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とが必要となります．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．

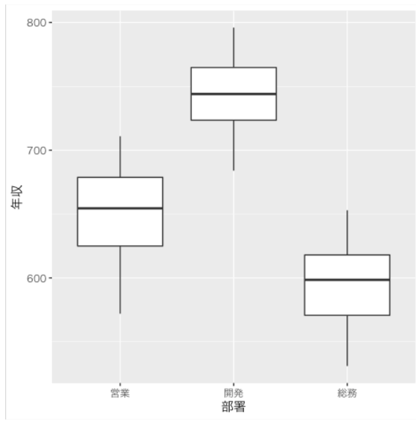
例えば，ある企業の社員100名分の部署と年収がペアになったデータのリストがあったとします．部署と年収のペアはランダムに並んでいるようで，眺めているだけではこのデータからはなにもわかりません．しかし，部署毎に年収をまとめ，箱ひげ図（後述）を描いてみると，部署によって年収が異なっており，開発 > 営業 > 総務の順に高額となっているらしいことが見えてきます．

部署	年収（万円）
営業	580
総務	653
開発	772
開発	792
総務	619
総務	594
開発	687
⋮	⋮
営業	711

部署毎にまとめる

部署	年収（万円）
営業	580
⋮	⋮
営業	711
開発	722
⋮	⋮
開発	687
総務	619
⋮	⋮
総務	653

箱ひげ図を描く



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°C」，「30°C」であるとき，30°C - 15°C = 15°Cと引き算することで札幌市と那覇市の気温差は15°C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．

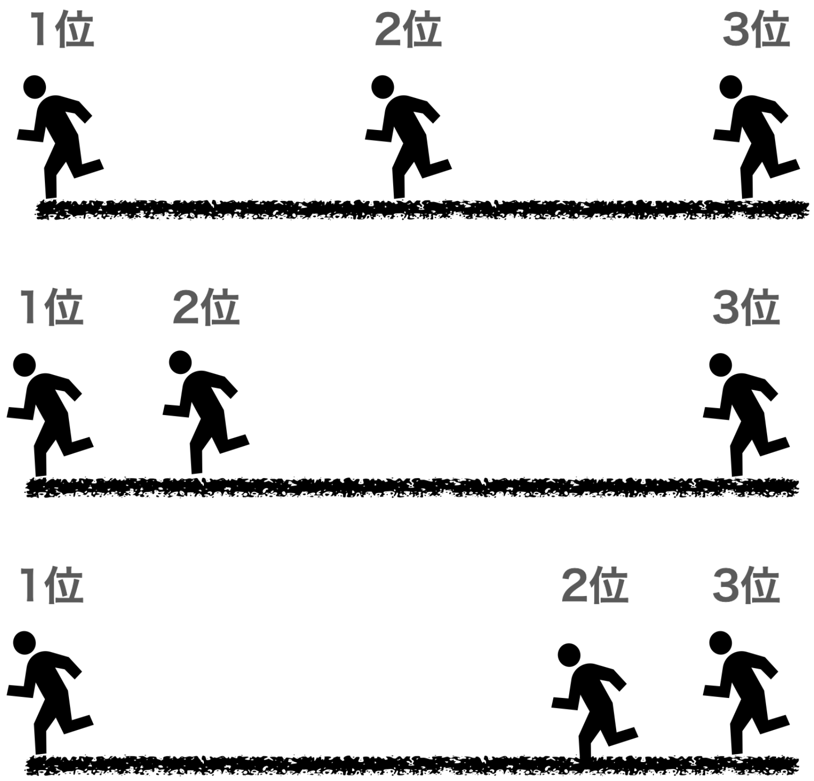
質的データおよび量的データに大別されたデータは表現する情報の性質を基準にして更に分類することができます．この基準のことを尺度と言います．尺度は一般的に「名義尺度」，「順序尺度」，「間隔尺度」，「比例尺度」の4つの水準に分類され[1]，質的データは名義尺度のデータと順序尺度のデータに分けられ，量的データは間隔尺度と比例尺度に分けられます．

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になります．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になります．

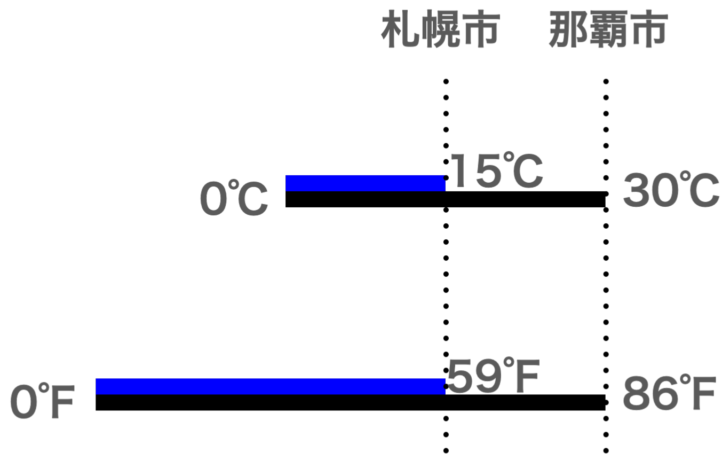
順序尺度とはデータの区別が可能で順序や大小の評価も可能な尺度です．順序尺度のデータは一般的に数字を使って表現されますが，名義尺度と同様，算術計算には意味がありません[2]．順序や大小には意味がありますが，順位や大小の差は，その間隔が一定であるとは限らないため意味を持ちません．

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す．しかし，順位の引き算を行い，(3位 - 2位) = (2位 - 1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3i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になります．



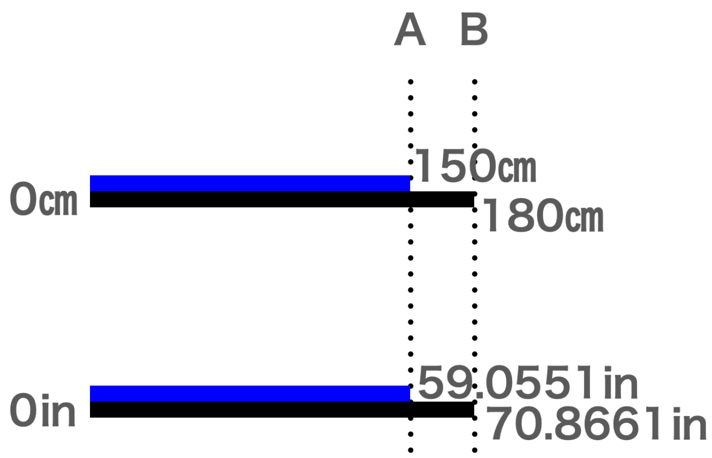
間隔尺度とはデータの順序の評価が可能で差や和の評価も可能な水準です．間隔尺度のデータの取り得る数値の間隔は等間隔で，足し算の結果や引き算の結果は意味を持ちます．しかし，0（ゼロ）が「無，何もない」と言う意味ではないため掛け算の結果や割り算の結果は意味を持ちません．

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°C」，「30°C」であるとき，30°C - 15°C = 15°Cと引き算することで札幌市と那覇市の気温差は15°C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°C（那覇市の気温）/ 15°C（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°F（那覇市の気温）/ 59°F（札幌市の気温） = 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言うことは気温の割り算には意味がない[6]と言うことです．よって，気温は間隔尺度と言うことになります．



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「150㎝」，「180㎝」であるとき，Bさんの身長はAさんの身長の1.2倍（= 180㎝ / 150㎝）と言うことができます．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言うことになります．



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．データ操作の自由度が高い水準はデータ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（0㎝以外のある）基準値，例えば150㎝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0㎝以上は「高い」0㎝未満は「低い」と高低を対応付ければ順序尺度になりますす．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」を対応させるかを決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．

要約統計量

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からは基準値が150㎝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から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． ↩