オフラインでも学べる！

初めて学ぶ統計
－ 公務員のためのオンライン講座 －
コースポイント集

＊ ＊

「初めて学ぶ統計　－ 公務員のためのオンライン講座　－」では上記の内容を学習しました。
このコースポイント集で全5章の学習のポイントをふりかえることができます。

＊ ＊

ぜひご活用ください。

第１章　統計とは

統計の定義

統計は「一定の条件で定められた集団について調べた結果を、 集計・加工して得られた数値」と定義されています。

統計を利用することの利点

全体の特徴を俯瞰的に捉えることができること、また、誰もが納得できる客観的な根拠を提示できることが統計を利用することの利点といえます。

統計を正しく利用するポイント

明確な定義に基づき、明確な条件で得られたデータであるか確認することが大切です。

Point 1：何を対象として集計しているかを知る

Point 2：統計の各項目は何を意味しているのかを正確に知る

Point 3：各数値は、何を調べて集計されたのかを正確に知る

Point 4：各数値は、いつの、どのような状況を表しているのかを正確に知る

公的統計の利用

「**公的統計**」は誰でも利用可能です。法令上で定められた利用や行政施策の立案、政策の評価における利用など、様々な場面で利用されます。

第２章　データの性質と代表値

データの分類

データは以下のように大別することができます。

●**質的データ**（数量で表すことができないデータ）

・**名義尺度**：順序や大小がないもの
（例）国籍、男女、血液型など

・**順序尺度**：何らかの順序が明確なもの
（例）テストの順位、検定試験の級、 満足度など

●**量的データ**（数量で表すことができるデータ ）

・**連続データ**：一定範囲であれば、その中のどの数値もとり得るもの
（例）気温、体重など

・**離散データ**：一定の値だけで、その間の数値はとり得ないもの
（例）世帯人員、コンビニエンスストアの数など

・**間隔尺度**：その数値やその間隔には共通認識があるが、ある値を別の値で割っても意味をなさないもの
（例）時刻、気温、偏差値など

・**比例尺度**：ある値と別の値の程度を比によって表すことができるもの
（例）経過時間、速度、年齢、体重など

なお、質的データは数値化して量的データに変換することにより、集計処理ができるようになり、統計に活用しやすくなります。

（例）男性を1、女性を 0 と二値変数に変換して集計、分析する等

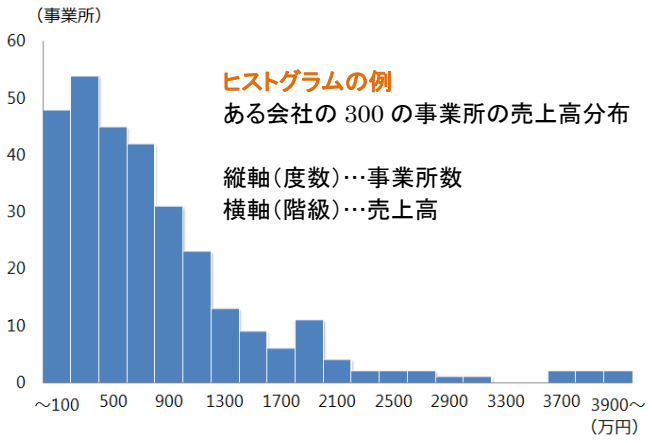
度数分布とヒストグラム

データ全体を区別する区分を「**階級**」、各階級に属するデータの個数を「**度数**」といいます。設定した各階級における度数の、全体の分布状況を「**度数分布**」といい、階級ごとの度数を、柱の面積で表したグラフのことを「**ヒストグラム**」といいます。ヒストグラムを活用することで階級毎の度数の分布状況が視覚化され、わかりやすくなります。

ヒストグラムにおいてデータが集中している箇所を「**峰（ピーク）**」とよびます。

ヒストグラムと棒グラフ

ヒストグラムは、横軸が必ず数値であり、量のつながり（連続性）を表現するために、柱同士の間隔はあけません。棒グラフが度数を「棒の長さ」のみで表すのに対し、ヒストグラムは「柱の面積、縦×横」で表します。



ヒストグラムの階級の幅（数）

ヒストグラムの階級の幅は広すぎると峰の位置が不明確になり、逆に狭過ぎると凹凸が激しく全体像が不明確

になります。

ヒストグラムの階級の数を決める一つの方法として「スタージェスの公式」というものがあります。

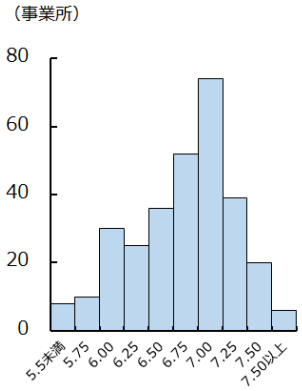
$$m = 1 + \log_2 n = 1 + \frac{\log_{10} n}{\log_{10} 2}$$

m：階級の数、n：データ数

ヒストグラムの対数変換

収入、貯蓄、資本金等の分布のように裾野が片方に大きく広がったヒストグラムの場合、各階級の値を「対数変換」するという方法があります。

対数変換したグラフでは、分布形が左右対称に近づきますが、対数変換した数値による階級は意味が分かりにくくなるので注意が必要です。



対数変換をすると・・・

80万円　→ $\log_{10}(80) \fallingdotseq 5.903$

2,000万円　→ $\log_{10}(2000) \fallingdotseq 7.301$

対数変換後の階級	売上高（万円）
～5.50	～31.6
5.50～5.75	31.6～56.2
5.75～6.00	56.2～100
6.00～6.25	100～177.8
6.25～6.50	177.8～316.2
6.50～6.75	316.2～562.3
6.75～7.00	562.3～1,000
7.00～7.25	1,000～1,778.3
7.25～7.50	1,778.3～3,162.3
7.50～	3,162.3～

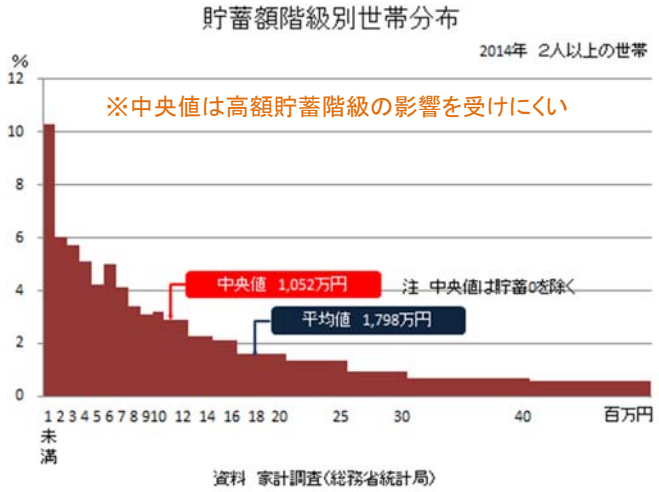
データの代表値　～平均値～

平均値には以下のようなものがあります。

- **算術平均**：データの総和をデータ数で割ったもの
- **幾何平均**：データの数値をすべて掛けて、データの個数による累乗根をとったもの
- **トリム平均**：両端のデータを除いて計算したもの
- **加重平均**：同じ値のデータの個数を重みとして計算したもの

データの代表値　～中央値～

中央値は、データ全体を順番に並べたときの真ん中の値です。平均値に比べて、外れ値（他の値から大きくはずれたもの）の影響を受けにくいのが特徴です。



データの代表値　～最頻値～

最頻値とは、最も度数が多い階級の値をいいます。「いくつ以上～いくつ未満」など幅をもって表現されている階級において、特定の値を「最頻値」として決めたい場合は、以下のような算出方法があります。

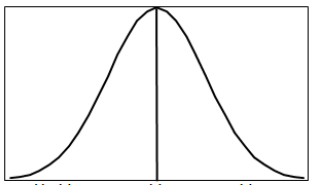
（例1）階級の真ん中の値を最頻値とする

（例2）最頻値を含む階級の度数と両隣の階級の度数の差の比で案分する

平均値と中央値と最頻値の違い

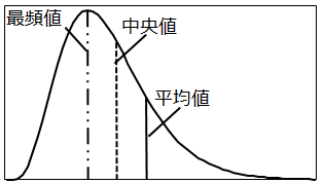
データの分布状態によって平均値、中央値、最頻値の関係に違いが生じてくるのでデータの特徴等を考慮し、最も的確な代表値を選びましょう。

左右対称分布

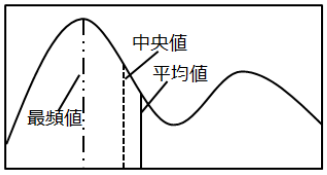


平均値＝中央値＝最頻値

右側の裾が長い分布



最頻値＜中央値＜平均値



双峰性の分布

第３章　データの分布と相関

データの散らばり

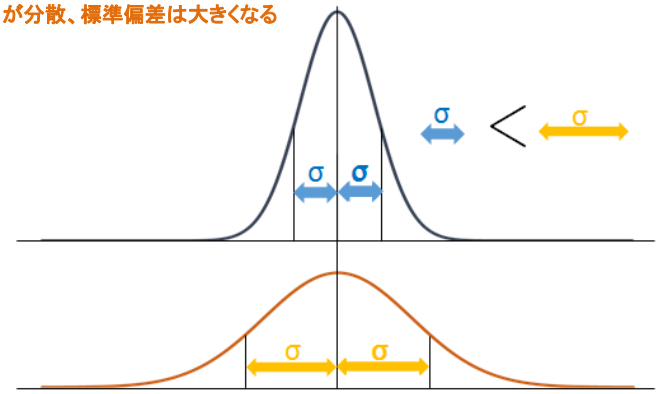
各データの平均からの差を「**偏差**」といい、各データの偏差を用いてデータ全体の散らばり「**分散**」を計算することができます。また「**標準偏差**」は分散の平方根をとったもので、それぞれの計算式は次のようになります。

偏差（i）＝{データ（i）－平均｝

$$分散(\sigma^2)=\frac{\{データ_{(1)}-平均\}^2+\cdots+\{データ_{(N)}-平均\}^2}{N}$$

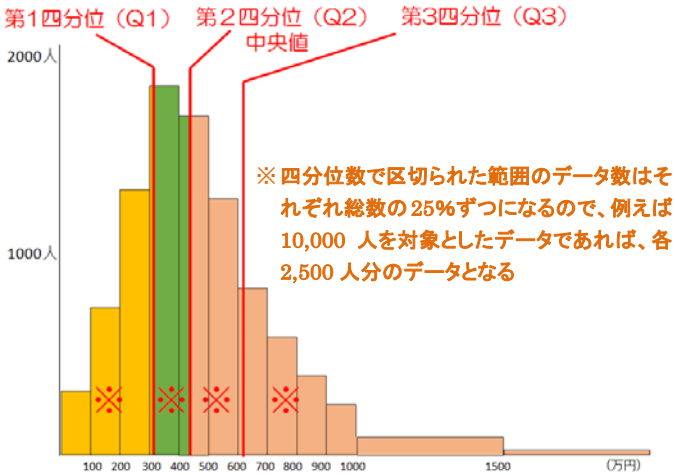
$$標準偏差(\sigma)=\sqrt{分散（\sigma^2）}$$

ばらつきの大きい下の分布の方が分散、標準偏差は大きくなる



四分位数、四分位範囲、四分位偏差

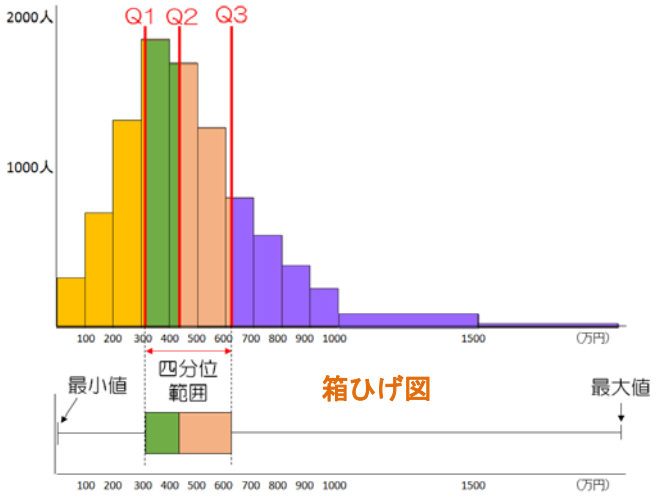
データを小さい方から順に並べ、中央値を**第2四分位数（Q2）**とし、第2四分位数（Q2）より小さい値の集団の中での中央値を**第1四分位数（Q1）**、第2四分位数（Q2）より大きい値の集団の中での中央値を**第3四分位数（Q3）**といいます。



第3四分位から第1四分位を引いた値を「**四分位範囲**」といい、四分位範囲を2で割ったものを「**四分位偏差**」といいます。ばらつきの大きい分布においては、四分位範囲、四分位偏差ともに大きくなります。

箱ひげ図

四分位範囲に記載した箱の第2四分位の値に線を引き、データの最大値と最小値まで線を引いたものを「**箱ひげ図**」といいます。サンプルサイズが異なる箱ひげ図を並べて見る際には箱ひげ図の幅を変えて表現できます。箱ひげ図は狭いスペースに複数の分布を並べて表現することが可能です。



パーセンタイル

四分位数はデータ数全体を 25%ずつで区切る値でしたが、この割合を任意で決めることができます。このデータ数を区切る値を「**パーセンタイル**」といいます。

（例）下位 10%のデータを区切る値…10 パーセンタイル

正規分布

様々な要因が積み重なって発生する誤差、成長など自然界でしばしば観察される釣鐘型の分布のことを「**正規分布**」といい、正規分布は、平均と標準偏差が決まれば、その形が決まります。

偏差値と標準化

テストの教科別得点による総合評価では、各教科における得点分布のばらつきの違いを考慮する必要があります。このような場合、評価対象者の得点が平均から標準偏差の何倍離れているかを指標化した「**偏差値**」という考え方をよく用います。

$$偏差値=\frac{得点-平均点}{標準偏差}\times 10+50$$

また、各データを以下の式に当てはめて変換することを「データの**標準化**」といい、標準化されたデータの平均は 0 となり、その標準偏差は 1 となります。

$$標準化データ=\frac{得点-平均点}{標準偏差}$$

標準正規分布

標準化されたデータが正規分布に近い分布と判断される場合、個々のデータが標準正規分布のどのくらいの位置（何％点）となっているかを割り出すことができます。

相関図

「身長が高い人は体重が重い」といったように、データの項目には相互に関係性があると思われるものがあります。このようなデータの項目間の関係性を見る際には、各項目を縦軸と横軸にとってデータをプロットします。こうして作成された図を相関図、又は散布図とよびます。

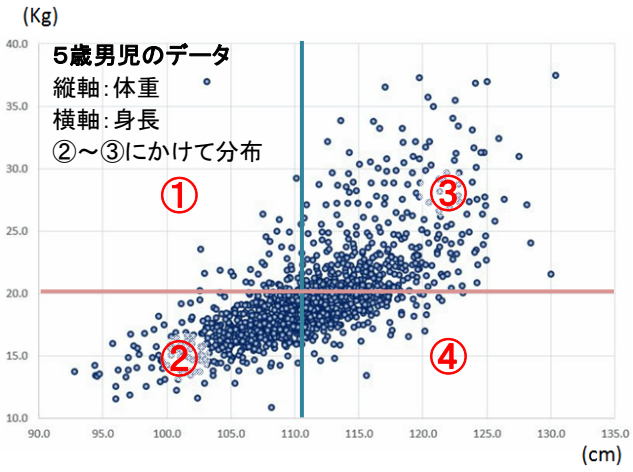
散布図の縦軸・横軸をデータの平均値で4つの領域に区切って相関の傾向を見ることができます。

相関のパターンには、以下の3パターンがあり、AとBは「データ間に関係性がある」、Cは「データ間に関係性はない」と推測できます。さらに一方の項目の増減と他方の項目の増減に直線的関係性があることを「相関がある」といいます。

A･･･②のエリアから③のエリアにかけて分布

B･･･①のエリアから④のエリアにかけて分布

C･･･①②③④すべてのエリアにまんべんなく分布



相関係数

変数同士の相関の強さは「**相関係数**」で表すことができます。相関係数は以下の計算式で求められます。

$$\text{項目}X\text{と}Y\text{の相関係数}=\frac{\{X_{(1)}\text{の偏差}\times Y_{(1)}\text{の偏差}\}+\cdots+\{X_{(n)}\text{の偏差}\times Y_{(n)}\text{の偏差}\}}{n}\Big/\left(\text{標準偏差}\ \sigma(X)\times\text{標準偏差}\ \sigma(Y)\right)$$

算出された数値が 0 より大きい場合は「**正の相関**」が、0 より小さい場合は「**負の相関**」があるといえます。ただし、いずれにおいても数値が 0 に近い場合は「**相関がない**」と考えた方がいいでしょう。相関係数と相関図には以下の特性があります。

- 相関係数の最小値は−1、最大値は+1
- 相関係数は−1や1に近いほど、相関図上では直線的な関係が強い
- 相関係数は0に近いほど相関図上では直線的な関係が弱い

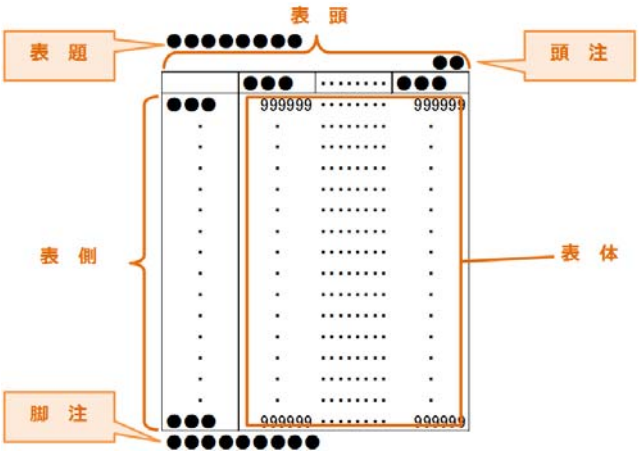
相関係数は極端なデータの存在に大きな影響を受けるので、そのような場合は、縦軸と横軸の変数をそれぞれの順位にして相関図、相関係数を見ることが有効なケースがあり、これを「**順位相関**」といいます。

また、2つの変数には直接的な関係がないにも関わらず、別の共通の要因によってもたらされた変化があたかも2変数間に関連があるように見せてしまうことを「**疑似相関**」といいます。

第４章　データの見方

統計表

統計表は以下のような構造になっています。



表題の記述には以下のルールがあります。

- 分類項目がクロスしている場合、カンマ(,)で結ばれる

男女, 年齢階級, 産業大分類別就業者数

	男	女
	産業大分類	産業大分類
年齢階級		

- 分類項目が並列の場合、なかてん(･)で結ばれる

産業大分類･職業大分類, 年齢階級別就業者数

	産業大分類	職業大分類
年齢階級		

表頭、表側の分類事項

表頭、表側の分類には、性別、産業、職業といった「**質的分類**」と、年齢、年間収入、従業者数といった「**量的分類**」があります。質的分類は各統計間で定義が異なると比較が困難になるので、「日本標準産業分類」や「日本標準職業分類」といった**標準統計分類**が設定されており、各統計はこれに基づいた分類で集計を行っています。

統計表の数値を理解するための注意点

統計表の数値を理解するためには、まず用語の定義と調査方法を理解することが重要です。

(例)「完全失業者」の定義

① 月末1週間に少しも仕事をしなかった

② 仕事があればすぐに就くことができる

③ 月末1週間に仕事を探す活動や事業を始める準備をしていた

以上の条件をすべて満たす者

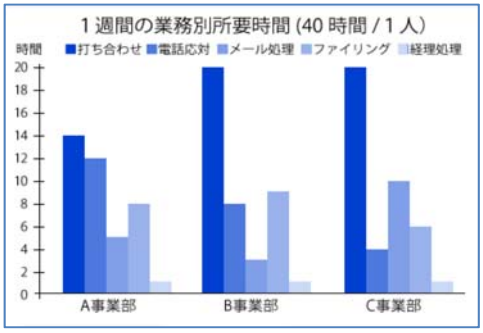
統計表の中の記号の意味

統計表の中で用いられる記号のそれぞれの意味は以下のとおりです。

0又は0.0	該当する数値は存在するが、表示する単位に満たない場合
－（バー）	定義上、該当する数値が存在しない場合
･･･（スリードット）	数値が得られない場合
X（エックス）	調査客体が特定できてしまうため秘匿されている場合

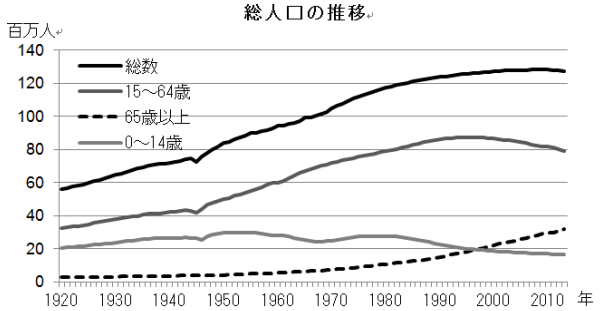
棒グラフ

数量の大小を比較する際に使用し、棒の高さや長さが数量を表します。棒を横向きにした横棒グラフ、何種類かの値を同時にグラフ化した複数系列の棒グラフもあります。



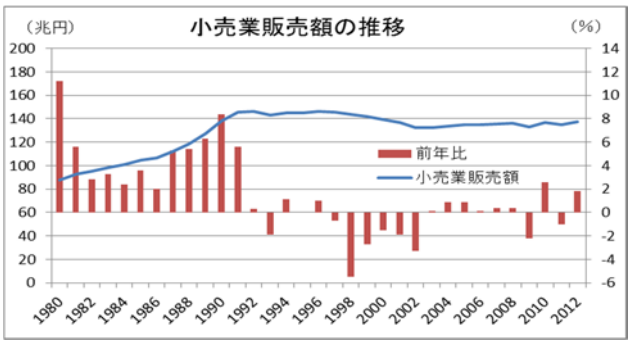
折れ線グラフ

時間とともに数量が変わる様子を折れ線の傾き方で表します。傾きが急な場合は大きく増加(減少)し、緩やかな場合は変化が少ないといえます。横軸は必ず目盛を等間隔に設定することが大切です。



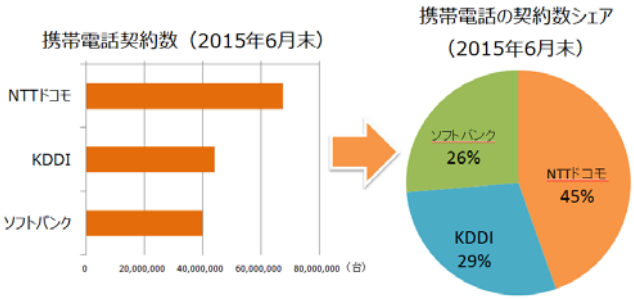
複合グラフ

棒グラフと折れ線グラフを一つにまとめたグラフが典型的なものです。



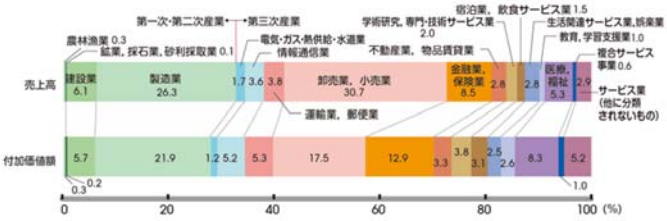
円グラフ

全体に対する割合を視覚的に表現するグラフで、扇形の中心角の大きさで各カテゴリーの割合を表します。



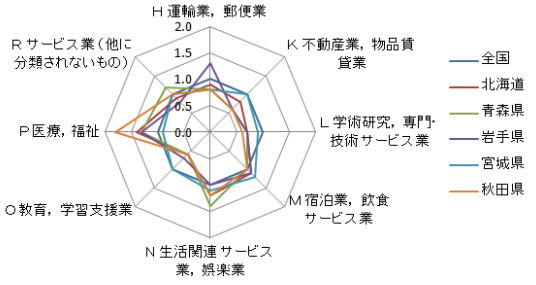
帯グラフ

全体を100%としたときのそれぞれの割合を帯の幅で表します。円グラフ同様に割合を表すグラフですが、総数の異なる二つのデータは、割合を計算し、帯グラフにして並べると比較をしやすくなります。



レーダーチャート

項目の数に合った多角形の形をしており、各頂点はそれぞれの項目の基準値に対する比率に対応させ、各頂点を線分で結びます。値が大きいほど外に広がり、小さいほど中心に集束し、また各項目の値のバランスが取れているほど正多角形に近い形となります。



ヒストグラム

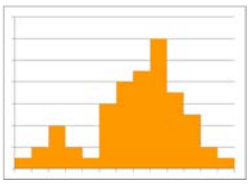
連続型の量的データの度数分布表を柱の面積で表したグラフで、横軸が必ず数値となっています。

量のつながり(連続性)を表現するために、柱同士の間隔はあけません。

ヒストグラムからは以下の特徴を読みとることができます。

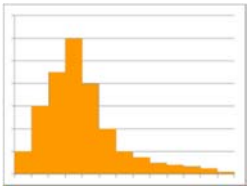
- **多峰性**

  ピークが2つ以上あり、異質な集団のデータが混在している可能性があるのでデータを分けて分析するなどの工夫が必要。



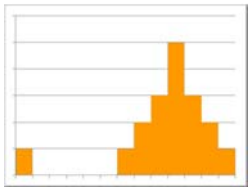
- **左右非対称**

  ピークが右や左に偏り、片側に長く裾を引く場合がある。代表値を見る場合には注意が必要。



- **外れ値**

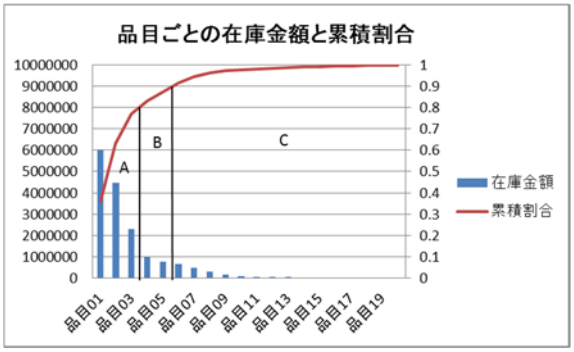
  異質なデータが混在している可能性がある。入力ミスや異質なデータが混在していないかの確認が必要。



パレート図

質的データの度数分布表をもとに度数を表す棒グラフと累積相対度数を表す折れ線グラフを合わせて表したグラフです。

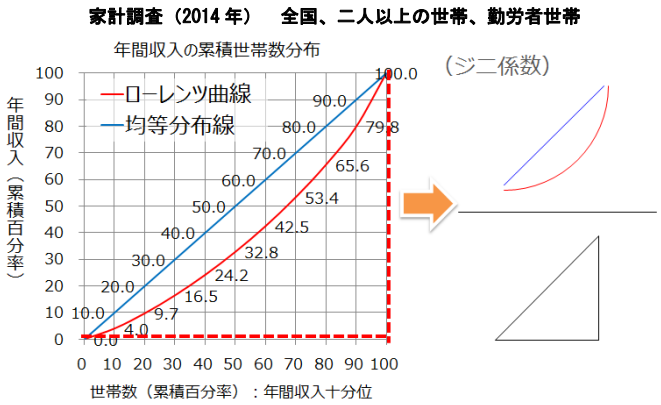
パレート図を用いて全体に占める度数の割合が大きい項目をA、中程度の項目をB、少ない項目はCと分類して、全体に占める割合の大きさごとに分析を行っていく分析手法を「**ABC分析**」と言います。この分析手法は品質管理等で活用されています。



ローレンツ曲線とジニ係数

データのばらつきの大きさ、分配の不平等度を表すものとして、「**ローレンツ曲線**」と「**ジニ係数**」があります。

次の図のように縦軸と横軸にそれぞれの値の累積百分率をとって 10%の世帯で全体の何パーセントの収入を得ているか、20%でいくつ、というようにグラフを描いたもので、この曲線が下方向に張り出すほど、不平等度が高いことを表します。ジニ係数とは、均等分布線と横軸と縦軸(右側)で囲まれた三角形の面積を分母に、均等分布線とローレンツ曲線で囲まれた弓形の面積を分子にとって計算したものです。



構成比と相対比

比率には、総数とその内訳の比率を表す「**構成比**」と、異なるデータを分子・分母に取った比率や単位当たりの量といった「**相対比**」があります。

- **構成比の例**

  15 歳未満人口割合＝(15 歳未満の人数)／(総人口)

  エンゲル係数＝(食料費)／(消費支出)

- **相対比の例**

  人口密度＝ (人口)／(面積)

  BMI＝(体重)／(身長 ^{2})

構成比を用いた地域間比較

地域の産業構造の特徴を見比べる際に実数だけで見比べると人口規模の違いにより、その特徴が見えにくくなることがあります。こうした場合には、構成比を用いて比較するとその特徴がより分かりやすくなります。

相対比を用いた地域間比較

構成比と同様に規模の影響を排除して比較する際に用いられ、一般的に分母に基準とする単位を取ることが多く、地域間比較では目的に応じて以下のような分母をとります。

- 近接性や利便性を見たい場合 ⇒ **面積を分母**
  - 温泉の数　等
  - コンビニエンスストアの数(利便性)
- 一人当たりの量を見たい場合 ⇒ **人口を分母**
  - 自家用車の保有数　等
  - コンビニエンスストアの数 (混雑率)

時系列データの種類

時間の順序で並べられたデータを「**時系列データ**」といい、一般的に時点の古い方から新しい方に向かってデータが並べられます。様々な観測頻度や区切りの時系列データがあるので、利用する際は注意が必要です。

時系列データは、ある一時点の状態をとらえた「**ストックデータ（静態データ）**」とある期間内の発生量や変化量を表した「**フローデータ（動態データ）**」があります。

- **ストックデータ（静態データ）の例**
  平成 27 年 10 月 1 日現在の人口
- **フローデータ（動態データ）の例**
  平成 26 年の 1 年間の出生数

経済データでは、フローデータの減少が先に発生し、その後ストックデータが減少に転じるという傾向が出ますので、経済の見通し等はフローデータで見て、普及状況等はストックデータで見るといった使い分けをします。

名目と実質

金額を扱う統計では「**名目**」と「**実質**」という考え方が用いられます。

名目はその時々の価格により表した金額で、消費実感に近い金額であり、実質はある基準となる時点の価格により表した金額です。実質は物価変動を排除して、購入量による金額変動を見たい時などに利用します。

時系列データにおける季節性

「**季節変動**」とは季節に関連する要因によって発生する変動です。

（例）夏にビール消費が増える、冬に灯油購入が増える
　　ボーナス時期に商品売上が増える　等

その年の傾向を季節性を排除してみる方法に「前年同月比」という考え方があります。

$$\text{前年同月比} = \frac{\text{当月の値}}{\text{前年の同月の値}}$$

※季節変動のパターンは毎年ほぼ一定と仮定した際に有効

季節調整法

季節性のあるデータで前月比動向を見たい場合には、様々な「**季節調整法**」が用いられます。

季節調整法では、時系列データ（**原系列 O**）を

- **傾向変動（T）**長期にわたる傾向的な変化
- **循環変動（C）**周期的に繰り返される１年周期ではない変動
- **季節変動（S）**１年周期の規則的な変動
- **不規則変動（I）**上記以外の不規則な変動

からなると考え、季節性を除去します。

季節調整法には、前後の数か月の値を平均した値をその月の値とみなし、不規則な変動をスムーズにならす「**移動平均法**」やアメリカセンサス局が開発した「X12-ARIMA」といったものがあります。

第５章　行政運営のための公的統計

公的統計の役割

統計は、現在の状態を客観的かつ正確に把握するためのものであり、現在の状態を客観的かつ正確に捉えるためのデータを計測し、目的に応じて集計・加工し、適切に記述します。代表的な公的統計調査である国勢調査は、「国内の人口や世帯の実態を明らかにするための調査」です。

行政機関、地方公共団体や独立行政法人等が作成する「**公的統計**」に対して、民間が実施する統計調査によって得られる統計を「**民間統計**」といいます。

基幹統計と一般統計

公的統計は「**基幹統計**」と「**一般統計**」に分けられます。

- **基幹統計**
  国勢統計、国民経済計算といった特に重要な統計のことで、回答者に報告義務を課している。
- **一般統計**
  一般統計調査においては、回答者の報告は任意となっている。

一次統計と二次統計

統計を作成する方法は「**調査統計**」、「**業務統計**」、「**加工統計**」の3つに分類することができます。

- **調査統計**
  統計調査を実施することによって得られる統計
  （例）国勢調査、経済センサス、農林業センサス等
- **業務統計**
  政府の業務で得られた行政記録から作成される統計
  （例）人口動態統計、貿易統計、建築着工統計等
- **加工統計**
  調査統計や業務統計を基に加工して作成される統計
  （例）国民経済計算、消費者物価指数、鉱工業指数等

調査統計と業務統計は「**一次統計**」、加工統計は「**二次統計**」と言われています。

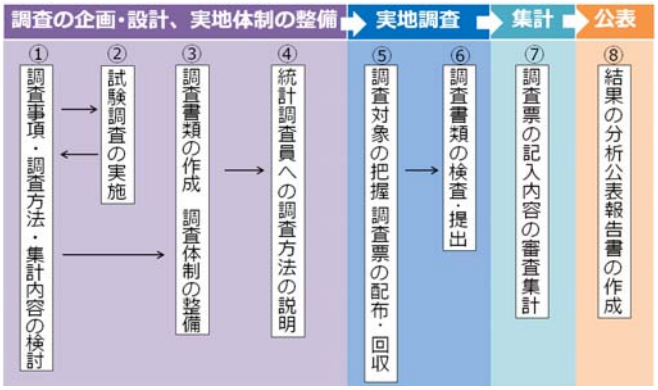
統計法

公的統計の体系的かつ効率的な整備及びその有用性の確保を図ることを目的として「**統計法**」が定められています。統計法のポイントは次のとおりです。

① 公的統計の整備に関する基本的な計画の策定
② 統計データの利用促進と秘密の保護
③ 統計委員会の設置

公的統計の作成

公的統計の企画から結果の公表までの流れは次のとおりです。



全数調査と標本調査

統計調査において、調べたい対象全体からなる集団のことを「**母集団**」、母集団から抽出された一部の集団のことを「**標本**」といいます。また、母集団のすべてを調べる調査のことを「**全数調査**」、母集団の一部の情報を基に母集団を推定するために実施する調査のことを「**標本調査**」といいます。

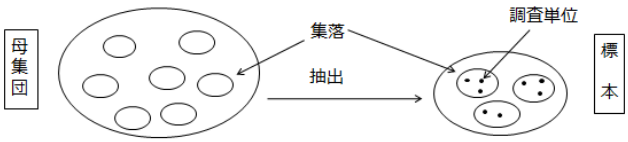
標本調査における対象の抽出方法には、調査対象を公平に選定できるように、無作為に抽出する「**無作為抽出**」と、母集団をよく代表していると考えられる調査対象を専門家の判断に基づいて抽出する「**有意抽出**」があります。無作為抽出された標本調査であっても回答には意図しない偏りが出ることもあるので注意が必要です。

様々な標本抽出方法

標本の抽出方法には次のようなものがあります。

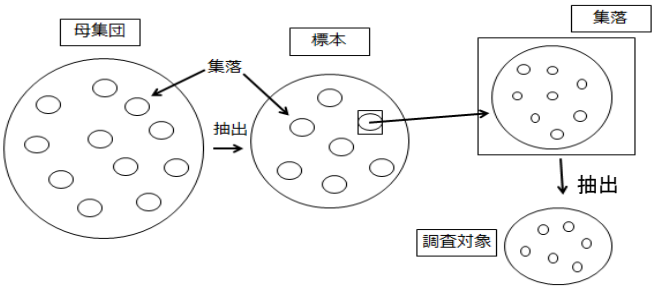
- **集落抽出法**

  調査対象の集まりである集落を無作為に抽出し、その集落内のすべての調査対象を調査する方法



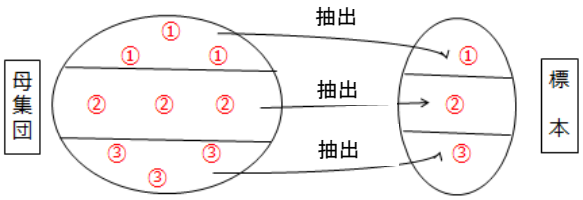
- **二段抽出法**

  1段目で調査地域を選び出し、2段目で調査地域内から調査対象を選び出すという2段階で標本を選ぶ方法



- **層別抽出法**

  調査対象を同質なグループに分け、グループごとに標本を無作為抽出する方法



結果の推定方法

標本調査では、母集団から一部の標本を抽出して、その標本の値を用いて、母集団の値を推定します。推定の方法は、標本理論に基づいた推定式によって求められます。推定式は、抽出方法によって異なります。

標本誤差と非標本誤差

標本調査の結果は、必ずしも母集団の値、つまり真の値とは一致せず、何らかの差があります。このように標本を無作為に抽出することによって生じる差のことを、「**標本誤差**」といいます。

また、標本調査の調査結果が確率的にばらつく幅を示す値を「**標準誤差**」といいます。

標準誤差は、近似的に次のように表すことができ、標本の大きさの平方根に反比例します。

$$\text{標準誤差} \fallingdotseq \frac{\text{母集団の標準偏差}}{\text{標本の大きさの平方根}}$$

標本誤差が一部の標本から母集団を推定することによって生じる誤差であるのに対して、調査や集計の不完全さによって生じる誤差のことを「**非標本誤差**」といいます。