 及び標準偏差 σ は、

$$\sigma^2 = \frac{\sum_{i=0}^{11} n_i (i - \mu)^2}{n} = \frac{81109.3}{11326} = 7.161 \tag{1.28}$$

$$\sigma = 2.676 \tag{1.29}$$

となります。ちなみに報告書における平均と標準偏差は、8.2 と 2.7 です。

1.5.4　最頻値、中央値

最後に、平均や分散以外で、データ群の特徴を示す指標としてよく使われるものを二つほど紹介します。

一つは最頻値で、これは、データが離散量でない場合は、度数分布表を作っていないと求めることができない値です。

まずは度数分布表を作り、その中で度数がもっとも大きい項を探します。これが最頻値と呼ばれる値です。例えば先ほどの岩手県における国語 B の正答数を例に挙げるなら、最頻値は 11 問、ということになります。

次に中央値ですが、これは、データ全体の中で、文字通り中央に属する値です。

求め方としては、まず、領域 j におけるデータ数 n_j の、領域 1 から j までの和（累積度数）が n の半分、もしくは領域 j での度数分布の領域 1 から j までの和（累積度数分布）が 0.5 をはじめて超える領域 j を探します。その領域に特定の値が対応しているならば、その値が中央値になり、またその領域が特定の範囲をしめしているなら、その値を抽出し、「その領域までの累積度数 $-n$ の半分」番目の値が中央値となります。

先の岩手県での「国語 B」の場合は、8 問までの累積度数が 4969、9 問までが 6748 であり、また n の半分が 5663 であるため、9 問が中央値となります。

このように、データをとっても、平均値と最頻値と中央値は一致しないことがほとんどですので、それぞれの値を注意して見たり、あるいは平均値にこだわっているような分析に対して「中央値は？最頻値は？」と疑うことも、時として必要となります。なお、この 3 値が珍しく一致する場合は、分布が完全に平均値を中心に左右対称であり、かつ平均値がそのまま最頻値である場合のみです。

第 2 章

相関関係

2.1 はじめに

前章では一つのデータの集合に対して、その性格（平均、分散など）を明らかにする作業を行いました。しかし、調査や研究によっては、複数のデータ間の関係性を調べる必要があることもあります。そのための作業として、データ間の相関関係を求めるという方法があります。

ただし相関関係の検討（と、次章で取り扱う回帰分析）は、度数分布表でできるものではなく、個々のデータセットが必要になります（度数分布のクロス集計表があれば簡易的に求めることはできますが）。また、二つのデータの関係性を相関係数を用いて表す場合、その元となったデータは（大概の場合は量それ自体が膨大なので）省略されることが多いです。というわけで本章では、相関係数の持つ「意味」を記述するにとどめておきます。

2.2 相関係数とは？

例えば、$\{(x_i, y_i)\}\cdots(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ というデータの組（データセット）があるとします。このとき、データ群 $\{x_i\}$ と $\{y_i\}$ がどのような関係になっているかの概形を検討するための指標が、相関係数です。

相関係数（一般に r と表記されます）は、-1 から 1 までの値で表されます。そして、相関係数 r が 1 に近づくほど、「x が大きくなると y も大きくなる」という関係が強くなり、また -1 に近づくほど、「x が小さくなると y も小さくなる」という関係が強くなります。そして r が 1 もしくは -1 になる場合は、x と y は完全に一次関数の関係を示します。つまり、相関係数とは、x と y の関係がどれだけ一次関数に近いかということを示す指標なのです。

相関係数はこのような性格を持つが故に、注意しなければならないこともあります。例えば、x が特定の値 x' に近づくほど y は大きくなるが、x' に離れるほど y は小さくなり、x と y の分布は近似的に $x = x'$ を対称軸とした二次関数の関係を持つと見なすとします。x と y が例えばこのような関係にあり、かつ x の平均が x' で、分散や歪度がそれほど大きくないとすると、x と y は二次関数に似た関係にあると見なすことができますが、相関係数をとると、かなり低い値が出る——すなわち、x と y に関係性はほとんどない、と見なされる可能性があるのです。そのため、相関係数を求めるかどうかについては、まずは x と y をプロットしてからにしても遅くはないでしょう。

そうといっても、社会学や経済学、あるいは教育学で用いられるデータの関係性については、一次関数の関係を想定するとかなりうまく説明できる（これについては、次章の「回帰分析」で説明します）ことが多いので、相関係数は、少なくとも社会科学の研究においては、二つのデータの関係性を示す有効な指標として用いられることが多いです。

2.3　相関係数の求め方

2.3.1　共分散

相関係数を求める前に、まずは共分散という考え方を知る必要があります。共分散とは、データセット群 $\{x_i, y_i\}$ があるとして、さらに $\{x_i\}$ の平均を μ_x、$\{y_i\}$ の平均を μ_y とするとき、$(x_i - \mu_x)(y_i - \mu_y)$ の平均のことを指します。この値はプラスにもマイナスにもなり得ますが、ここで中学の数学「正負の数」で習ったことを思い出してみましょう。

二つの実数 a, b の積を求めるとします。この場合、積がプラスになるためには、a, b の両方がプラスであるか、あるいは両方がマイナスである必要があります。逆に積がマイナスの場合は、a, b のどちらかがプラスであり、またどちらかがマイナスである必要があります。

ここで、a, b を、$x_i - \mu_x, y_i - \mu_y$ に置き換えてみましょう。$(x_i - \mu_x)(y_i - \mu_y)$ の結果がプラスになるときは、どのようなときでしょうか。この場合は、x_i, y_i 共にそれぞれの平均値より大きい場合、もしくはそれぞれの平均値より小さい場合です。そしてそれぞれの結果は、「x が大きい（小さい）と、y も大きい（小さい）」という関係を支持するデータと見なします。

逆に、$(x_i - \mu_x)(y_i - \mu_y)$ の結果がマイナスだとどうでしょうか。この場合は、x_i と y_i のどちらかが平均より大きく、もう片方が平均より小さいことを示します。従って、「x が大きい（小さい）と、y は小さく（大きく）なる」という関係を支持するデータとなるのです。

そして、$(x_i - \mu_x)(y_i - \mu_y)$ の平均が共分散と呼ばれます。この値がプラスであれば、x と y の関係は、「x が大きい（小さい）と、y も大きい（小さい）」というものに近くなり、またマイナスならば「x が大きい（小さい）と、y は小さく（大きく）なる」という関係に近くなるのです。

これを数式で表すと、

$$Cov(x, y) = E((x - \mu_x)(y - \mu_y)) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y - \mu_y) \tag{2.1}$$

と表されます。$Cov(X, Y)$ とは、$\{x_i\}$ と $\{y_i\}$ の共分散を示す記号です。

なお、$Cov(X, Y)$ については、以下のように変換することもできます。

$$\begin{aligned} Cov(x, y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \mu_x y_i - \mu_y x_i + \mu_x \mu_y) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\mu_x}{n}\sum_{i=1}^{n} y_i - \frac{\mu_y}{n}\sum_{i=1}^{n} x_i + \frac{\mu_x \mu_y}{n}\sum_{i=1}^{n} \\ &= E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_x \mu_y \\ &= E(XY) - \mu_x \mu_y = E(XY) - E(X)E(Y) \end{aligned} \tag{2.2}$$

以上より、「共分散 = 積の平均 − 平均の積」として求めることも可能です。

2.3.2　相関係数

とはいえ共分散は、x や y のデータの幅によってかなり左右されるので、x と y の関係性を示す決定的な指標とはなりにくいです。そこで登場するのが相関係数です。相関係数とは、x と y の共分散を、x と y の、それぞれの標準偏差（分散の正の平方根）で割った値となります。