MAT220 第2回講義：データの可視化

Contents

1 提示としてのデータ可視化 **2**

1.1 提示は科学というよりも芸術である 2

2 量的データの表現 **3**

2.1 度数表とヒストグラム 4

2.2 相対度数分布 5

2.3 ヒストグラムの悪用：ビン幅の変更 5

2.4 分布の形 6

2.5 ヒストグラムから分かること 8

3 カテゴリーデータの表現 **8**

3.1 棒グラフ 8

3.2 例：アメリカ独立宣言 9

3.3 円グラフ 10

4 時間依存データの表現 **11**

4.1 線形変化 12

4.2 非線形変化 12

4.3 尺度と物語 13

5 対になったデータの表現 **14**

5.1 例：通学時間と距離 14

5.2 相関 15

5.3 相関は因果関係ではない 16

6 ケーススタディ：日本のコンビニ **16**

6.1 地図による表現 17

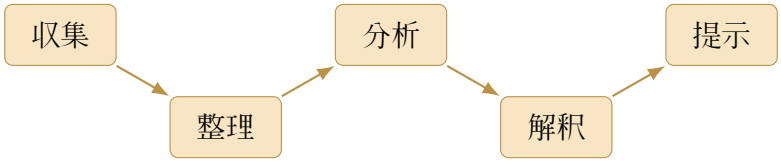
6.2 色による表現 18

6.3 データ可視化に関するいくつかの観察 21

第1回講義では、統計学の基本的な言葉を導入しました。すなわち、データ、変数、母集団、標本、そして統計的研究の流れです。この講義では、データの視覚的な提示に焦点を当てることで、記述統計の議論を始めます。この段階では、まだ公式は導入していません。その代わりに、今日はコミュニケーションという考え方を学びます。つまり、データセットを、他の人が素早く眺めて理解できるような視覚的な要約に変える方法です。前回の講義の重要な点と同じく、異なるデータセットは異なる構造を持つことがあり、そのため「正しい」提示方法はその構造に依存します。

1　提示としてのデータ可視化

前回の講義では、基本的なレベルでは、統計的研究を次の5つの重要な要素に分けられることを見ました。



私たちは、この流れが依存関係の鎖のようなものだと強調しました。どれか一つの箱にそれぞれの仕方で欠陥があると、それは「妥当性の鎖を壊す」ようなものであり、研究の結果は信頼できなくなります。例えば、そもそも不正確なデータや偏ったデータを収集してしまえば、どれほど統計分析が優れていても、その結果は最初のデータと同じくらい信頼できません。あるいは、正確なデータを偏りなく収集しても、それに対して誤った公式を適用してしまえば、研究全体は疑わしいものになります。

この授業の後半では、統計的研究の分析と解釈の段階について多くの時間をかけて議論します。これは統計的実践において非常に重要な要素です。どの公式を正しく適用すべきかを知るだけでなく、それらの公式がなぜ正しいのか、そして何を意味するのかを理解する必要があるからです。しかし、今日の講義では、こうしたより深い考察はすでに確立されているものと仮定し、研究過程の最後の段階である提示へと進みます。

定義

データの可視化とは、表、グラフ、チャート、図、地図などの視覚的対象を用いてデータを提示することである。 その目的は、データに含まれる重要な構造を見やすくすることである。

もちろん、統計的研究には非常に深い洞察を与える可能性があります。しかし、私たちの洞察は、それをどのように表現するかに大きく左右されます。データの提示が悪ければ、そこにある教訓や物語は見つけられないかもしれません。言い換えれば、正確なデータであっても、その中にあるパターンを明確に見られなければ役に立ちません。

1.1　提示は科学というよりも芸術である

一般に、データセットを可視化する完璧な方法はありません。これから見るように、異なるデータセットには多くの異なる表現方法があります。したがって、データを提示する私たち

は、データのどの要素を強調したいのかについて、情報に基づいた選択をしなければなりません。異なる提示方法は聞き手に異なる効果を与え、そのため少し異なるメッセージを伝えます。この理由から、データの提示は厳密な科学というよりも芸術に近いものです。最良のデータ提示を作るために適用できる「公式」はなく、代わりに実践的な判断と美的感覚を組み合わせる必要があります。これから見るように、適切な組み合わせは、しばしばデータ、文脈、そして聞き手に依存します。

重要な点

データの提示は、科学というよりも芸術である。 グラフは単なる中立的な情報の入れ物ではない。むしろ、提示者が途中で行った選択を含んでおり、その選択は聞き手が何に気づくかに影響を与える。

もちろん、これは何でもありという意味ではありません。これから見るように、グラフは本当に悪かったり、見苦しかったり、誤解を招いたり、あるいは明らかに不誠実だったりすることがあります。今日の講義を通して、良いデータ提示のための重要な指針を少しずつ確認していきます。その過程で、データの基本的な表現方法をいくつか概観します。扱う内容は次の通りです。

1. 量的データ、
2. カテゴリーデータ、
3. 時間依存データ、
4. 対になったデータ。

最後に、この講義のために用意したケーススタディで締めくくります。同じデータをいくつかの異なる方法で提示できることを確認しましょう。

2 量的データの表現

データセットは、算術が意味を持つ数値的な測定値や個数から成るとき、「量的」と呼ばれます。数値データには数学的構造（順序、算術など）があるため、データを視覚的に表現するときには、この追加の構造を意識すると役立つことが多いです。

一般に、大量の量的データをそのまま直接提示することは、あまり有用ではありません。これを説明するために、次のデータ表を考えましょう。これは、42人の靴のサイズをセンチメートルで測定したものです。

42人の靴のサイズ					
25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

上の表は単に生データを表示しているだけです。この提示方法は非常に開かれていて正直です

が、データの有用なパターンを簡単に見ることができないため、特に洞察的ではありません。例えば、数値だけを見ても、次のような有用な問いに素早く答えることはできません。

- どの靴のサイズが最も多いか。
- この集団の靴のサイズはおおよそどのように分布しているか。
- 外れ値はあるか。

これらの問いの答えは、データの中のどこかには含まれていますが、隠れています。したがって、こうした隠れた構造を別の表現によって明らかにする方がはるかに良いでしょう。

2.1 度数表とヒストグラム

量的データを扱うときには、データを同じ大きさの区間に分けると便利なことがよくあります。これらの区間はしばしば階級と呼ばれ、より一般にはビンとも呼ばれます。このグループ分けから、下に示すような度数表を作ることができます。分かりやすさのために、同じ靴のサイズの生データを、そのグループ化された度数表と並べて再掲します。

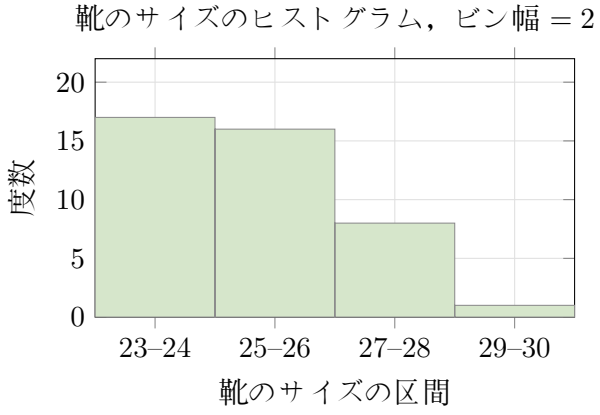
42人の靴のサイズ					
25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

→

階級区間	度数
23–24	17
25–26	16
27–28	8
29–30	1

ここでは、42個のデータ点を幅2のビンにまとめています。これは、データ全体にわたる長さ2の区間です。見て分かるように、度数表はすでに改善になっています。なぜなら、多くのデータが23から26の範囲に集まっていることがすぐに分かるからです。

次に、度数表の情報を使ってグラフを描きます。x軸には区間を表示し、y軸には個数を表示します。この種類の図はヒストグラムと呼ばれます。



定義：ヒストグラム

ヒストグラムとは、量的データをビンと呼ばれる区間に分けて表すグラフである。各棒の高さは、そのビンに入るデータ値の個数を表す。

ヒストグラムは、度数表に含まれているものと同じ基本情報を表示します。しかし、度数どうしの関係を素早く見ることができるため、その情報ははるかに見やすくなります。

2.2 相対度数分布

度数とは、あるビンにいくつの観測値が入るかを表すものです。しかし、ときには生の度数そのものよりも割合の方が重要になることがあります。幸い、こうした割合は簡単な手順で計算できます。

定義：相対度数

ある階級の**相対度数**は

$$\text{相対度数} = \frac{\text{ビンの度数}}{\text{全体の個数}}.$$

これは、そのビンの中にデータセットのどの割合が含まれているかを表す。

例えば、靴のサイズのデータでは、最初の階級の度数は17であり、全体では42人います。したがって、その相対度数は

$$\frac{17}{42} \approx 0.405 = 40.5\%.$$

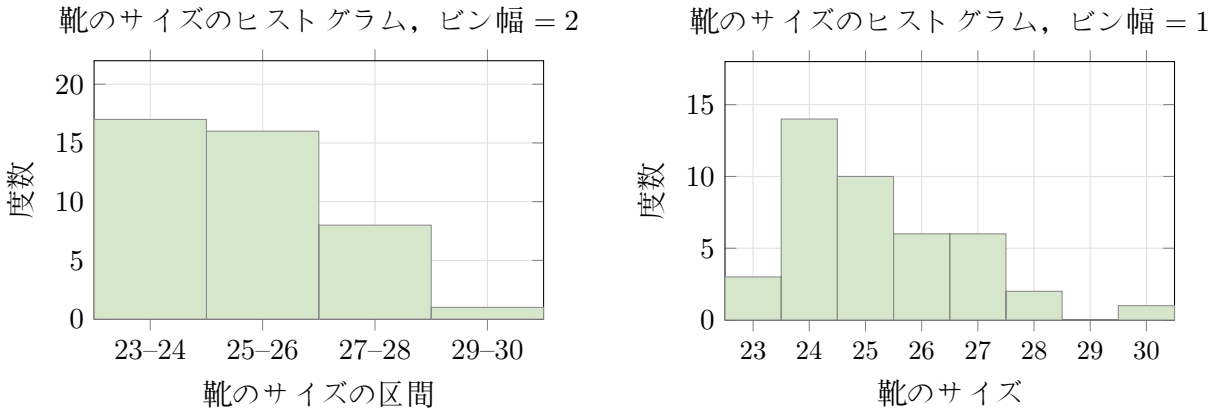
下の表は、相対度数表と呼ばれるものです。

階級区間	度数	相対度数
23–24	17	$17/42 \approx 40.5\%$
25–26	16	$16/42 \approx 38.1\%$
27–28	8	$8/42 \approx 19.0\%$
29–30	1	$1/42 \approx 2.4\%$

相対度数は、大きさの異なるデータセットを比較するときに役立ちます。 例えば、あるクラスに42人の学生がいて、別のクラスに120人の学生がいる場合、生の個数は直接比較しにくいことがあります。 割合を使うと、比較がより明確になることが多いです。

2.3 ヒストグラムの悪用：ビン幅の変更

ヒストグラムは便利ですが、とても操作しやすいものでもあります。その主な方法はビン幅を変えることです。ビン幅を変えると、データが異なるビンに振り分けられ、異なる形が現れることがあります。例えば、靴のサイズのデータは、幅1のビンに数えることもできます。その結果、かなり違った見た目のヒストグラムになります。



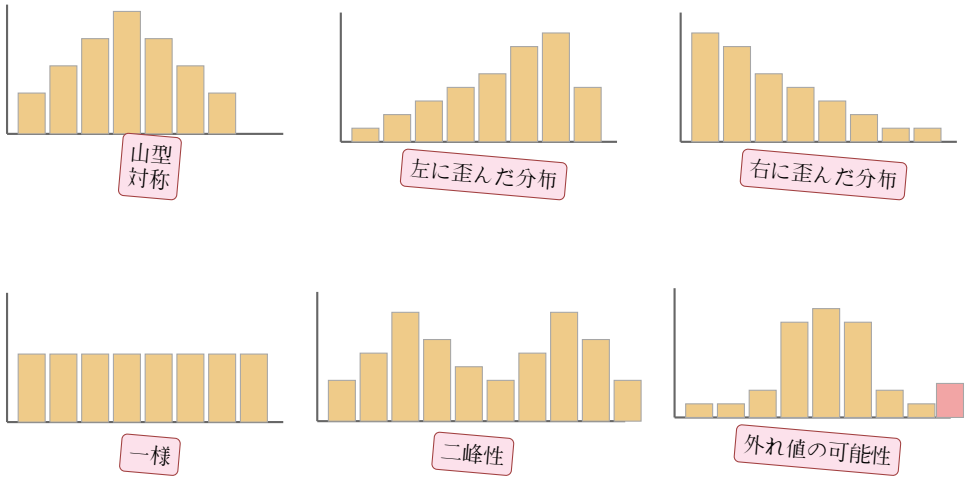
見て分かるように、右側のヒストグラムはデータをより高い解像度で説明しています。ここでは、各値ごとの分布を見ることができます。ビン幅2を使うと、このデータの一部が隠れていたことに注意してください。靴のサイズの23–24のビンの場合、最初のヒストグラムは単にそこに17個の値があると教えてくれるだけです。しかし、右側のヒストグラムは、サイズ23の人よりもサイズ24の人がはるかに多いことを伝えています。この追加情報は、ビン幅を2に広げると完全に失われます。

重要な警告

ビン幅を変えると、ヒストグラムの視覚的な物語が変わることがある。 責任ある提示者は、望ましい結論を誇張したり隠したりするビンではなく、データの構造を明らかにするビンを選ぶべきである。

2.4 分布の形

便利なことに、ヒストグラム全体の形は、データがどのように分布しているかについて教えてくれます。よく見られるヒストグラムの形を下に示します。



これから、これらの重要な形を一つずつ説明します。

2.4.1 山型対称分布

山型対称分布とは、ヒストグラムが中央に向かって上がり、その後おおよそ釣り合った形で下がっていく分布のことです。左右が完全に同一である必要はありませんが、多くの場合はかなり近い形になります。主な考え方は、中心部が分布の中で最も多くのデータを含み、データがその中心から両方向に広がっているということです。

山型対称分布

山型対称分布とは、ヒストグラムが中央に山を持ち、左側と右側がおおよそ釣り合っている分布である。

成人の身長のように自然に変動する測定値の多くは、しばしばおおよそ山型になります。しかし、ランダムに抽出されたすべてのデータがこの形を持つわけではありません。

2.4.2 歪んだ分布

歪んだ分布とは、一方の尾がもう一方よりも長い分布です。つまり、そこには本質的な非対称性があります。紛らわしいことに、歪みは長い尾の方向にちなんで名付けます。つまり、長い尾が分布の左側にある場合、そのグラフを「左に歪んだ」と呼びます。

歪んだ分布

一方の側の尾がもう一方よりも長いとき、その分布は**歪んでいる**という。

- **右に歪んだ分布：** 長い尾が右側にある。
- **左に歪んだ分布：** 長い尾が左側にある。

長い尾には、しばしば非常に大きい観測値や非常に小さい観測値が含まれます。 例えば、所得データはしばしば右に歪んでいます。多くの人は通常の範囲の収入を得ていますが、少数の人が非常に大きな収入を得ているからです。

2.4.3 一様分布

一様分布とは、すべての階級が同じ度数を持つことで、ヒストグラムが平ら、または長方形のように見える分布です。

一様分布

一様分布とは、階級がほぼ等しい度数を持つ分布である。

完全な一様分布は、観察された実生活のデータではあまり一般的ではありません。しかし、理論的には重要であり、特に後で確率論を議論するときに重要になります。一様分布は、どの区間も他の区間に比べて特別ではない状況を表します。

2.4.4 二峰性分布

二峰性分布には、目立つ山が二つあります。これはしばしば、データセットが二つの異なる母集団を混ぜていること、あるいは二つの異なる「説明要因」が母集団に作用していることを示

します。考え方は単純です。グラフが、対称な山型分布を二つ並べたように見えるのです。一つ一つの山は、データ内の下位集団、仕組み、または繰り返されるパターンを示唆することがあります。二峰性分布の場合には二つの山があり、これはデータに対して二つの異なる性質や力が作用していることを意味する場合があります。

二峰性分布

二峰性分布とは、二つの主要な山を持つ分布であり、通常は少なくとも一つの低度数の階級によって分けられている。

例えば、新宿のような人気のある都市部の駅を通過する歩行者の交通量を測定するとします。その（相対）度数分布には二つのはっきりした山が現れるでしょう。一つは午前8時ごろを中心とし、もう一つは午後6時ごろを中心とします。これは、通勤者が仕事へ向かう時間と仕事から帰る時間に、それぞれ交通量のピークがあるからです。

2.5 ヒストグラムから分かること

ヒストグラムや相対度数ヒストグラムは、量的データがどのようにおおよそ分布しているかを示すことができます。ヒストグラムを見ることで、しばしば次のことを判断できます。

- データ分布が対称、歪み、一様、二峰性のどれであるか、
- 外れ値の可能性があるかどうか、
- どのデータ区間に最も多くのデータが含まれているか、
- データがどの程度広がっているか。

3 カテゴリーデータの表現

カテゴリーデータとは、ラベル、名前、グループ、カテゴリーから成るデータです。カテゴリーデータには、必ずしも利用できる数学的構造が備わっているわけではありません。例えば、カテゴリーが血液型であった場合、A型マイナスとO型プラスを意味のある形で「足す」ということはできません。カテゴリーデータに付随する数学的構造は非常に少ないため、私たちにできることはかなり限られます。主な操作は、度数を数えることと、割合を比較することです。これら二つの操作から、カテゴリーデータの二つの一般的な表現が得られます。

- **棒グラフ**：度数やパーセントを表す。
- **円グラフ**：全体に対する割合を表す。

これから、これらを別々に説明します。

3.1 棒グラフ

棒グラフは、ヒストグラムに似た形で、個々のカテゴリーを別々の棒によって表すグラフです。各棒の長さ、または高さは、通常、度数やパーセントなどの値を表します。

定義：棒グラフ

棒グラフとは、各カテゴリーが別々の棒で表されるカテゴリーデータ用のグラフである。棒の高さまたは長さは、そのカテゴリーの度数、パーセント、または値を表す。

良い棒グラフには、次の三つの基本的な規則があります。

1. 棒は縦向きでも横向きでもよい。
2. 棒の幅と間隔はそろえるべきである。
3. すべての棒に対して同じ測定尺度を使うべきである。

棒と棒の間の隙間は重要です。 ヒストグラムは数値尺度上の区間を表します。 棒グラフは別々のカテゴリーラベルを表すため、棒は通常、視覚的に分けられます。

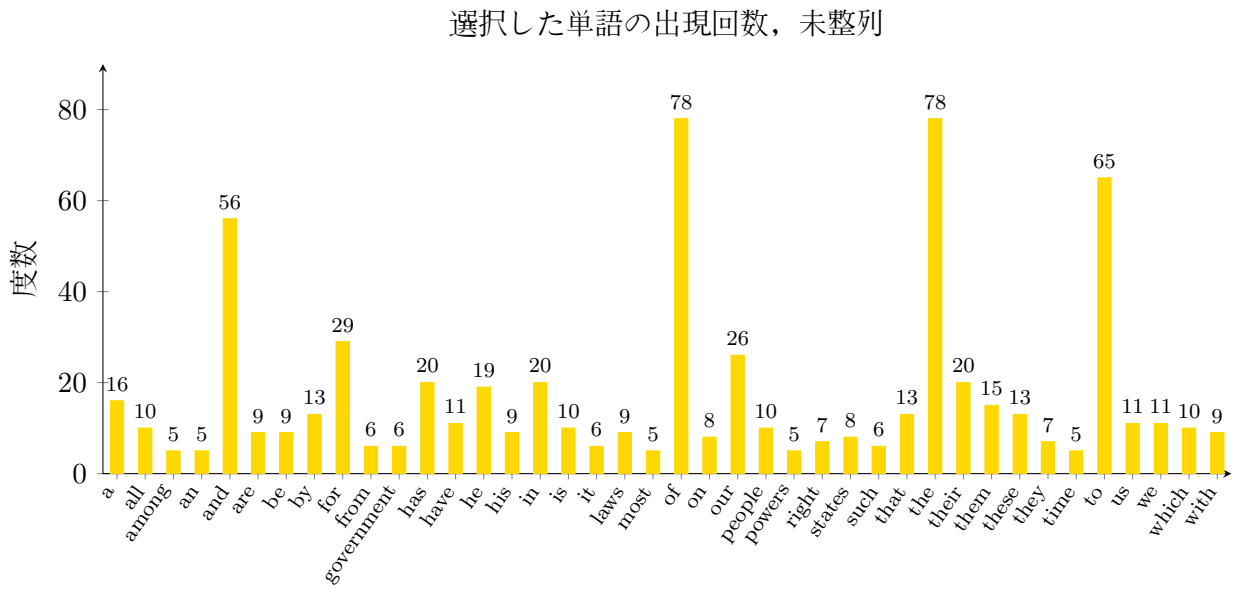
パレート図とは、棒が降順に並べられた棒グラフです。 最も多いカテゴリーが左に置かれ、最も少ないカテゴリーが右に置かれます。

定義：パレート図

パレート図とは、棒が大きいものから小さいものへと並べられた棒グラフである。 最も頻繁に現れるカテゴリーを素早く特定したいときに有用である。

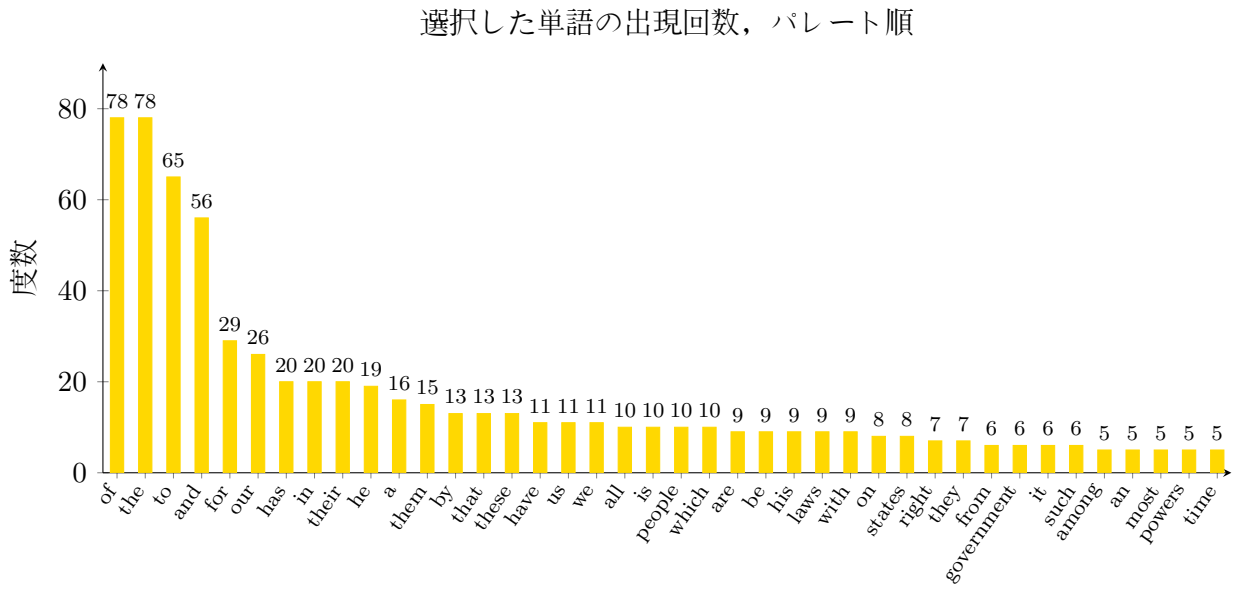
3.2 例：アメリカ独立宣言

アメリカ独立宣言に頻繁に現れる単語を数えるとします。この特定の単語数では、本文には1333語が含まれているため、すべての単語の度数を含む棒グラフはかなり混み合ってしまいます。そこで、少なくとも5回現れるすべての単語を棒グラフとして描くことにします。



見て分かるように、このグラフは読めますが、それでもかなり混み合っています。個々の棒が非常に多いため、「どの単語が最も頻繁に現れるのか」といった問いに素早く答えるのが難し

くなります。これは、代わりにパレート図を使うことで改善できます。



パレート図には同じ個数が含まれていますが、順序を変えることで、支配的なカテゴリーがはるかに見つけやすくなります。言い換えれば、パレート図の順序と調和は目にとって見やすいものになっています。

3.3 円グラフ

円グラフは、割合を円の扇形として表します。円グラフを作るには、まず各カテゴリーが何回現れるかを数え、それを全体の個数で割ることで割合を計算します。

定義：割合

割合とは、部分と全体の比較である。

$$\text{割合} = \frac{\text{度数}}{\text{全体の個数}}.$$

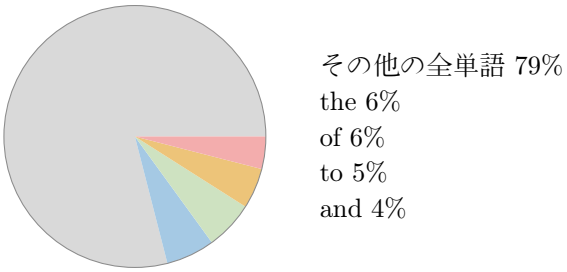
例として、前の節のデータを使います。単語“the”はアメリカ独立宣言の中で78回現れ、全体の単語数は1333語です。したがって、単語“the”の割合は

$$\frac{78}{1333} \approx 0.0585 = 5.85\% \approx 6\%,$$

つまり、“the”はおよそ6%の頻度で現れます。

定義：円グラフ

円グラフとは、カテゴリーを円の扇形として表すグラフである。 各扇形の大きさは、全体の中でそのカテゴリーが占める割合を表す。



読みやすい円グラフには、大きな扇形が少数だけある。

重要な警告

円グラフは、扇形が少数だけの場合に最もよく機能する。 円グラフに小さな扇形がたくさん含まれていると、聞き手はカテゴリーを正確に見たり比較したりできないことがある。

練習問題

ある40人のクラスで、好きなコンビニエンスストアを尋ねたとします。 結果は次の通りです。

セブン-イレブン : 18, ファミリーマート : 12, ローソン : 8, その他 : 2.

(a) セブン-イレブンを選んだ学生の割合は何か。
(b) このデータセットに対して円グラフは妥当か。
(c) このデータセットに対してパレート図は妥当か。

解答

(a) 割合は $18/40 = 0.45 = 45\%$ である。
(b) はい。カテゴリーが四つしかないため、円グラフは十分読みやすい。
(c) はい。特に最も人気のある店を素早く示したい場合には、パレート図も妥当である。

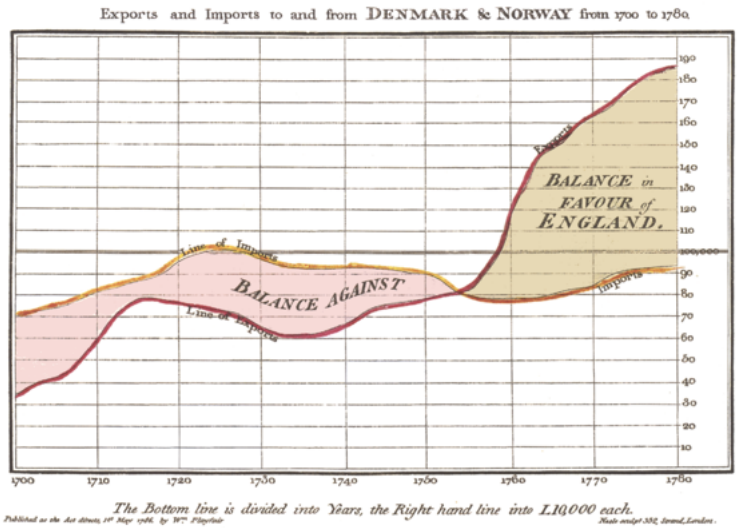
4 時間依存データの表現

同じ対象を繰り返し観測することでデータを集める場合があります。これにより、ある性質の時間的な変化を追跡できます。例えば、毎週体重を測ったり、ヒマワリの伸びる高さを毎日記録したり、毎晩の勉強時間を記録したり、多くの年代にわたってある単語の人気を追跡したりすることがあります。このようなデータは時間依存データまたは時系列データと呼ばれます。

定義：時系列グラフ

時系列グラフとは、時間依存データを時系列順にプロットしたグラフである。通常、横軸は時間を表し、縦軸は追跡している量を表す。

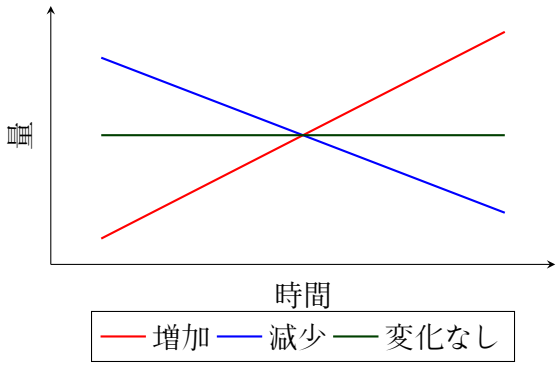
統計的な時系列グラフの初期の有名な例の一つは、18世紀末にウィリアム・プレイフェアによって作られました。



見て分かるように、時系列グラフは非常に直感的です。時間とともに量がどのように変化するかを、かなり明確に見ることができます。ここで、いくつかの有用な変化の種類を簡単に確認します。

4.1 線形変化

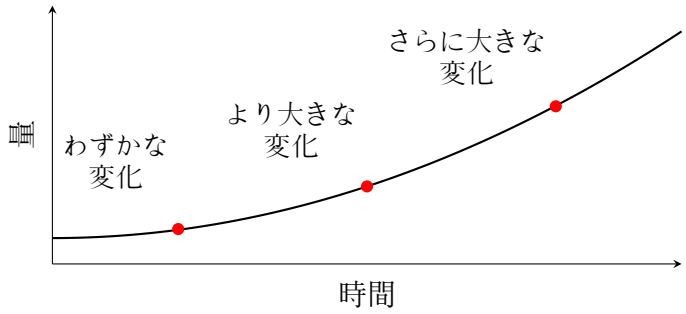
時間に伴う最も単純な変化は線形です。これは、時系列グラフにプロットしたとき、データセットが直線を描くことを意味します。幾何学的には、変化の割合は直線の傾きによって表されます。



増加する直線は、時間とともに量が増えることを示しています。 減少する直線は、時間とともに量が減ることを示しています。 水平な直線は、変化がないことを示しています。 いずれの場合も、基本的な視覚的考え方は同じです。直線の形が、変化の方向と速さを教えてくれます。

4.2 非線形変化

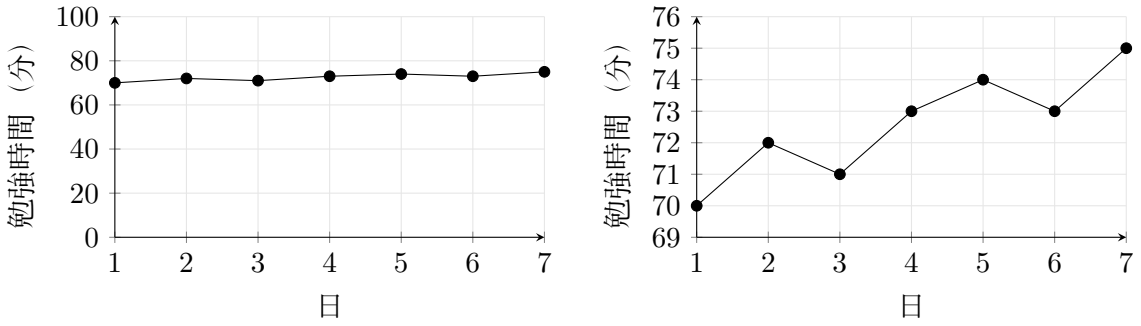
多くの場合、変化は一定の割合で起こるわけではありません。むしろ、測定している過程に応じて、変化そのものが速くなったり遅くなったりします。次を考えてみましょう。



上のグラフでは、時間とともに変化の割合そのものが増加しているため、加速の例になっています。

4.3 尺度と物語

時系列グラフは尺度に特に敏感です。縦軸を変えると、小さな変化が劇的に見えたり、大きな変化が普通に見えたりすることがあります。これを説明するために、次の二つのグラフを考えます。



上の二つのグラフは、どちらも同じ基礎データセットから作られています。しかし、縦軸の尺度を調整することで、データの見え方を大きく変えることができます。尺度の選択は、さりげなく異なる物語を含みます。最初のグラフはデータにほとんど変化がないことを示唆し、二番目のグラフはデータがかなり変動していることを示唆します。この観察は、次の点を動機づけます。

重要な警告

異なる尺度は異なる物語を作ることがある。 時系列グラフを読むときは、物語を解釈する前に必ず軸を確認すること。

これは、データの可視化が人々を誤解させる最も一般的な方法の一つです。 グラフは技術的には正しくても、効果を誇張するために尺度が選ばれている場合、視覚的には誤解を招くことがあります。

5　対になったデータの表現

私たちが集めるデータは、自然に対になっていることがあります。これは通常、同じ対象について同時に異なる性質を集める場合、または各観測について二つの性質を比較したい場合に起こります。例としては、次のようなものがあります。

- 同じ人の左足と右足の長さ。
- 学生がキャンパスへ通うときの移動時間と距離。
- 東京に住む人々の身長と体重。

定義：対になったデータ

対になったデータとは、各観測が二つの関連する測定値から成る場合に生じる。 二つの変数の間に関係があるかどうかを調べるために、対になったデータを扱うことが多い。

5.1　例：通学時間と距離

次の12人の学生のデータセットを考えます。

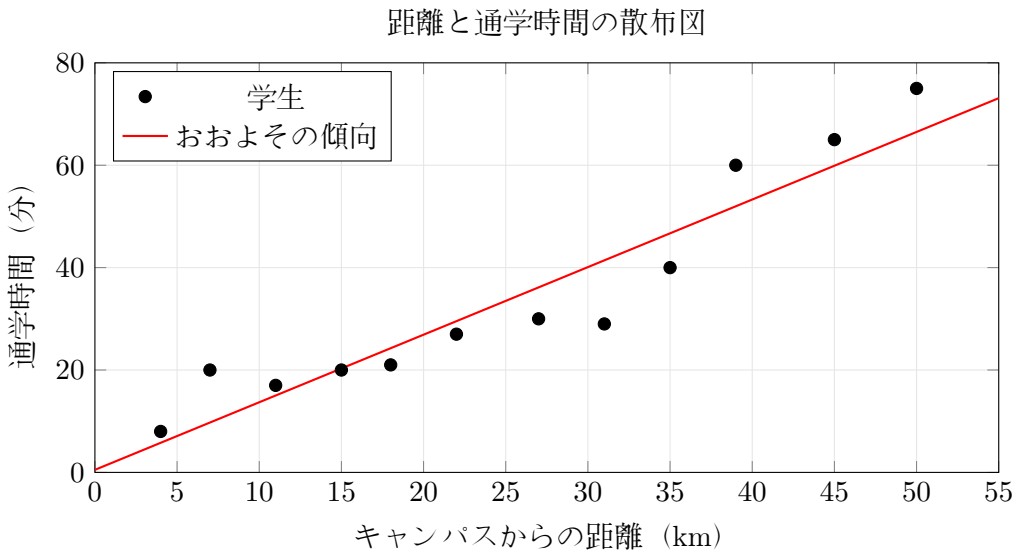
学生	キャンパスからの距離（km）	通学時間（分）
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75

このデータセットは、各行が同じ学生について二つの結びついた情報を含んでいるため、自然に対になっています。この対を二つの変数に分けることができます。第一の変数はキャンパスからの距離であり、第二の変数は通学時間です。この状況に自然なグラフは散布図です。

定義：散布図

散布図とは、対になった量的データのためのグラフである。 各観測は一点として表され、横座標が一方の変数を、縦座標がもう一方の変数を表す。

散布図を作るには、一方の変数をx軸に、もう一方の変数をy軸にプロットします。

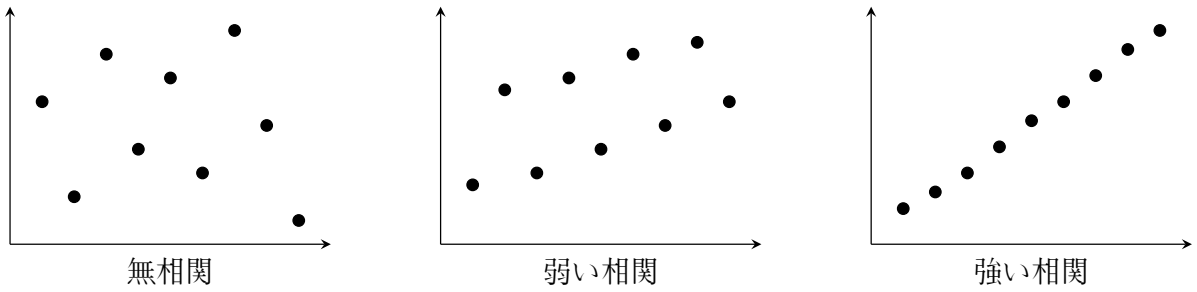


5.2 相関

上の散布図は、二つの変数の間に明確な正の関係があることを示しています。つまり、キャンパスから遠くに住んでいる学生ほど、通学時間が長くなる傾向があります。もちろん、これは完全な関係ではありません。通学時間は距離だけで決まるわけではなく、他にも多くの要因に依存するからです。しかし、それでも二つの変数の間には明確な一般的パターンがあります。対になったデータが一般的なパターンを示すとき、その変数は相関していると言います。

定義：相関

相関とは、二つの変数がともに変化する傾向があることを意味する。 散布図は、対になったデータが無相関、弱い相関、強い相関のどれであるかを示すことができる。



この授業の後半では、相関を数値的に記述する方法を学びます。しかし、今のところは、基本的な考え方を単に認識できれば十分です。

- 点が明確なパターンを作らない場合、変数は無相関かもしれない。
- 点が大まかに一方向に動くが、ばらつきが大きい場合、変数は弱く相関しているかもしれない。
- 点が直線に近い場合、変数は強く線形相関しているかもしれない。点が曲線に近い場合、変数は強く関連しているかもしれない。

負の相関は、下向きの傾向として現れます。

通学時間の例では、データを通るおおよその直線を描きました。 最もよく当てはまる直線を見つける作業は、**線形回帰**と呼ばれます。 これについては、後でより正式に学びます。

線形回帰

線形回帰とは、二つの量的変数の関係を要約する直線を見つける方法である。 この講義では、視覚的に使うだけである。 正式な方法は後で扱う。

5.3 相関は因果関係ではない

第1回講義で議論したように、相関は自動的に因果関係を意味するわけではありません。 散布図は二つの変数が関連していることを示唆できますが、それだけでは一方の変数がもう一方の変数を引き起こしていることを証明できません。

例。 キャンパスからの距離と通学時間が強く相関しているとします。 遠くに住んでいることが長い通学時間に寄与する、と考えるのは妥当です。 しかし、正確な通学時間は、鉄道路線、乗換駅、自転車の利用可能性、交通状況、歩く速さ、そして学生が便利な駅の近くに住んでいるかどうかにも依存します。

良い散布図は因果関係を示唆するかもしれませんが、その解釈には依然として文脈が必要です。

6 ケーススタディ：日本のコンビニ

ある会社があんぱんを製造しているとします。その会社は日本のどこかのコンビニエンスストア（コンビニ）で自社商品を販売したいと考えており、意思決定のために市場調査を行おうとしています。会社は気の毒なインターンのジェニーに、日本全国のコンビニの分布について調査し、その結果を提示するよう命じます。簡単に言えば、ジェニーの課題は次の通りです。

- コンビニが日本国内でどのように分布しているかを調べること、
- そのデータを上司たちに提示すること。

そこで、ジェニーは都道府県別のコンビニエンスストアのデータセットを使います。

都道府県	店舗数
北海道	3003
佐賀	360
山梨	459
福井	337
東京	7405
愛媛	583
宮城	1161
岐阜	854
茨城	1432
広島	1205
福島	904
千葉	2701

都道府県	店舗数
青森	610
香川	406
愛知	3642
岡山	797
栃木	921
山口	565
富山	485
鹿児島	665
秋田	449
山形	447
群馬	900
京都	1077

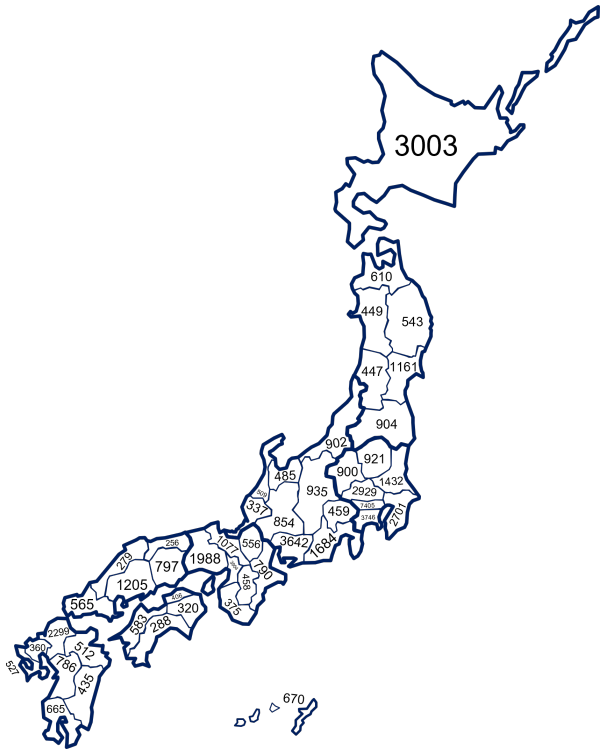
都道府県	店舗数
静岡	1684
高知	288
鳥取	256
島根	279
沖縄	670
新潟	902
長野	935
宮崎	435
大分	512
和歌山	375
熊本	786
神奈川	3746

都道府県	店舗数
石川	509
長崎	527
岩手	543
埼玉	2929
福岡	2299
滋賀	556
大阪	3954
兵庫	1988
三重	790
奈良	458
徳島	320

これから、このデータを視覚的に提示するいくつかの方法を考えます。

6.1 地図による表現

上に書かれたデータは正確ですが、解釈しやすくはありません。見て分かるように、関連する情報はすべて含まれています。しかし、データに含まれる基礎的なパターンを理解するのは非常に難しいです。仮にこれを日本地図上に正直にプロットしたとしても、結果は分かりにくいものになります。



見て分かるように、この地図はこのデータを表す良い方法ではありません。良い考えの兆しはありますが、主に三つの問題があります。

1. 数字が小さすぎて読みにくい。
2. それでもデータのパターンを見るのが難しい。
3. たとえすべての数字を読めたとしても、その正確な値が実際には重要でないかもしれない。

この問題を解決する一つの方法は、都道府県ではなく地方ごとにデータをまとめることで、地図の解像度を下げることです。まず、店舗数を日本の主要な地方に書き直すことができます。

地方	コンビニ店舗数
北海道	3003
東北	4114
関東	20034
中部	9807
関西	9198
中国	3102
四国	1597
九州	5584
沖縄	670

見て分かるように、このデータはより明確な物語を伝えます。関東はコンビニの数が圧倒的に多く、その次に中部と関西が続きます。しかし、データを読みやすくする過程で詳細は失われています。これを次に改善します。

6.2 色による表現

この提示の文脈では、ジェニーが調査の正確な数値データを提示する必要はおそらくありません。言い換えれば、生データは本当に有用であるには情報が多すぎる可能性があります。量を色のグラデーションで符号化することで、パターンを保ちながらこの情報を抑えることができます。下の表では、店舗数が多いほど青に近く、少ないほど赤に近く、中央付近は白になるように色付けしています。

都道府県	店舗数
北海道	3003
佐賀	360
山梨	459
福井	337
東京	7405
愛媛	583
宮城	1161
岐阜	854
茨城	1432
広島	1205
福島	904
千葉	2701

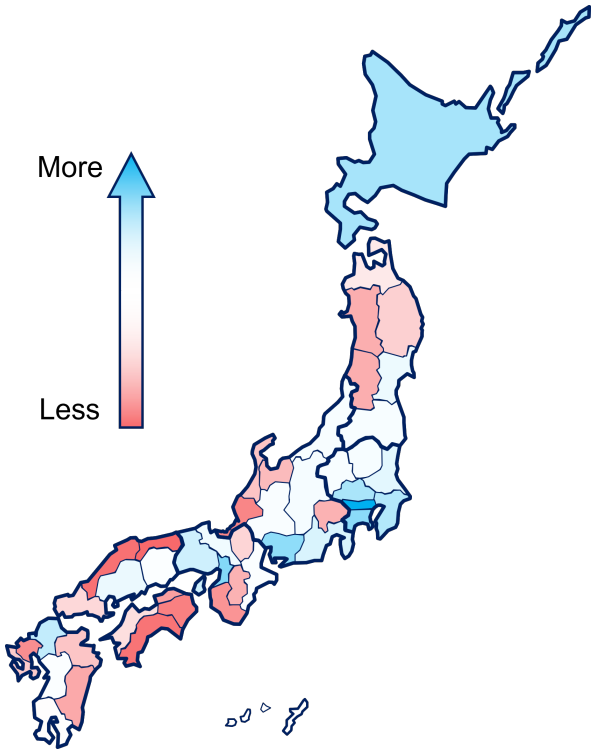
都道府県	店舗数
青森	610
香川	406
愛知	3642
岡山	797
栃木	921
山口	565
富山	485
鹿児島	665
秋田	449
山形	447
群馬	900
京都	1077

都道府県	店舗数
静岡	1684
高知	288
鳥取	256
島根	279
沖縄	670
新潟	902
長野	935
宮崎	435
大分	512
和歌山	375
熊本	786
神奈川	3746

都道府県	店舗数
石川	509
長崎	527
岩手	543
埼玉	2929
福岡	2299
滋賀	556
大阪	3954
兵庫	1988
三重	790
奈良	458
徳島	320

このデータを取り、それぞれの都道府県を対応する色で塗ることができます。もちろん、これ

は生の店舗数をそのまま提示する場合と比べると情報の損失です。しかし、その見返りは大きいです。データの地理的なパターンをすぐに見ることができるからです。



見て分かるように、このデータのパターンは、先ほどの地方別地図よりも詳細です。ここでは、都市部の周辺、つまり東京、名古屋、大阪、そしてある程度は仙台、広島、福岡の周辺に、明確な集中のホットスポットがあることが分かります。逆に、人口の多い地域の外では、コンビニの度数が大きく落ち込んでいることが分かります。島根、鳥取、四国、そして東北北部は、平均よりも低い店舗数を示しています。これら二つの一般的な観察に加えて、北海道が明らかな外れ値であることにも気づきます。

上の表現は以前のグラフよりも良いものですが、それでも改善できる点があります。ジェニーとその会社の文脈では、都市部には全国平均よりも多くのコンビニがあることを、経営陣はすでに知っている可能性が高いです。したがって、この調査はそれほど新しいことを示していないかもしれません。しかし、だからといって他に何も隠れていないというわけではありません。

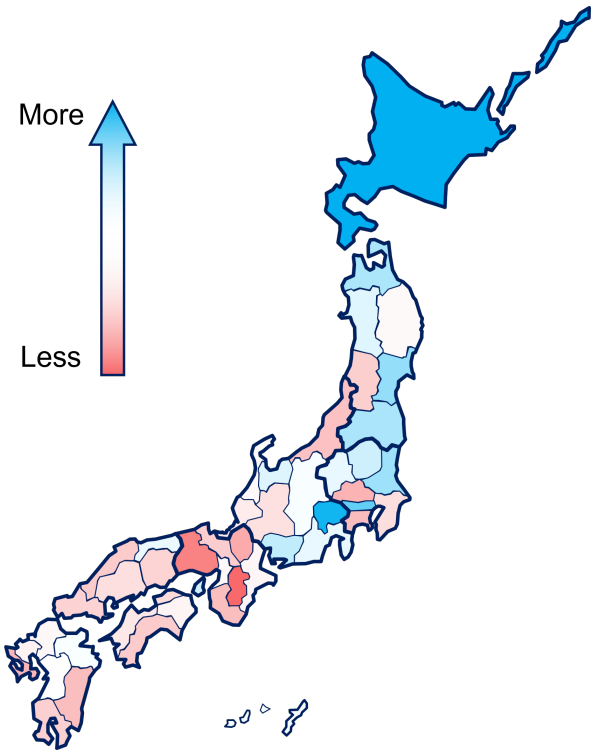
このデータにおける別のパターンを示すために、今度は店舗数を人口で割り、さらに100,000を掛けます。これにより、実質的に「人口10万人あたりのコンビニ数」という尺度を作ります。下の表では、この追加情報を再び色のグラデーションで示しています。

都道府県	店舗数	10万人あたり
北海道	3003	57.43
山梨	459	56.64
東京	7405	52.65
宮城	1161	50.40
茨城	1432	49.92
福島	904	49.29
青森	610	49.24
愛知	3642	48.26
栃木	921	47.62
富山	485	46.83
秋田	449	46.77
群馬	900	46.38
静岡	1684	46.33
鳥取	256	46.22
沖縄	670	45.63
長野	935	45.62

都道府県	店舗数	10万人あたり
大分	512	45.53
熊本	786	45.19
石川	509	44.91
岩手	543	44.83
福岡	2299	44.74
大阪	3954	44.72
三重	790	44.60
徳島	320	44.46
佐賀	360	44.33
福井	337	43.91
愛媛	583	43.65
岐阜	854	43.14
広島	1205	43.01
千葉	2701	42.96
香川	406	42.69
岡山	797	42.18

都道府県	店舗数	10万人あたり
山口	565	42.07
鹿児島	665	41.85
山形	447	41.83
京都	1077	41.75
高知	288	41.62
島根	279	41.54
新潟	902	40.96
宮崎	435	40.65
和歌山	375	40.63
神奈川	3746	40.54
長崎	527	40.13
埼玉	2929	39.87
滋賀	556	39.31
兵庫	1988	36.35
奈良	458	34.56

この表は、生の総店舗数とはまったく異なる物語を伝えます。いくつかの都道府県、特に山梨や東北北部の県は、突然赤から青へ変わります。これは、それらが突然、過剰に代表されるように見えることを意味します。逆に、以前は都市部として目立っていた地域の一部は、人口で調整されることで赤に移ります。しかし、東京は依然として高く表されており、北海道も依然として外れ値です。さらに興味深いのは、これを日本地図上に描いたときに現れるパターンです。下のグラフを、生の店舗数の地図と比較してみましょう。



人口あたりのコンビニ数に関して、日本には明確な東西の分断が見られます。日本の東側は一般に高い値を取り、日本の西側は一般に低い値を取ります。ジェニーの発表の文脈では、会社

が店舗密度、総店舗数、あるいは別のもののどれを重視するかに応じて、彼女は日本の東側をさらに調査することを提案するかもしれません。

6.3 データ可視化に関するいくつかの観察

最後に、データ可視化に関する重要な観察で締めくくります。

重要な点

データの可視化は、洞察を得るために行われる、意図的な情報の損失であることが多い。誠実さと正確さに加えて、提示は有用で興味深く、理想的には聞き手が他の方法では気づかなかったことを示すべきである、という原則に従うべきである。

データの提示は常に文脈に依存します。この講義で見てきたように、提示者の選択によって物語が歪められる可能性は常にあります。今後データを見るときには、次の問いを自分に投げかけるとよいでしょう。

1. **データの種類：** そのデータはどのような構造を持っているか。量的、カテゴリー的、時間依存、対になったデータ、あるいはそれらの組み合わせか。
2. **目的：** そのグラフはどの問いに答えようとしているのか。
3. **尺度：** 軸や尺度は誠実で、読みやすいか。
4. **グループ化：** データがグループ化されている場合、ビンやカテゴリーは公平で妥当か。
5. **情報の損失：** どのような詳細が隠されたり単純化されたりしているか。
6. **利益相反：** 誰がこのデータを作成したのか。また、なぜこのように提示している可能性があるのか。
7. **物語：** その可視化は、聞き手にどのような結論を導くよう促しているか。
8. **別の見方：** 別の表現方法なら、異なる、しかし同じくらい妥当な物語を語るだろうか。