# 確率統計（１）

## 概要

1. 統計学の基礎
    - 1次元のデータ
    - 2次元のデータ
    - 相関
2. 確率の基礎
    - 確率変数と確率分布
    - 中心極限定理
    - 独立性
3. 仮説検定
    - t検定
    - カイ二乗検定
4. 推定
    - 点推定・区間推定
    - 最尤推定

# 確率統計を学ぶ意義

- 不確実性を持った現象を分析をする
  （信号処理や画像処理などでも利用される）
- 全数調査ができない時に限られた全てのデータから分析をする
  （アンケート調査やデータから傾向を発見する）
- 統計的機械学習などに繋がる
  （将来予測や自動化への応用など）

# 統計学の基礎

平均（Mean）

- データを合計したものをデータ数で割った値
- 平らに均した値で、物理的には重心に相当し全体のバランスが取れる位置を示す

中央値（Median）

- データを大きさ順に並べて、中央に位置するデータの値
- 外れ値（異常に大きい値など）がある場合、平均は外れ値の影響を受けるが、中央値は影響を受けにくい

最頻値（Mode）

- 最も多く出現する値。
- 外れ値の影響を受けにくいが、データのサンプル数が少ない場合には一意に決まらない。

○○パーセンタイル

- データを小さい順に並べたとき、初めから数えて全体の○○パーセントに位置する値
- 中央値は50パーセンタイル値となる

# 統計学の基礎

偏差
- 各データの値と平均値の差
- 各データが平均値からどれくらいずれているかを表す

平方和
- 偏差の2乗の和

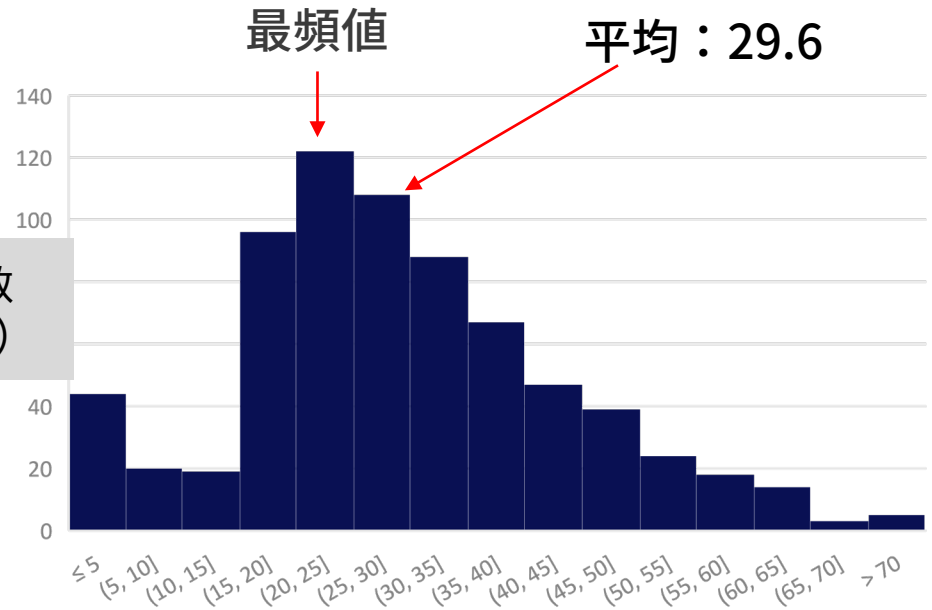
分散
- 偏差の2乗の平均値
- 平方和は標本数に影響されるので、これを調整したもの

標準偏差
- 分散の平方根で、データの単位とばらつきを表す単位が同じとなる
- データのばらつきを表す指標として重要

偏差値
- 平均を50、標準偏差を10に固定して、その中での相対位置を測る指標
- 難易度の異なるテストの得点の比較はできないが偏差値なら比較可能

# １次元のデータ

タイタニック号の乗客データ（https://www.kaggle.com/c/titanic/data, train）の年齢について調べると以下のようなグラフを書くことができる




縦軸：度数（例：人数）

横軸：変数の値（例：年齢）

変数の値を一定間隔(ここでは5歳ごと)でカテゴライズした階級

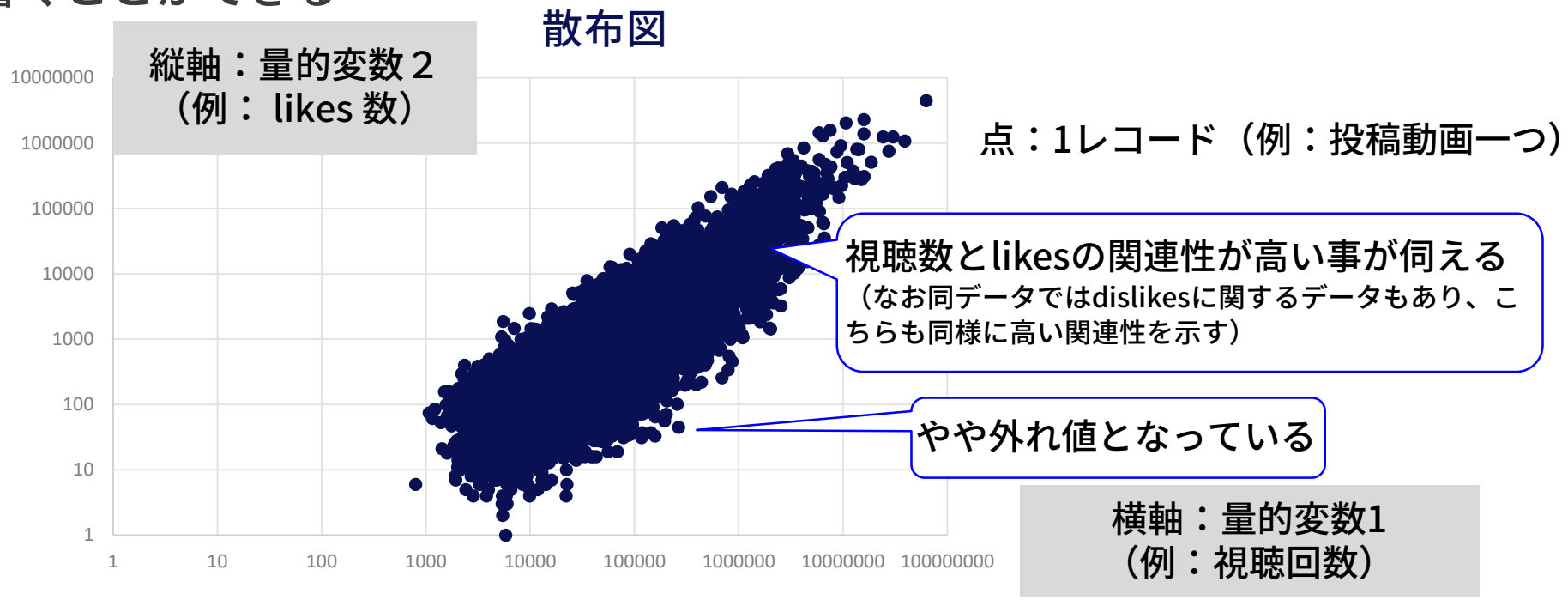
横軸の広がりがばらつきを表す：分散や標準偏差で数値化できる

このような表示方法をヒストグラム（histogram）と呼ぶ

横軸に変数の値をとり、縦軸にその頻度をとる方法は確率分布（第2回確率統計：e-learning にて解説）の考え方とも関連が深い

# ２次元のデータ：散布図

YouTubeの動画視聴数データ（https://www.kaggle.com/datasnaek/youtube-new, JPvideos.csv）の視聴回数とlikesの数ついて調べると以下のようなグラフを書くことができる




- 縦軸、横軸に異なる量的変数を設定、両対数グラフを使用している
  （両対数グラフについては「可視化：講義」にて解説）
- 1つの点は、1レコードを表す
- 因果関係が仮定される場合は、横軸を要因系、縦軸を結果系とする事が多い

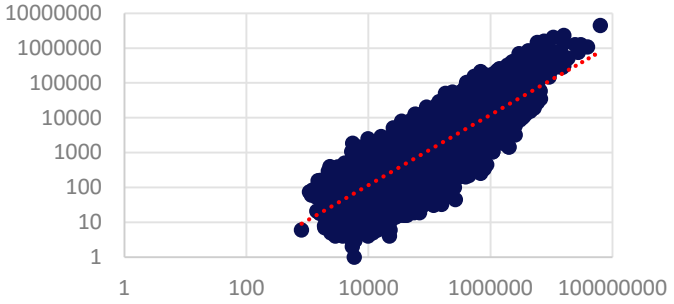
# 相関関係

■ 相関係数とは、散布図において、縦軸変数（ｘ）と横軸変数（ｙ）の直線的な関連の程度を表すもので、$-1 \leq r_{xy} \leq 1$を満たす

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

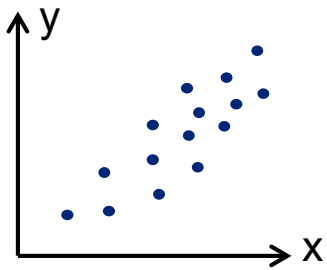
- $r_{xy}$が1に近いほど「正の相関が強い」：
  xが増加すればyも増加する
- $r_{xy}$が-1に近いほど「負の相関が強い」：
  xが増加すればyは減少する
- $r_{xy}$が0に近いときは相関がない（無相関）と考える
- $r_{xy} \fallingdotseq 0$だからといって、xとyに関連がないとは限らない（次のページで具体例を示す）

YouTubeの動画視聴数データ

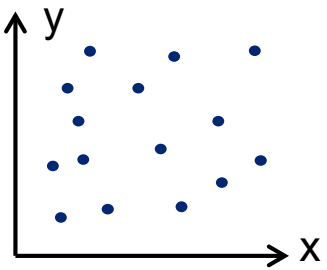



先ほどの例では、
視聴数とlikes数の相関係数は0.82と計算され、正の相関があることがわかる
（重複や０を除く処理を行ったデータから計算）

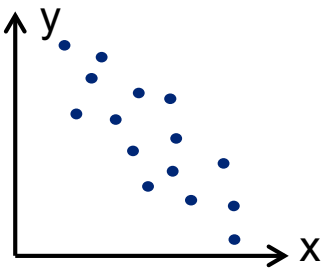
正の相関




相関なし

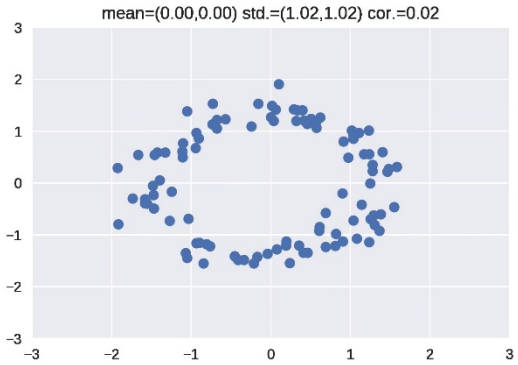



負の相関

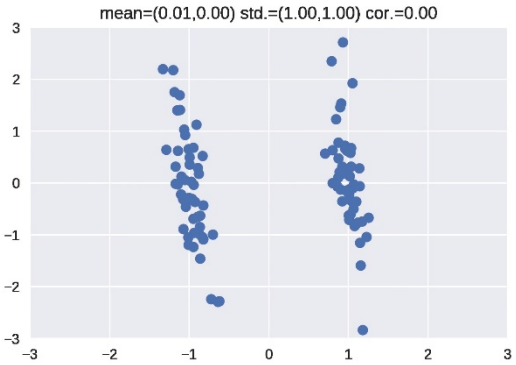


# 相関係数$r_{xy} \fallingdotseq 0$　平均$\fallingdotseq 0$　分散$\fallingdotseq 1$のデータの例

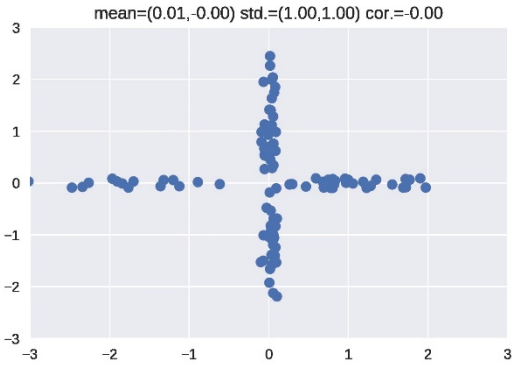
円形



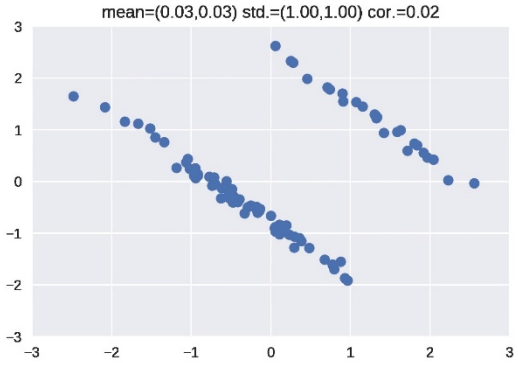

左右



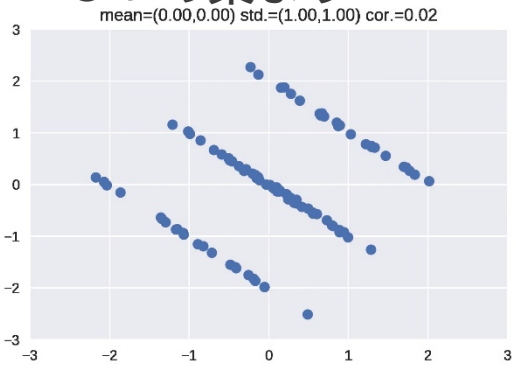

十字



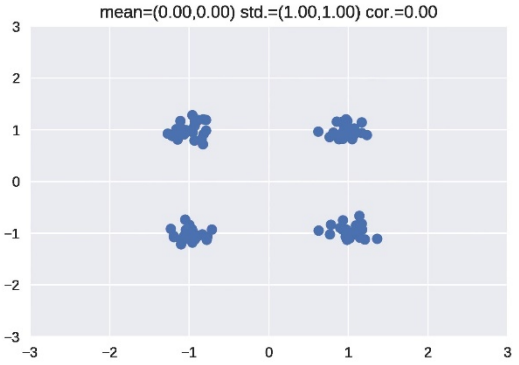

２つの集まり




３つの集まり




4つの集まり（点に集中）




単に、相関係数$r_{xy} \fallingdotseq 0$　平均$\fallingdotseq 0$　分散$\fallingdotseq 1$、の場合でも、
実際のデータは様々なばらつき方をしている可能性がある。
よく分からないデータの場合はプロットしてみることが重要

# 相関の値が大きい場合の留意点

相関関係 ≠ 因果関係

- 相関関係

  どちらか一方が増加すると、他方が増加または減少する、という2つの変数の線形的な関係性

- 因果関係

  一方が要因となって他方を増加または減少させる、原因と結果の関係

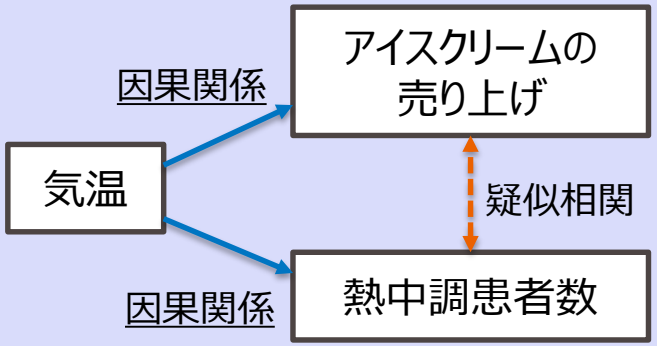
2つの事象に因果関係がないにもかかわらず、相関があり、関係があるかのように見える相関を「疑似相関」という

（例1）アイスクリームの売上と熱中症患者数は高い相関があるが・・・

気温の上昇→アイスクリームの売上高の増加

気温の上昇→熱中症患者数の増加

因果関係ではなく、擬似相関




相関関係は相関係数から算出できるが、そこに因果関係が存在しているのか判断するには、問題の背後にある現象を理解し、慎重に判断する必要がある

# 相関係数が使えないようなデータ

## 変数の種類に注意（変数の種類については「可視化：講義」にて解説）

住所（東京、大阪、名古屋、…）や職業（会社員、自営業、学生、…）などの変数は値の大小を定義できないため、相関係数のような指標を定義することができない

## 分割表

2変数以上の関係を記録するもの

以下の例では、100人の母集団のうち、20歳未満の男性は30人、20歳未満の女性は20人、…存在していることを表す

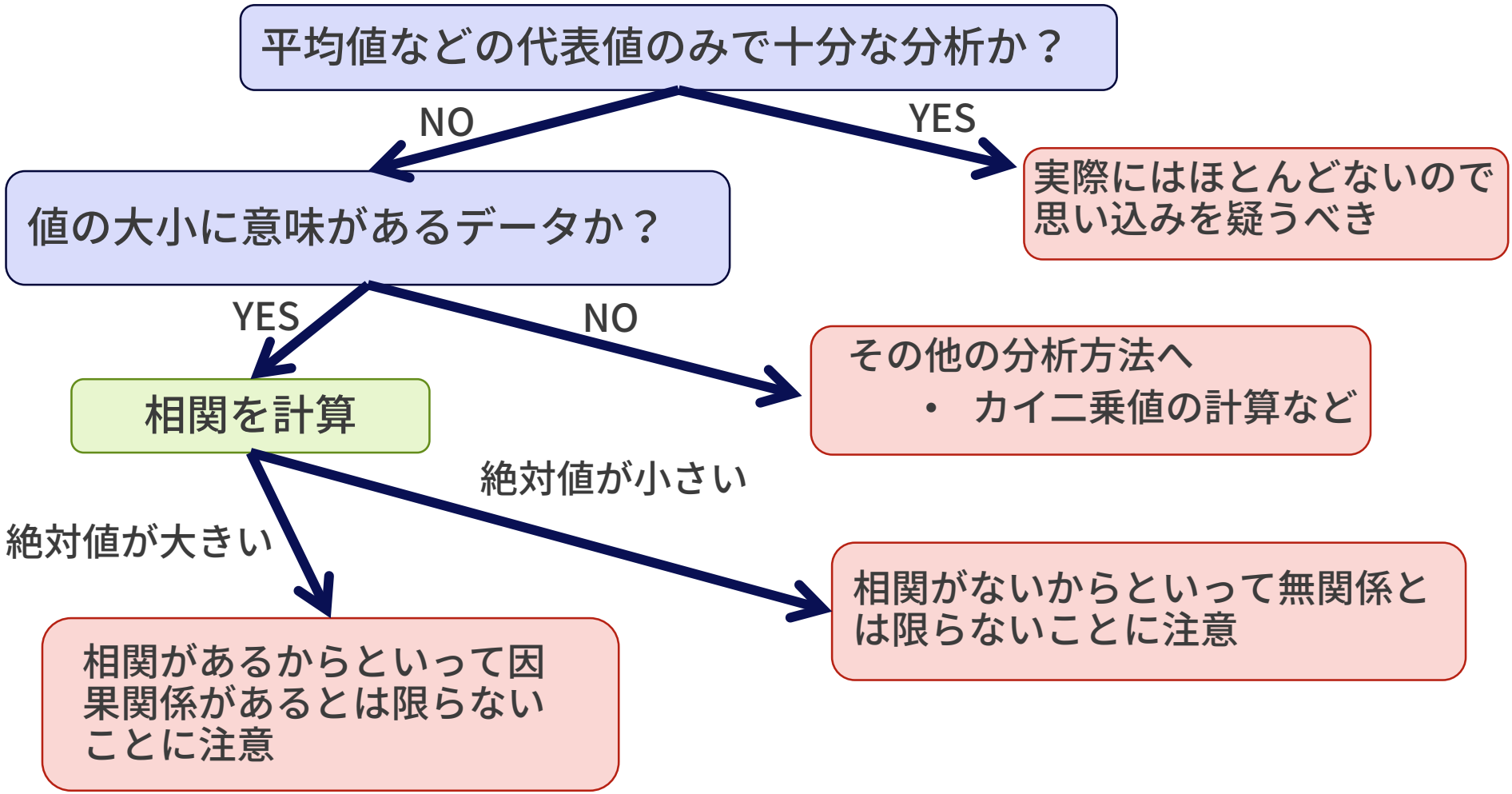
| | | Y: 年齢 | | |
|---|---|---|---|---|
| | | 20歳未満 | 20歳以上 | 合計 |
| X: 性別 | 男性 | 30<br>(N(男性,20歳未満)=30) | 10 | 40<br>p(男性)=0.40 |
| | 女性 | 20 | 40 | 60<br>p(女性)=0.60 |
| | 合計 | 50<br>p(20歳未満)=0.50 | 50<br>p(20歳以上)=0.50 | 100<br>(N=100) |

このような分割表が与えられた時に、

性別（男性、女性）と年齢（20歳未満、20歳以上）の間の関係性の強さを測りたい

→カイ二乗値（「第3回確率統計：e-learning」にて解説）

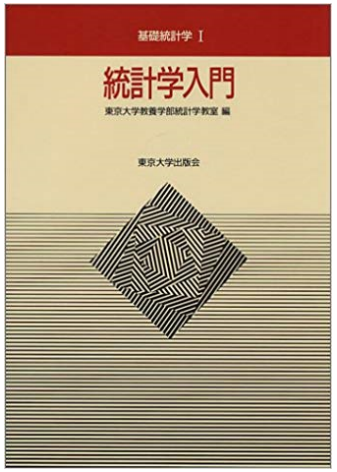
## 相関を扱うまとめ：



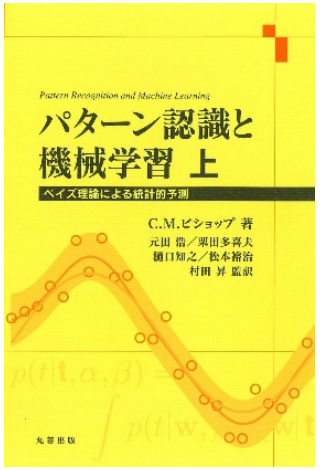

正しく理解して、過大解釈・過少評価にならないようにすることが大切

# より学びたい人への参考書籍

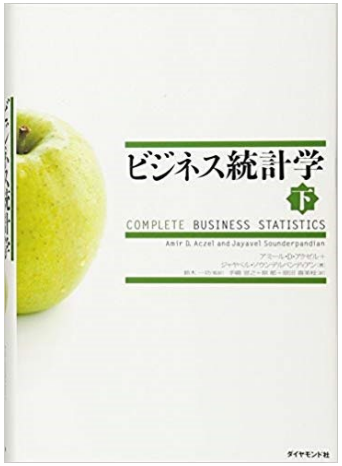
## 統計学一般の話～基礎的な話








## より実務よりの話

# ここまでのeラーニングでわからない人向けの外部のeラーニング授業動画・教材の紹介

【YouTubeチャンネル】大手前大学通信教育部

http://www.youtube.com/user/OteUniv/

- アニメーションで学ぶ統計学(1) - テストの結果から偏差値を示してみると

  https://www.youtube.com/watch?v=uTmDaiN6PyY

予備校のノリで学ぶ「大学の数学・物理」

https://www.youtube.com/channel/UCqmWJJolqAgjIdLqK3zD1QQ

- 【大学数学】推定・検定入門①(母集団と標本)【確率統計】

  https://www.youtube.com/watch?v=Bj8fkq533Dc&t=364s