1-3 データ観察

東京大学 数理・情報教育研究センター

2021年4月21日

概要

- 本節では、詳細なデータ分析に進む前段階として、収集したデータを俯瞰的に観察するための基本的な手法について学び、起きている事象の背景や意味合いを理解することを目標とします。

- また、データを観察していく上で注意すべきいくつかの点についても学びます。データの観察を効果的に行う上で、データがどのような背景から得られたものなのかを正しく理解することも重要です。

本教材の目次

データの種類

データとは、物事の推論の基礎となる事実、また参考となる資料・情報です。また、コンピューターでプログラムを使った処理の対象となる記号化・数値化された資料、とデジタル大辞泉にて説明されています。データを集計する際に、データの種類として大きく分けて「量的変数」と「質的変数」の2種類があることに注意します。

量的変数：数量で表すことができ、さらに以下のように分類することもできます。

比率データ：四則演算すべて意味がある。例：体重、年収、長さ

間隔データ：和や差はできるが、積や除算には意味がない。
例：西暦年、温度（「温度70%減」とはいわない）

質的変数：数量で表すことが困難であるもので、さらに以下のように分類することもできます。

名義尺度：同じ値か否か
例：名前、性別、職業、既婚／未婚

順序尺度：大小関係あり
例：ランキング、成績の五段階評価

データの項目はデータによって異なります。

右の例では、「地域コード」「都道府県」「市」「世帯人員」「米」「食パン」「他のパン」等の項目があり、「米」以降の項目はそれぞれ一世帯あたり年間支出金額を示しています。

まずは、比較対象の設定を的確に行うことが重要です。たとえば、各食品の支出金額と「世帯人員」との関連性に焦点を当てるのか、それとも食品支出金額の間の関連性に興味があるのか、といったことをはっきりさせます。

地域コード	都道府県	市	世帯人員	米	食パン	他のパン	・・
R01100	北海道	札幌市	2.96	30994	8496	18942	・・
R02201	青森県	青森市	2.98	23773	7777	17336	・・
R03201	岩手県	盛岡市	3.15	25867	8270	20622	・・
R04100	宮城県	仙台市	3.00	20207	7972	18989	・・
R05201	秋田県	秋田市	2.88	19508	6461	17978	・・
R06201	山形県	山形市	3.19	26733	7781	18735	・・
R07201	福島県	福島市	3.00	24612	7077	18422	・・
R08201	茨城県	水戸市	2.90	19367	8495	17673	・・
R09201	栃木県	宇都宮市	2.85	22135	9053	19055	・・
R10201	群馬県	前橋市	2.81	25322	7652	22129	・・
R11100	埼玉県	さいたま市	3.04	24816	9350	22858	・・
R12100	千葉県	千葉市	3.00	22629	10092	22679	・・
R13100	東京都	東京都区部	2.93	22412	11064	24885	・・
R14100	神奈川県	横浜市	2.84	24983	10722	23457	・・
・	・	・	・	・	・	・	
・	・	・	・	・	・	・	

「都道府県庁所在市別・家計消費データ」を加工して作成
(https://www.nstac.go.jp/SSDSE/)

クロス集計表

2種類の項目を組み合わせて、合計、平均、標準偏差等を集計したものを「クロス集計表」とよび、データの全体像が把握しやすくなります。

右のクロス集計表は、都道府県ごとに一世帯当たりの年間支出金額を食品大分類ごとの合計を示したものです。前出データの「米」、「食パン」、「他のパン」は、右のクロス集計表ではすべて「穀物」に分類されています。

	穀類	魚介類	肉類	乳卵類	野菜・海	果物	・・
北海道	81474	79328	83095	41262	104045	36067	・・
青森県	71992	90933	83349	38677	106830	38863	・・
岩手県	80203	78310	76514	51711	118250	42415	・・
宮城県	70942	87815	84141	48489	120474	43636	・・
秋田県	68139	84401	80686	42682	117898	41537	・・
山形県	79598	74850	93770	48522	118255	47970	・・
福島県	73184	76986	76085	48264	110835	49477	・・
茨城県	67318	68527	75129	48425	100131	41079	・・
栃木県	74050	72694	82490	46981	114270	42331	・・
群馬県	77456	71940	67833	44449	105257	42123	・・
埼玉県	80828	73940	88061	49560	121177	42751	・・
千葉県	78500	80770	88786	50870	123233	45467	・・
東京都	81177	79327	95859	50541	125815	44229	・・
神奈川県	82257	83487	97515	49915	127908	46446	・・
・	・	・	・	・	・	・	
・	・	・	・	・	・	・	

「都道府県庁所在市別・家計消費データ」を加工して作成
(https://www.nstac.go.jp/SSDSE/)

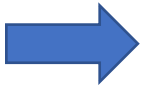
前述のクロス集計表はデータを要約する上で効果的ですが、それらの表の値を見ているだけでは、データがどのように、どの程度ばらついているかを把握するのは困難です。

まずデータの値を適当な範囲で区切って、それぞれの区間に入るデータ数を表にします。これを度数分布とよび、データ数が多くても全体の傾向がわかりやすくなります。

各階級の上限と下限の差を階級幅、それらの中央値を階級値とよびます。たとえば「2.72～2.76」の階級の階級幅は0.04（人）、階級値は2.74（人）となります。

〈データ〉

都道府県	世帯人員
北海道	2.96
青森県	2.98
岩手県	3.15
宮城県	3.00
秋田県	2.88
山形県	3.19
福島県	3.00
茨城県	2.90
栃木県	2.85
群馬県	2.81
埼玉県	3.04
千葉県	3.00
東京都	2.93
神奈川県	2.84
・	・
・	・



〈度数分布〉

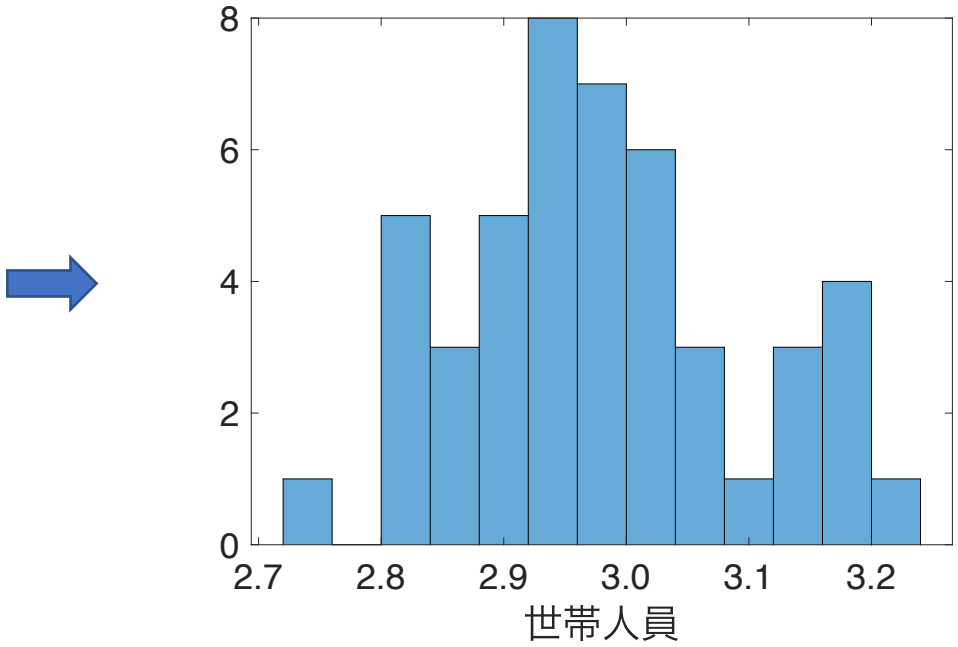
世帯人員	都道府県数
2.72～2.76	1
2.76～2.80	1
2.80～2.84	5
2.84～2.88	4
2.88～2.92	5
2.92～2.96	6
2.96～3.00	7
3.00～3.04	6
3.04～3.08	3
3.08～3.12	1
3.12～3.16	3
3.16～3.20	4
3.20～3.24	1

ヒストグラム

度数分布の最もデータが集中する階級が「データの中心」の一つの目安になります。「世帯人員」においては、「2.92～2.96」が最頻階級で、データのほぼ真ん中に位置しています。

度数分布表の値を棒グラフにしたものをヒストグラムとよびます。
視覚化することによりデータのばらつきの傾向がわかりやすくなります。

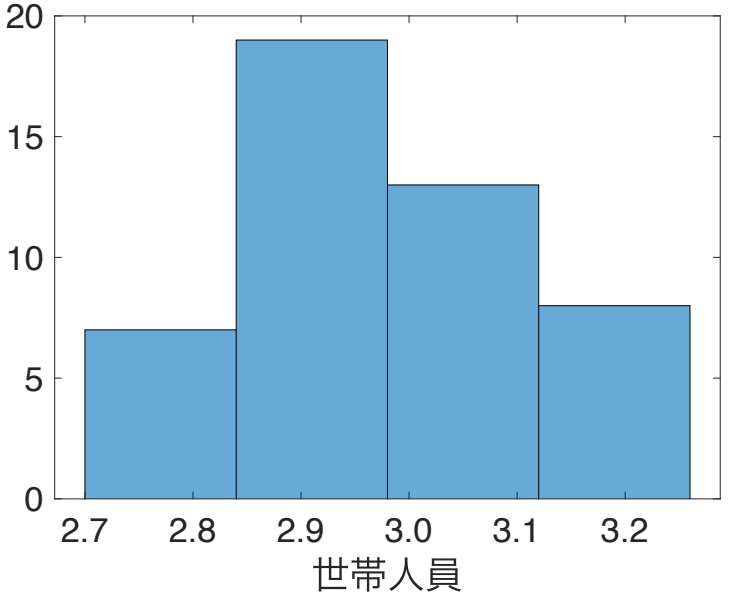
世帯人員	都道府県数
2.72～2.76	1
2.76～2.80	0
2.80～2.84	5
2.84～2.88	3
2.88～2.92	5
2.92～2.96	8
2.96～3.00	7
3.00～3.04	6
3.04～3.08	3
3.08～3.12	1
3.12～3.16	3
3.16～3.20	4
3.20～3.24	1

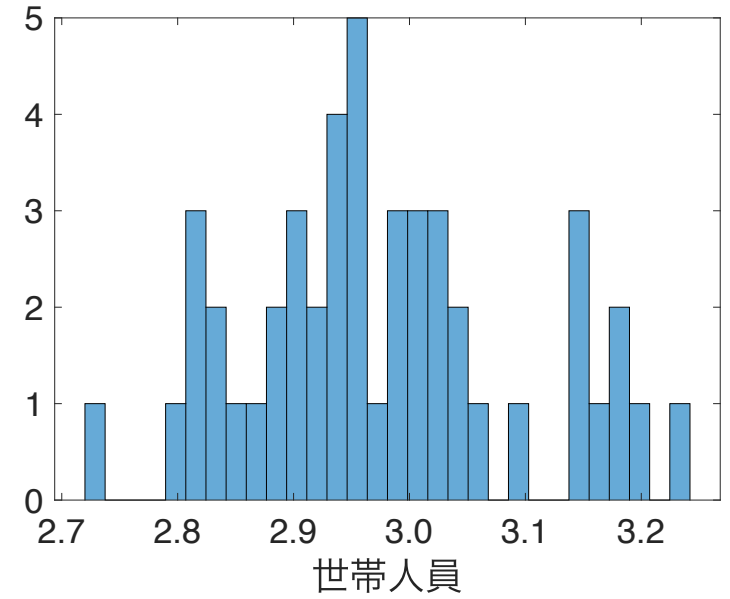


データを区切る範囲は分析者が任意に指定することになりますが、その選び方によっては形状が大きくかわったり、また傾向を掴みづらくなることがあるため注意が必要です。

まず区間設定は粗くしすぎても、細かくしすぎてもいけません。粗くしすぎると、上のヒストグラムのように大雑把になりすぎて、データのばらつきをあまり的確に把握できなくなります。逆に細かくしすぎると、下のヒストグラムのように、データが一つも該当しない区間が多数見受けられ、あまり適切であるとはいえません。

区間設定を適切に行うために、ある程度の試行錯誤が必要になります。





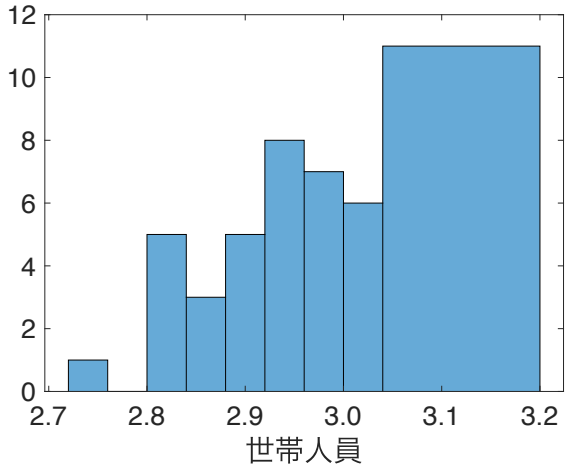
階級幅を等間隔にとるのが理想的ですが、そのようにデータが与えられない場合もあります。

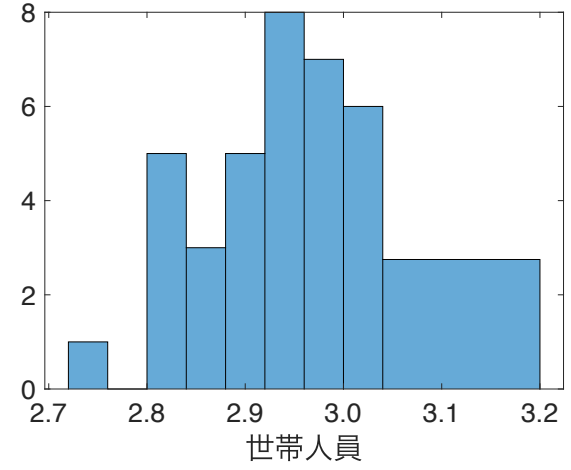
たとえば右の度数分布表は、はじめの8つの階級は階級幅が0.04で統一されていますが、最後の階級だけ階級幅が0.20と5倍に延伸されています。

上のヒストグラムのように最後の階級にそのまま度数11をプロットすると、その階級が極端に頻度が高いと誤解を与えてしまいます。

よって、下のヒストグラムのように、柱（長方形）の面積が度数に対応するように高さを調節する必要があります。

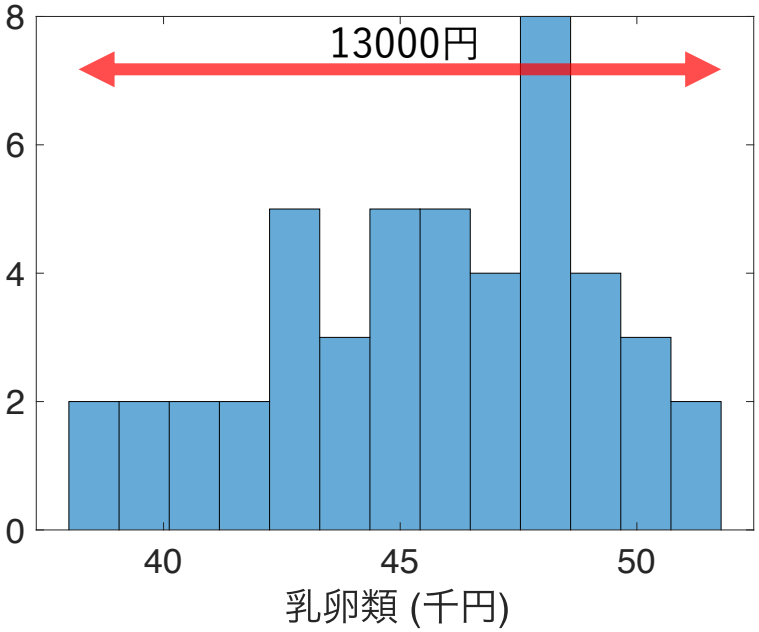
世帯人員	都道府県数
2.72～2.76	1
2.76～2.80	1
2.80～2.84	5
2.84～2.88	4
2.88～2.92	5
2.92～2.96	6
2.96～3.00	7
3.00～3.04	6
3.04～3.24	11

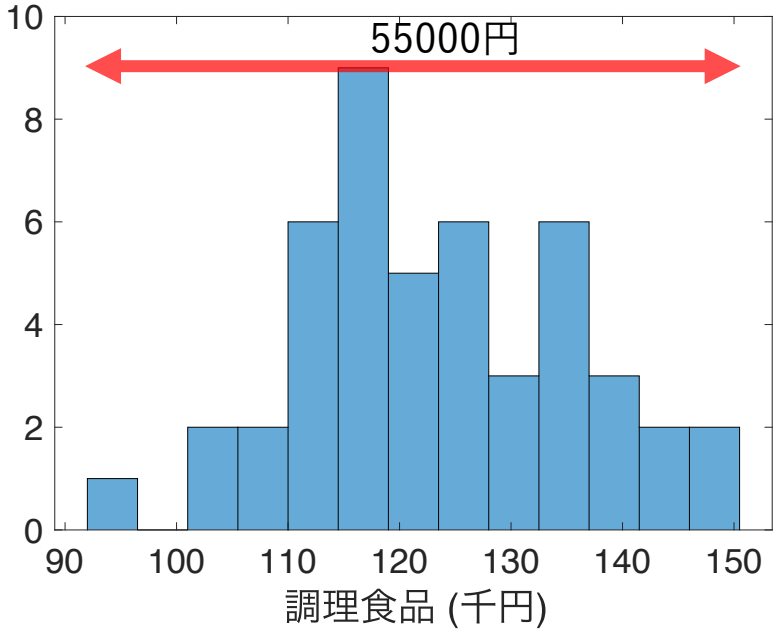




複数のヒストグラムを並べることで、データの相違性や傾向性を視覚的に認識できることがあります。

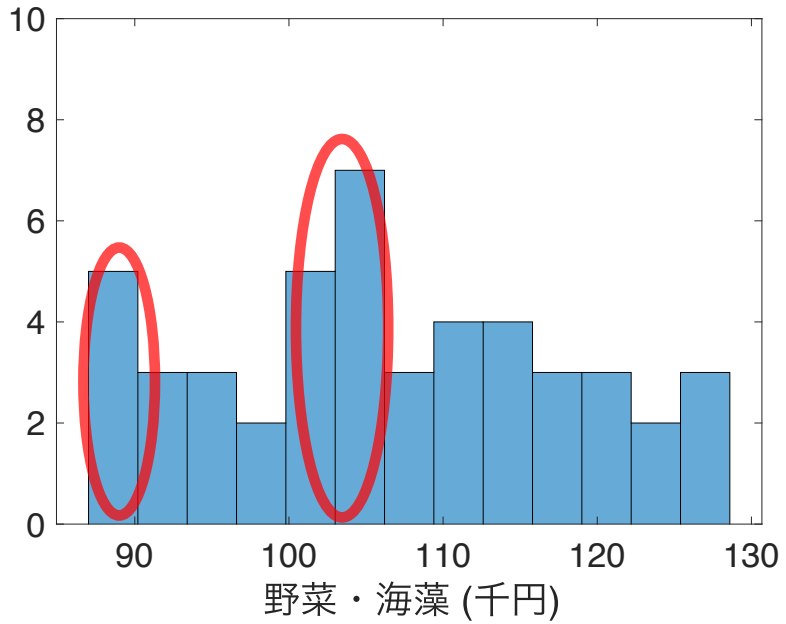
たとえば、「乳卵類」の年間支出金額は都道府県間であまり大きくばらつかない傾向にあるようですが、「調理食品」の支出額には大きなばらつきがある、という相違を把握できます。

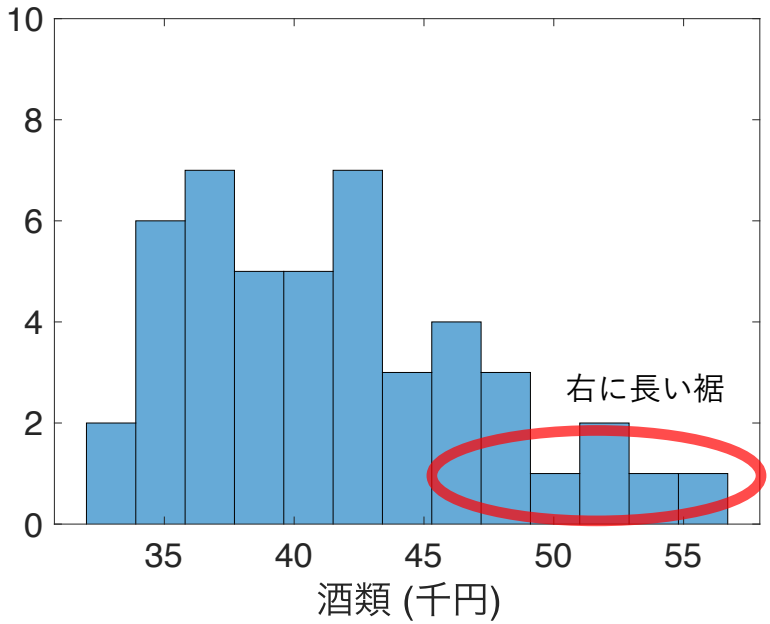




ヒストグラムの観察を通して、データのばらつきのおおまかな形状も捉えることができます。

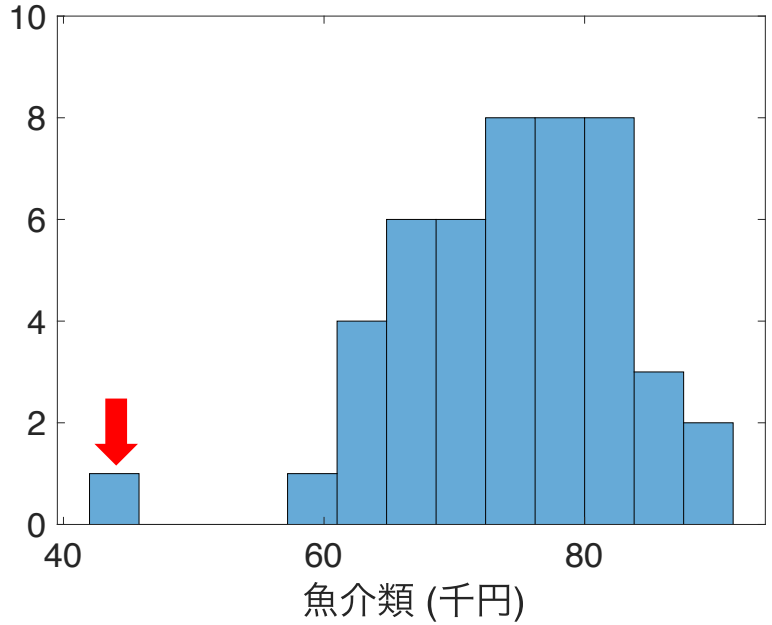
たとえば、「野菜・海藻」のヒストグラムは左右ほぼ対象ですが、峰が少なくとも2つ（90周辺と105周辺）確認できます。また、「酒類」のヒストグラムはやや左寄りになっていて、年間支出金額の少ない都道府県が多い傾向があることがわかります。

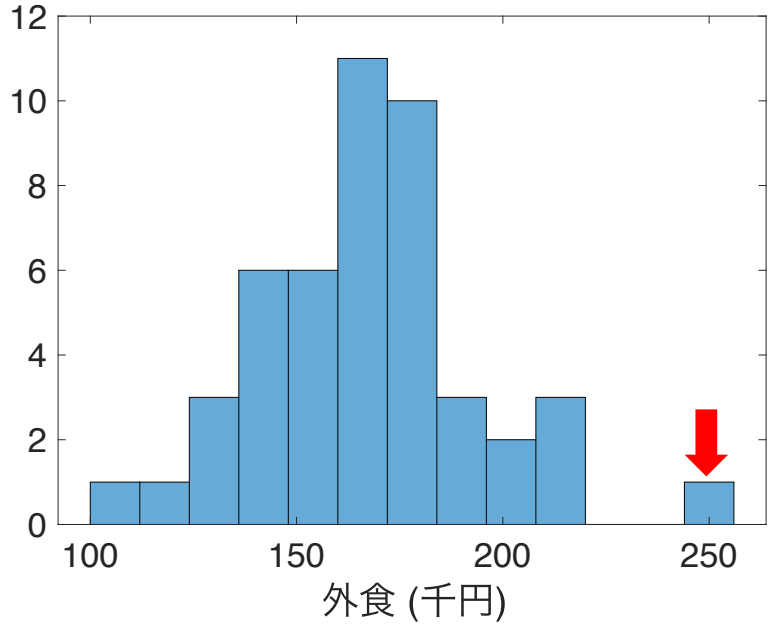




ヒストグラムによってデータの特異点を明確に捉えることができることがあります。

たとえば、「魚介類」と「外食」の年間支出金額をみると、それぞれ一件だけ突出する都道府県が存在することがわかります。それらを外れ値とよびます。もとのデータに立ち戻ってみると、「魚介類」の外れ値は沖縄県、「外食」の外れ値は東京都にあたるということを見つけることができます。

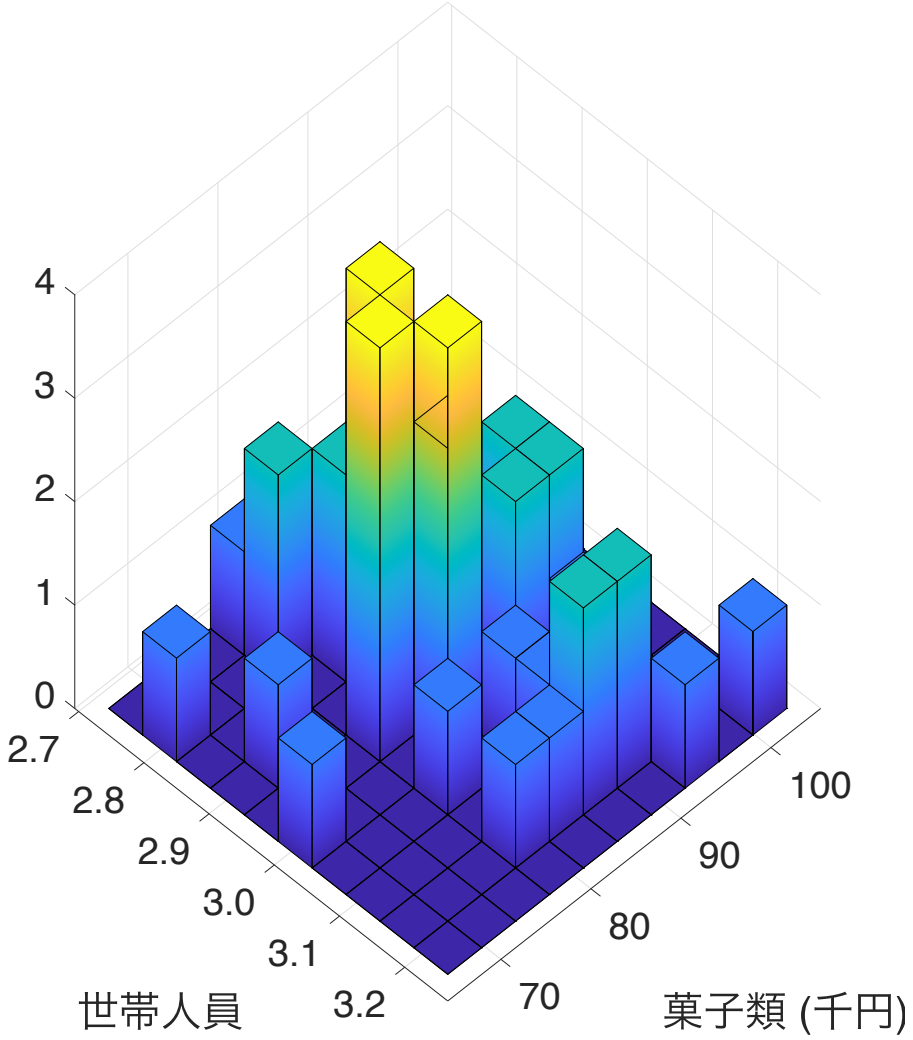




二項目の関連性に興味がある場合にもヒストグラムを利用することができますが、あまり見やすいとはいえない場合がほとんどです。

たとえば右の例では、高い山の裏側を見ることはできません。

したがって、ヒストグラムは単項目のプロットにとどめておくのが典型的です。

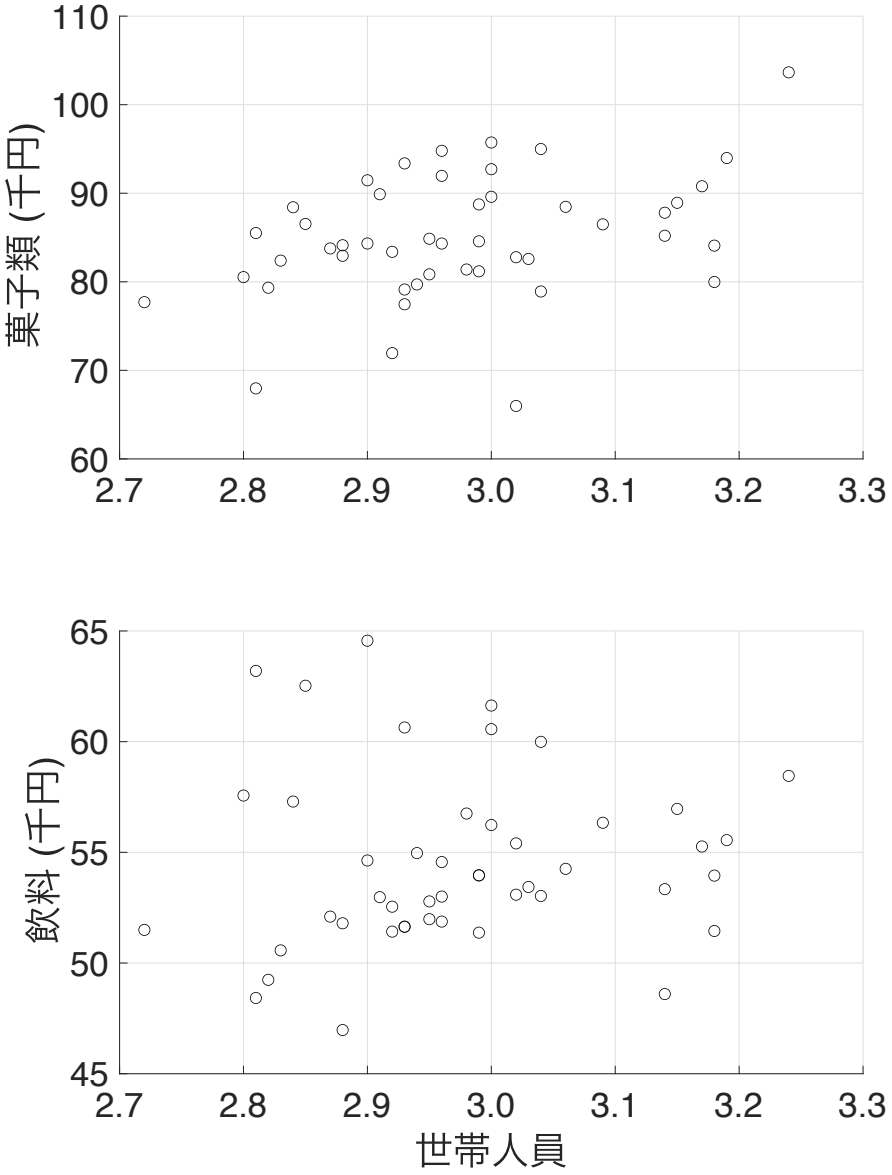


散布図

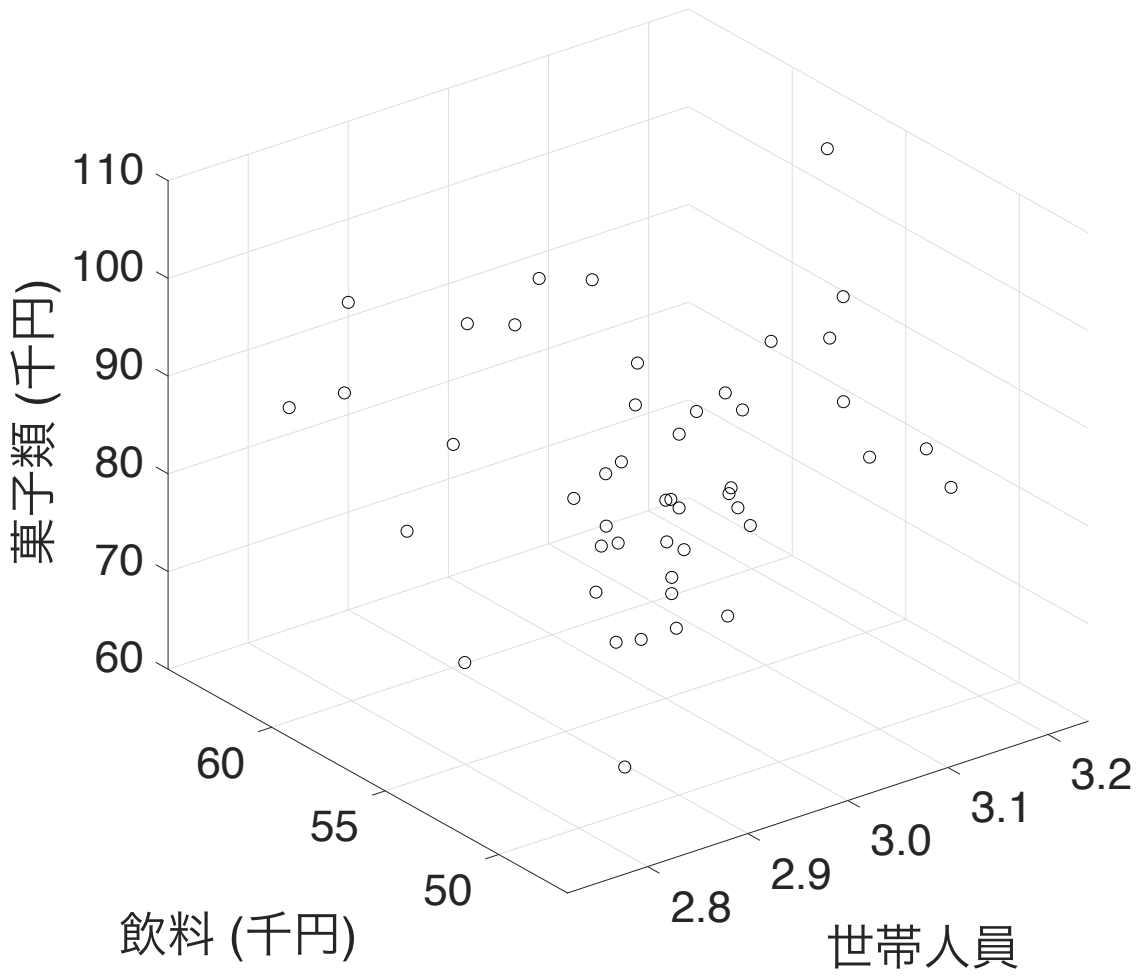
そこで二項目の値をそのまま二次元にプロットしたものを散布図とよびます。散布図に現れる点の位置や形により二項目の関連性を読み取ることができます。

世帯人員が増えると菓子類の年間支出金額も緩やかに上がっていく傾向、すなわち正の関連性が見えます。

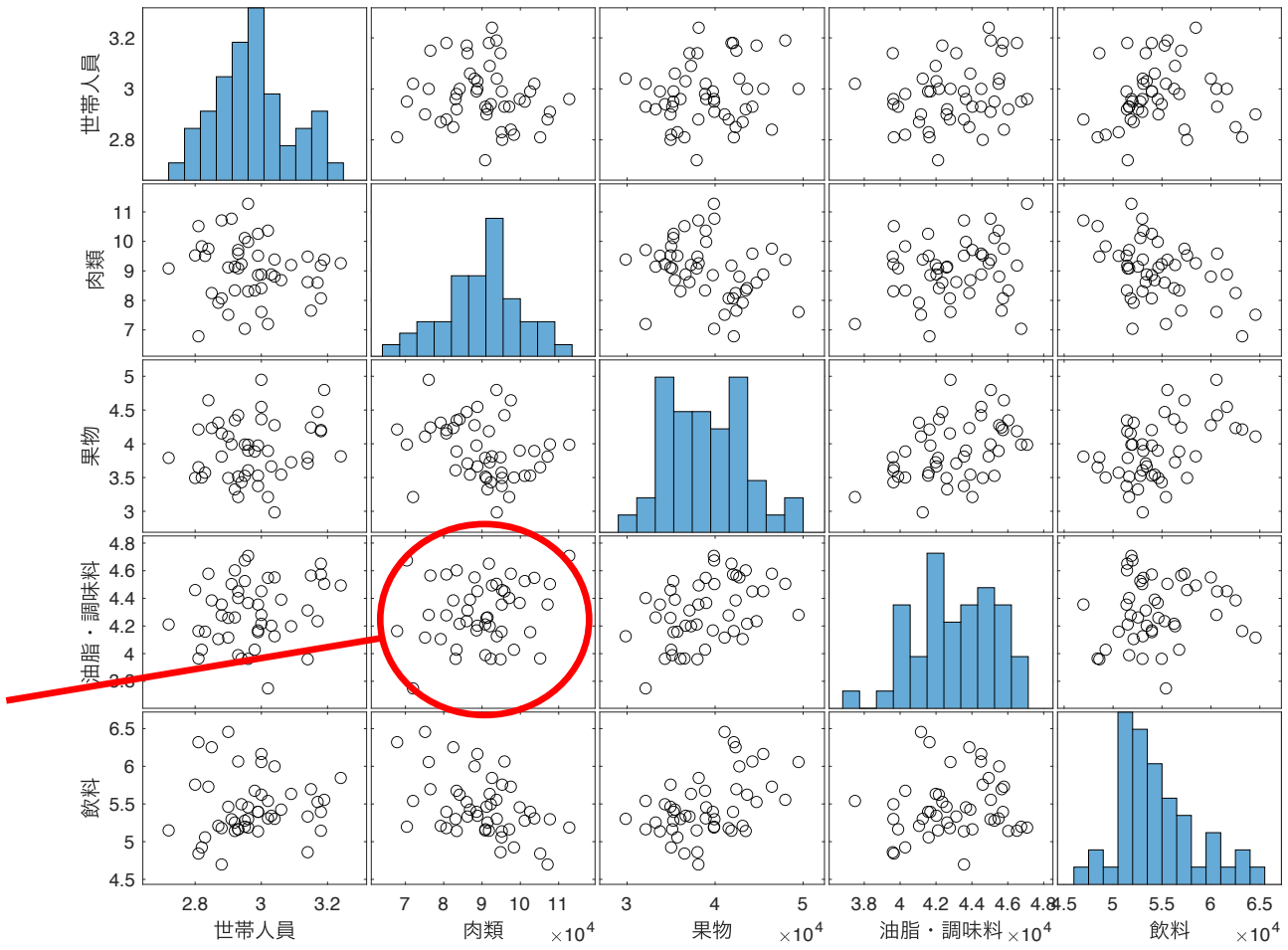
その一方で、世帯人員が多くても飲料の年間支出金額が増えたり減ったりという明確な関連性はないようです。



三次元の散布図をプロットすることで、 三項目を同時に観察することもできますが、ここから関連性を見い出すのは困難である場合がほとんどです。



データの全ての項目に対して、任意の二種類の散布を行列で表示したものを散布図行列とよびます。



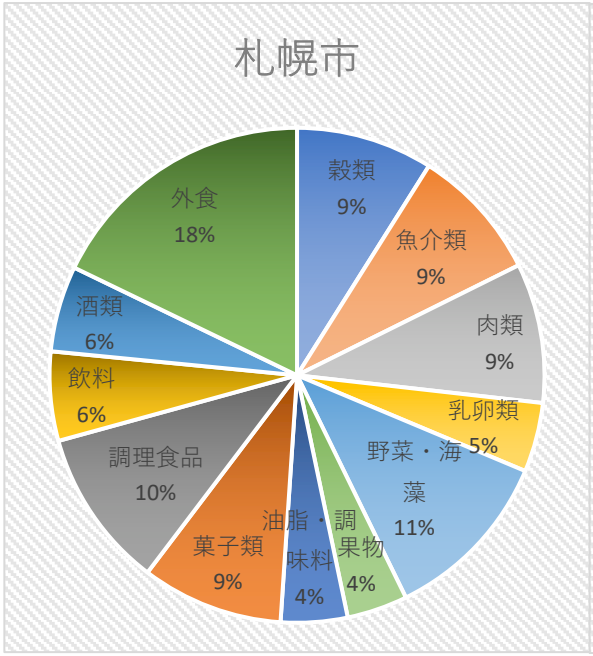
「油脂・調味料」と「肉類」の散布図をあらわします。

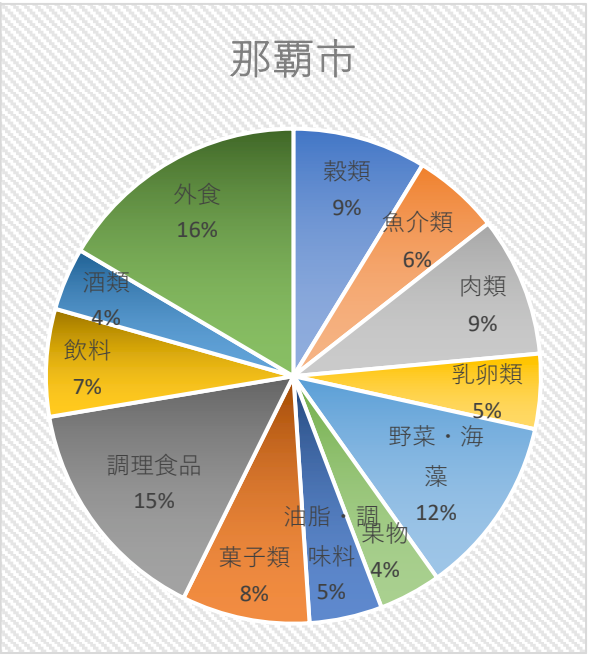
その他のデータ観察手法

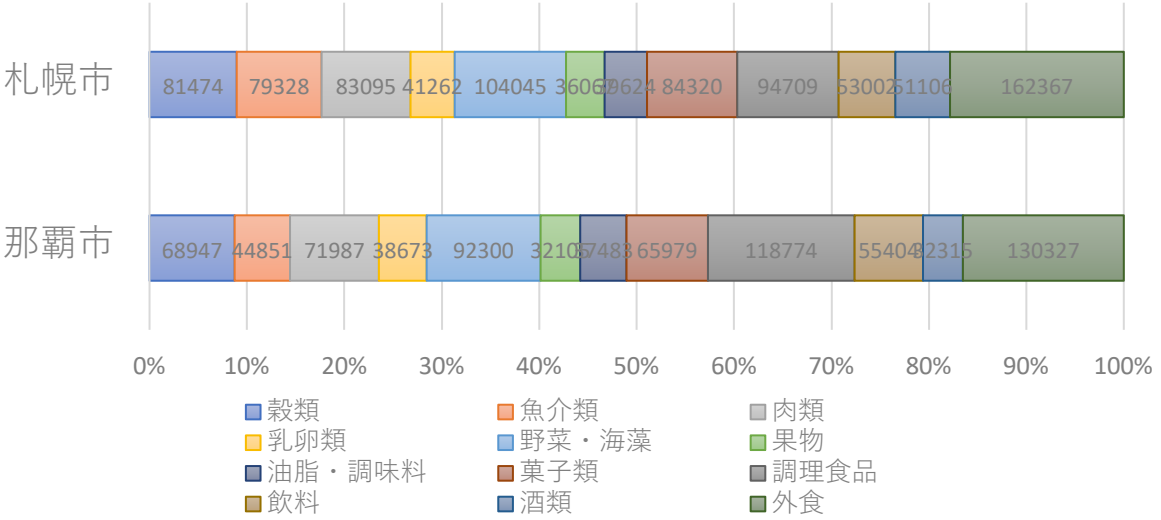
ヒストグラムや散布図以外にもデータ観察に適した方法があります。

たとえば、各項目の割合に焦点を当てたい場合には、右のような円グラフや帯グラフが適しています。

ここでは、札幌市と那覇市における食品年間支出金額そのものではなく、各項目が占める割合を、円グラフと帯グラフを用いて視覚的に表現しています。南北の二都市間に大きな相違性は見られないようです。

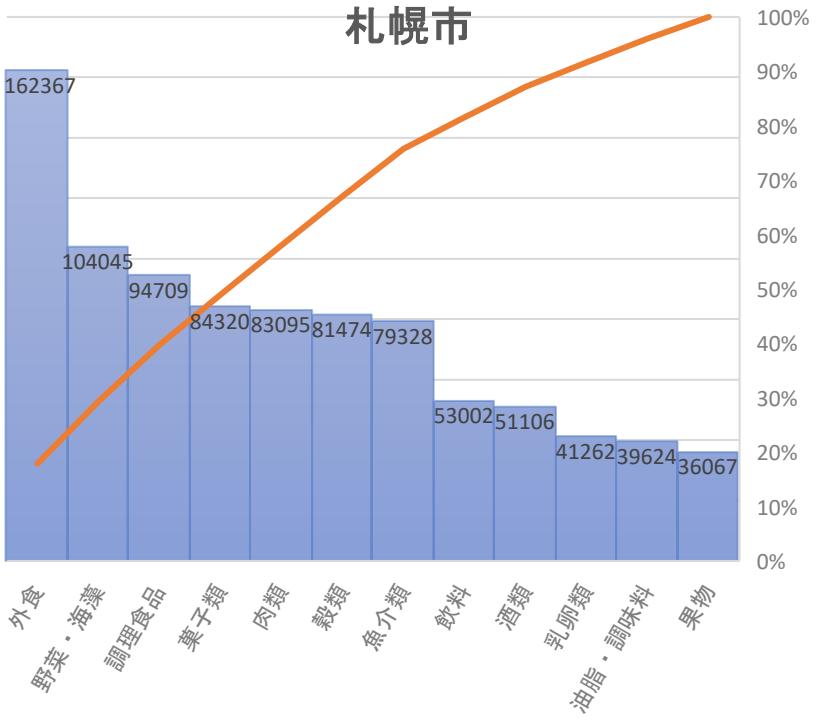


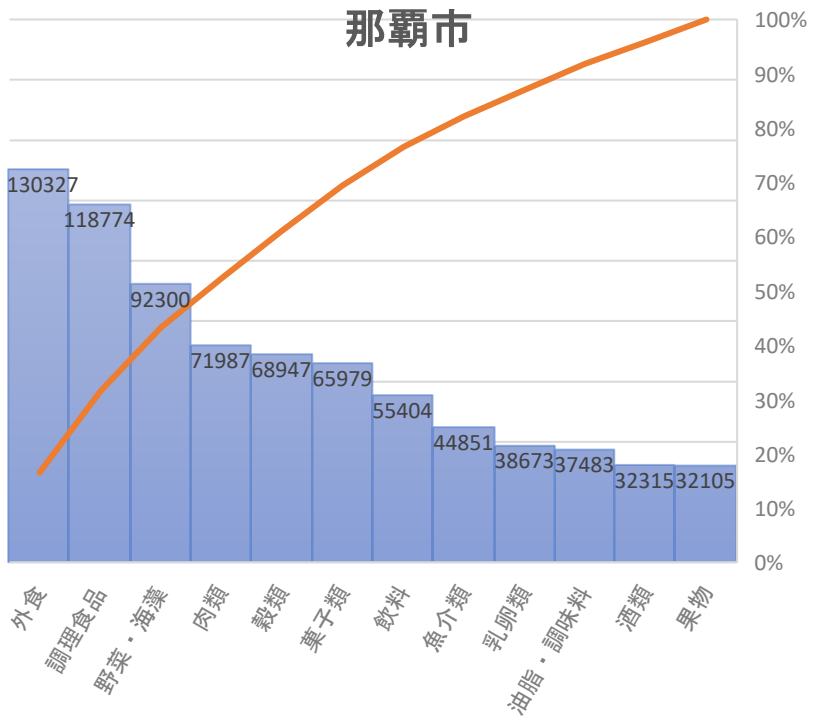




数量と割合を同時に観察する際には、棒グラフと折線グラフを一つにまとめた複合グラフが便利です。

ここでは、各都市の食品年間支出金額が大きい順に並べられたヒストグラムに重ねて、累積相対比率がプロットされていて、このようなグラフを特にパレート図とよびます。

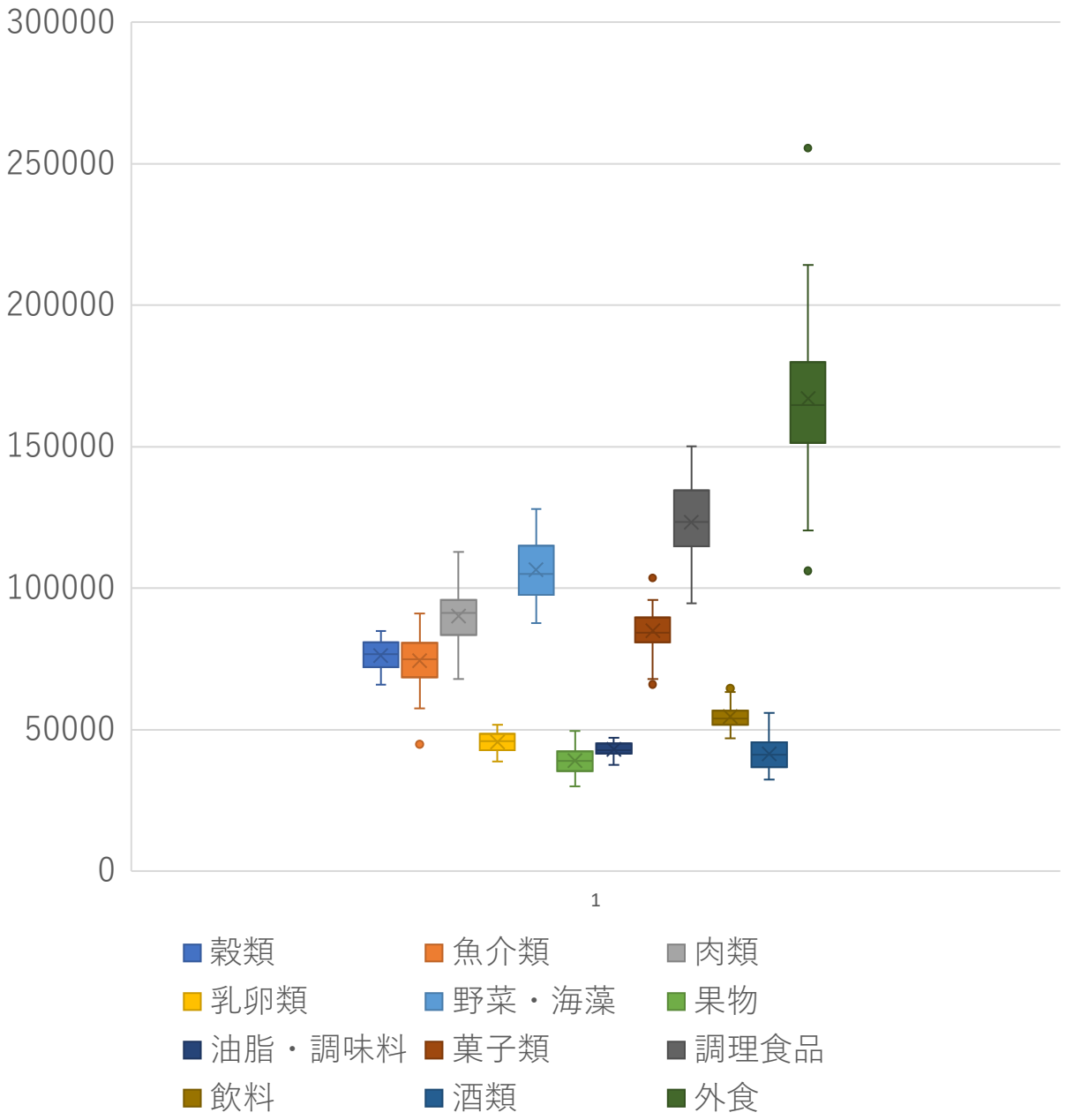




右のグラフは箱ひげ図とよばれ、「平均値」「四分位数」「最大値」「最小値」「外れ値」といった情報の観察に便利です。

このように複数項目の箱ひげ図を一括してプロットすることで、それらの情報の項目間比較をすることも可能です。

円グラフ、帯グラフ、箱ヒゲ図等のデータ観察手法については、「1-5. データ可視化」の節で詳しく学習します。



時間に沿って集計されるデータを時系列データとよびます。横軸に時点、縦軸に対象となる項目を折線グラフで表すと、時間に沿ってデータがどのように変化するかを見やすく表現することができます。下の時系列グラフは4年毎に開催される夏季オリンピックにおける男子100mの各メダルのタイム（秒）で、ゆるやかに記録が更新されていくのがわかります。時系列については「1-4. データ分析」、折線グラフについては「1-5. データ可視化」の節で詳しく学習します。

開催年	開催地	金メダル	銀メダル	銅メダル
2016	リオデジャネイロ	9.81	9.89	9.91
2012	ロンドン	9.63	9.75	9.79
2008	北京	9.69	9.89	9.91
2004	アテネ	9.85	9.86	9.87
2000	シドニー	9.87	9.99	10.04
1996	アトランタ	9.84	9.89	9.90
1992	バルセロナ	9.96	10.02	10.04
1988	ソウル	9.92	9.97	9.99
1984	ロサンゼルス	9.99	10.19	10.22
1980	モスクワ	10.25	10.25	10.42
1976	モントリオール	10.06	10.08	10.14
1972	ミュンヘン	10.14	10.24	10.33
1968	メキシコ	9.95	10.04	10.07
1964	東京	10.0	10.2	10.2
1960	ローマ	10.2	10.2	10.3
1956	メルボルン	10.5	10.5	10.6
1952	ヘルシンキ	10.4	10.4	10.4
1948	ロンドン	10.3	10.4	10.4