

社会人のためのデータサイエンス演習

第4週:ビジネスにおける予測と分析結果の報告

第4回:分析結果の報告(解釈の注意点)

講師名:高橋 範光

第4週の内容紹介

第1回

- 回帰分析による予測

第2回

- モデル評価と予実評価

第3回

- 分析結果の報告（記述/可視化方法）

第4回

- **分析結果の報告（解釈の注意点）**

第5回

- 予測・分類等代表的手法と活用場面

第6回

- ビジネスシーンにおける「統計的検定」とその活用例

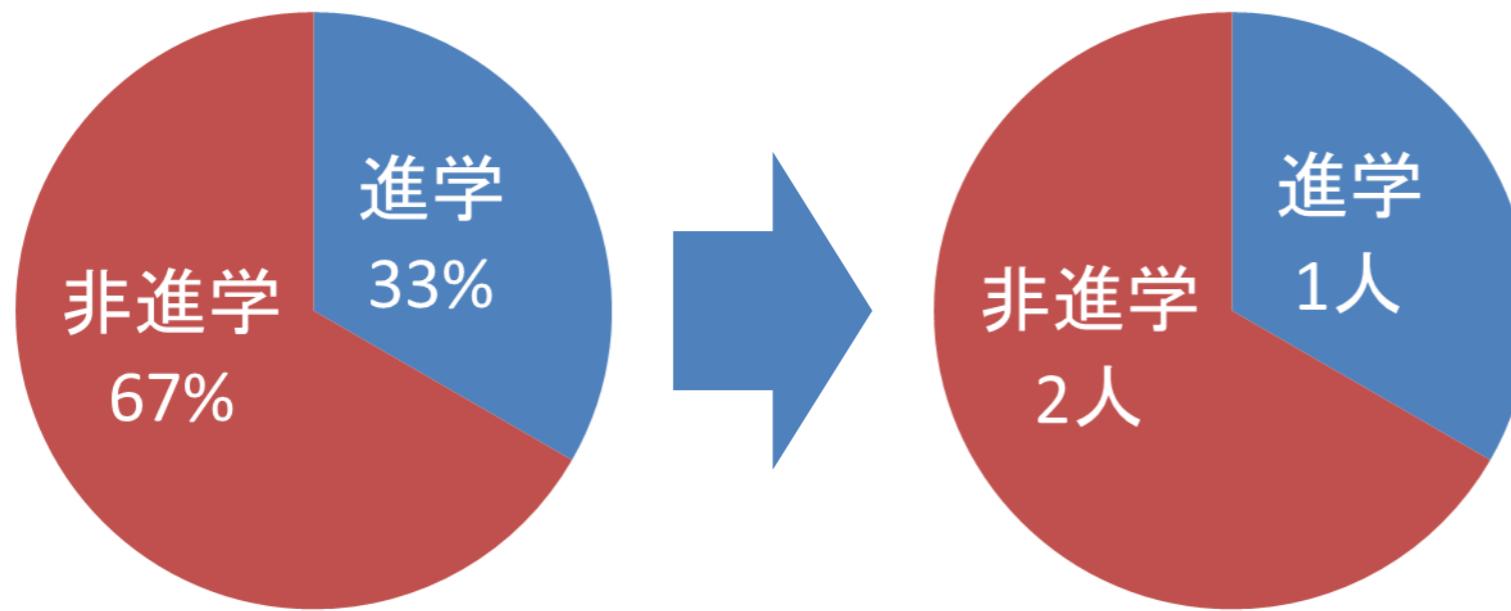
結果解釈の罠

- 分析結果の解釈には知識と注意が必要であり、熟練した人でも間違えることがある。
- より良い判断を行うためには、自身の分析に注意を払うだけでなく、他者の分析結果を見る目も養うべき。

結果解釈において陥りがちな悪例を知り、
結果を読む力を身につけよう

情報の偏り

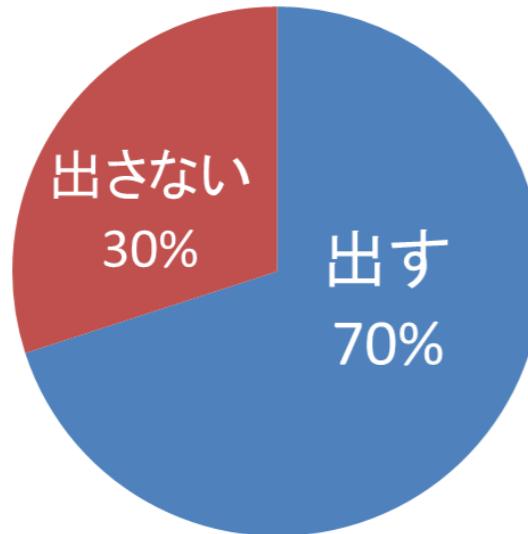
- 学部の女性卒業生の3割以上が大学院進学
 - 学部の女性卒業生は3人しかおらず、そのうち1人が大学院に進学



情報の偏り

- 年賀状を出す人は70%

- 平日昼間の固定電話による世論調査だったため、回答が高齢者に偏り

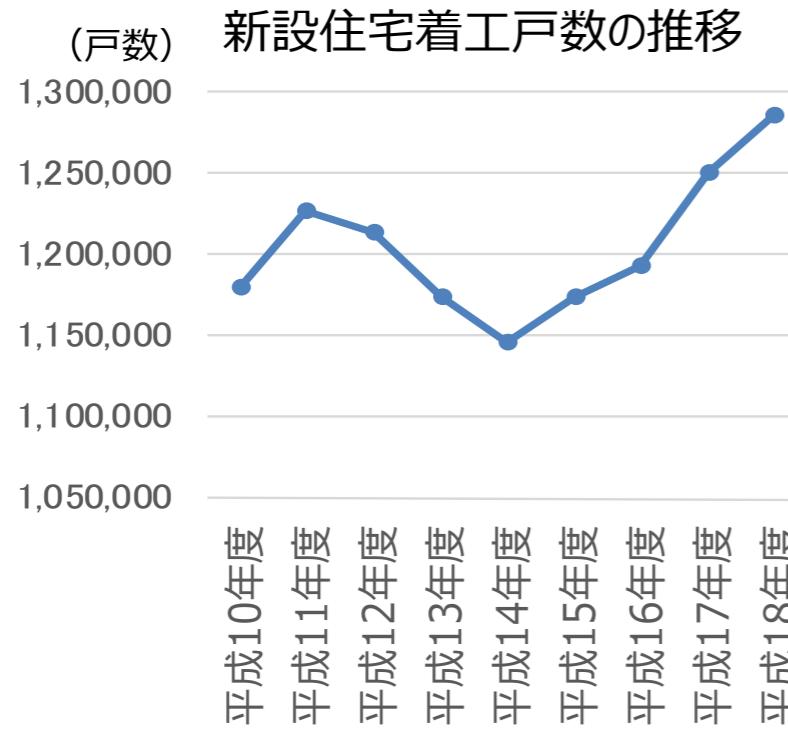


いずれのケースもデータの取り方が分かれば納得

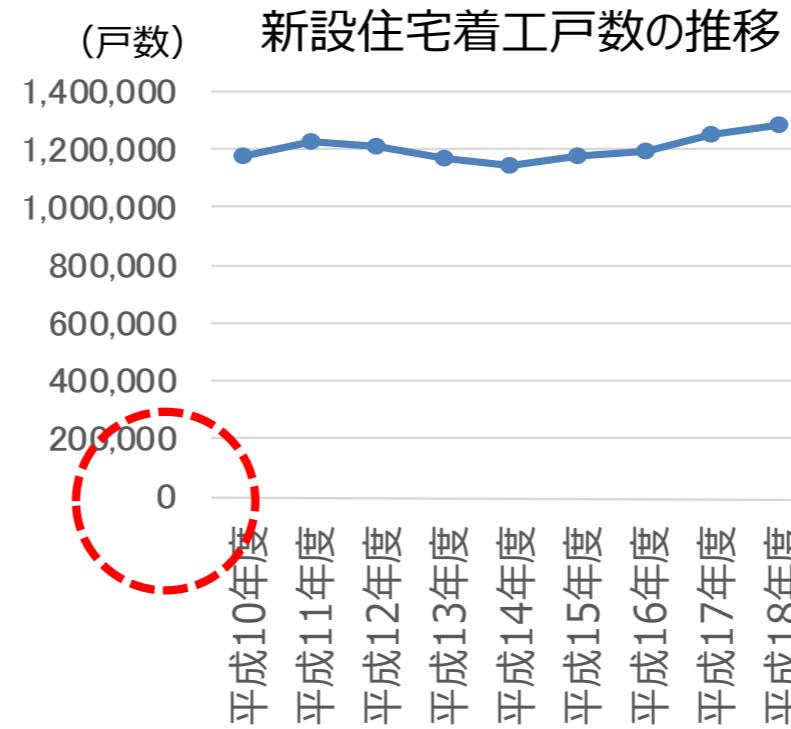
抽出条件やデータの取得方法に注意

グラフのウソ

● 住宅の着工戸数は回復しているのか



平成18年度にかけて大幅に回復？



基準点を0にしたらほぼ横ばいに
⇒軸の操作で変化が誇張されていた

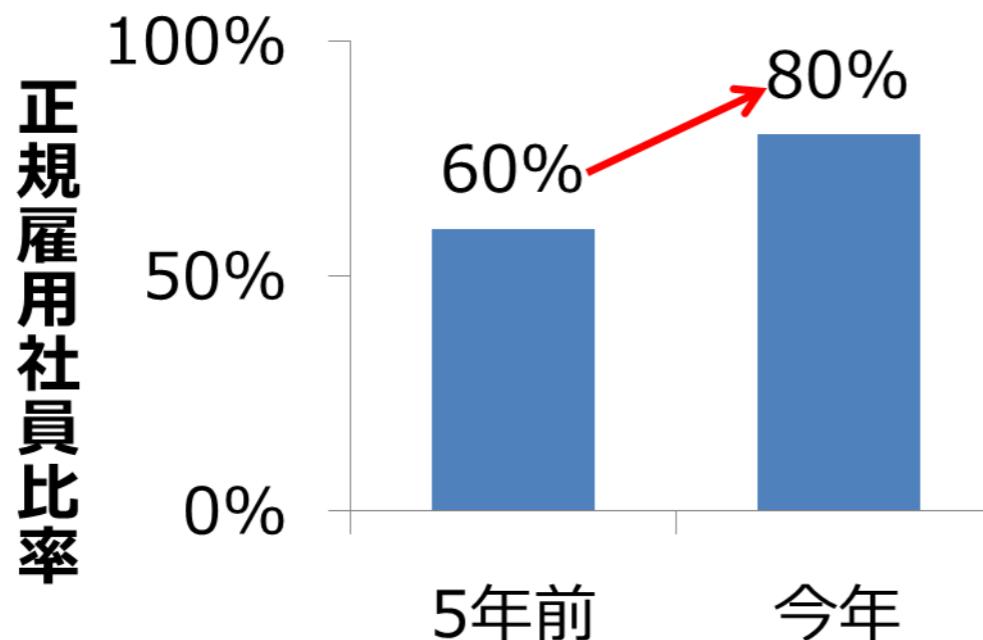
基準点や単位、期間の取り方などに注意

出典：国土交通省 建築着工統計

http://www.mlit.go.jp/sogoseisaku/jouhouka/sosei_jouhouka_tk4_000002.html

ロジック展開のウソ

- A社の正規雇用社員比率は60%から80%に增加了。
- A社は非正規雇用社員のキャリアアップに力を入れている模範的な企業だ。



ロジック展開のウソ

- 非正規雇用社員が大量に退職して、比率が上がっただけだった。

	正規 雇用社員	非正規 雇用社員	計
5年前	60人 (60%)	40人 (40%)	100人 (100%)
今年	60人 (80%)	15人 (20%)	75人 (100%)



分析結果の論拠に注意

錯覚・思い込み

- IウイルスよりもEウイルスの方が致死率が高い。
- 小学校でもEウイルス対策を優先して行う必要がある。

	Iウイルス	Eウイルス
致死率	0.1%	80%

錯覚・思い込み

- Eウイルスの罹患率に比べて、Iウイルスの罹患率は非常に高い。
- 日本の健康な児童が罹患して死に至る確率は、Iウイルスの方が高い。

	Iウイルス	Eウイルス
致死率	0.1%	80%
	×	×
罹患率	50%	0.001%
罹患して死に至る確率	0.05%	0.0008%

前提を飛ばした思い込みに注意

様々な結果解釈上のミスと注意点

不適切なサンプル	母集団を代表していない標本。偏りのある標本、少ない標本など。
グラフの作為	基準点、単位、期間の異なる比較。視覚的な錯覚を生じる表現など。
定義の違い	定義の違いを無視して比較する場合など。
ヒューリスティックス	感覚や経験的知識によるバイアス。記憶や想像のしやすさによる利用可能性ヒューリスティックス、典型事例を全体像として錯覚する代表ヒューリスティックスなど。
確証バイアス	自分を正当化する情報にしがみつくバイアス。

次回のテーマ

次回は

「予測・分類等代表的手法と活用場面」

お疲れ様でした！

社会人のためのデータサイエンス演習

第4週:ビジネスにおける予測と分析結果の報告

第5回:予測・分類等代表的手法と活用場面

講師名:矢島 安敏

第4週の内容紹介

第1回

- 回帰分析による予測

第2回

- モデル評価と予実評価

第3回

- 分析結果の報告（記述/可視化方法）

第4回

- 分析結果の報告（解釈の注意点）

第5回

- **予測・分類等代表的手法と活用場面**

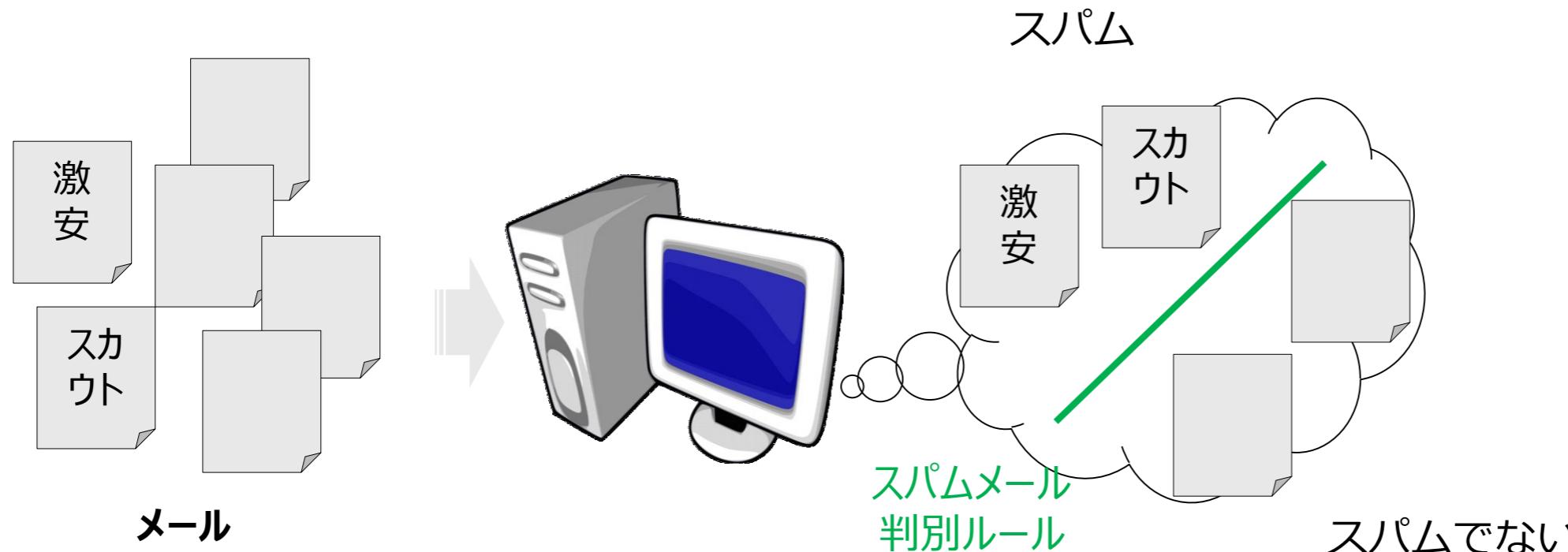
第6回

- ビジネスシーンにおける「統計的検定」とその活用例

機械学習の身近な応用例

● スパムフィルタ

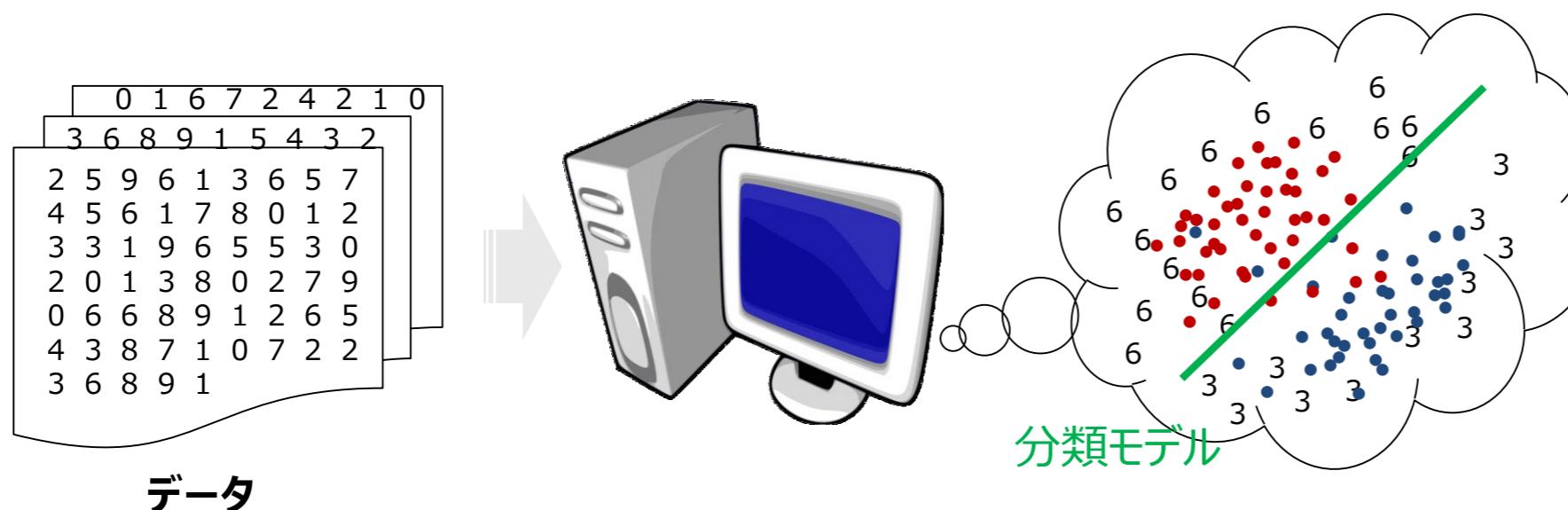
- メールに書かれた文言から、機械的に「スパムメールか」「スパムメールでないか」を判別している



このルールを機械的に学習する

機械学習の概要①

- 機械学習は、過去のデータから法則性を「コンピュータで自動的に」導き出す
 - 統計学がデータを“説明”することにより重きを置くのに対し、機械学習は、データから“予測”することにより重きを置く



機械学習の概要②

● なぜ今、機械学習が脚光を浴びているのか？

- コンピュータの発達により、スピーディに、かつ、自動的に、ビッグデータからパターンやルールを見つけ出し、予測を行うことが可能になった
- 実際に、機械学習の利用分野は、金融、ウェブから、様々な分野に急速に拡がっている



機械学習の応用例①

● 商品 recommendation

- ある顧客の購買履歴と製品在庫目録から、それらの在庫製品のうちその顧客が興味を持って購入しそうなものを識別し、顧客に商品を推奨し購入を促している

⇒有用な販売戦略に寄与

顧客1と顧客2の
購買傾向が近い!
顧客2に商品Cも
レコメンドしよう



	商品A (購買個数)	商品B (購買個数)	商品C (購買個数)	商品D (購買個数)
顧客1	10	5	7	5
顧客2	11	5		5
顧客3	1		20	1

機械学習の応用例②

● 画像タグ付

- ユーザーがSNSに投稿した写真に対し、画像の判別を行うことで「誰がどこに映っているのか」自動でタグ付を行う

⇒ユーザー利便性向上に寄与

過去に投稿された
画像からすると、
左がHanna、
右がAnny



次回のテーマ

次回は

**「ビジネスシーンにおける
「統計的検定」とその活用例」**

お疲れ様でした！



総務省統計局

社会人ためのデータサイエンス演習

第4週:ビジネスにおける予測と分析結果の報告

第6回:ビジネスシーンにおける「統計的検定」とその活用例

講師名:菅 由紀子

第4週の内容紹介

第1回

- 回帰分析による予測

第2回

- モデル評価と予実評価

第3回

- 分析結果の報告（記述/可視化方法）

第4回

- 分析結果の報告（解釈の注意点）

第5回

- 予測・分類等代表的手法と活用場面

第6回

- ビジネスシーンにおける「統計的検定」とその活用例

仮説検定とは

- 仮説に対して、正しいか否かを統計学的に検証する手法
- 統計的な判断として、差に意味があるかを明らかにする

有意差とは

立てた仮説と結果の差について、統計的に意味があるものを「有意差」という。統計調査などによって得られた2つの値の差が、統計的に信頼できるものか、偶然のものかどうかを判定する方法を有意差検定という。

ミュージカル鑑賞の1年間の平均支出金額

男性・女性それぞれ100人に対してのアンケート



男性 1500円



女性 3000円



男性 1500円



女性 1550円

男女の結果に差があるといえる

男女の結果差はある・・・？

帰無仮説と対立仮説

- 仮説検定は「差がないこと」を検証する
- 帰無仮説は「差がない」、対立仮説は「差がある」

帰無仮説

男性より女性の方が
ミュージカル鑑賞に支出する



対立仮説

帰無仮説が棄却された際に
採択される仮説。

最初に導いた仮説を否定するも
のを設定。帰無仮説が否定され
れば、対立仮説が採用される。

ミュージカル鑑賞の1年間の支出
金額平均に男女の差はない

ミュージカル鑑賞の1年間の支出金額
平均に男女の差はある

統計量を計算し、2つの仮説のどちらを採用するか
検証を行う

代表的な仮説検定

t検定

χ^2 (カイ) 2乗検定

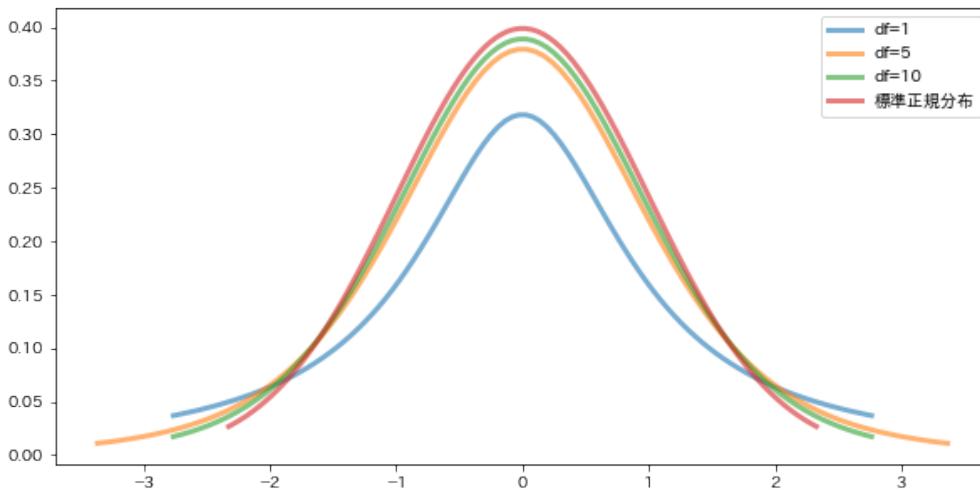
t検定とは

- サンプルの標本平均や標本標準偏差から2群の母平均が等しいと言えるかをp値によって調べる方法。

回帰分析の結果の解釈にもt検定が用いられる

t分布とは

t分布は、自由度fというパラメータをもち、自由度が大きいと標準正規分布に近づく。サンプルサイズが小さく平均が不明（未知）のときの検定・推定の計算に用いられる。



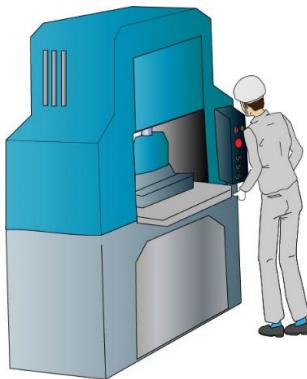
P値（有意確率）とは

帰無仮説において、その結果以上または以下ができる確率を指す。母集団から無作為抽出したサンプルの標本平均や標本標準偏差から、それらの母平均が等しいと言えるかを判定する。

t検定の活用例と特徴

t検定は母集団の分散が未知の場合に用いられ、多くの現場で活用されている検定手法。Excelの分析ツールなどで活用できるのがメリット。

製造工場における品質管理



製造工程における製品の完成度合いのばらつきを調査し、検品を行う。

医薬品の治験における効果測定



治験によって得られたデータをもとに、医薬品の効果の有無を判定する。

アンケート調査の結果判定



アンケートの集計結果におけるデータの差について、有意差があるかどうかを判定する。

カイ2乗検定 (χ^2 検定) とは

- 観測されたデータが予測される確率どおりかどうかを調べる
- 2つの出来事について、関連性があるかどうかを調べる

統計モデルを構築した際、データとモデルの適合度の検定などに使用

適合度の関係

観測値が理論値に当てはまるか、観測値と理論値のズレを算出して検定する。

例) 3000回サイコロを振った時、各数字が出る回数は500回

帰無仮説：当てはまる

対立仮説：当てはまらない

独立性の関係

2つの分類基準について、関連性があるかどうかを判定する。

例) 好きな教科と性別には関係があるか？

帰無仮説：関係がない

対立仮説：関係がある

カイ2乗検定の活用例と特徴

カイ2乗検定は質的データを対象とした検定手法。
独立の検定ともいわれ、得られた結果同士の関連性を知ることができる。

食品メーカーにおけるニーズ調査



新商品開発のための事前アンケートにおいて、年代と嗜好品の調査を実施。

年代によって嗜好品の傾向に違いがあるかどうかを判定。

飲食店における売上傾向分析



特定のメニューの売上が、店舗によって差があるかどうかを判定し、店舗の特徴を把握したり売上傾向を知る。

次回のテーマ

次回は

「各週のおさらい」

お疲れ様でした！