

2-1: データの記述

データの整理

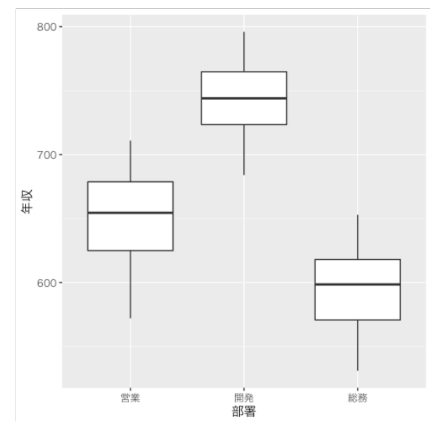
データから有益な情報を引き出すためには、まず最初にデータを体系的な方法に則り整理することが必要となります。整理をしなければデータはただの数字や文字などの羅列であり、そこから傾向や特徴を読み取ることが困難な為です。

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、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。

![[データから情報へ]{#fig:1}

部署	年収 (万円)		部署	年収 (万円)
営業	580		営業	580
総務	653		⋮	⋮
開発	772		営業	711
開発	792	部署毎にまとめる	開発	722
総務	619	➡	⋮	⋮
総務	594		開発	687
開発	687		総務	619
⋮	⋮		⋮	⋮
営業	711		総務	653

➡ グラフを描く



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。

質的データ/ 量的データ

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。

尺度水準

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

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

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。

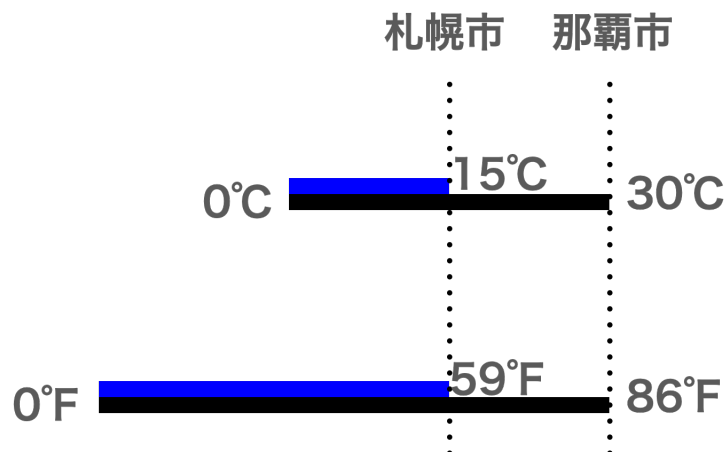
順序尺度のデータの例として100m走の順位を挙げることができます。ある大会でAさんが1位でBさんは2位、Cさんは3位だったとします。このとき、Bさんは、Aさんより遅くCさんより速い言うことはわかります。しかし、順位の引き算を行い、 $(3\text{位} - 2\text{位}) = (2\text{位} - 1\text{位})$ なのでAさんとBさんの走力の差はBさんとCさんの走力の差と同じ、とはなりません。Bさんは1位のAさんと僅差の2位だったのかも知れませんし、あるいは3位のCさんと殆ど同タイムの2位だったのかも知れません。順位の算術計算には意味がないことがわかります。よって、順位は順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



間隔尺度

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。

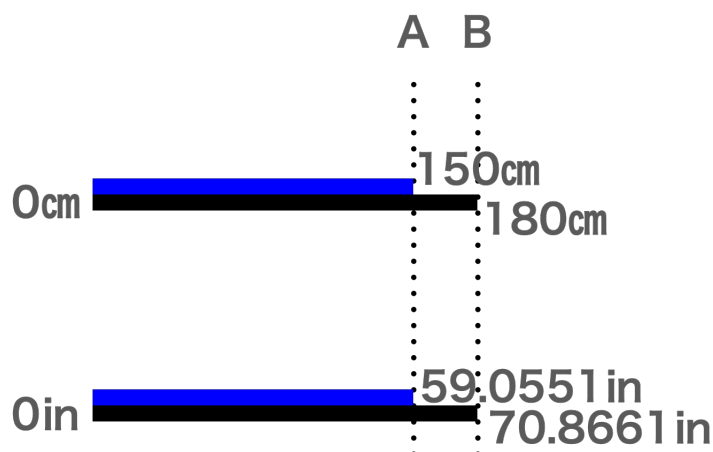
間隔データの例として気温を挙げることができます。ある日の札幌市、那覇市の気温がそれぞれ「15°C」、 「30°C」であるとき、 $30^{\circ}\text{C} - 15^{\circ}\text{C} = 15^{\circ}\text{C}$ と引き算することで札幌市と那覇市の気温差は15°Cであることがわかります。引き算により意味のある結果を得ることができます。引き算の逆演算³である足し算の結果も意味を持ちます。しかし、 30°C （那覇市の気温） \div 15°C （札幌市の気温） $= 2$ となるから那覇市の気温は札幌市の気温の2倍である、とはなりません。この割り算を摂氏ではなく華氏⁴で行うと 86°F （那覇市の気温） \div 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

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。

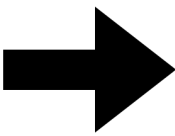
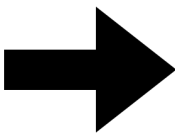
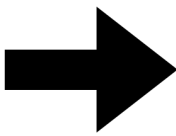
比例尺度の例として身長を挙げることができます。Aさん、Bさんの身長がそれぞれ「150cm」、
「180cm」であるとき、Bさんの身長はAさんの身長の1.2倍（ $= 180\text{cm} / 150\text{cm}$ ）とすることができます。気温の時とは異なり、身長の単位をセンチメートルからインチに替えて⁷計算しても $70.8661\text{in} / 59.0551\text{in} = 1.2$ 倍となります。身長は割り算の結果も意味を持つので比例尺度とすることになります。



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「比例尺度」>「間隔尺度」>「順序尺度」>「名義尺度」となっています。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? されることもあります?）。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でできる値が飛び飛びとなるデータのことです。一方、連続量とは気温や身長のように取ることでできる値が連続しているデータのことです。

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イメージ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。

分布の中心

要約統計量のうちデータの分布の中心的位置を表す統計量には平均値、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統計量は**代表値**とも呼ばれます。

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）を指しますが、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挙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は 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¹²⁾。

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要素の逆数の算術平均値の逆数です。データ要素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。

東京ー福岡900kmを最初の300kmは時速60km/hで、次の300kmは時速120km/hで、最後の300kmは時速150km/hで移動したとします。300kmずつを時速60km/h、時速120km/h、時速150km/hで移動するので東京から福岡までの移動時間を T とすると

$$T = \frac{300}{60} + \frac{300}{120} + \frac{300}{150} = 9.5$$

となります。また、移動速度の算術平均値を A 、調和平均値を H とすると

$$A = \frac{60 + 120 + 150}{3} = 110$$

$$H = \frac{3}{\frac{1}{60} + \frac{1}{120} + \frac{1}{150}} = 94.73\dots$$

となります。

「東京から福岡までの移動速度の平均値は M である」とすれば、東京から福岡までは速度 M で移動し続けたとして東京から福岡までの移動時間 T を計算することができます。この時、 M 、 T には以下の関係が成立します。

$$T = \frac{900}{M}$$

$M = A$ および $M = H$ としてそれぞれ T を計算すると

$$T = \frac{900}{A} = \frac{900}{110} = 8.18\dots$$

$$T = \frac{900}{H} = \frac{900}{94.73\dots} = 9.5$$

となります。東京から福岡までの移動時間は9.5時間なので算術平均値を速度の平均値としてしまうと誤った数値を導いてしまいます。この場合の平均値は調和平均値が妥当であることがわかります。

算術平均値、幾何平均値、調和平均値のいずれの計算にも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 $X_0 = (55, 56, 57, 58, 59, 60)$ の算術平均値 A 、幾何平均値 G 、調和平均値 H は以下のようになります。

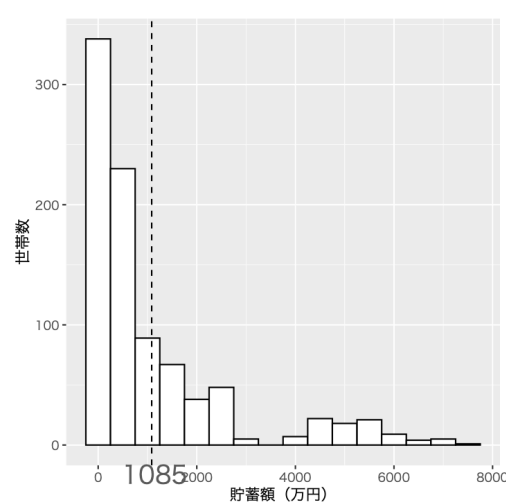
$$A = 57.5, G = 57.47\ldots, H = 57.44\ldots$$

データ X_0 に外れ値として10が含まれたデータ $X_1 = (10, 55, 56, 57, 58, 59, 60)$ の A 、 G 、 H は以下のようになります。

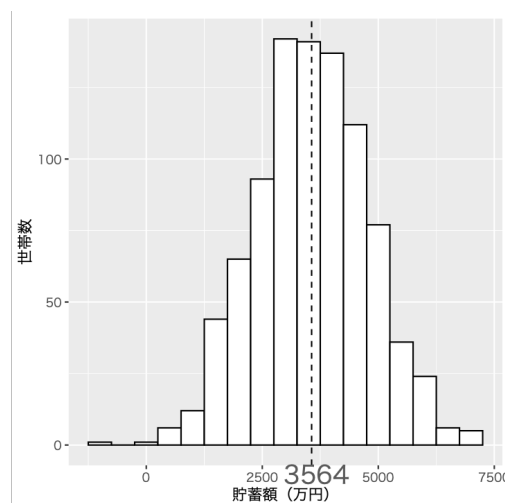
$$A = 50.71\ldots, G = 44.76\ldots, H = 34.23\ldots$$

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, 60] 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。」¹³ 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。



偏りのある分布



対称な分布

左の図は「平成28年 国民生活基礎調査の概況」（厚生労働省）¹⁴ 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¹⁵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¹⁶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。

中央値

中央値とはデータを昇順あるいは降順に並べた時、中央になる値です。データの要素数が無限の場合、中央値は存在しません。

データの要素数を N 、データを昇順に並べた時の i 番目の要素を a_i とすると中央値 Med は以下の式で求められます。

$$Med = \begin{cases} a_{\frac{N+1}{2}} & (N: \text{奇数}) \\ \frac{a_{\frac{N}{2}} + a_{\frac{N}{2}+1}}{2} & (N: \text{偶数}) \end{cases}$$

データの要素数が奇数の時はデータを昇順に並べた時の中央の要素の値、データの要素数が偶数の場合はデータを昇順に並べた時の中央に近い2つの要素の値の算術平均値となります。データを昇順に並べるためデータの要素間で大小の評価が必要となります。よって、中央値は順序尺度以上のデータに対して意味を持ちます。

上の中央値を求める式からも判るように、中央値（の値?）は中央の1つの要素あるいは中央に近い2つの要素のみから影響を受けます。そのため平均値に比べ中央値は外れ値の影響を受けにくい代表値となっています。

先に平均値における外れ値の影響を確かめた際に使ったデータ X_0 とデータ X_1 についてそれぞれの中央値を Med_0 、 Med_1 とすると

$$Med_0 = \frac{57 + 58}{2} = 57.5$$
$$Med_1 = 57$$

となります。中央値は平均値と比べ外れ値の影響が少なく、外れ値に対して頑強であることが判ります。

一方で、1つあるいは2つの要素の値という限られた値のみを利用しているため、データ全体の変化を中央値では捉えられないこともあります。

ある企業の部署毎の売り上げを昇順にしたデータが前年度は(10億円, 15億円, 30億円, 35億円, 40億円)で今年度は(15億円, 20億円, 25億円, 45億円, 50億円)だったとします。前年度の売上総額を S_0 、算術平均値を A_0 、中央値を Med_0 、今年度の売上総額を S_1 、算術平均値を A_1 、中央値を Med_1 とすると、

$$S_0 = 130, A_0 = 26, Med_0 = 30$$
$$S_1 = 155, A_1 = 31, Med_1 = 25$$

となります。前年度に比べ今年度の売上総額および算術平均値は増えていますが¹⁷、中央値だけは減っています。中央値の変化だけを見ていると前年度に比べ今年度の売り上げが約1.2倍に伸びたことに気づかないだけでなく、逆に業績が悪化したと判断してしまいかねません。

データのうち1つまたは2つの要素の値しか反映しないという中央値の特徴が弱みとなった例です。

最頻値

最頻値とはデータの中で出現する回数が最も多い値です。値同士の区別ができればその値の出現回数を数えることができるので、最頻値を求めるために算術計算や大小評価をする必要はありません。従って、最頻値はすべての尺度水準のデータに対して意味を持ちます。

最頻値は名義尺度のデータに対しては唯一の有効な代表値となります。日本国内に住民票のある居住者全員の住所の都道府県名のデータは(北海道, 東京, 東京, 大阪, …)と言うようになります。このデータは名義尺度であるため算術計算をすることができず平均値を求めることができません。また、大小の評価をすることもできないので中央値を求めることもできません。しかし、「北海道と東京は別の値である」と区別をすることはできるためデータの中で最も出現回数の多い値が「東京」であると決定することができます。最頻値は求めることができます。

最頻値は最も出現回数の多い値だけに依存して決定されます。また、外れ値の出現回数が他の値の出現回数よりも多く、外れ値が最頻値となることは非常に特殊なケースを除いてはありません¹⁸。従って、最頻値は殆どの場合、外れ値の影響は受けず、外れ値に対して頑強であると言えます。

また、先に平均値におけるデータの分布の偏りの影響を確かめた貯蓄額毎の世帯数分布のように極端に分布に偏りのあるデータに対しては、最頻値が代表値として妥当である場合があります。

最頻値は一意に定まるとは限りません。ある政党に対する支持を問うたアンケートに対する120名の回答が以下の様だったとします。この場合、最頻値は「強く支持する」と「どちらとも言えない」の2つとなっています。

1つのデータに対して平均値や中央値が2つ以上存在するということはありませんが、最頻値はこのように2つ以上存在することもあります。このことは最頻値の代表値としてのデメリットと言えます。

	回答数
強く支持する	30
支持する	10
どちらとも言えない	30
支持しない	25
まったく支持しない	25

また、このアンケート結果から政党を支持しているのかあるいは支持していないのかをはっきりさせるため、「強く支持する」の回答数と「支持する」の回答数をまとめ、新たに「支持する」の回答数とし、「まったく支持しない」の回答数と「支持しない」の回答数をまとめ、新たに「支持しない」の回答数として結果をまとめ直すと以下の様になります。

	回答数
支持する	40
どちらとも言えない	30
支持しない	50

まとめ直したアンケート結果の最頻値は「支持しない」となり、元のアンケート結果とは異なる印象を与えます。最頻値は度数分布の階級幅に大きく影響を受けることがわかります。このことも最頻値の代表値としてのデメリットと言えます。

適用可能な代表値と尺度水準の対応は以下のようになります。

	名義尺度	順序尺度	間隔尺度	比例尺度
算術平均値	—	—	○	○
幾何平均値	—	—	—	○
調和平均値	—	—	—	○
中央値	—	○	○	○
最頻値	○	○	○	○

平均値はデータが有している情報をすべて反映していますが、中央値や最頻値はデータが有している情報を切り捨てています。データが有している情報を反映している程度は平均値 > 中央値 > 最頻値となっています。従って、原則的には代表値としては平均値を使うことが好ましいと言えます。

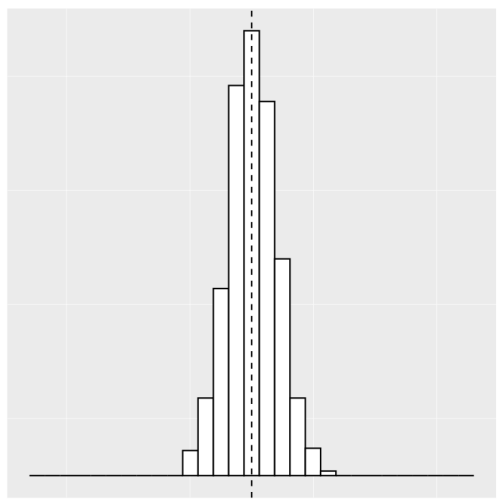
しかし、対象とするデータの尺度水準に適用可能な代表値でなければ、その代表値は意味を持ちません。また、ここで述べたように平均値、中央値、最頻値にはそれぞれメリット、デメリットがあります。データに対して代表値使う時には、これらを踏まえて使用することが重要です。

分布の散らばり

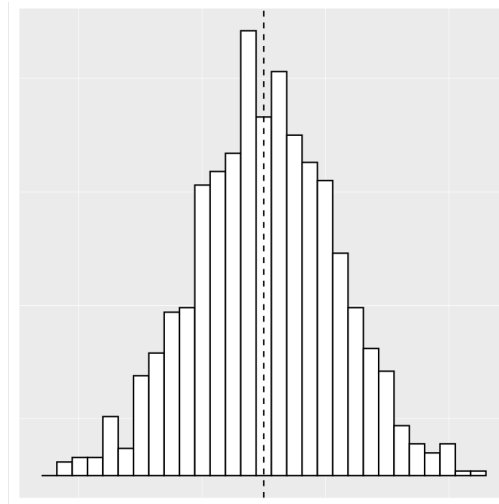
要約統計量のうちデータの分布の散らばり具合を表す統計量には分散、標準偏差、範囲、四分位範囲などがあります。データの分布の散らばり具合を表すこれらの統計量は**散布度**とも呼ばれます。

分散

分散とはデータの分布が算術平均値に集中している程度により散らばり具合を指標化した数値です。具体的にはデータの算術平均値とデータ要素の距離の二乗の平均値として定義されます。よって、それぞれのデータ要素が平均値から離れているほど分散は大きくなります。従って、分散は平均値への集中度合いを表す散布度となります。



分散の小さい分布



分散の大きい分布

要素数が n 個のデータ (x_1, \dots, x_n) からなる母集団の算術平均値を μ とすると、この母集団の分散である**母分散** σ^2 は以下の式で求められます。

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

要素数が n 個のデータ (x_1, \dots, x_n) からなる標本の算術平均値を \bar{x} とすると、この標本の分散である**標本分散** s^2 は以下の式で求められます。

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

多くの場合、データ分析では母集団の性質を知ることが目的となりますが、母集団のデータをすべて入手できることは殆どありません。そのため、母分散 σ^2 も直接計算することはできません。しかし、標本から母分散 σ^2 を推定することができます。

この標本に基づき推定される母分散 σ^2 の推定量は**不偏分散**と呼ばれ、一般的に標本分散 s^2 よりも若干小さな値になることが知られています。

要素数が n 個のデータ (x_1, \dots, x_n) からなる標本の算術平均値を \bar{x} とすると、不偏分散 $\hat{\sigma}^2$ は以下の式で求められます¹⁹。

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

標本分散 s^2 と不偏分散 $\hat{\sigma}^2$ の計算式から判るように、データのサンプルサイズ n が大きいほど、標本分散 s^2 は不偏分散 $\hat{\sigma}^2$ に近づきます。

あるクラスを対象に数学のテストを実施し、全受験者の中からランダムに選んだ5名の点数が(60, 65, 70, 80, 90)だったとします。この5名の点数の算術平均値を \bar{x} 、標本分散を s^2 、今回のテストの点数の不偏分散 $\hat{\sigma}^2$ とすれば

$$\bar{x} = \frac{60 + 65 + 70 + 80 + 90}{5} = 73$$

$$s^2 = \frac{1}{5} \{ (60 - 73)^2 + (65 - 73)^2 + (70 - 73)^2 + (80 - 73)^2 + (90 - 73)^2 \} = 116$$

$$\hat{\sigma}^2 = \frac{1}{5-1} \{ (60 - 73)^2 + (65 - 73)^2 + (70 - 73)^2 + (80 - 73)^2 + (90 - 73)^2 \} = 145$$

となります。

また、標本に基づき推定される推定量の期待値が母集団のそれに等しい時、この推定量は不偏推定量と呼ばれますが、不偏分散の期待値は母分散に一致するので分散の不偏推定量となっています。即ち、母集団から標本抽出を k 回行い、 i 回目の標本から計算される不偏分散を $\hat{\sigma}_i^2$ とすると、

$$\sigma^2 = \lim_{k \rightarrow \infty} \left(\frac{1}{k} \sum_{i=1}^k \hat{\sigma}_i^2 \right)$$

となります²⁰。

データ分析では母分散や標本分散よりも頻繁に不偏分散を扱います、このためデータ分析において分散とは一般的に不偏分散を意味します。

標準偏差

標準偏差とは分散の正の平方根として計算される散布度です。データが母集団の時の標準偏差を**母標準偏差**、データが標本の時の標準偏差を**標本標準偏差**呼びます。

要素数が n 個のデータ (x_1, \dots, x_n) からなる母集団の算術平均値を μ とすると、この母集団の母標準偏差 σ は以下の式で求められます。

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

要素数が n 個のデータ (x_1, \dots, x_n) からなる標本の算術平均値を \bar{x} とすると、この標本の標本標準偏差 s は以下の式で求められます。

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

母分散の不偏推定量である不偏分散 $\hat{\sigma}^2$ の平方根である $\hat{\sigma}$ を不偏標準偏差と呼びます（呼ぶことがあります?）。 $\hat{\sigma}$ の期待値は母集団の標準偏差と等しいとは限らず、 σ は母集団の標準偏差の不偏推定量ではありませんが、母集団の標準偏差の推定量として利用されます。

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

分散も標準偏差も平均値を基準にしたデータ分布の散らばり具合を表していますが、分散は平均値と単位が異なります。他方、標準偏差は平均値と単位が同じになります。例えば、cmを単位として測定したデータの平均値と標準偏差の単位はcmですが、分散の単位は cm^2 となります。

このため平均値と標準偏差は直接比較することができますが、平均値と分散を直接比較することは自然ではありません。データ分布を代表値と散布度で記述する場合も代表値が平均値の時は、散布度として分散でなく、標準偏差を採用することが一般的です。

レンジ

レンジとはデータの存在している範囲によりデータの散らばり具合を指標化した数値です。データの要素の最大値が x_{max} 、最小値が x_{min} である時、このデータのレンジ R は以下の式で求められます。

$$R = x_{max} - x_{min}$$

データに外れ値が含まれている場合、レンジは外れ値の影響を受けます。

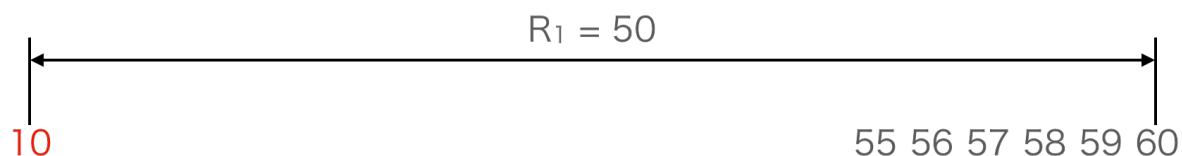
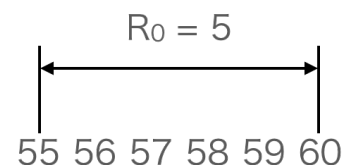
データに外れ値が含まれていれば、（非常に特殊な場合を除き？）データの要素の最大値あるいは最小値のうち少なくともどちらかは外れ値です。よって、レンジは外れ値の影響を直接受けます。

先に平均値における外れ値の影響を確かめた際に使ったデータ X_0 とデータ X_1 についてそれぞれのレンジを R_0 、 R_1 とすると

$$R_0 = 60 - 55 = 5$$

$$R_1 = 60 - 10 = 50$$

となります。レンジが外れ値の影響を受けることが確認できます。



レンジはデータの要素の最大値と最小値の2つの値により決定され、この2つの値以外からは一切影響を受けません。よって、サンプルサイズが変化してもレンジに対する外れ値の影響は変化しません。一方、データのサンプルサイズが計算式に含まれる分散や標準偏差はサンプルサイズが大きくなると外れ値の影響は小さくなります。

このため、レンジは分散や標準偏差よりも外れ値に対して敏感です。

四分位範囲

四分位範囲とはレンジと同様、データの存在している範囲によりデータの散らばり具合を指標化した数値です。レンジは全データの存在する範囲として最大値と最小値から求められますが、四分位範囲はデータの中央部の約50%が存在する範囲として四分位数を使って求められます。

四分位数とはデータを昇順に並べ4等分にした時、その4等分した位置にある値のことです。小さい方から第1四分位数、第2四分位数、第3四分位数と言います。第2四分位数は中央値と等しくなります。

第1四分位数を Q_1 、第3四分位数を Q_3 とすると四分位範囲は以下の式で求められます。

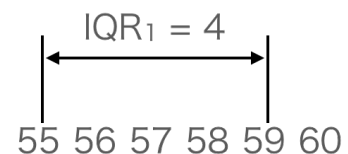
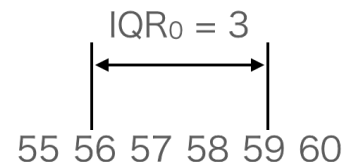
$$IQR = Q_3 - Q_1$$

データに外れ値が含まれていれば、外れ値は（非常に特殊な場合を除き？）最小値から第1四分位数の間または第3四分位数から最大値の間に存在しています。しかし、（レンジは最大値と最小値を使って求められましたが、？）四分位範囲は最小値から第1四分位数までのデータと第3四分位数から最大値までのデータを切り捨てて求められます。従って、四分位範囲は外れ値に対して頑強です。

先に平均値における外れ値の影響を確かめた際に使ったデータ X_0 とデータ X_1 についてそれぞれの第1四分位数を Q_{1_0} 、 Q_{1_1} 、第3四分位数を Q_{3_0} 、 Q_{3_1} 、四分位範囲を IQR_0 、 IQR_1 とすると

$$\begin{aligned} Q_{1_0} &= 56, Q_{3_0} = 59 \\ IQR_0 &= Q_{3_0} - Q_{1_0} = 59 - 56 = 3 \\ Q_{1_1} &= 55, Q_{3_1} = 59 \\ IQR_1 &= Q_{3_1} - Q_{1_1} = 59 - 55 = 4 \end{aligned}$$

となります。四分位範囲がレンジとは異なり、外れ値に対して頑強であることが確認できます。



10

データ分布を代表値と散布度で記述する場合に代表値を中央値とした時は散布度として標準偏差ではなく四分位範囲を採用することが一般的です。標準偏差は平均値に対してのデータの散らばりを指標化した散布度なので、中央値とペアで使うことが不自然なためです。

四分位数の具体的な求め方については複数の手法が提唱されていますが²¹、どの手法が最も優れているかについては合意が形成されていません。以下では中央値を利用した四分位数の求め方について説明します。

1. データを昇順に並べ中央値を求めます。この中央値が第2四分位数 Q_2 となります。
2. データを Q_2 より小さい部分と大きい部分に分けます。
3. 小さい部分の中央値を第1四分位数 Q_1 、大きい部分の中央値を第3四分位数 Q_3 とします。

先程の四分位範囲における外れ値の影響を確かめた例ではこの手法で第1四分位数および第3四分位数を求めています。

散布度は間隔尺度以上のデータに対して意味を持ちます。

分布の形状

要約統計量のうちデータの分布の形状を表す統計量には歪度、尖度があります。

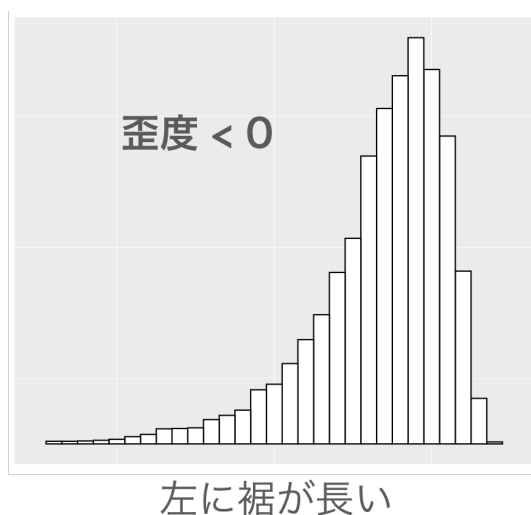
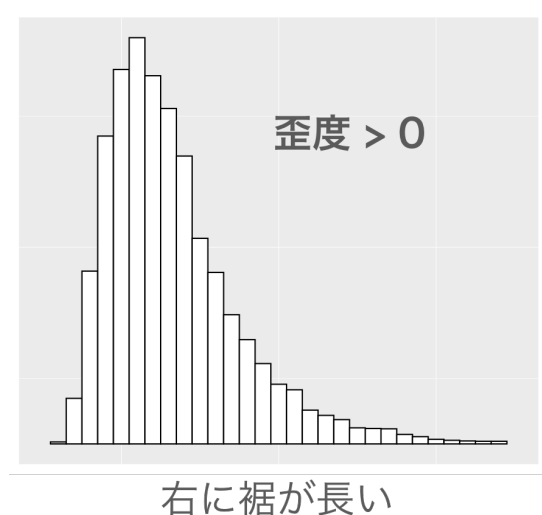
歪度

歪度とは分布の歪み具合を指標化した統計量です。要素数が n 個のデータ (x_1, \dots, x_n) からなる標本の算術平均値を \bar{x} 、不偏標準偏差を $\hat{\sigma}$ とすると、母集団の歪度の推定量 b_1 は以下の式で求められます。ただし、 b_1 は不偏推定量ではありません。

$$b_1 = \frac{m_3}{\hat{\sigma}^3}$$

ただし、 $m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$ とする

歪度が正となる時、分布は右に裾の長い形状となります。歪度が負となる時、分布は左に裾の長い形状となります。



尖度

尖度とは分布の裾の重さを指標化した統計量です。要素数が n 個のデータ (x_1, \dots, x_n) からなる標本の算術平均値を \bar{x} 、不偏標準偏差を $\hat{\sigma}$ とすると、母集団の尖度の推定量 b_2 は以下の式で求められます。ただし、 b_2 は不偏推定量ではありません。

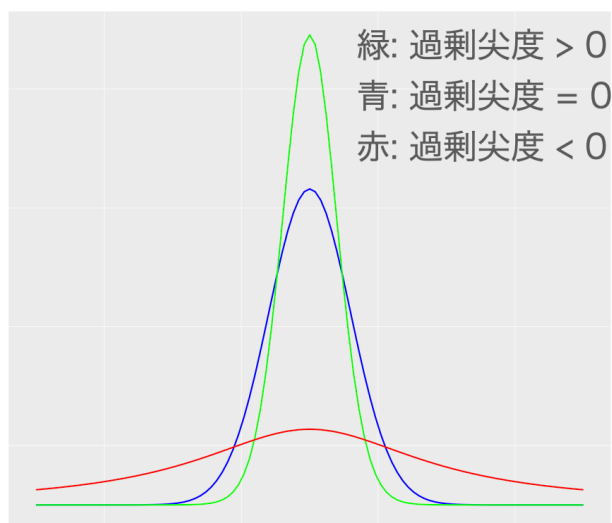
$$b_2 = \frac{m_4}{\hat{\sigma}^4}$$

ただし、 $m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$ とする

尖度 b_2 はデータ分布が正規分布²²に従う時、 $b_2 = 3$ となるので、正規分布を基準とした評価を行いたい時には以下の式で与えられる g_2 を尖度として使います。 g_2 は過剰尖度と呼ばれます。

$$g_2 = b_2 - 3$$

過剰尖度が正となる時、分布の裾は正規分布の裾よりも軽くなります。過剰尖度が負となる時、分布の裾は正規分布の裾よりも軽くなります。分布が正規分布に従う時は過剰尖度は0になります。



上図では青色の曲線が正規分布です。過剰尖度を調べることで極端な値を取るデータの割合が、正規分布と比べて大きいかを判断することができます。

歪度および尖度は間隔尺度以上のデータに対して意味を持ちます。

関係を表現する

変数間の関係を調べることは、データ分析の大きな目的です。変数間の関係の表現方法は、その変数がどの尺度水準なのかによって変わってきます。

分割表

分割表とは、質的変数間の関係を分析する際に、その前段階として作成される変数間の関係を集計した表のことです。 n 個の変数を対象とすれば n 次元の分割表が作成されます。

3次元以上は紙面（平面?）での表現が困難なため、ここでは2つの変数を対象とした2次元分割表について説明します。

関心のある2つの変数のうち、一方の変数の値を行名、もう一方の変数の値を列名にします。行に対応させた変数 A のデータ中に現れた値が (A_1, A_2, \dots, A_r) の r 種類、列に対応させた変数 B のデータ中に現れた値が (B_1, B_2, \dots, B_s) の s 種類ならば分割表は以下の様に $(r + 1)$ 行 \times $(s + 1)$ 列なります。この表は $r \times s$ 分割表と呼ばれます。

i 行 j 列のセルには、変数 A の値が A_i で変数 B の値が B_j となっているデータ（サンプル?）の出現回数を記入します。最下行には各列の合計値を、最右行には各行の合計値を記入します。表の右下に記入される数値はデータのサンプルサイズとなっています。

	B ₁	B ₂	...	B _s	計
A ₁	N ₁₁	N ₁₂	...	N _{1s}	N _{1.}
A ₂	N ₁₂	N ₂₂	...	N _{2s}	N _{2.}
⋮	⋮	⋮	⋮	⋮	⋮
A _r	N _{r1}	N _{r2}	...	N _{rs}	N _{r.}
計	N _{.1}	N _{.2}	...	N _{.s}	N

仮に想定したある大学の今年度の学部卒業生のデータを例にして分割表の作成の仕方を見てみます。

学籍番号 学部 進路 出身

100020230 工学部 就職 愛知県

100030447 文学部 不明 大阪府

100040268 理学部 進学 東京都

⋮ ⋮ ⋮ ⋮

このデータから卒業学部と進路を変数とした分割表は以下の手順で作成します。

1. データ中に現れる学部の値と進路の値を抽出します。今回は学部の値が(理学部, 工学部, 経済学部, 文学部)の4種類, 進路の値が(進学, 就職, 不明)の3種類たったとします。
2. 5×4 の表を作成し行名を「理学部」, 「工学部」, 「経済学部」, 「文学部」, 列名を「進学」, 「就職」, 「不明」とします。
3. 行名と列名のクロスしたセルには, データからその行名と列名の組み合わせに該当する卒業者をカウントし, その数を記入します。学部が理学部で進路が就職となっているデータ (サンプル?) が80件であれば, セル [理学部 - 就職] には80と記入します。
4. 最下行には列ごとの合計値を, 最右行には行ごとの合計値を記入します。

	進学	就職	不明	計
理学部	123	80	3	206
工学部	152	146	2	300
法学部	26	147	7	180
文学部	15	154	13	182
計	316	527	25	868

手作業での分割表の作成はかなり大変な作業となりますが, RやPythonであれば1つの関数 (コマンド?) で作成することができます。

この分割表から理系学部における進学割合と文系学部における進学割合には差がありそうだと
言うことがわかりますが、実際に学部の違いが進学割合に影響を与えているかを客観的に判断するた
めには検定をする必要があります²³。

相関係数

2-2: データの可視化

棒グラフ

X軸が名義尺度または順序尺度

ヒストグラム

散布図

箱ひげ図

折れ線グラフ

ネットワークグラフ

ヒートマップ

-
1. 尺度水準 [↗](#)
 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FによりAがBになるとき、BをAにするような演算GのことをGをFの逆演算と言います。 [↗](#)
 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。数値を丸めずに計算すれば元データの値と一致します。 [↗](#)
 13. 三省堂大辞林第三版 [↗](#)

14. 「平成28年 国民生活基礎調査の概況」（厚生労働省） [↩](#)

15. 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。 [↩](#)

16. 単峰性の分布とはピークが1つの分布のことです。 [↩](#)

17. 幾何平均値も調和平均値も前年度に比べ今年度の値は大きくなります。 [↩](#)

18. データの要素数が極端に少ない場合には外れ値が最も多く出現するケースも皆無ではありません。 [↩](#)

19. 不偏分散の計算式の妥当性は数学的に導くことができます。参考文献 [↩](#)

20. 不偏分散の期待値が母分散と一致することも数学的に導かれます。 [↩](#)

21. [Quartile](#) [↩](#)

22. 正規分布は社会科学分野でも自然科学分野でも頻繁に用いられる確率分布です。詳細は○章○節参照。 [↩](#)

23. χ^2 検定（カイ二乗検定）など。○章○節 [↩](#)