# 標準得点

* データを位置(b), 尺度(a)に一次変換する
  + (平均)
  + (分散)
  + (標準偏差)
* 特に、a=1/σ、b=-μ/σの時、平均は0(各データ値から平均を引くため)、標準偏差は1(各データ値から標準偏差を割るため)に揃ったことになる。このzをxの標準化、標準得点(Z得点)と呼ぶ。
* さらに、標準得点に一次変換(平均を50、標準偏差を10)となるように変換したものを偏差値得点(T得点)と呼ぶ

# 平均差・ジニ係数

* (平均差)
* (ジニ係数)

# ピアソンの積率相関係数(相関係数)

* + を偏差積と呼ぶ。共分散は偏析差の全データの平均である。
  + σxはxの標準偏差、σyはyの標準偏差、cov[x,y]はx、yの共分散
* 相関関係は因果関係(どちらか一方がもう片方に影響を与える)ではない
* 相関関係には、見かけ上の相関の場合があり得る
  + 例)ある統計調査において、飲食店の数と金融機関の店舗数に強い正の相関関係がみられたが、実は(昼間)人口を間に挟んで強い正の相関関係が成立する。
  + 見かけ上の相関関係の場合、編相関係数を用いた方がよい。

|  |
| --- |
|  |

## 偏相関係数

## 相関係数の範囲の証明

* xi, yiの標準化変量を考える
* 上の時、rxyはz、wの共分散に等しい
* このことから、以下式は左辺が常に０以上のため、-1<=相関係数rxy<=1が成立する。

## 相関係数を指標とする際の注意点

* 見かけ上の相関
* 層別

## 順位相関係数

* 量的データではなく、質的データ間の相関を示す指標
* スピアマンの定義、ケンドールの定義が用いられることが多い

### スピアマンの定義

* RとR’に通常の積率相関係数を適用したもの

### ケンドールの定義

* 正順、つまりRi>Rj, Ri’>Rj’またはRi<Rj, Ri’<Rj’の場合+1
* 逆順、つまりRi>Rj, Ri’<Rj’またはRi<Rj, Ri’>Rj’の場合-1
* +1となった対の数Gと-1となった対の数Hの大小関係を観察する
* 対の全数n(n-1)/2の中で割合を取る

# 時系列データの因果関係

* 同じデータ系列でも、異時点間の相関関係を分析する
* 指標として、自己相関係数、系列相関係数がある
* グラフとしてコレログラムがある

## 遅れhの自己相関係数

* 遅れ(lag)1
* 遅れ(lag)h
* hが大きくなる＝時間の隔たりが大きい2時点間での相関係数を見ることであるから、hと共にrhは0へ近づく。ただし、季節性を含む月別の時系列データなどではr12は大きくなる場合がある。

# 直線への当てはめ

* 2変数x、yの間に、因果関係(x->y)がある場合、xを独立(説明)変数、yを従属(被説明)変数という。

## 回帰(最小二乗法)

* 2乗和を最小にするa,bの値を考える
* Lはa,bの2次式であるから、最小を求めるためにa,bで偏微分して極小値におけるa,bを求めると最終的に下正規方程式となる。
* これをa,bの方程式として解くと
  + a,bによる1次式をyのx上への回帰方程式、あるいは回帰直線と呼ぶ
    - aは傾き、または偏回帰係数と呼ぶ
    - bは切片と呼ぶ
* 回帰直線は以下式でも表される
* 回帰直線によって算出された当てはめ値(予測値)を下式で表現できる
* 実際の被説明変数と予測値の差が少ないほど当てはまり(予測)が良い
* (偏)回帰係数を変形すると以下式になる
  + 相関係数rの定義からaとrには、下式の関係が成立する
  + - これは、x,yの相関関係は直線関係の当てはまりの良さという意味を持つ  
      (σy/σxが1に近いほど(σx、σy同じほど)、aとrは一致していく)
  + よって、相関係数rは回帰直線への当てはまりの良さの尺度である。
    - r^2が1に近いほど(rが±1に近いほど)、yiはypに近づき、r^2=1なら正確にy=ax+bが成立し、yはxから完全に決定できる。つまりr^2はxがyを決定する強弱の度合いを示している。r^2を決定係数と呼ぶ
  + 被説明変数yのばらつきをSS0、回帰後のyのばらつきをSSEとすると、下式のようになるため、yのばらつきはSS0からSSEへr^2だけ減少したことになる。
  + 減少高をSSRとおくと、r^2が大きいほど回帰の効果も大きいと言える

# 平面のあてはめ

# 確率論

* 例)さいころを一回投げたとき、可能な結果は1-6であり、これを標本点(ω)と呼び、その全体の集合を標本空間または全事象(Ω)と呼ぶ。
* 上記の例では標本点は有限であるが、長さ、重さ、温度など区間内のすべての値を取り得る場合、標本点は無限個存在する。
* 標本点の発生を事象(event)と呼び、標本空間の部分集合(subset)として定義される。
  + 事象には、標本空間自身や、また標本点を一つも含まないような決して起こり得ないこと(空集合、φ)も事象と見なす
  + 例)コインを投げて、表を0、裏を1とした場合
    - 標本空間はΩ=(0,1)
    - 事象は({0,1}, 0, 1, φ)
  + 上記例で、{0,1}は複合事象と呼び、0と1に分解可能である。一方0または1を根元事象と呼び分解不可能である。
  + 例)電源の寿命の標本空間(購入時)は、Ω=(0,∞)で連続的な標本点として表す。100時間後の電源の寿命の標本空間は、Ω=(100,∞)で定義する。
* n個からr個とる順列の数(permutation)
* n個からr個とる組み合わせの数(combination)
  + 二項係数と呼ばれる

## スターリングの公式

## 事象の演算

* 和事象(A∪B、OR)
* 積事象(A∩B、AND)
  + AとBが排反事象であるとき、A∩B=φ
* 分配の法則
  + (A∪B)∩C=(A∩C)∪(B∩C)
  + (A∩B)∪C=(A∪C)∩(B∪C)
* Aが起こらないという事象をAの補事象と呼びAcで表す
* ド・モルガンの法則

## ラプラスの定義

* 確立とは、思考の根源事象が全部でN回あって、それらは同程度に確からしいとする
* この時、事象Aにとって都合の良いような根源事象、すなわち、それが出ればその事象の起こるような根源事象の数がR個あれば、事象Aの確率はP(A)=R/Nで定義できる。
* 利点として、確率が標本点の個数、つまり起こり方の場合の数の数え上げに帰することであり、順列・組合せの諸定理が使える。

## 確立の公理

* 以下を満たす事象は確立として認められる
  + 全ての事象Aに対し0≦P(A)≦1
  + P(Ω)=1
  + P(A1∪A2∪A3…)=P(A1)+P(A2)+P(A3)+…

## 加法定理

|  |
| --- |
|  |

* 事象AとBが排反事象(A∩B=φ)の場合、公理によって以下が成立する
* 事象AとBが背反事象でない場合、上図から、以下の通りになる
* 独立である場合、２つの積事象の確率は、確率の積としてあらわすことができる。
* ３つの事象ABCの独立性は、以下で定義できる。

## ベイズの定理

* 例)いくつかのツボに黒と白の玉が混ざって入っているものとし、このツボのどれからかいくつかの玉が抜き出されたとする。ただし、どのツボからは知られていない。  
  抜き出した玉という結果から、どのツボから取り出されたかという原因を推定したい。
* Aを結果、H1、H2、･･･Hkを原因とする。この時、ベイズの定理によって、P(Hi|A)を計算できる。
* H1、H2、･･･Hkは互いに背反で、かつH1∪H2∪･･･∪Hk＝Ωのごとくすべての場合を知り尽くしているとする。この時、
  + P(Hi)はHiの事前確立、P(Hi|A)は事後確率と呼ばれる。Aの事象が起こる事前と事後である。事前確立にはしばしば主観確率が用いられる。

# 確率変数

## 離散型確率変数

* Xのように大文字を用いて表す。さいころであれば
  + P(X=1)=1/6(p1), P(X=2)=1/6(p2), … P(X=6)=1/6(p6)
* 以下が成り立つ
  + p1≧0, p2≧0, … p6≧0
  + p1+p2+…+p6=1
* 一般に可算集合{x1, x2, …}の中の値を取る確率変数Xは離散型と呼ばれる
* 離散型確率変数Xの確率をXの確率分布と呼ぶ
* 離散型確率分布は以下を満たす

連続型確率変数

* 確率変数Xが関数f(x)によって以下を満たす時、Xは連続型の確率分布を持つという
* 連続型確率分布は以下を満たし、f(x)をXの確率密度関数と呼ぶ
* 連続型確率分布は以下の特徴を持つ
  + a=bとすると、P(X=a)=0で、一点の確率は0となる。
* 数学的に扱いやすい連続型の確率分布として、指数分布がある。
  + f(x)≧0の時、以下式で規格化されており、確率分布の性質を満たす
  + 定積分すると、一般に待ち時間(ある事象(偶然的な)が起こってから、次の事象(偶然的な)が起こるまでの時間Xなどの確率変数)が求まる。
* 一様分布は基本的な確率分布である
  + <https://bellcurve.jp/statistics/course/8015.html>
  + 確率変数Xがどのような値であっても、その時の確率密度関数f(x)が一定の値を取る分布のこと

## 累積分布関数

* 確率変数Xに対して、xを実数とする時x以下の確率F(x)=P(X≦x)をXの累積分布関数と呼ぶ。例えば、連続型の場合F(x)は密度関数fの定積分で表される
* 指数分布の累積分布関数は以下で表される
* 一様分布の累積分布関数は以下で表される
* 離散型の場合は以下で表される
* 累積分布関数は以下の性質を持つ
  + 広義単調増加
    - x1<x2ならば、F(x1)≦F(x2)
  + 範囲
    - x=∞のとき、F(x)=1
    - x=-∞のとき、F(x)=0
  + 右連続

## 期待値

* 期待値はE(X), μなどで表される。
* 離散型
* 連続型
* 存在しない場合φも定義する
* 指数分布の期待値
* 一様分布の期待値
* 期待値の性質
  + <https://bellcurve.jp/statistics/course/6714.html>
    - X,Yは確率変数、cは定数

## 分散

* 離散型
* 連続型
* 上述の式は、以下のように書き換え可能である
* 分散の性質
  + <https://bellcurve.jp/statistics/course/6718.html>