コンピュータ演習

AIリテラシー　06講　データを読み、説明し、扱う

目次

第06講 データを読み、説明し、扱う

データの種類を知る

データの種類を知って正しく扱う

扱い方に気をつけましょう。

> 平均年齢が20代の婚活パーティ

に出掛けてみたら、

- 0歳児
- 50代

しかいないかもしれません。

連続データと離散データ

> - 連続データ：アナログデータ
> - 離散データ：デジタルデータ

時計の針で考えてみよう。

- アナログ時計では針の角度は連続に可変している(一部、秒針が360/60=6度ずつ変わるものもあるが)
- デジタル時計では数字で表されているので、**12:5.5:30**のような表記にはならず、整数で飛び飛びとなります。

質的データと量的データ

データに数字がついている場合、以下のような分類方法があります。

- 質的データ
- 量的データ

質的データ

番号がついているが、足したり引いたりすることに意味がないデータ

> - 名義尺度：性別(男：1, 女：2)等、番号がラベルでしかないもの
> - 順序尺度：アンケート(好き：1, どちらでもない:2, 嫌い：3)等、順番が関係はあるもの

量的データ

測定できる、単位があるなど、数値として意味があり、計算ができるもの

> - 間隔尺度：湿度・日付けなど、加減に意味があるもの
> - 比例尺度：身長・体重・など四則演算に意味があり、0は何もないことを表すもの

Column:データを扱うときの注意点

自分やその周りの環境を普通だ、と受け入れることは自然なことですが、注意が必要です。

- 日本で「貧しい」と思う人が、世界的に見れば中央値にある暮らしをしている

先入観なくデータを読み解くことが重要です。

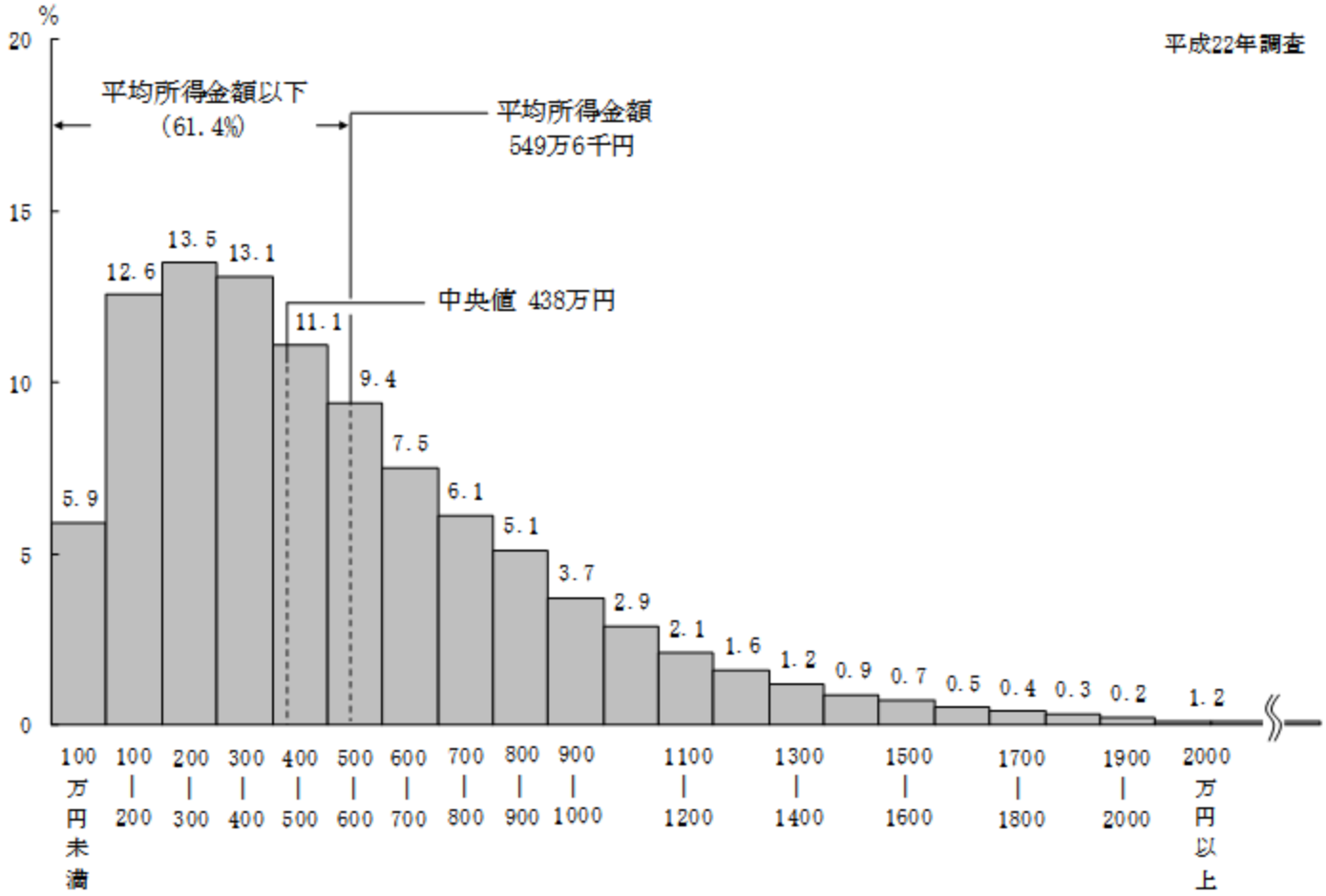
基本統計量でデータの特徴をつかむ

基本統計量とは

> 基本統計量とは、データの基本的な特徴を表す値のことで、代表値と散布度に区分できる。代表値とは、データを代表するような値のことで、例えば、平均値、最大値、最小値などがある。散布度とは、データの散らばり度合いを表すような値のことで、例えば、分散、標準偏差などがある。

「など」と書いてあるように、上記以外のもの中央値・第一四分位数....など色々あるので注意

参考グラフ：所得の分布状況



https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa10/2-2.html

基本

- 横軸に区間(階級)
- 最小値は0円
- 最大値はグラフが小さすぎて読み取れない

年収はとても範囲が広いデータである

データの真ん中を表す指標

平均549万、中央値438万、最頻値200〜300万

どれもデータの真ん中を表す指標だが、

> - 平均：全部のデータを足して、データの個数で割った値
> - 中央値(メディアン)：データを大きさ順に並べたときに、ど真ん中にくる値のこと
> (99個のデータなら50番目)
> - 最頻度(モード)：データの個数が最もたくさんある値のこと

どれが真ん中にふさわしいですか？

指標の使い分け

グラフの形状が釣鐘型であれば3つの値は近くになるのですが、グラフの形が歪んでいると、中央値・最頻値の方を見た方が良いでしょう。

データの散らばり具合を見る

> 分布：データの散らばり具合
> ヒストグラム：度数分布表すグラフ

- 【1分統計学】分散・標準偏差ってなに？

Photoshopでもウィンドウ-ヒストグラムにて利用されます。

もととなるデータを集める

母集団と標本

> 全数調査：データを収集するときに、対象となる対象全てを調査する方法
>
> - メリット：正確
> - デメリット：手間・お金がかかる

に対して

> 標本調査(サンプル調査)：一部の対象だけを調査

がある。

> - 母集団：全体
> - 標本：調査される一部の対象

標本誤差

標本を2回選べば結果は全く同じになることはほぼない。

> 標本誤差：標本の選び方による誤差

誤差が大きければ、母集団の推測は難しくなるが、誤差を限りなく小さくすれば、母集団の分析に利用可能。

無作為抽出

誤差を小さくするには母集団から標本を抜き出す時に限りなく完全に
ランダム(でたらめ)
にすると良いことが知られている。

集めたデータを集計する

クロス集計

2つ以上のデータを掛け合わせて表にまとめる集計方法。

- Excel クロス集計はピボットテーブルが便利

Excelの回でピボットテーブルは範囲外でしたが、こういうことができることを覚えておきましょう。

相関関係と因果関係

> 相関関係：一方が増加すると、他方が増加(正の相関)または減少する(負の相関)、二つの変量の関係

因果関係は原因と結果がある関係で、相関関係とは異なる考え方です。

- 因果関係があれば相関関係はある
- 相関関係があっても因果関係があるとは限らない

相関関係があって因果関係がない例

- 暑いとアイスの消費量は増える
- 暑いと溺死者が増える

しかし、

- アイスを食べるから溺死者が増える

わけではありません。

これを疑似相関と呼びます。

> 疑似相関：2つの事柄が無関係なのに、第三の要素によって意味のある関係を結んでいるかのように見えてしまうこと

地図上の可視化

5講でも多少触れてる...

地図上でデータをプロットすることで可視化して集計することも可能と言いたいのでしょうか...

誤読しないデータの読み方、データの比較方法

16回目にて詐欺グラフを紹介しました。チャートジャンクとも呼ぶそうです。

- あなたの知らない「詐欺グラフ」の世界

column: 新規技術との付き合い方について

(略)
新しい技術が登場した時に、その技術の良い面だけに目を向けて、闇雲に導入するようなこともまた、避けなければなりません。皆さんは是非、技術との上手な付き合い方を身につけてください。