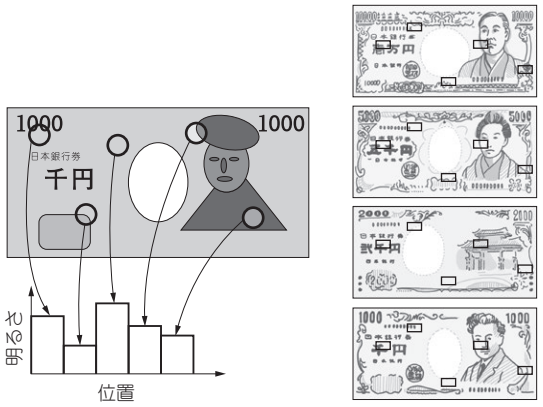
わりとよく使われるタイプは動かしてガッテン！

ダウンロード・データあります

人工知能アルゴリズム探検隊

第28回 紙幣の種類判定データの妥当性を統計的「t検定」で確かめる

牧野 浩二，石田 和義



(a) お札の5カ所の明るさを測定　(b) 1000円札以外のお札も使う

図1　お札の種類判定データの妥当性を統計的「t検定」で確かめる

前回はお札の分類を行いました（**図1**）．お札の分類には多数の明るさデータを使いましたが，ここである疑問が生じます．それは，「これらのデータにはちゃんとした差があるのか」ということです．

そこで今回は，データ同士の差を調べる手法の1つである「t検定」を紹介します．t検定そのものは人工知能のアルゴリズムではありませんが，有用な解析手法の1つです．

最初に，筆者が作成したデータからExcelを用いてt検定を体験してもらおうと思います．Excelには，t検定の計算を行える関数も用意されています．

次に，前回（第27回，2019年5月号）お札の分類で紹介した，幾つかのお札データについて「データ同士に差があるかどうか」をt検定を使って確認します．

本稿では，t検定をイメージしやすくするために，あえて簡単な言葉に置き換えています．インターネットなどで調べやすくするために，専門用語も示すことにします．他の書籍でしっかり勉強したい方は，これらの言葉が専門用語に置き換わることをご理解ください．また，t検定にはいろいろな種類がありますが，本稿では「対応のない2種のデータのt検定」を対象とします．

統計的解析手法「t検定」とは

t検定は，データの検証でさまざまなところで使われています．その使われ方の1つとして，2組のデータに「差があるかどうか」を調べることに利用されています．2組のデータとは，例えば山梨県の桃の重さと福島県の桃の重さの両データとなります．

ここで，「差があるかどうか」と書きましたが，実際にはもう少し厳密な定義があります（後に詳しく解説する）．まずは，t検定の概要をイメージすることから始めましょう．

● t分布という確率分布に従うと仮定する

t検定では，データの確率分布はt分布に従っていると仮定しています．確率分布とは，横軸のある値を取るときの確率を表しています．数学的な仮定を置くと説明できますが，ここではt分布に従っているとします．

なお，t検定を行うときには「片側検定」と「両側検定」という2つの検定方法があります．興味のある方はコラムを参考にしてください．

● 差が有効かの判断基準を決める

2つのデータの差が有効かどうかのことを「有意差」と呼びます．有意差を結論付けるには，「p値」という値が必要です．このp値が，「0.05以下ならば有意な差がある」と言えるからです．p値に関して，詳しくは後述します．

● t検定が活躍している分野[注1]

t検定は以下の分野で使われています．

- 工場の品質管理[注2]
- 薬の効果[注2]
- アンケート分析（マーケティング）[注3]
- 電車の中刷り広告にあるトクホ食品の効果グラフ

注1：データサイエンス研究所（https://www.statweb.jp/method/t-kentei）のホームページにはさまざまな事例が載っています．

第4回　パンダ/トラ/ゾウ…パラメータから分類「主成分分析」（2016年12月号）
第5回　市場調査に使われる多数データのグループ分け「クラスタ分析」（2017年2月号）
第6回　少数データを丁寧に分けられる「階層型クラスタ分析」の基本原理（2017年3月号）