確率と統計　第 2 回

データ分析への誘い

4 月 20 日

Table of contents

統計学とは

統計学とは

どのようにデータを収集し, 分析し, 結論を出すか?

の方針を与えるもの.

統計学とは

どのようにデータを収集し, 分析し, 結論を出すか?

の方針を与えるもの. ここでデータとは, 観察や実験で得た観測値.

新薬の病気抑制効果の検証

新薬の病気抑制効果の検証

研究課題: 新薬がある病気の発症を抑えられるかを検証する.

新薬の病気抑制効果の検証

研究課題: 新薬がある病気の発症を抑えられるかを検証する.
方法: 病気のリスクのある患者 500 名を集めて, ランダムに処理群と対照群に分ける.

新薬の病気抑制効果の検証

研究課題: 新薬がある病気の発症を抑えられるかを検証する.
方法: 病気のリスクのある患者 500 名を集めて, ランダムに処理群と対照群に分ける.

処理群: 新薬を与える. (250 名)

対照群: 偽薬 (プラセボ) を与える. (250 名)

新薬の病気抑制効果の検証

研究課題: 新薬がある病気の発症を抑えられるかを検証する.
方法: 病気のリスクのある患者 500 名を集めて, ランダムに処理群と対照群に分ける.

処理群: 新薬を与える. (250 名)

対照群: 偽薬 (プラセボ) を与える. (250 名)

半年間で病気を発症したかを記録する.

新薬の病気抑制効果の検証

研究課題: 新薬がある病気の発症を抑えられるかを検証する.
方法: 病気のリスクのある患者 500 名を集めて, ランダムに処理群と対照群に分ける.

処理群: 新薬を与える. (250 名)

対照群: 偽薬 (プラセボ) を与える. (250 名)

半年間で病気を発症したかを記録する.

どのように分析を行うか?

収集

	処理群/対照群	発症あり/なし
1	処理	あり
2	処理	なし
⋮	⋮	⋮
500	対照	あり

分析

	発症あり	発症なし	発症率
処理	50	200	…$50/250 = 0.20$
対照	40	210	…$40/250 = 0.16$
合計	90	210	

処理の発症率 > 対照の発症率

結論

結論

新薬に病気の発症を抑える効果はない.

結論

新薬に病気の発症を抑える効果はない.
むしろ逆に病気の発症を促進している.

結論

新薬に病気の発症を抑える効果はない.
むしろ逆に病気の発症を促進している.

発症率の差が本物か偶然かを判断するには統計学が必要.

データ行列

データ行列

まずは行を事例 (あるいは観測単位), 列を変数とするデータ行列の形にデータを集める.

データ行列

まずは行を事例 (あるいは観測単位), 列を変数とするデータ行列の形にデータを集める.

事例 (or 観測単位): ひとつひとつの事象

データ行列

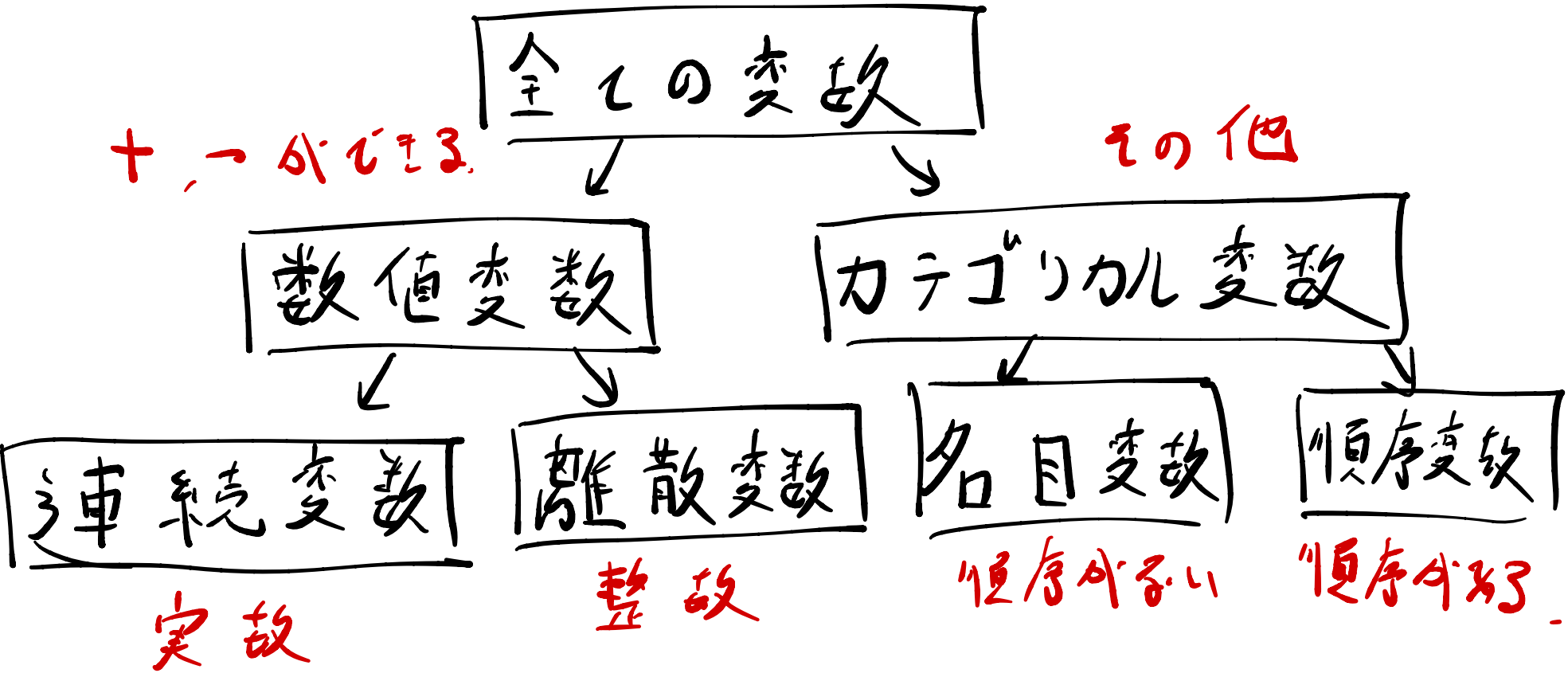
まずは行を事例 (あるいは観測単位), 列を変数とするデータ行列の形にデータを集める.

事例 (or 観測単位): ひとつひとつの事象
変数: ひとつひとつの事例が持つ特性値

消費者金融の顧客データの行列

事例	貸付金額	金利	返済期間	グレード	持ち家
			変数		
1	8000	5	36	B	あり
2	100	3	3	A	なし
3	1000	5	12	B	なし
⋮	⋮	⋮	⋮	⋮	⋮
50	200	10	24	C	なし

変数の分類


全ての変数
＋，－が定義できる
その他
数値変数
カテゴリカル変数
連続変数
離散変数
名目変数
順序変数
実数
整数
順序がない
順序がある

問題

例の 5 つの変数 (貸付金額, 金利, 返済期間, グレード, 持ち家) はそれぞれ数値変数, カテゴリカル変数のどちらか?

問題

例の 5 つの変数 (貸付金額, 金利, 返済期間, グレード, 持ち家) はそれぞれ数値変数, カテゴリカル変数のどちらか?
さらに連続変数, 離散変数, 名目変数, 順序変数のどれか?

説明変数と目的変数

説明変数と目的変数

データ分析では, データの中のある変数 X とある変数 Y の関係を調べることが多い.

説明変数と目的変数

データ分析では, データの中のある変数 X とある変数 Y の関係を調べることが多い.

ある変数 X がある変数 Y に影響していると考えるとき, X を説明変数, Y を目的変数という.

説明変数と目的変数

データ分析では, データの中のある変数 X とある変数 Y の関係を調べることが多い.

ある変数 X がある変数 Y に影響していると考えるとき, X を説明変数, Y を目的変数という.
(ただし, 本当に相関関係や因果関係が二つの変数の間にあるとは限らない)

例: 日焼け止めと皮膚がん

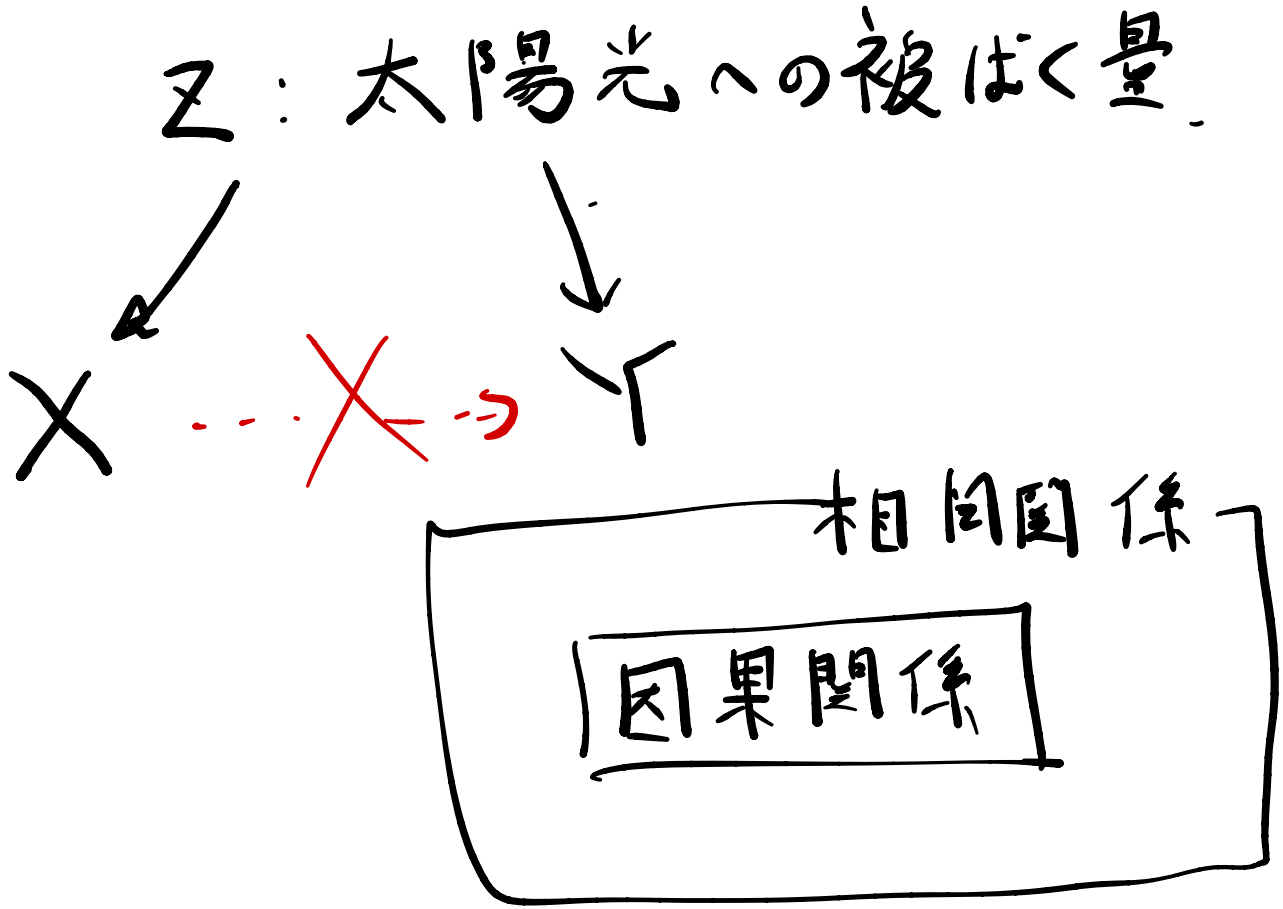
例: 日焼け止めと皮膚がん

日焼け止めの使用量 X と皮膚がんの発生率 Y の間に正の相関があったとする.

例: 日焼け止めと皮膚がん

日焼け止めの使用量 X と皮膚がんの発生率 Y の間に正の相関があったとする.
さて X と Y の間に因果関係があるといえるか?

因果関係があるとはいえない



観察研究と実験研究

観察研究と実験研究

研究の方法には大きく分けて二つの方法がある.

観察研究と実験研究

研究の方法には大きく分けて二つの方法がある.

観察研究: すでにある事例からサンプリング調査を行う

観察研究と実験研究

研究の方法には大きく分けて二つの方法がある.

観察研究: すでにある事例からサンプリング調査を行う

実験研究: 自分で事例を生成しランダム化された統計的実験を行う.

観察研究と実験研究

研究の方法には大きく分けて二つの方法がある.

観察研究: すでにある事例からサンプリング調査を行う

実験研究: 自分で事例を生成しランダム化された統計的実験を行う.

観察研究では, 説明変数と目的変数の間の相関関係までしか検証できない.

観察研究と実験研究

研究の方法には大きく分けて二つの方法がある.

観察研究: すでにある事例からサンプリング調査を行う

実験研究: 自分で事例を生成しランダム化された統計的実験を行う.

観察研究では, 説明変数と目的変数の間の相関関係までしか検証できない.
一方で実験研究では, 説明変数と目的変数の因果関係を検証できる.

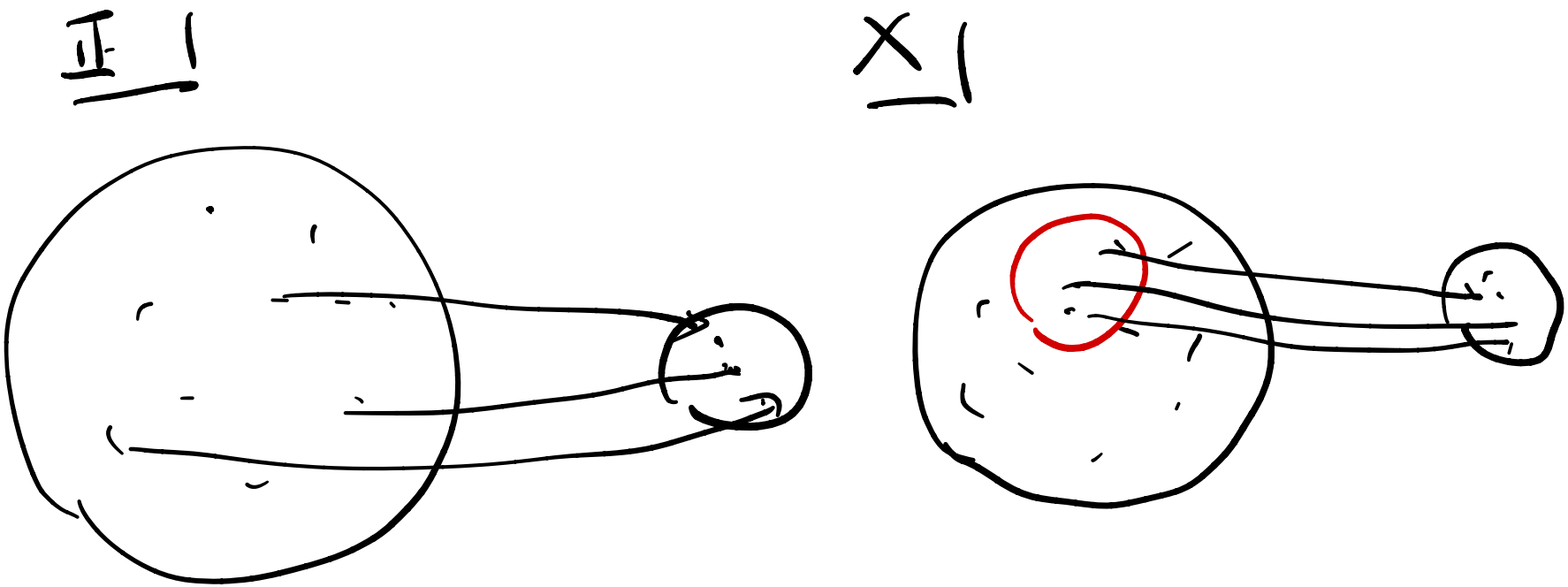
サンプリングについての注意

サンプリングについての注意

サンプリングとは, 母集団から標本(サンプル)を抽出すること.

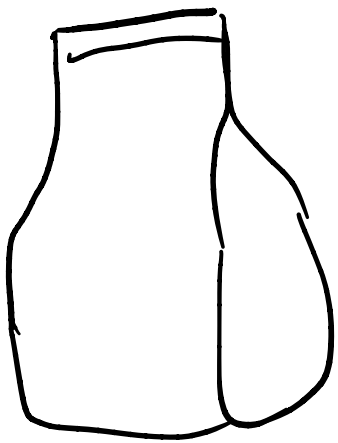
サンプリングについての注意

サンプリングとは, 母集団から標本(サンプル)を抽出すること.サンプリングは偏りなく無作為に行う必要がある.



例: 商品のレビュー

ある商品



★★★★★

レビューの95%が5つ★

しかし、よくみると、

★1のレビューに大量の"参考になった"がついている。

このレビューは購入者全体の正しい標本なのか？