

データを読む

データの整理

データから有益な情報を引き出すためには、まず最初にデータを体系的な方法に則り整理する必要があります。整理をしなければデータはただの数字や文字などの羅列であり、そこから傾向や特徴を読み取ることが困難な為です。

例えば、ある企業の社員100名分の部署と年収がペアになったデータのリストがあったとします。部署と年収のペアはランダムに並んでいるようで、眺めているだけではこのデータからはなにもわかりません。しかし、部署毎に年収をまとめ、グラフを描いてみると、部署によって年収が異なっており、開発 > 営業 > 総務の順に高額となっているらしいことが見えてきます。



このようにデータを整理することが、データから情報を読み取る足がかりとなるのです。

データの種類

データの整理の仕方はデータの種類によって異なってきます。この節ではデータの種類について見ていきます。

データは質的データと量的データに大別されます。質的データとは分類を表すデータで、算術計算することはできないデータのことです。一方、量的データとは数値で表されたデータで、算術計算することができるデータのことです。

質的データの例としては学籍番号や氏名、順位などが挙げられます。数字で表されたデータであっても計算結果に意味がないデータであれば、それは質的データとなります。Aさん、Bさん、Cさんの学籍番号がそれぞれ「10010」、「10020」、「20030」であるとき、 20030 (Cさんの学籍番号) = 10010 (Aさんの学籍番号) + 10020 (Bさんの学籍番号) となりますが、「CさんはAさんとBさんを足し合わせたような人物」とはなりません。計算結果には意味がありません。よって、学籍番号は数字で表されていますが質的データとなります。

量的データの例としては気温や西暦、身長などが挙げられます。量的データは数字により表現され、計算結果には意味が伴います。ある日の札幌市、那覇市の気温がそれぞれ「15℃」、「30℃」であるとき、 $30^{\circ}\text{C} - 15^{\circ}\text{C} = 15^{\circ}\text{C}$ と引き算することで札幌市と那覇市の気温差は15℃であることがわかります。計算により意味のある結果を得ることができます。よって、気温は量的データとなります。

質的データおよび量的データに大別されたデータは表現する情報の性質を基準にして更に分類することができます。この基準のことを尺度と言います。尺度は一般的に「名義尺度」、「順序尺度」、「間隔尺度」、「比例尺度」の4つの水準に分類され¹、質的データは名義尺度のデータと順序尺度のデータに分けられ、量的データは間隔尺度と比例尺度に分けられます。

名義尺度とはデータの区別（だけ）が可能な尺度です。名義尺度のデータには便宜的に数字を割り振ることができますが、割り振った数字の算術計算には意味がありません。また、大小関係にも意味がありません。しかし、区別はできなければなりませんので等しいか等しくないかには意味があります。

名義尺度のデータの例としては学籍番号を挙げることができます。学籍番号は学生一人一人に数字を割り振った記号として捉えることができます。この学籍番号が等しければ同じ学生を、等しくなければ異なる学生を示していることとなります。しかし、先に見たように学籍番号の算術計算には意味がありませんし、Aさんの学籍番号はBさんの学籍番号よりも小さいのでAさんの成績はBさんの成績よりも下位であると言うようなことはありません。大小関係には意味がないのです。よって、学籍番号は名義尺度のデータと言うこととなります。

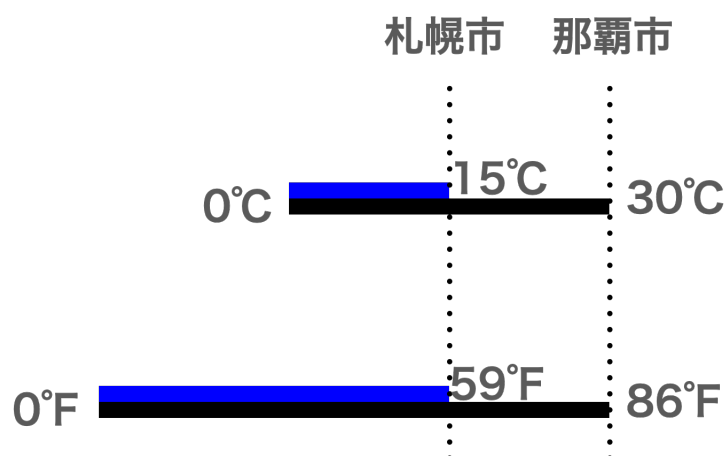
順序尺度とはデータの区別が可能で順序や大小の評価も可能な尺度です。順序尺度のデータは一般的に数字を使って表現されますが、名義尺度と同様、算術計算には意味がありません²。順序や大小には意味がありますが、順位や大小の差は、その間隔が一定であるとは限らないため意味を持ちません。

順序尺度のデータの例として100m走の順位を挙げることができます。ある大会でAさんが1位でBさんは2位、Cさんは3位だったとします。このとき、Bさんは、Aさんより遅くCさんより速い言うことはわかります。しかし、順位の引き算を行い、 $(3\text{位} - 2\text{位}) = (2\text{位} - 1\text{位})$ なのでAさんとBさんの走力の差はBさんとCさんの走力の差と同じ、とはなりません。Bさんは1位のAさんと僅差の2位だったのかも知れませんし、あるいは3位のCさんと殆ど同タイムの2位だったのかも知れません。順位の算術計算には意味がないことがわかります。よって、順位は順序尺度と言うこととなります。



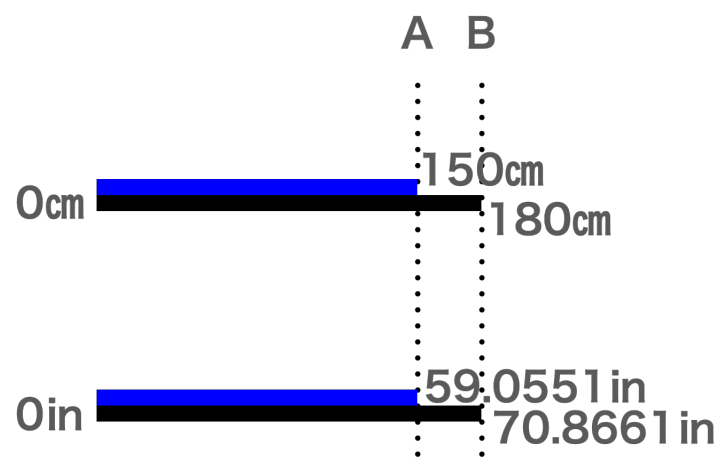
間隔尺度とはデータの順序の評価が可能で差や和の評価も可能な水準です。間隔尺度のデータの取り得る数値の間隔は等間隔で、足し算の結果や引き算の結果は意味を持ちます。しかし、0（ゼロ）が「無、何もない」という意味ではないため掛け算の結果や割り算の結果は意味を持ちません。

間隔データの例として気温を挙げることができます。ある日の札幌市、那覇市の気温がそれぞれ「15℃」、 「30℃」であるとき、 $30^{\circ}\text{C} - 15^{\circ}\text{C} = 15^{\circ}\text{C}$ と引き算することで札幌市と那覇市の気温差は15℃であることがわかります。計算により意味のある結果を得ることができます。引き算の逆演算³である足し算の結果も意味を持ちます。しかし、 $30^{\circ}\text{C} \text{ (那覇市の気温)} / 15^{\circ}\text{C} \text{ (札幌市の気温)} = 2$ となるから那覇市の気温は札幌市の気温の2倍である、とはなりません。この割り算を摂氏ではなく華氏⁴で行うと $86^{\circ}\text{F} \text{ (那覇市の気温)} / 59^{\circ}\text{F} \text{ (札幌市の気温)} = \text{約}1.5$ と摂氏による計算とは異なった結果となります。比率は無次元数⁵なので、気温の単位の取り方により結果が異なるということは気温の割り算には意味がない⁶ということです。よって、気温は間隔尺度ということになります。



比例尺度とはデータの比率の評価が可能な水準です。比例尺度のデータでは0（ゼロ）は「無、何もない」ことを意味し、掛け算や割り算の結果も意味を持ちます。

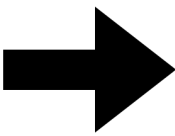
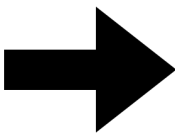
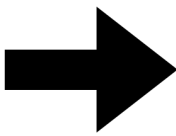
比例尺度の例として身長を挙げることができます。Aさん、Bさんの身長がそれぞれ「150cm」，「180cm」であるとき、Bさんの身長はAさんの身長の1.2倍（= 180cm / 150cm）とすることができます。気温の時とは異なり、身長の単位をセンチメートルからインチに替えて⁷ 計算しても70.8661 in / 59.0551 in = 1.2倍となります。身長は割り算の結果も意味を持つので比例尺度ということになります。



以上をまとめた表です。ここで見てきたようにデータに対して意味のある算術計算が尺度の水準によって異なり、データ操作（データに対する数学的操作?）の自由度は「名義尺度」 < 「順序尺度」 < 「間隔尺度」 < 「比例尺度」の順で高くなっています。データ操作（データに対する数学的操作?）の自由度が高い水準はデータ操作（データに対する数学的操作?）の自由度の低い水準の性質を含んでいるので、高い水準のデータは低い水準のデータに変換して扱うことができます。

データの種類	尺度水準	尺度の意味	算術計算	大小比較	差	比	データの例
質的データ	名義尺度	区別できる	不可	－	－	－	学籍番号, 氏名, 天気
	順序尺度	順序, 大小がある	不可	○	－	－	順位, 学年, 満足度
量的データ	間隔尺度	間隔が等しい	加法, 減法	○	○	－	気温, 時刻, 日付
	比例尺度	0（ゼロ）に意味がある	加法, 減法, 乗法, 除法	○	○	○	身長, 経過時間, 絶対温度

比例尺度である身長のデータを（0cm以外のある）基準値、例えば150cmからの差に変換したデータは間隔尺度のデータとなります⁸。このデータに対して0cm以上は「高い」0cm未満は「低い」と高低を対応付ければ順序尺度になります。更に、このデータに対して「高い」は「A」，「低い」は「B」とアルファベットを対応付ければ名義尺度となります⁹。

比例尺度		間隔尺度		順序尺度		名義尺度
150cm		0cm		高い		A
180cm		30cm		高い		A
145cm	対応付ける	-5cm	対応付ける	低い	対応付ける	B
148cm		-2cm		低い		B
⋮		⋮		⋮		⋮
172cm		22cm		高い		A

しかし、名義尺度のデータである「A」と「B」を順序尺度の「高い」と「低い」に変換しようとしても「A」と「B」のどちらに「高い」を対応させてどちらに「低い」を対応させるかを定めることができません。「A」に「低い」を対応付けし、「B」に「高い」を対応付けてしまうと元データとは異なるデータとなってしまいます。

このように高い水準のデータは低いデータの水準に変換することができますが、低い水準のデータを高い水準のデータに変換することはできません。

数値で表されたデータ（量的データ?）については尺度とは異なる視点から、離散量と連続量に区別されます（することもできます?）。離散量とはサイコロの出目や都道府県の人口など取ることのできる値が飛び飛びとなるデータのことで、一方、連続量とは気温や身長のように取ることのできる値が連続しているデータのことで、

飛び飛びの値を取っていても、そのデータを必ずしも離散量と見做すとは限りません。例えばテストの得点は通常、83点、76点、92点、・・・と言うように整数で表され離散的な値しか取りません。しかし、テストの得点が表している受験者の学力は連続的に変化するものと考えられるので、83点は82.5点～83.4点に対応する学力を表していると解釈し、テストの得点データは連続量として扱うことが一般的です。

要約統計量

データの分布はそのデータの持つ傾向や特徴を探る重要な手がかりとなります。要約統計量はこのデータの分布を概括して表現してくれる数値です。要約統計量からはデータの分布の中心的位置や散らばり具合、形状を知ることができます。よって、与えられたデータの要約統計量を把握する(?)ことはデータの全体像を把握するヒントとなります。ただし、データの尺度水準により可能な数学的操作が異なるため適用できる要約統計量もデータの尺度水準に依存します。

要約統計量のうちデータの分布の中心的位置を表す数値には平均値、中央値、最頻値があります。データの分布の中心的位置を表す数値（これらの数値は?）は代表値とも呼ばれます。（平均値、中央値、最頻値にはそれぞれメリット、デメリットがあるのでそれらを踏まえて使用することが重要です。?）

平均値

多くの場合、平均値と言えば算術平均値（相加平均値）を意味しますが、算術平均値ではなく幾何平均値（相乗平均値）や調和平均値などが平均値として扱われる場合もあります。

算術平均値（相加平均値） 要素数が n 個のデータ $\{a_1, \dots, a_n\}$ の算術平均値 A は以下の式で求められます。

$$A = \frac{a_1 + a_2 + \dots + a_n}{n} = \frac{1}{n} \sum_{i=1}^n a_i$$

算術平均値はデータの要素すべてを足し合わせた値をデータの要素数で割った値です。足し算を行うので間隔尺度以上のデータに対して意味を持ちます。

算術平均値は数学や物理学、工学などの理系分野だけではなく言語学や経済学、社会学などの文系分野でも広く利用されています。

幾何平均値（相乗平均値） 要素数が n 個のデータ $\{a_1, \dots, a_n\}$ の幾何平均値 G は以下の式で求められます。

$$G = \sqrt[n]{a_1 \times a_2 \times \dots \times a_n} = \left(\prod_{i=1}^n a_i \right)^{\frac{1}{n}}$$

幾何平均値はデータの要素すべてを掛け合わせた値の n 乗根¹⁰です。掛け算を行うので比例尺度以上のデータに対して意味を持ちます。また、データはすべて正の数である必要があります。

幾何平均値はデータの要素同士の掛け算が有益な（有意義な？）データの代表値として採用されます。応用例として比率の平均値を上げることができます。

日本の移動通信の2011年から2020年の各年3月の月間平均アップロードトラフィックの推移は(9.9Gbps, 23.4Gbps, 44.2Gbps, 80.0Gbps, 123.3Gbps, 184.5, 249.0, 335.9, 404.6, 442.3Gbps)となっています¹¹。よって、月間平均アップロードトラフィックの前年度比は(2.36, 1.89, 1.81, 1.54, 1.50, 1.35, 1.35, 1.20, 1.09)となります。前年度比の算術平均値を A 、幾何平均値を G とすると

$$A = \frac{2.36 + 1.89 + 1.81 + 1.54 + 1.50 + 1.35 + 1.35 + 1.20 + 1.09}{9} = 1.565\dots$$

$$G = \sqrt[9]{2.36 \times 1.89 \times 1.81 \times 1.54 \times 1.50 \times 1.35 \times 1.35 \times 1.20 \times 1.09} = 1.524\dots$$

となります。

「2011年から2020年までの前年度比の平均値が M である」とすれば、2011年から2020年までは毎年 M の比率で月間平均アップロードトラフィックが増加または減少しているとして2011年の月間平均アップロードトラフィックの値 T_{2011} から2020年の値 T_{2020} を計算することができます。この時、 M 、 T_{2011} 、 T_{2020} には以下の関係が成立します。

$$T_{2020} = T_{2011} \times M^{(2020-2011)} = T_{2011} \times M^9$$

2011年の月間平均アップロードトラフィックの実測値9.9を T_{2011} に代入し、 $M = A$ および $M = G$ としてそれぞれ T_{2020} を計算すると

$$T_{2020} = 9.9 \times A^9 = 9.9 \times 1.565^9 = 557.5\dots$$

$$T_{2020} = 9.9 \times G^9 = 9.9 \times 1.524^9 = 439.0\dots$$

となります。2020年の月間平均アップロードトラフィックの実測値は442.3なので算術平均値を前年度比の平均値としてしまうと2020年の値を大きく見積もりすぎてしまいます。この場合の平均値は幾何平均値が妥当であることがわかります。

調和平均値要素数が n 個のデータ $\{a_1, \dots, a_n\}$ の調和平均値 H は以下の式で求められます。

$$H = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

調和平均値はデータの逆数の算術平均値の逆数です。データの逆数を利用するので比例尺度以上のデータに対して意味を持ちます。また、データはすべて正の数である必要があります。

調和平均値は幾何平均値と同様、比率の平均値として採用されます。応用例として速度の平均値を挙げることができます。

調和平均は、典型的には率や比に対する平均を考える場合に適切である。例えば速度の平均を計算することを考えると、乗り物がある距離を時速 60 km で走りそれから同じ距離を時速 40 km で走った場合、全体の走行時間と走行距離から求められる平均速度は調和平均の値である時速 48 km であって、算術平均によって求められる時速 50 km を平均とするのは適切ではない。もっとも、調和平均が適切な場合でもしばしば誤って算術平均が用いられる

平均値の計算にはすべてのデータが使われます。これは平均値にはすべてのデータが影響を及ぼしているということであり、分布の全体像を概括するという要約統計量として優れている点だと言えます。

一方で、すべてのデータを利用しているが為に、データに他の値とは大きく異なる値である外れ値が含まれていると、平均値はその影響を大きく受けます。

外れ値の影響を確かめてみます。データ $X_0 = (55, 55, 57, 58, 59, 59)$ の平均値は以下の様になります。

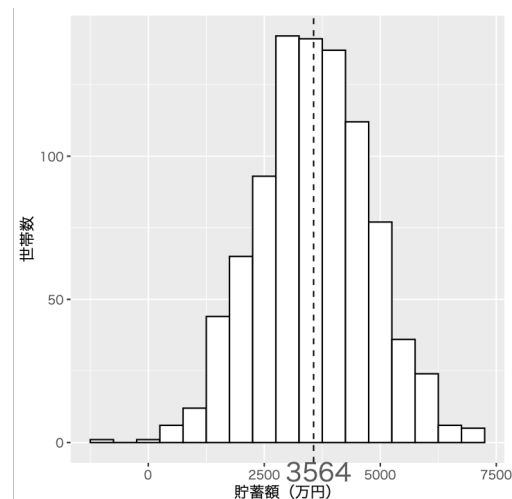
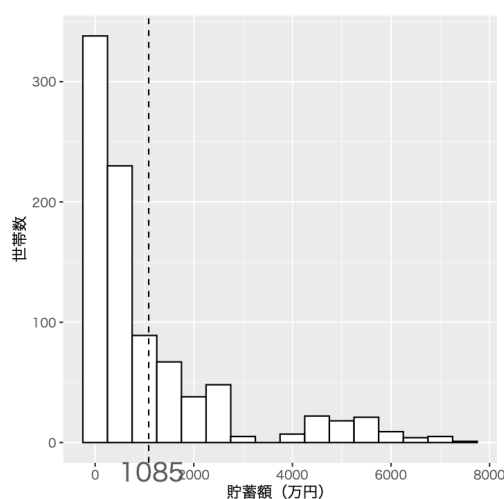
算術平均値 = 57.17, 幾何平均値 = 57.14, 調和平均値 = 57.12

データ X_0 に外れ値として10が含まれたデータ $X_1 = (10, 55, 55, 57, 58, 59, 59)$ の平均値は以下の様になります。

算術平均値 = 50.43, 幾何平均値 = 44.55, 調和平均値 = 34.14

外れ値が1つ含まただけでデータ X_1 の3種類の平均値はいずれもデータ X_0 の区間（範囲？）[55, 59]には含まれない値となっています。これは平均値が外れ値に対して頑健ではないことを示しています。

また、平均的とは「その同類全体の中で最も一般的であるさま。普通程度であるさま。」¹¹を意味しますが、データの分布に偏りがあると平均値は平均的な値とはなりません。



左の図は「平成28年 国民生活基礎調査の概況」（厚生労働省）¹²に掲載されている各種世帯の貯蓄額階級別世帯数の割合の表をベースに仮想的に作成した貯蓄額毎の世帯数の分布をグラフにしたものです。データの分布に偏りがあります。このデータから計算される貯蓄額の平均値¹³は1,085万円となりますが、この貯蓄額が平均的であるということには無理がありそうです。

右の図は左右対称になるように作成した貯蓄額毎の世帯数の分布をグラフにしたものです。このデータから計算される平均値は3,564万円となります。偏りがある分布とは異なり、単峰性¹⁴で偏りのない分布では平均値が平均的な値であることがわかります。

ここで見たように、外れ値が含まれるデータや分布に偏りのあるデータの平均値は代表値としては適切ではない場合があるので注意が必要です。

要約統計量のうちデータの分布の散らばり具合を表す数値には（不偏）分散、標準偏差、範囲、四分位偏差、四分位範囲（など？）があります。データの分布の散らばり具合を表す数値（これらの数値は？）は散布度とも呼ばれます。

要約統計量のうちデータの分布の形状を表す数値には歪度、尖度（など？）があります。

- 分布の中心を表す数値（代表値）
- 代表値
 - 平均値
 - 算術平均、幾何平均、調和平均、調整平均
 - 中央値
 - 最頻値
- 散布度
 - 分散
 - 標準偏差
 - 範囲
 - 四分位偏差
 - 四分位範囲
- 形状

- 歪度
- 尖度

平均

中央値

最頻値

分散

標準偏差

範囲

分位値

歪度

尖度

最大値

最小値

要素数

データの分布を概観する時に役立つのが要約統計量です。

データの分布は適切なグラフ（ヒストグラム？）を描けば視覚的に捉えることができますが、それだけでは定性的な把握に止まってしまいます。定量的に比較したり

尺度水準

また、データは連続なのか、非連続なのかで

離散値（計数データ） - 連続量（計量データ）

離散値: 個数や枚数など数えることができる

連続量: 計る（測る、量る）ことはできるが数えることができない、測る

データの取り得る範囲に対してデータの最小間隔が十分に小さい場合は離散量であっても計量データと見做すこともあります。例えば、人数は離散値ですが、世界の人口を対象として人数の

本来、連続量のデータが計測の都合で離散的に表現されている場合は計量データとして扱う。

例えば、

最頻値: 佐藤

データは表現する性質を基準にして更に分類することができます。この基準のことを尺度と呼び、4つの水準「名義尺度」、「順序尺度」、「間隔尺度」、「比例尺度」に分類されます（尺度水準）。

-
1. 尺度水準 <https://science.sciencemag.org/content/sci/103/2684/677.full.pdf> [↵](#)
 2. データに対して間隔は一定であると言う前提条件を設定することで算術計算に意味を持たせるようにすることもあります。 [↵](#)
 3. ある演算によりAが [↵](#)
 4. 華氏温度をF, 摂氏温度をCとすると $F = \frac{9}{5}C + 32$ となります。 [↵](#)
 5. 単位に依存しない数を無次元数と呼びます。長方形の縦横比（アスペクト比）や比重などが無次元数の例です。 [↵](#)
 6. 摂氏を単位とした場合と華氏を単位とした場合では足し算や割り算の結果も数値は異なりますが、足し算や引き算の結果は無次元量ではなく単位を伴った数値であるため、（那覇市の気温） - （札幌市の気温） = 15°C（30°C - 15°C） = 27°F（86°F - 59°F）という等式が成り立ちます。 [↵](#)
 7. センチメートルで測定した長さをM, インチで測定した長さをFとすると、 $F = \frac{M}{2.54}$ となります。 [↵](#)
 8. 与えられた間隔尺度のデータだけでは基準値が150cmであるとはわからないので、データ同士の差は身長差として意味を持ちますが、データ同士の比は意味を持ちません。 [↵](#)
 9. 与えられた名義尺度のデータだけでは高低とアルファベットの対応の仕方がわからないので、身長によるグループ分けでAに属しているのかBに属しているのかはわかりますが、AとBの高低を比較することはできません。 [↵](#)
 10. xのn乗がyになるとき、xをyのn乗根と言います [↵](#)
 11. 我が国の移動通信トラヒックの現状 [↵](#) [↵](#)
 12. 元データの値442.3Gbpsと一致していない理由は前年度比の数値やGの値を丸めて計算しているためです [↵](#)
 13. ここでの平均値は算術平均値です。貯蓄額が0の世帯もデータに含まれるため幾何平均値や調和平均値は計算できません。 [↵](#)
 14. 単峰性の分布とはピークが1つの分布のことです。 [↵](#)