1章2節　分析設計

東京大学 数理・情報教育研究センター

2021年4月17日

概要

- データ分析をどう進めていくべきかを考えるために必要となる基本的事項について学ぶ
- 実際に分析方法を定めてゆく際にどんな数理・情報関連知識が対応して必要となるかについて理解する

本教材の目次

1. データ分析の進め方

1-1. 仮説検証サイクル

データ分析の目的は大きく分けて次の2つにあります：

1. 傾向や仮説の候補の**発見**
2. 仮説の**検証**

発見しただけでは傾向や仮説が正しいかどうかが言えず、 **検証**のためには良い仮説が無いといけません。また、2 の結果によっては仮説の候補を再検討する必要があります。
そこで通常は 1→2→1→2 ... という反復（仮説検証サイクル）を満足な仮説とその検証が得られるまで 繰り返すことになります。

1. データ分析の進め方

1-2. 分析目的の設定

データ分析の各ステップにおいては、そこでの主要な目的が**発見**にあるのか**検証**にあるのか明確にしておいた方が良いでしょう。おおざっぱに言って、**発見**のためには可視化や特徴量の比較、クラスタリングなどの手法が、**検証**には相関分析や回帰、仮説検定などの手法が関係してきます。

どちらの場合も、分析対象や分析データの特徴、検証しようとしている仮説について良く考えることが必要です。

AIや深層学習への期待で誤解されがちですが、**「よく考えないままにデータを投げ入れると答えを出してくれる」**と思って良いということはありません。

3. データの収集、加工、統合

分析に進む際に、扱うデータがそれぞれどんな性質のものかについて把握しておく必要があります。データの種類には大別して以下があります：

- **量的変数**
  自然に数値で扱うことができるもの（金額、人数、長さ、視聴率、…）
- **質的変数**
  大きさを表す数値として扱えないもの。さらに次の2つに大別できます：
  - **順序尺度**：区分の間に順番はつけられるもの
    （Jリーグの年間順位、和牛の等級、癌のステージ、…）
  - **名義尺度**（カテゴリー変数）：順番もつけられない、互いに違うもの
    （居住地区、職業、支持政党、「上司にしたい芸能人」、…）
    → カテゴリー変数は、便宜上整数を割り当てて扱うことがあります。
    ダミー変数と呼ばれるこの数には当然、大きさも順序も意味がありません

3. データの収集、加工、統合

データには欠損値、外れ値、データの間での書式の不統一、例外、…などがあることが普通です。計算機で処理する際にはこれらの値を統一の書式に揃えたり統合したりする必要があります。

このような加工や統合には手間がかかることが多く、場合によってはデータ収集や入力の仕方の再考をした方が良いこともあるでしょう。

また、どのようなデータを取るべきかや欠損や例外の処理の仕方は仮説によって変わることがあります。

3. データの収集、加工、統合

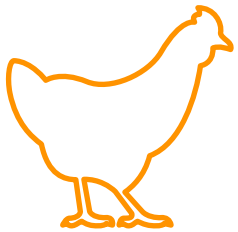
餌と鶏の成長についてのデータの例

データ行	鶏体重[g]	日齢	鶏のID	餌の種類
1	42	0	1	1
2	51	2	1	1
3	59	4	1	1
4	64	6	1	1
5	76	8	1	1
6	93	10	1	1
7	106	12	1	1
8	125	14	1	1
9	1490	16	1	1
10	171	18	1	1
11	199	20	1	1
12	205	21	1	1
13	40	0	2	1
14	NA	2	2	1
15	58	4	2	1

明らかに
おかしい値

測定できなかった等

量的変数　量的変数　名義尺度のダミー変数



(https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ChickWeight.html を加工)

4. 基本的なデータ可視化手法：ヒストグラム

量的変数データの可視化の方法として、**ヒストグラム**があります **(参照→1-3)**。
これは、変数を適当な範囲で区切ってその範囲に入る数を集計した**度数分布表**を棒グラフで表したものです。範囲の幅は、基本的には分析者が決めます。
ヒストグラムに見られるようなデータの分布の位置と広がり具合については、それぞれ**代表値**と**標準偏差**によって定量的に表すことができます（後述）。

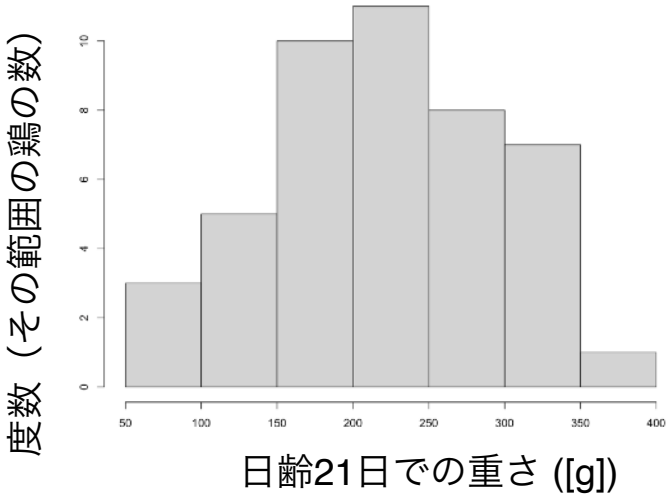
（日齢21日の鶏体重の元データ）

重さ[g]	日齢	鶏のID	餌の種類
205	21	1	1
215	21	2	1
202	21	3	1
157	21	4	1
223	21	5	1
157	21	6	1
305	21	7	1
98	21	9	1
124	21	10	1
175	21	11	1
205	21	12	1
96	21	13	1
266	21	14	1
142	21	17	1
157	21	19	1
117	21	20	1
331	21	21	2
167	21	22	2
175	21	23	2
74	21	24	2
265	21	25	2
251	21	26	2
192	21	27	2
233	21	28	2
309	21	29	2
150	21	30	2
256	21	31	3
305	21	32	3
147	21	33	3
341	21	34	3

度数分布表

重さの階級	度数
0g～50g	0
50g～100g	3
100g～150g	5
150g～200g	10
200g～250g	11
250g～300g	8
300g～350g	7
350g～400g	1
400g 以上	0

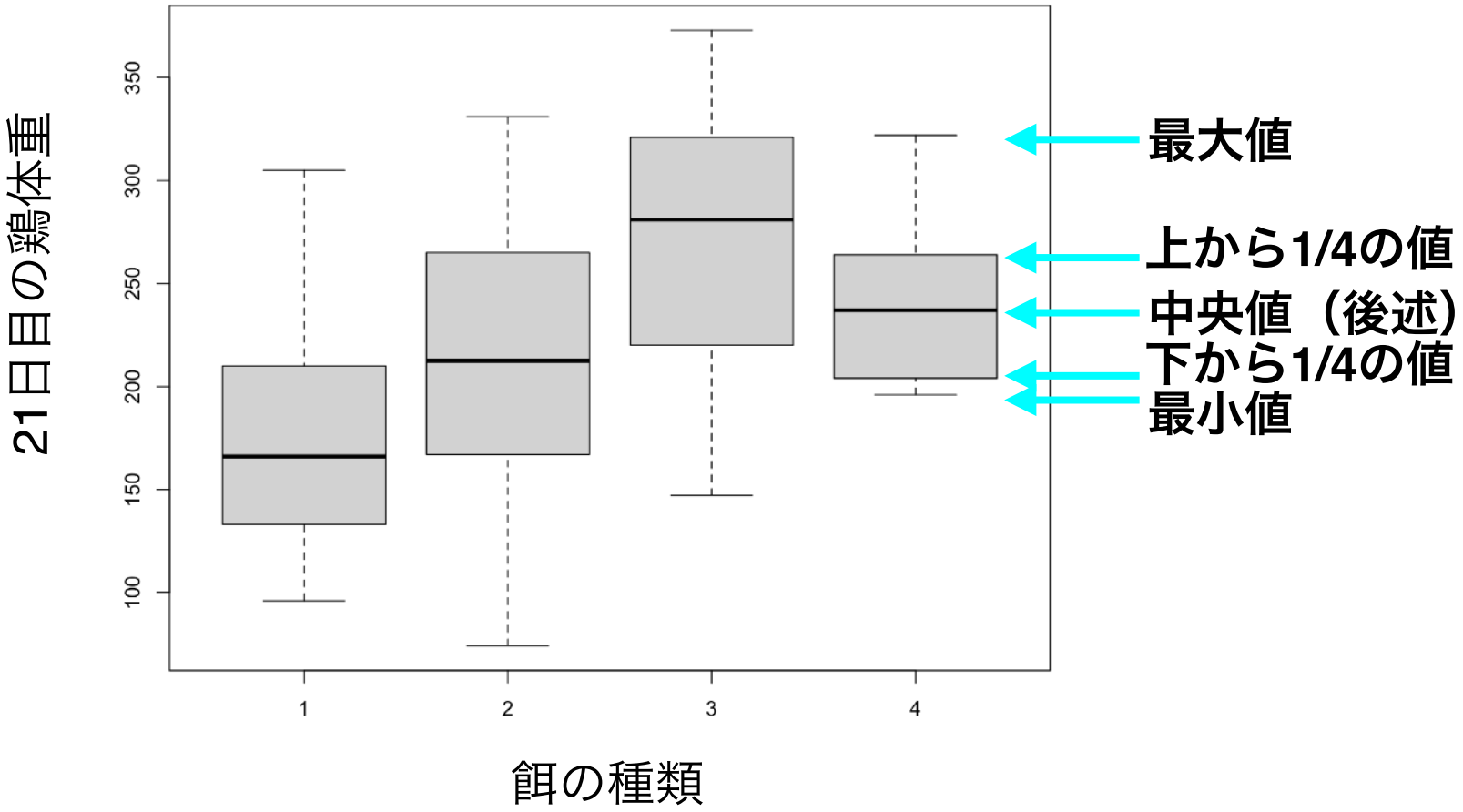
ヒストグラム



21日目の鶏の重さが大体どれくらいで、
どのように散らばっているかが把握できる

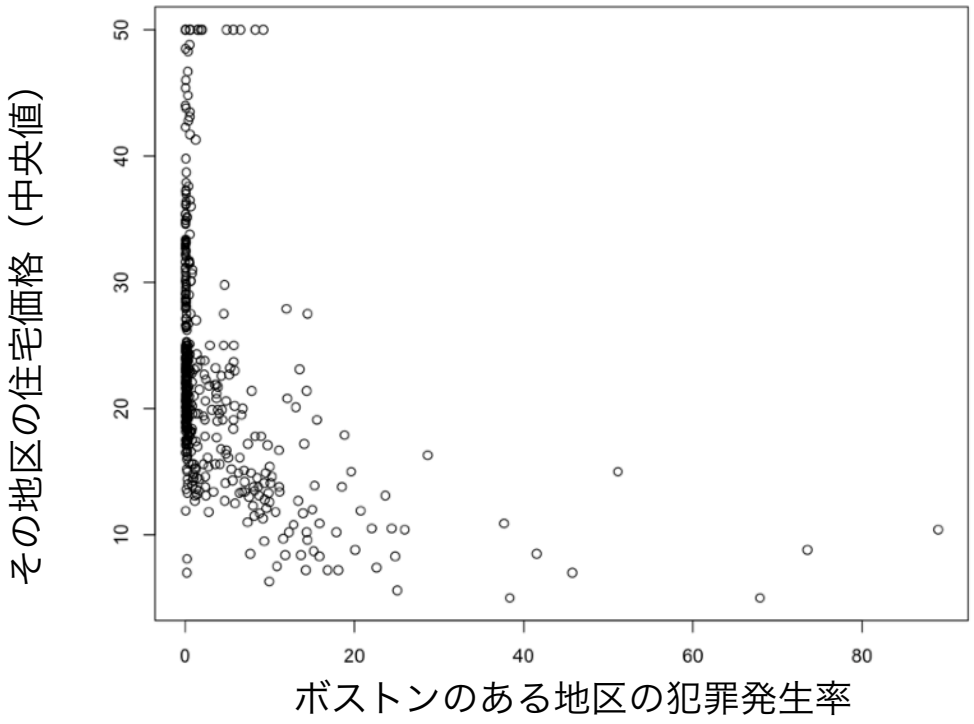
4. 基本的なデータ可視化手法：箱ひげ図

分布同士を比較するのには、**箱ひげ図**が便利です。 **(参照→1-5)**
これら各条件間に違いがあるといって良い（有意である）かどうかについて検証するには、**区間推定**や**仮説検定**の考え方を理解する必要があります **(参照→1-6)**。



4. 基本的なデータ可視化手法：散布図

量的変数項目同士の関係の可視化の方法として、**散布図**があります。これは関係を見たい二つの項目の値をそれぞれの軸にとってプロットしたものです。**(参照→1-3, 1-5)** 例えば下の図の例では犯罪発生率が高い地区ほど住宅価格が低い傾向が見て取れますが、このような**相関**は後述の**相関係数**などで定量的に評価できます。



（元データ）Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics and Management* **5**, 81–102.

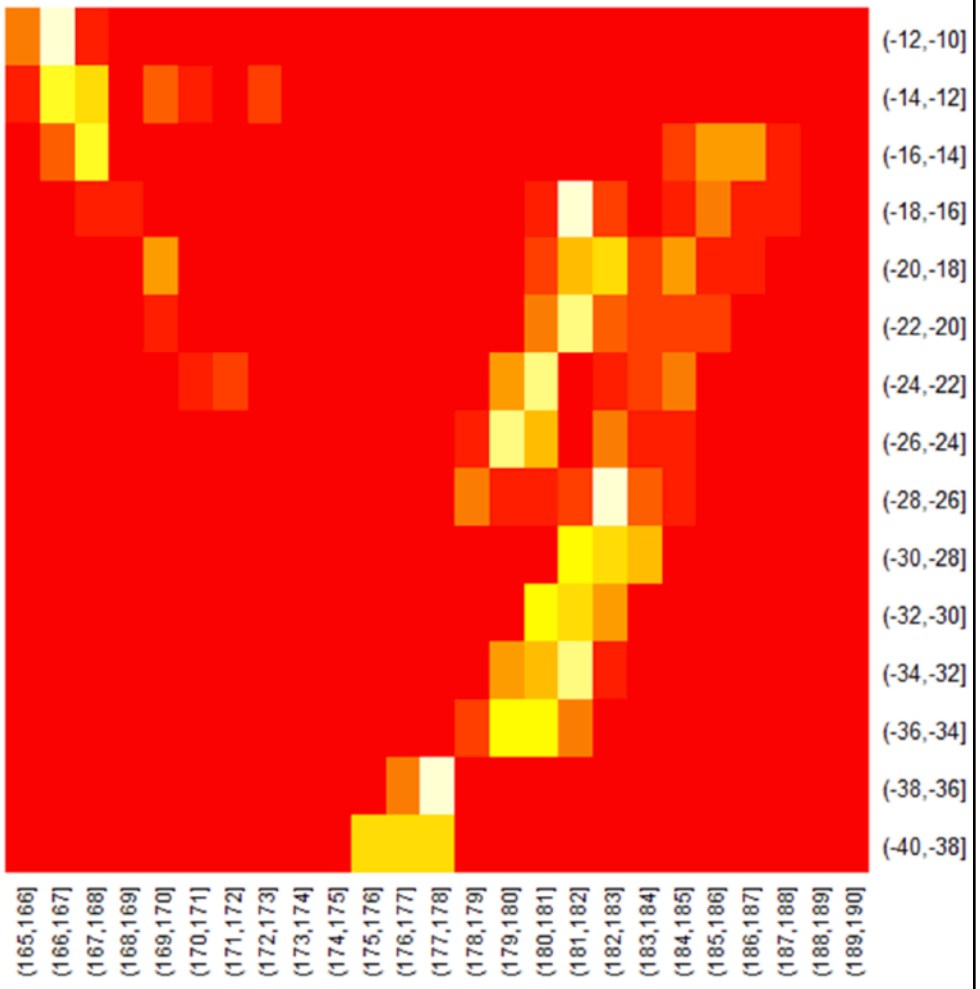
Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

4. 基本的なデータ可視化手法：

ヒートマップ **(参照→1-5)**

- 散布図においてデータの密度を色を変えて表現したものをヒートマップといいます。
- 右図はフィジーの地震データを緯度と経度をもとに、地震の発生頻度を色を分けて表現しています。

  ※白に近い色がより地震の頻度が高いです。



4. 基本的なデータ可視化手法：分割表

質的変数項目同士の関係性をつかむためには分割表とクロス集計表が便利です。

（参照→1-3）

- データの２種類の項目について、項目の値のペア毎にデータの個数を数え、表にまとめたものが「分割表」、さらに２種類の項目の値のペア毎（例えば右下表で「文系」と「2組」のペア等）に別の項目（右下表では「点数」）の合計、平均、標準偏差等を集計したものが「クロス集計表」と呼ばれます。

＜データ＞

（クラス・文理別の数学の点数）

文理	クラス	点数
理系	2組	93
理系	１組	48
文系	３組	41
文系	３組	28
理系	３組	75
文系	３組	68
・	・	・
・	・	・
・	・	・

＜分割表＞

	１組	２組	３組	全体
文系	44	39	34	117
理系	38	46	36	120
全体	82	85	70	237

各項目ペアに対してデータの個数を数える

＜クロス集計表＞（点数の平均値を集計）

	１組	２組	３組	全体
文系	38.3	39.8	43.0	40.2
理系	52.4	50.9	57.8	53.5
全体	44.8	45.8	50.6	46.9

各項目ペアでの平均等をそれぞれ計算

5. 基本的なデータ分析手法：代表値

データの分布の位置を表す基本的な量として、可視化の説明で既に用いた**平均値**と**中央値**に加えて**最頻値**があり、これらは**代表値**と呼ばれます。

- **平均値**： $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$

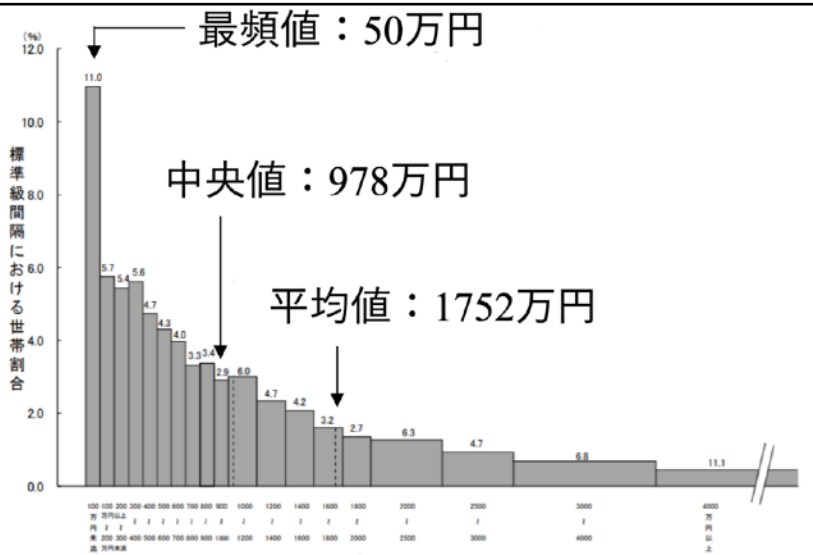
- **中央値**： データを大きい順もしくは小さい順に並べた時に真ん中に位置する値（データ数が偶数なら真ん中2つの値の平均）

- **最頻値**：データの中で最も多く現れた値。変数が連続値の場合や、整数でもデータ数に比べて範囲が広い場合などは度数が最大の範囲の中央の値。

5. 基本的なデータ分析手法：代表値

分布が平均値まわりで大体左右対称な山をつくっているような形状の場合は三つの代表値は同じような値になりますが、下の例のように現実の分布などではそれぞれ**異なる場合があることに注意が必要**です。はじめて解析するデータについては、代表値の違いが大きいかどうかや分布の形状そのものを確認してみましょう。
前述の箱ひげ図は、代表値に加えてこのような分布形状の情報を残すための可視化手法と言えます。

- 右図は2018年の全国の二人以上の世帯の貯蓄額のヒストグラムです。
- 一部の裕福な世帯の影響を受け、平均値は中央値よりもかなり高い値になっています。（３分の２の世帯が平均を下回る）



「貯蓄現在高階級別世帯分布（二人以上の世帯）」（総務省統計局）を加工して作成
(https://www.stat.go.jp/data/sav/sokuhou/nen/pdf/2018_gai2.pdf)

5. 基本的なデータ分析手法：分散と標準偏差

データの分布の広がりを表す量としては、**分散**と**標準偏差**が基本です。

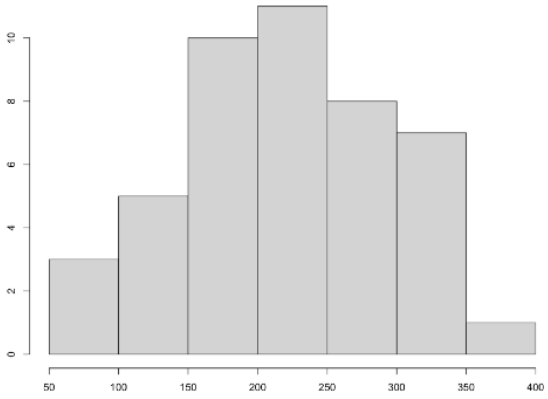
- **分散**：$\sigma^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \dfrac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$
- **標準偏差**：$\sigma = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$

標準偏差は元のデータの次元の量で、ヒストグラムがひと山の形の場合は右下の鶏体重の例のように、代表値まわりの分布の幅に対応します。
分布が多数の山を持つ場合や、貯蓄金額の例のように
裾が広い形の場合はこの限りではありません。
また、分布の中でのデータの相対的位置の指標として

- 偏差値：$50 + 10\left(\dfrac{X_i - \bar{X}}{\sigma}\right)$ が使われることがあります。

(鶏の重さの例)

$\sigma = 71.5$



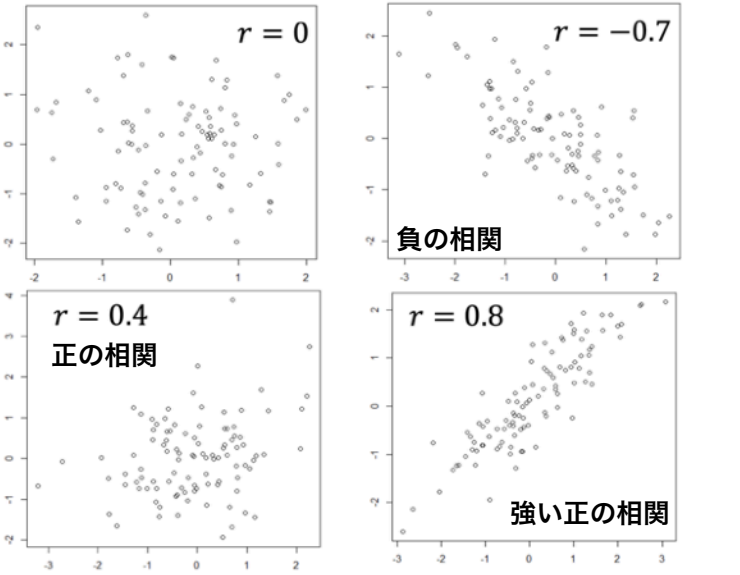
5. 基本的なデータ分析手法：相関と相関係数

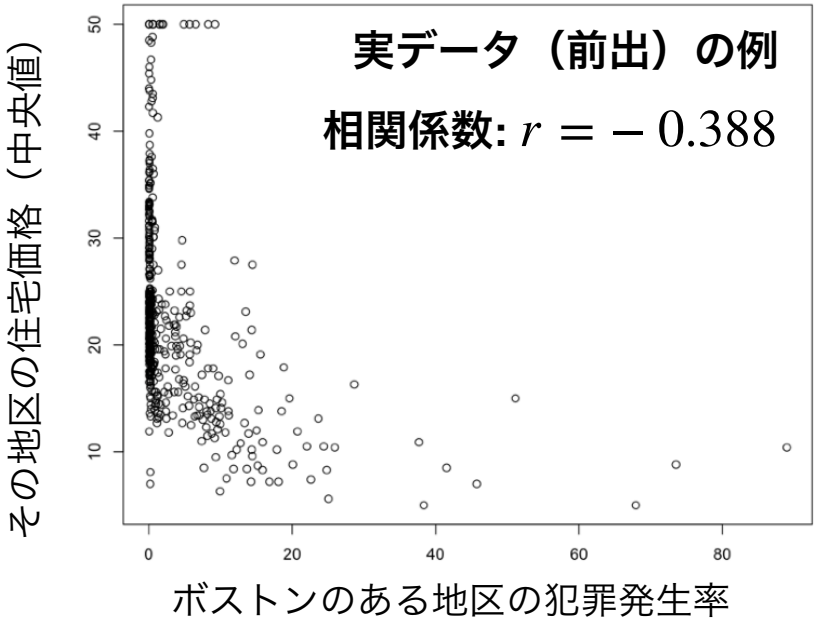
散布図で可視化したような二つの項目の相関関係を表す量として、

相関係数（ピアソンの相関係数）：$C_{XY} = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$

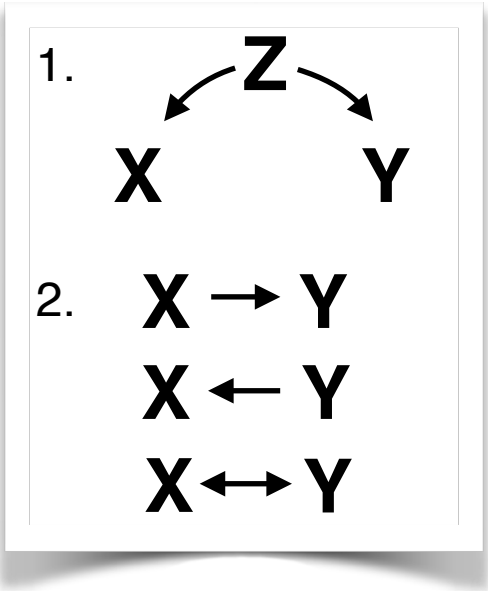
が使われます。この相関係数は-1から1までの値をとり、正の相関（散布図で右肩上がりの関係）のとき正の値、負の相関（散布図で右肩下がり）のとき負の値をとります。**（参照→1-6）**

相関係数と散布図での見え方の対応例





5. 基本的なデータ分析手法：相関と因果

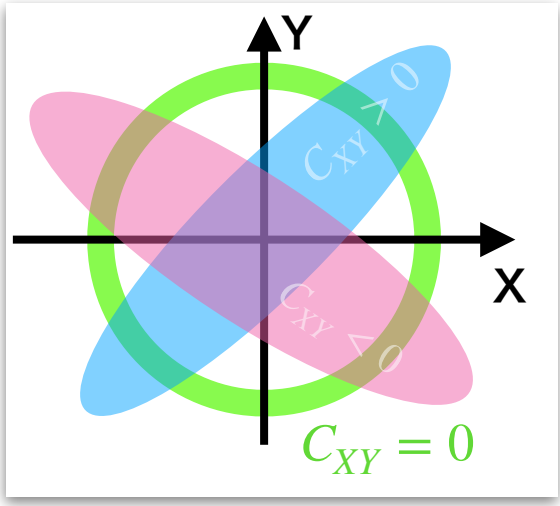


2つの量に**<u>相関があるからといってそれらの間に因果関係があるとはいえない</u>**ことに注意が必要です。

一方が直接的に影響を及ぼしているとは限らず、

(1.) 共通の要因 Z（**<u>交絡要因</u>**）があるのかもしれません。そのような場合の相関を擬似相関といいます。

(2.)また、相関係数では影響の方向を測ることはできません。影響の方向を決定するには他の基準やよく設計された実験が必要です。

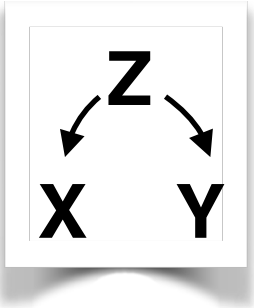


さらに、相関係数が0に近いからといって、関係性が無いとも言えないことに注意しましょう。左図の例のように、XYの間に明らかな関係性があっても、線形に相関していない場合は相関係数は0になります。 **(参照→1-6)**

（*発展的）左図のような関係性は、**相互情報量**によって検出することができます

5. 基本的なデータ分析手法：偏相関係数

交絡要因 Z について心当たりがある場合、その交絡による擬似相関を取り除いた上での2変数データ X と Y の相関関係を**偏相関係数**によって評価することができます：



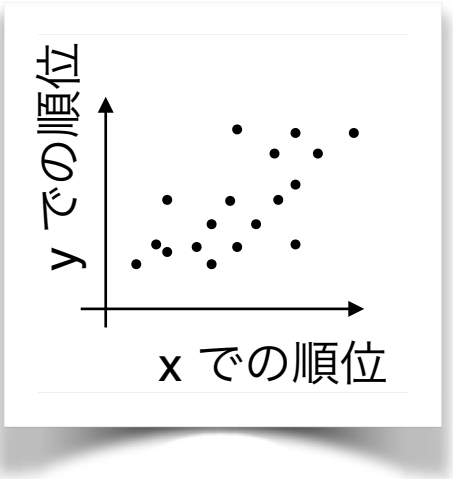
$$C_{XY|Z} = \frac{C_{XY} - C_{XZ} \cdot C_{YZ}}{\sqrt{1 - C_{XZ}^2}\sqrt{1 - C_{YZ}^2}}$$

0でない偏相関係数が得られたとしても、想定していない**交絡要因がまだ他にあるかもしれない**という点には注意が必要です。
また、偏相関係数も、因果性や XY 間の影響の方向についての情報とはなりません。

5. 基本的なデータ分析手法：順位相関係数

相関係数 C_{XY} は X と Y が順序尺度の場合にも同様に活用することができます。

また、**元の変数が量的変数である場合にも**、その量同士では無く各データにおける順位の相関を見ることがしばしば有効です（**スピアマンの順位相関係数**）。



$$C_{xy}^{S} = \frac{\sum_{i=1}^{n}\left(\mathrm{rank}(x_i) - \overline{\mathrm{rank}(x)}\right)\left(\mathrm{rank}(y_i) - \overline{\mathrm{rank}(y)}\right)}{\sqrt{\left[\sum_{i=1}^{n}\left(\mathrm{rank}(x_i) - \overline{\mathrm{rank}(x)}\right)^2\right]\left[\sum_{j=1}^{n}\left(\mathrm{rank}(y_i) - \overline{\mathrm{rank}(y)}\right)^2\right]}}$$

$$\left(\overline{\mathrm{rank}(x)} \text{ は順位の平均を表し、同順位が無ければ } = \frac{n(n-1)}{2}\right)$$

$$= 1 - \frac{6}{n(n^2-1)}\left(\sum_{i=1}^{n} d_i^2\right) \quad \left(d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i) \ \left(\text{順位の差}\right)\right)$$

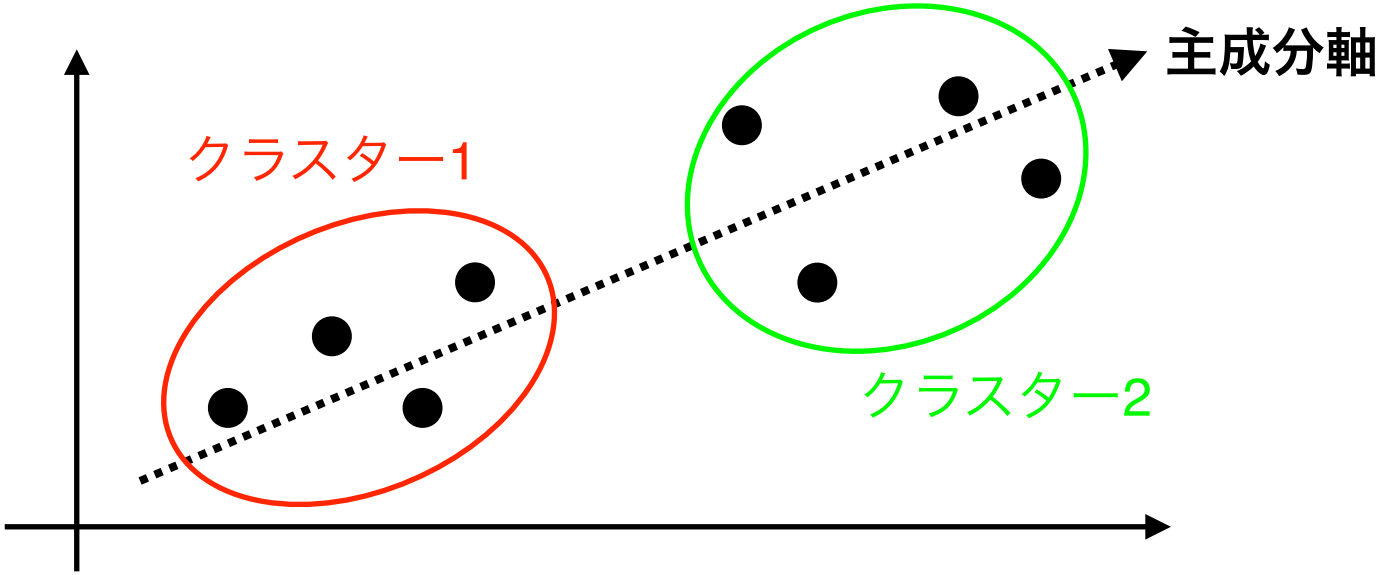
（例）前出の世帯貯蓄額と、世帯の他の項目との相関をとると貯蓄額が非常に大きい世帯の特徴のみによって相関係数が大きく左右されてしまいます。貯蓄額の順位との相関をとった場合はこのような不都合を避けられます。

5. 基本的なデータ分析手法：

主成分分析とクラスタリング **(参照→1-4)**

データの各項目間に関係性がある場合には、より少ない情報でデータが良く要約できるかもしれません。このような隠れた構造を抜き出すための代表的手法として**主成分分析**と**クラスタリング**（クラスター分析）があります。

主成分分析ではデータのばらつきをもっとも良く表現するための少数の軸方向（下図点線）を求め、（各種の）クラスタリングではデータの自然なグルーピング（下図色分け）を求めることができます。



5. 基本的なデータ分析手法：回帰分析 **(参照→1-4)**

X が Y に影響を及ぼしていると**仮定して**、
Y（目的変数）の変動を X（説明変数）の変動で表す良いモデル関係式を求めることを**回帰分析**と言います。最も基本となるのは線形回帰：

$$y \sim \alpha + \beta x$$

です。線形回帰については、この関係式と実際のデータとのあいだの誤差

$$\sum_{i=1}^{n} \left[y_i - (\alpha + \beta x_i) \right]^2$$

を最小にする係数 (α_*, β_*) を簡単に求めることができます。

説明変数が複数の場合：$y \sim \alpha + \sum_{i=1}^{n} \beta_i x_i$ にも同様に、最適な解は簡単に（見通し良く、少ない計算量で）求めることができます。

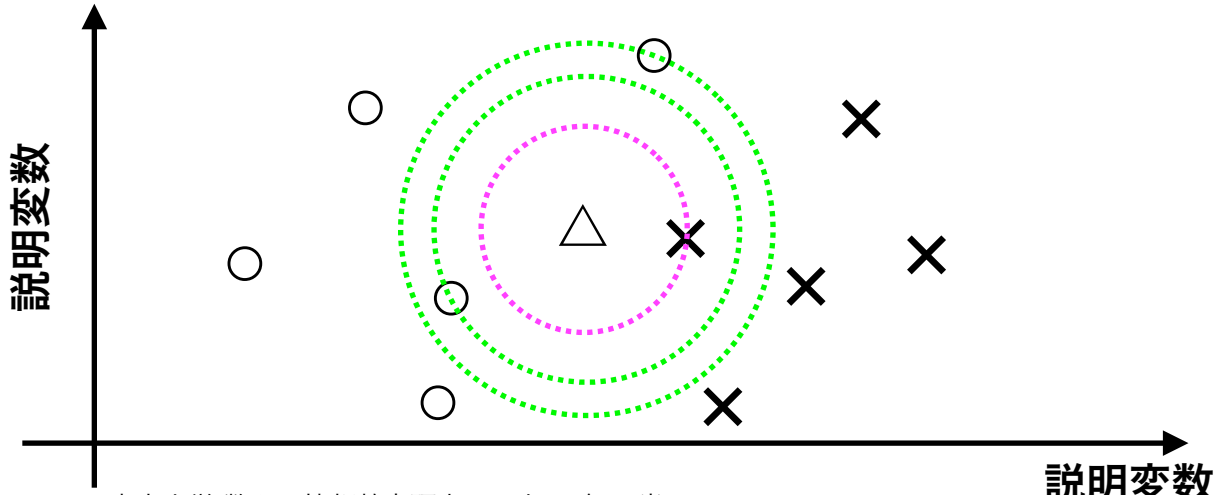
5. 基本的なデータ分析手法：判別

X により説明したい目的変数 Y が名義尺度である場合（「健康診断の数値から再検査が必要かどうか判定する」、「手書き文字を読み取る」、など）を**判別問題**といいます。

代表的な判別手法として、前述の回帰と同じ原理に基づく**線形判別分析**（回帰による目的変数の値が一定値以上かどうかで判別する）や、

K近傍判別分析（説明変数の空間での近傍の点の正解を参考にしてその点を判別する）などがあります。

（K 近傍法）データ△について、k=1 なら×, k=3 なら○、と判別



5. 基本的なデータ分析手法：仮説検定 **(参照→1-6)**

データ分析で発見したずれが統計的に意味があるかどうかを考える標準的枠組みとして**仮説検定**があります。仮説検定では、

1. 検証したいずれや傾向が「無い」とした仮説を帰無仮説として立てます
2. この帰無仮説が正しいと仮定した際に、ある確率以下でしか起こらない事象を 1 つ選びます（この「ある確率」を**有意水準**と呼びます）
3. 実際のデータにおいてその事象が起きているかを検証します

   → 起きている場合は帰無仮説は誤っている（＝比較した量は有意にずれている）と結論します。これは、「仮説を棄却する」「ずれは有意である」などと言います

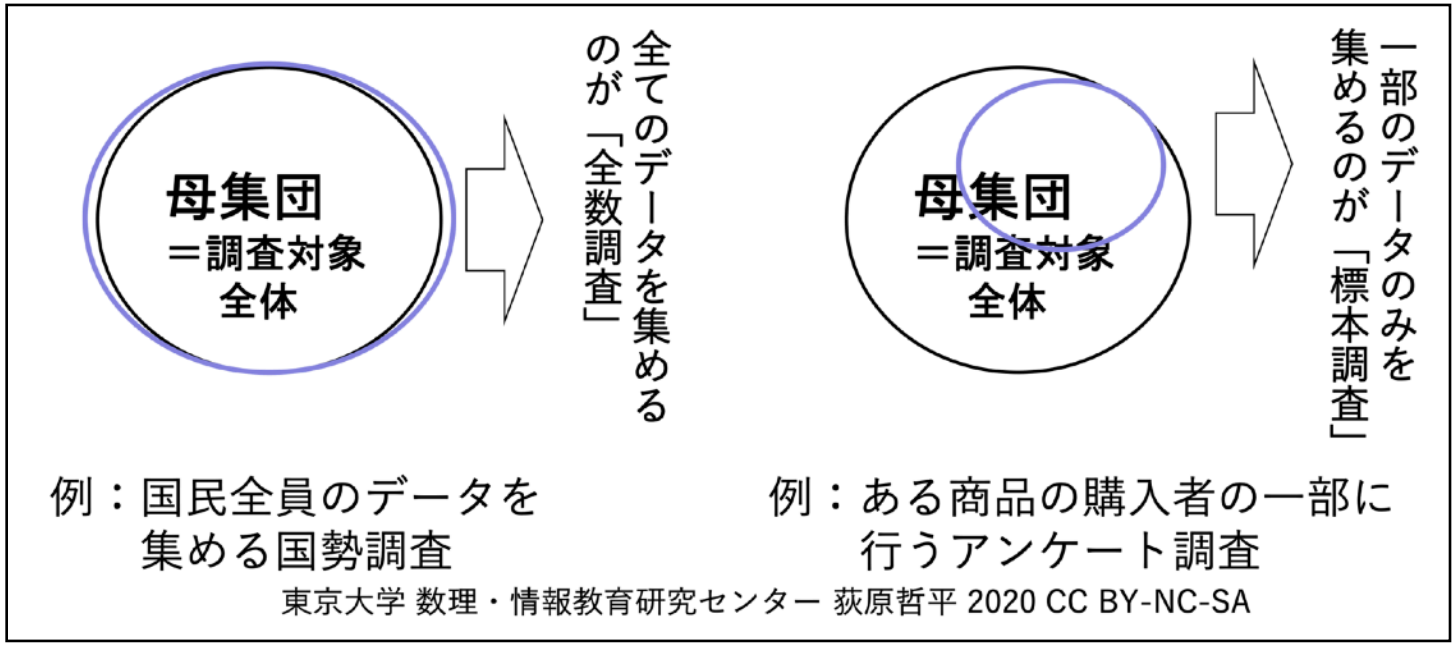
有意水準の取り方は分析の目的によって異なりますが、多くの場合は 5% または 1% にとるようです。

また、**仮説が棄却されるぎりぎりの有意水準は p 値と呼ばれ、**有意度の大きさの指標として使われます。

6. 分析目的に応じた適切な調査の設計 (参照→1-6)

データ解析を行う際には、調査したい対象のデータが全て手に入ることは稀です。対象の一部のデータのみに対して解析を行う場合、調査対象全体を「**母集団**」と呼び、入手可能な一部のデータを**標本**といいます。調査対象全体からすべてのデータを集める時、「**全数調査**」、一部の データのみを集める時、「**標本調査**」といいます。

（注）既知のデータ全てを使って未知のデータを予測しようとする場合も、起こりうる事象という母集団のうちの一部のデータについて調査するので標本調査の枠組みを用いることになります



6. 分析目的に応じた適切な調査の設計

標本調査から母集団の性質を推測する前提として、**偏りが生じないような標本の抽出が非常に重要です**。代表的な抽出方法として以下のような方法が挙げられます。

- **無作為抽出**：母集団の中からランダムに標本を抽出します。
- **層別抽出**：母集団を属性ごと(例えば性別・年代・業種等)にいくつかに分け、各層から必要数の標本を抽出することで属性の偏りを避けます。
- **多段抽出** ：母集団をいくつかのグループに分け、まずグループをランダムに選びます。選んだグループをさらに小グループに分けてその中からランダムに選ぶことを繰り返し、データを抽出します。

（例）まず全国から都道府県をランダムに選んで、その中から地域をランダムに選んで、…選ばれた小さな地域からランダムに人を選んでデータを抽出する。全国民からのランダムサンプルにするためには、各地域を選ぶ確率をその地域の人口に応じて定める必要があります。

6. 分析目的に応じた適切な調査の設計

偏りが生じないような標本の抽出ができているかということについては、最大限の配慮と色々な工夫が不可欠です。

（例）選挙の出口調査では、複数で出てきた場合、調査をお願いする対象を「向かって右端の人」などと定めていることが多い。「グループに呼びかけた場合に率先して答えてくれる人」が母集団のランダム抽出にならない恐れがあるため。

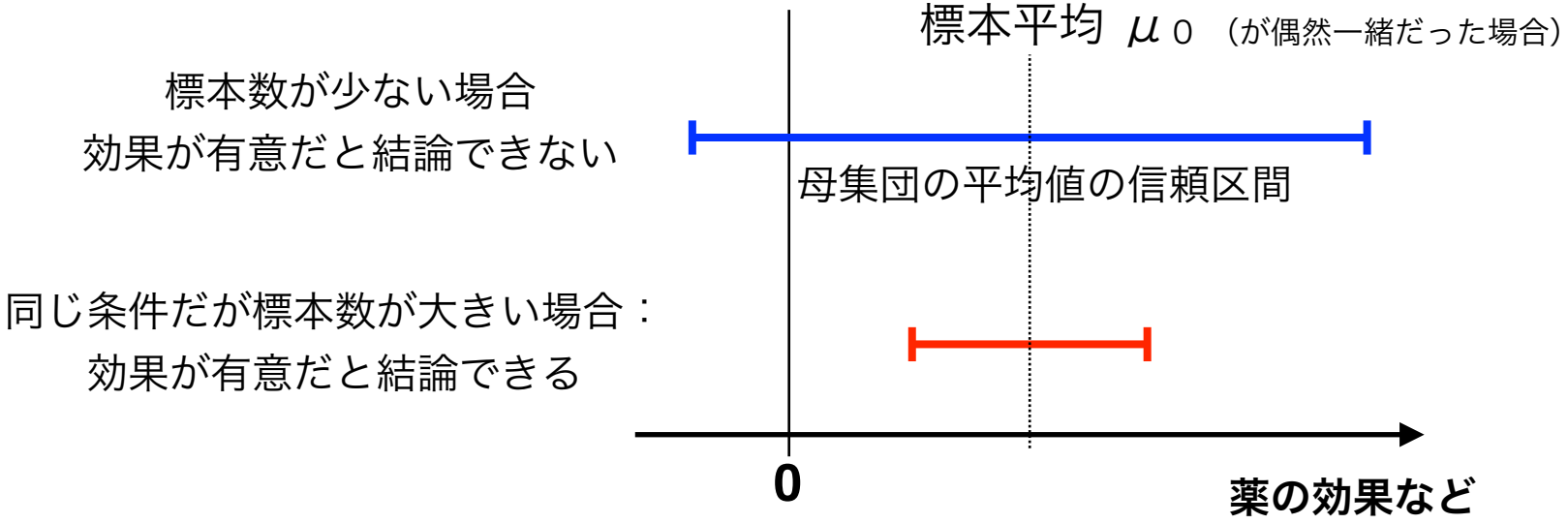
また「相関と因果」の部分で見たように、**隠れた因子の影響を極力排除するためにも、比較するサンプル間でできる限り条件を揃える工夫が必要**です。

（例1）投薬の効果を見る際には、試験薬を与えるグループは、偽薬（プラシーボ）を与えたグループと比較しなければいけません。これは、試験の対象はランダムに選んでいたとしても、「投薬された」という患者の認識が一般に結果に大きく影響するためです。

（例2）さらには、観察結果の報告自体にも観察者の事前知識が影響しうるため、どの患者に偽薬を与えたのか医師にも分からなくする（**二重盲検法**）ことがより理想的です。このような**観測・分析側の主観を除くための工夫**もされた分析は**ランダム化比較試験**と呼ばれることがあります。

6. 分析目的に応じた適切な調査の設計 **(参照→1-6)**

調査規模（標本数）の決定：
偏りの無い標本抽出ができているという仮定のもとで、母集団の性質についての推定ができます。この場合、母集団の代表値などの推定の幅は大雑把にいって標本数を n として $\frac{1}{\sqrt{n}}$ に比例します（この理解には**大数の法則**と**区間推定**を学んでください）。推定の幅を 1/10 にしたければ 100 倍のサンプル数が要るというこの関係と現実的に可能な調査規模を勘案して、標本数や検証する仮説の設計をする必要があります。



6. 分析目的に応じた適切な調査の設計

調査規模（標本数）の決定：

母集団の平均値などの推定量から“説明因子はなんの効果が無い”としたモデルを否定できない（帰無仮説が棄却されない）という結果を得た場合にも、

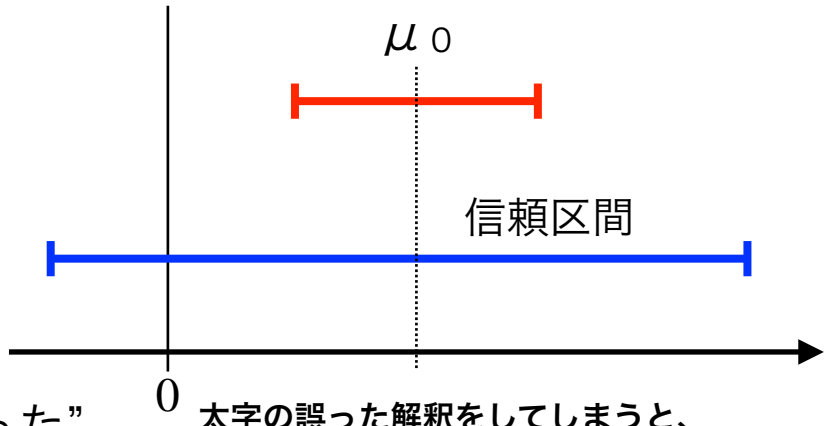
「“その説明因子が効果が無い”という結論が得られた」わけでは無いということに注意が必要です。

また逆に、**標本数を無限に大きくしていけばどんな集団のどの項目間にも有意な差異を見つけてしまえる**ということにも留意が必要です。例えば、対象人数を非常に大きくとればどのような項目についても男女差を見出すことは可能です。見つかった差に意味があるかどうかについては適宜検討しましょう。

分析１：**“効果あり”の報告**
（標本数が多い）

分析２：**“効果なし”の報告**
（標本数が少ない）

→**正しくは:** “効果があるとは示せなかった”



太字の誤った解釈をしてしまうと、
見ている現象は同じなのに結論が異なってしまう

6. 分析目的に応じた適切な調査の設計

実験計画法：

目的の量（鶏の体重など）に影響すると思われる因子（餌の種類など）が複数あって、それら複数の因子の量の最適な組み合わせを知りたいという場合が良くあります。この場合、理想的には全ての因子の量の組み合わせについて実験/調査する必要がありますが、そのような組み合わせの数は非常に大きくなってしばしば非現実的です。

（例）：検討したい因子が10個ある場合、各因子の量について「有り/無し」の2値で調べるとしてもその組み合わせの数は 2^{10} ～ 千通り、「多い/少ない/無し」の3値で調べるとすると 3^{10} ～ 6万通りになります。さらにこれら全てのケースのそれぞれについて、比較するためには前述のように単一では無く複数の標本が必要です。

このような場合の実験・調査の設計のために、多数の因子の効果を「それぞれの因子の効果（**主効果**）の和」と、「そこからのずれ（**交互作用**）」とに分けて考えることが有効です。

主効果の和でうまくデータが説明できると仮定すると、最適な因子量の組み合わせの比較的少数の組を定めることができます（直交表）。また、この組についての観測から、交互作用の効果が無視できない領域を見出すこともできます。このような手法は**実験計画法**と呼ばれます。

