KYOTO UNIVERSITY

統計的モデリング基礎②
～確率分布・2変量間の関係・相関係数～

鹿島久嗣
（情報学科 計算機科学コース）

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

第2回目の今回は、前回にひきつづき、初歩的な内容をみていきます。

参考書

実証分析のための計量経済学
ー正しい手法と結果の読み方
／山本 勲（2015）

A5判/260頁
ISBN 978-4-502-16811-6

「経済学」と銘打っているが、技術的には統計的なデータ分析を扱っている。回帰分析を中心に、因果分析等についてもカバーしている

（かなり細かいところまで書いてある）参考書

現代数理統計学の基礎

久保川 達也 著・新井 仁之・小林 俊行・斎藤 毅・吉田 朋広 編

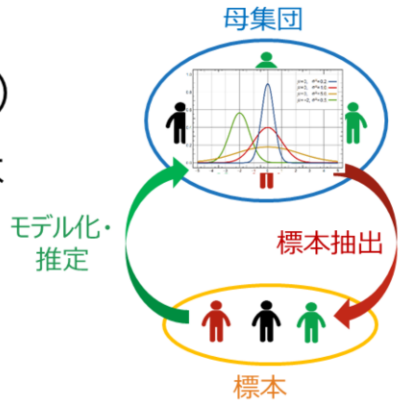
シリーズ名	共立講座 数学の魅力　全14巻+別巻1 【11】巻
ISBN	978-4-320-11166-0
判型	A5
ページ数	328ページ
発売日	2017年04月11日
本体価格	3,200円

理論的な背景含め基礎的事項がきちんと説明されている

統計的モデリングの考えかた

統計モデリングの考え方：部分から全体について知る

- 母集団：
  - 興味のある集合のすべての要素
  - 確率分布
    （分布のクラスやパラメータで指定される）
- 標本：母集団からの無作為抽出あるいは確率分布に従った抽出
  - 確率変数：確率的に値が決まる変数
- 標本から母集団について推測する
  （標本抽出の逆）



前回の復習です。母集団、標本について確認しておきます。

母集団が、我々が興味ある対象を全て含んだ集合、あるいはこれを表す確率分布とします。

確率分布は通常いくつかのパラメータを持ち、パラメータを決めると、その形が決まります。

我々は母集団を直接観測することはできず、ここから抽出された一部、すなわち標本のみが観測されます。

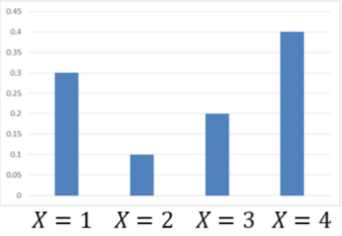
通常、標本は母集団から無作為に一様ランダムに抽出されると仮定します。あるいは、母集団が確率分布の場合には、この確率分布にしたがって、互いに独立に抽出されるとします。

我々の目的は、標本、すなわち母集団のごく一部から、全体、すなわち母集団について知ることです。

これはつまり、標本抽出の逆をするということであり、これこそが本講義の中心的なテーマです。

離散型確率変数の代表的な確率分布：離散分布、ベルヌーイ分布と2項分布

- 離散分布 $P(X=k)=f(k)$（但し $\sum_{k\in\mathcal{X}} f(k)=1, f(k)\geq 0$）
- ベルヌーイ分布：$\mathcal{X}=\{0,1\}$上の離散分布
- 二項分布
  - ベルヌーイ試行： 1が出る確率pのベルヌーイ分布からn回 独立に抽出する
  - 二項分布：ベルヌーイ試行において1がk回出る確率を与える



$$P(X=k\mid p)=\binom{n}{k}p^k(1-p)^{n-k}$$

n回の試行中のどこでk回の1が現れるかの場合の数

  - モデルパラメータpによって分布の形が一意に決定される

私たちは、データを前にしたとき、母集団が従うであろう確率分布を仮定します。これを確率モデルとよびます。

どのような確率モデルを考えるかは、データの形式や、私たちが対象について知っていることなどから決めます。

たとえば、サイコロの目のように、離散的な値をとる対象のモデルとしては、1か6までの整数のそれぞれをとりうる確率を指定した、離散分布を考えるのが自然です。

離散分布のパラメータは、確率変数がとりうる値の数になります。ただし、確率であるためには、全てのとりうる値の確率を足し合わせたものが1とならないといけないので、実質的なパラメータ数としては、ここから1引いたものになります。

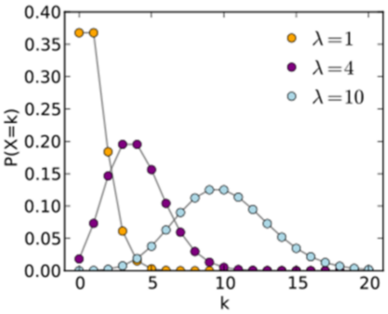
サイコロの場合は6個、実質的には5個になります。

離散分布の特殊な場合として、0か1の2値をとる場合には、ベルヌーイ分布と呼ばれます。ベルヌーイ分布は、コイントスで表が出る確率でイメージできます。パラメータは、表が出る確率pです。

2項分布とは、ベルヌーイ分布からの抽出、すなわちベルヌーイ試行をn回行ったときに、1がk回現れる確率を表すものです。パラメータはnとpになります。式を確認しておいてください。

離散型確率変数の代表的な確率分布：ポアソン分布（2項分布の極限）、その他

- ポアソン分布：$P(X = k \mid \lambda) = \frac{\lambda^k}{k!}\exp(-\lambda)$
  - 比較的稀な事象が何回起こるかを表現
    - 1分あたりのWebサーバアクセス数
    - ロットあたりの不良品数
  - パラメータ $\lambda > 0$
    - 2項分布のパラメータ（n, p）がない
    - 2項分布で$np = \lambda$として、$n \to \infty, p \to 0$とするとポアソン分布になる
- ほか、離散型の確率分布には幾何分布、負の2項分布などがある



https://en.wikipedia.org/wiki/Poisson_distribution#/media/File:Poisson_pmf.svg

離散的な値をとる確率変数に対する代表的な確率分布の別の例としては、0以上の整数をとる確率変数のモデルである、ポアソン分布があります。

ポアソン分布は、比較的稀な事象が起きる回数をモデル化するのによく使われます。

たとえば、Webサーバにおける1分あたりのアクセス数や、工場で生産される製品の、ロットあたりの不良品数などが例として挙げられます。

ポアソン分布は、唯一のパラメータλを持ちます。λは分布の平均を表します。図はλを変えたときの分布の形を示しています。

ポアソン分布は、二項分布において、コインの表が出る確率を表すパラメータpを0に近づけたときの極限になっています。

ポアソン分布が、比較的稀な事象が起きる回数をモデル化するのに適しているのは、これが理由です。

その他の離散型確率分布の代表例としては、幾何分布や負の2項分布などもあります。

連続型確率変数の代表的な確率分布：確率密度関数で指定される

- 連続分布は確率密度関数$f(x)$で指定される
  - 確率 ＝ 確率密度の積分
    $[a,b]$内の値をとる確率：$P(a \le X \le b) = \int_a^b f(x)dx$
  - 連続変数がある特定の値をとる確率：$P(X = a) = 0$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
- 一様分布：閉区間$[a,b]$上の一様分布は

$$f(x) = \begin{cases} \dfrac{1}{b-a} & (a \le x \le b) \\ 0 & (\text{その他}) \end{cases}$$

8　KYOTO UNIVERSITY

確率変数が連続値をとるとき、その分布はしばしば確率密度関数で指定されます。

確率密度関数を用いて、確率変数がある区間内の値をとる確率は、確率密度関数の積分として与えられます。

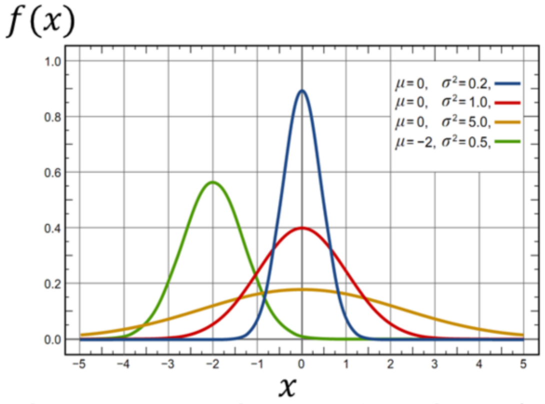
ある特定の値をとる確率はゼロであることに注意してください。確率密度関数を全域で積分すると1になります。

連続型の確率分布の代表例として、一様分布があります。これは、ある区間内での確率密度関数が一定になるというものです。

他の例として、次に紹介する正規分布があります。

連続型確率変数の代表的な確率分布：正規分布

- 正規分布：$f(x) = N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - パラメータ：平均μと分散σ^2



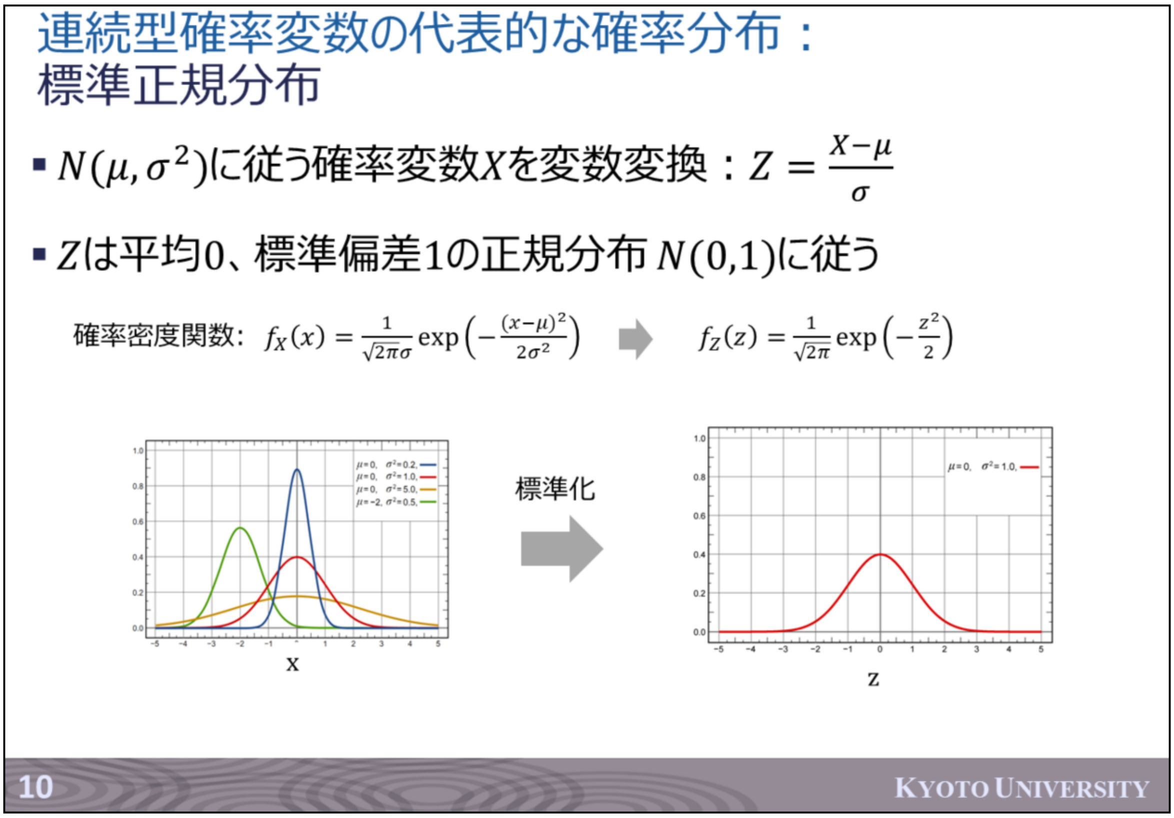
- 他、t分布、カイ2乗分布、ガンマ分布、ベータ分布、指数分布など

連続型の確率分布のもうひとつの代表例が、前回の講義でも紹介した、正規分布です。

正規分布の確率密度関数の形を、式と図で確認しておいてください。

正規分布には、平均μと分散σ^2の2つのパラメータがあります。

その他の連続型の確率分布の代表例としては、t分布、カイ2乗分布、ガンマ分布、ベータ分布、指数分布などがあります。このうちいくつかは後の講義でも出てきます。



正規分布は非常によく使われる確率モデルです。

しばしば扱いを簡単にするために、変数変換をして、平均が0、分散が1になるように正規化したものが用いられます。

これを標準正規分布と呼びます。

変数変換では、元々の確率変数から、その平均を引いて、標準偏差で割っています。

確率分布の特性値：期待値は確率分布の代表値

- 確率変数Xの関数$g(X)$の期待値：確率での重みづけ平均

$$
E[g(X)] = \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x & \text{(連続型確率変数)} \\ \displaystyle\sum_{x \in \mathcal{X}} g(x) f_X(x) & \text{(離散型確率変数)} \end{cases}
$$

- さまざまな関数$g(X)$に対する期待値によって分布の特性を捉える
- 性質：
  - 線形性：$E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$
  - イェンセンの不等式：$E[g(X)] \geq g(E[X])$ (ただしgは凸関数)

前回の講義で、ヒストグラムの代表値として、平均や分散などを紹介しました。

確率分布の場合も、これを特徴づける代表値があります。

これを考えるために、まず「期待値」と呼ばれるものを定義します。

期待値は、ある確率変数Xの関数g(X)について定義され、関数gの値を、確率で重みづけて足し合わせたものになります。

連続型の確率分布の場合には積分で定義され、離散型の確率分布の場合には和で定義されますが、気持ちは同じです。

定義を確認してください。

この期待値を使って、様々な関数gに対する期待値を考えることで分布の特徴を捉えます。

期待値の良く知られた性質として、線形性や、イェンセンの不等式があります。これらは非常によく出てきますので、確認しておいてください。

さまざまな期待値：平均と分散

- 平均 $\mu = E[X]$：Xの期待値 （分布の“真ん中”）
- 分散 $\sigma^2 = \mathrm{Var}(X) = E[(X-\mu)^2]$：平均からの二乗偏差の期待値 （分布の“幅”）
  - $\mathrm{Var}(X) = E[X^2] - E[X]^2$
  - 標準偏差σ：分散の正の平方根
    - 正規分布なら$\mu \pm \sigma$：68%，$\pm 2\sigma$：95%，$\pm 3\sigma$：99.7%
- より一般的にはモーメント$E[X^k]$
- 例：厳密なサイコロ $P(X = i) = \frac{1}{6}$ の平均、分散を求めよ

我々の興味の対象である確率変数の特徴を、様々な関数に対する期待値を考えることで捉えるといいましたが、具体的に見ていきます。

まず、もっとも代表的な代表値として、平均が挙げられます。平均は、期待値の定義中における関数gが確率変数Xそのものである場合に相当します。これはまさにヒストグラムにおける標本平均とおなじく、分布の「真ん中」を示す代表値になっています。

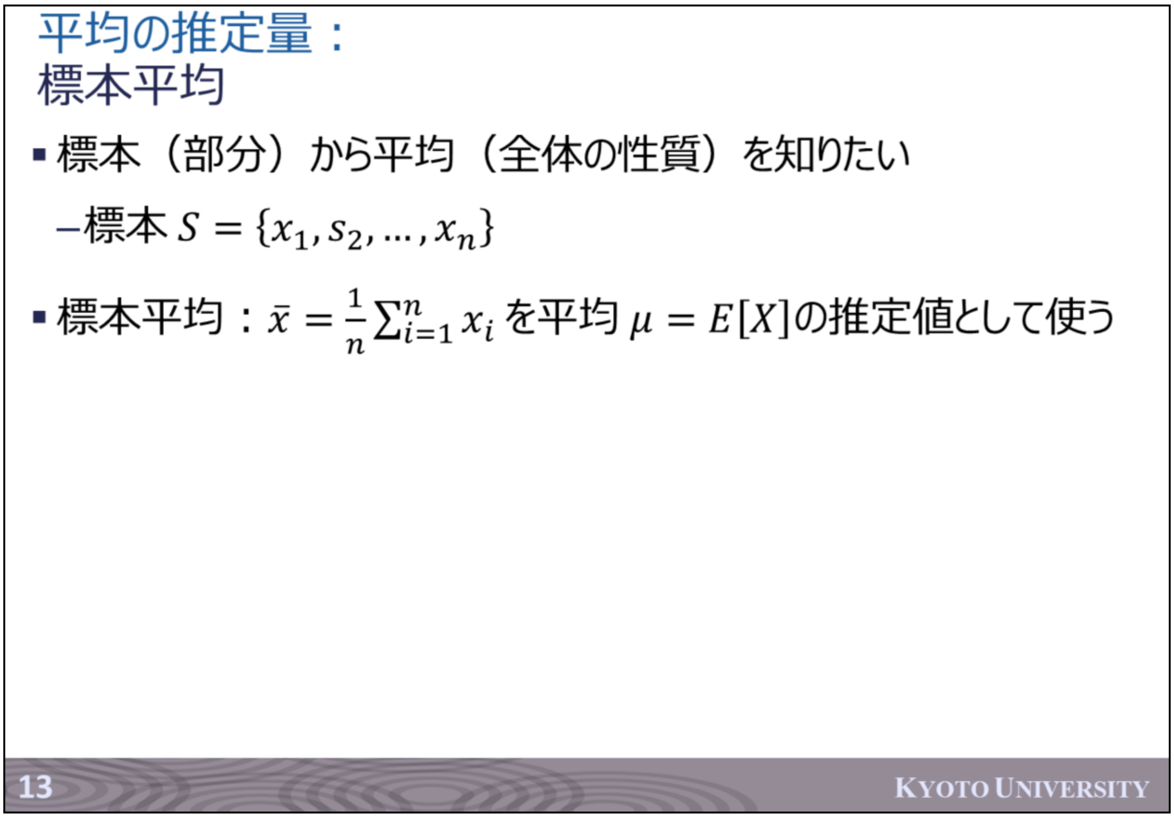
つぎに、平均からの広がりを表す代表値として、平均からの二乗偏差をgとしてとったときの期待値として定義します。これを分散と呼びます。

標準偏差は分散の平方根のことです。通常、標準偏差をσ、分散をσ^2と書きます。

さて、平均はXに、分散はX^2に関係する量であることに着目すると、gを3次以上の関数にすると、さらに分布の細かい形を捉えることができそうです。

これを見るのがモーメントです。モーメントはX^kの期待値として定義されます。

一般に、より高次のモーメントまで見ていくと、より詳細に関数の形を捉えることができます。



さて、私たちの興味は母集団をあらわすモデルそのものにあるが、実際にはこれを直接見ることは出来ないので、手元の標本（データ）からこれを伺い知るというのが私たちの立場でした。

たとえば、私たちは母集団の平均に興味があるが、これを直接知ることはできません。そこで、データからこれを推定することになります。

では、平均はどのように推定できるでしょうか？ 前回の講義で私たちは標本平均を見ました。

直感的には、標本平均は、平均の推定値として妥当な気がしますが、しかし、他にも候補はいくらでも考えられます。

たとえば、最後に観測したxでもいいし、適当に3つ選んだ値の真ん中の値でも構いませんが、直感的にはあまり良くなさそうです。

この「良い」とか「良くない」というのは、どのように評価できそうでしょうか。

推定量としての標本平均の好ましさ：標本平均は一致性と不偏性をもつ

- 標本平均は平均の推定値として好ましいか？
- 不偏性 $E_S[\bar{X}] = \mu$：標本平均の期待値は母集団の平均に一致する
  - E_Sは標本についての期待値 （何度も標本をとり直して、何度も標本平均を求めたときの、それらの平均）
- 一致性：標本サイズが大きくなるほど母集団の平均μに近づく
  - 標本平均の分散 $\mathrm{Var}_S[\bar{X}] = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$ （大数の法則）

14 KYOTO UNIVERSITY

標本平均が、平均の推定量として好ましいということを、どのように表すことができるでしょうか。

推定量の好ましさを表す性質として、いくつかの性質があります。

たとえば、不偏性は推定量の好ましさを表す性質として代表的なもののひとつです。不偏性とは、推定量の期待値が、その真の値に一致するという性質です。

標本平均は、不偏性をもちます。つまり、標本平均の期待値は、母集団の平均に一致します。

これは、推定量が平均的なふるまいとしては、大きい値でも小さい値でもなく、正しい値をとる、つまり、偏りがないということを言っています。

別の性質としては、一致性があります。これは、標本サイズが大きくなるほど、推定量が真の値に近づいていくというものです。

これは、たくさんデータをとれば、それだけ推定精度が上がっていくということで、これも好ましい性質です。

他にも、効率性、つまり推定量の分散が小さいほうが良いといったものもあります。

推定量としての標本平均の好ましさ：標本平均はBLUE（最良な線形不偏推定量）

- 効率性：推定値の分散が小さいこと
  - 標本平均$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$の代わりに最初の値を使う$\tilde{x} = x_1$とする
  - 標本平均のほうが効率的
    - 標本平均の分散$\frac{\sigma^2}{n}$ <最初の値の分散σ^2
- BLUE（最良な線形不偏推定量）：加重平均で表されるすべての不偏推定量のなかで、最も分散が小さい（効率的）なもの
  - 加重平均による推定量 $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} a_i x_i$

15 KYOTO UNIVERSITY

さて、私たちは、標本平均は平均の推定値として好ましいかということが知りたいのでした。

でも、標本平均のほかにも、最後に観測した値でもよいし、適当に3つとってきた値の真ん中の値でもよいですね。

これらは、スライドの一番下のような加重平均によって統一的に書くことができます。

aの値を変えると、上記の推定量は全てこの形で表せることがわかると思います。

実は、標本平均は、この形で表せるすべての不偏推定量の中で、最も分散が小さい、つまり、効率的であることが知られています。

このような推定量のことを、最良な線形不偏推定量（BLUE: Best Linear Unbiased Estimator）と呼びます。

「標本平均はBLUEです」

分散の推定量：
不偏分散

- 標本分散：$\frac{\left(x^{(1)}-\bar{x}\right)^2+\cdots+\left(x^{(n)}-\bar{x}\right)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 不偏性をもたない： $E_S\left[\frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}-\bar{X}\right)^2\right] = \frac{n-1}{n}\sigma^2$
- 不偏分散：$\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 不偏性をもつ：期待値が母集団の分散に一致する
- どちらも一致性はもつ：
  - 標本サイズが大きくなるほど母集団の分散に近づく
  - nが大きいところではnも$n-1$も大した違いはない

16 KYOTO UNIVERSITY

前回の講義で出てきた不偏分散について思い出してみましょう。

不偏分散とは、標本平均からの二乗誤差を足して、n-1で割ったものでしたが、この-1がちょっと気持ち悪い感じでした。

実は、不偏分散はその名のとおり、n-1で割ることによって不偏性を持ちます。言い換えると、nで割ると、分散を過小評価してしまうということになります。

一方で、nで割っても、n-1で割っても、一致性は保証されます。

一致性とは、データサイズが大きいときの性質ですので、nが大きいときには、nもn-1もそれほど変わらないと考えれば納得ですね。

2変量データの解析

2つ以上の変量のデータ分析：
変量間の関係を調べることでより深い分析が可能

- 前回は、1変数の単純分析について考えた
- 2つ（もしくはもっと多く）の変数の関係に興味があることが多い
- ２変量（あるいはさらに多く）の間の関係を調べることで、より積極的なデータ利活用が可能になる
  - ある属性をもった人は、ある商品を買いやすいのか？
  - ある薬を飲むと、ある病気に効果があるのか？
- 変数の種類によって、さまざまな分析手法がある
  - 量的変数：散布図、相関、回帰
  - 質的変数：クロス表、リスク差・比、オッズ比

18 KYOTO UNIVERSITY

これまでは、1変数の分析について扱ってきましたが、実際のデータ分析において、2つ、もしくはより多くの変数の関係に興味がある場合が少なくありません。

たとえば、ある属性を持った人が、ある商品を買いやすいのかといった、マーケティング的な分析や、ある薬は、ある病気に効果があるのがといった、医療的な分析など、いずれも2つの変数の関係について調べようとしています。

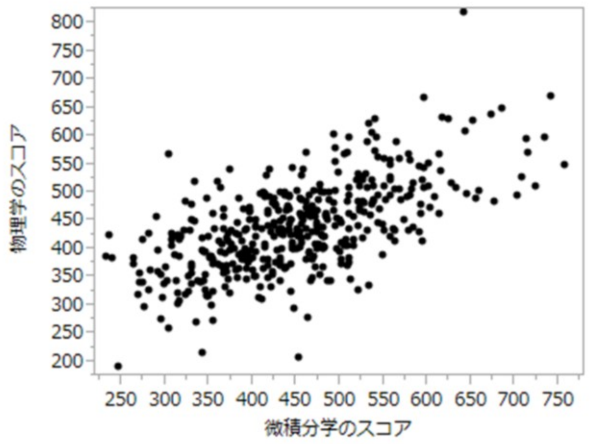
複数の変数の関係を調べることで、より有益なデータ分析を行うことが期待できます。

変数が量的であるか質的であるかに応じて、様々な分析法がありますが、以下では、2つの量的変数の関係をまずは見ていきます。

2変量の単純な分析：散布図による視覚化

- 例：微積分の点数と物理の点数の関係

	微積分学のスコア	物理学のスコア
1	441.4	470.7
2	632.16	508.44
3	361.56	412.75
4	479.39	425.47
5	476.32	408.27
6	446.92	400.99
7	394.2	390.62
8	645.76	496.97
9	329.75	367.39
10	496.07	453.41
11	487.91	498.97
12	403.82	441.7
13	480.21	400.41
14	460.33	460.72
15	303.72	259.66
16	463.01	278.04
17	428.98	396.21
18	412.12	380.53
19	366.84	355.72



JMPサンプルデータ

量的変数が1つの場合、ヒストグラムを描いて視覚的にデータの分布を捉えるのが最初のステップでした。

2つの量的変数の関係を調べようと思ったときにまず描くのは散布図です。

散布図は2つの変数の組合せを2次元上の点で描いたものです。

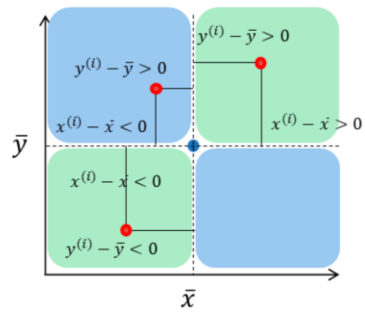
この図は、学生の数学のテストと物理のテストの点数を視覚化したものですが、このように散布図を描くことで、両者の間にはともに大きい、あるいは、ともに小さくなるといった、右肩上がりの関係があることが見て取れます。

つまり、数学の点が高い人は物理の点も高く、その逆も然りということがわかります。

さて、ヒストグラムと標本平均の関係と同じく、やはり、散布図から読み取れる傾向を何らかの数値的指標として得たい場面もあるでしょう。

2変数間の関係の指標：共分散と相関

- 一方が増えたときに他方が増える（減る）関係性を表す指標
- 共分散(covariance)：$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$
  - ただし、$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$， $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
  - 偏差積の平均 (データのバラツキを表現)
    - 偏差 $(x^{(i)} - \bar{x})$ と偏差 $(y^{(i)} - \bar{y})$ の符号が一致する（緑領域）なら正の値をとる
    - 偏差 $(x^{(i)} - \bar{x})$ と偏差 $(y^{(i)} - \bar{y})$ の符号が不一致である（青領域）なら負の値をとる



- ただし、x, y の単位やスケールに影響されるため共分散の絶対的な大きさのみでは関係の強さを評価できない

20 KYOTO UNIVERSITY

2つの量的確率変数の関係を表す指標のひとつが、相関係数と呼ばれるものです。
相関係数と、その元となる、共分散について見ていきましょう。

共分散は上記の式で定義されます。

見た感じは、不偏分散に似たところがありますが、不偏分散が平均からの偏差の二乗によって定義されるのに対して、共分散は2つの異なる変数のそれぞれの平均からの偏差の積によって定義されます。

同一の変数の共分散を考えると、不偏分散に一致することがわかります。

さて、共分散の定義を眺めてみると、2つの変数の増減の向きが概ね一致するときには正の大きな値を、逆に増減の向きが正反対のときには負の大きな値をとることがわかります。

つまり、2つの変数の関係が右肩上がりのとき共分散は正の値、関係が右肩下がりのときに共分散は負の値をとることになります。

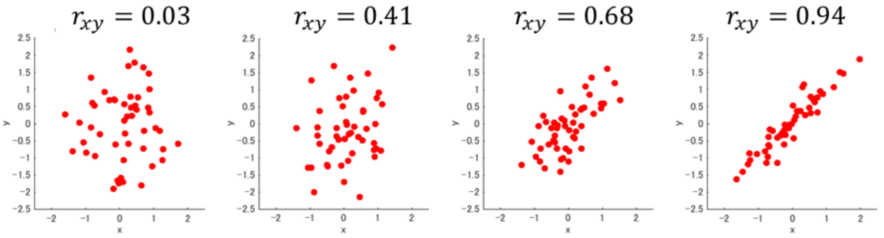
このような関係がないとき、共分散は0に近い値となります。

基本的には、共分散を見ることで、2つの変数の増減の関係を調べることができますが、このままだと、もとの変数のスケールに影響を受けてしまいます。

たとえば、片方の変数を単純に10倍すると、共分散も10倍となるため、共分散の絶対値そのものの大きさは、かならずしも関係の強さを表しません。

2変数間の関係の指標：共分散と相関

- 相関 (correlation)：$r_{xy} = \dfrac{\sum_{i=1}^{n}(x^{(i)}-\bar{x})(y^{(i)}-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x^{(i)}-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y^{(i)}-\bar{y})^2}}$
  - $r_{xy} > 0$：正の相関　$r_{xy} < 0$：負の相関　$r_{xy} = 0$：無相関
  - $-1 \leq r_{xy} \leq 1$の値を取る



そこで、共分散のスケールを調整して、-1から1の間の値をとるようにしたものが相関係数です。

定義式から分かるように、相関係数は共分散をそれぞれの変数の標準偏差で割ったものです。

図を見ると、相関係数が0に近いときは、2変数の関係は見られませんが、だんだん1に近づくにつれ右肩上がりの関係が見えてくるのが分かります。

ヒストグラムを代表する値として、標本平均や中央値、不偏分散などがあったのに対して、散布図を代表する値として相関係数をみることで、その傾向を大まかにつかむことができます。

たとえば、年収と消費の間に関係があるのかといったことや、薬の投与量と検査値の間の関係を調べるといったことができます。

相関についての注意：
相関関係と因果関係はイコールではない

- 相関関係 (correlation) があるからといって必ずしも因果関係 (causality) があるわけではない
  - 体重と身長の相関は高いが片方が他方を決めるともいえない
  - 因果関係を示すことは難しい
- 見かけ上の因果関係に注意
  - 背後に共通原因が存在する場合もある
  - 例：「明かりをつけたまま眠る子供は近視になりやすい」？
    - 両者に「親が近視」という別の原因がある
  - そのほか、原因と結果が逆、互いに一方が他方の原因になっている、といったケースあり

22 KYOTO UNIVERSITY

最後に、注意事項を

相関関係は2つの変数の大小がお互いにどのように関係しているかを示すものですが、これは必ずしも因果関係、すなわち、片方の変数を操作することで、もう片方の変数に影響を与えることができるということを意味しません。

たとえば、体重と身長の間には、ある程度の正の相関がみられますが、かといって、カロリーの高い食事をして体重を増やせば、身長が伸びるとは思えません。

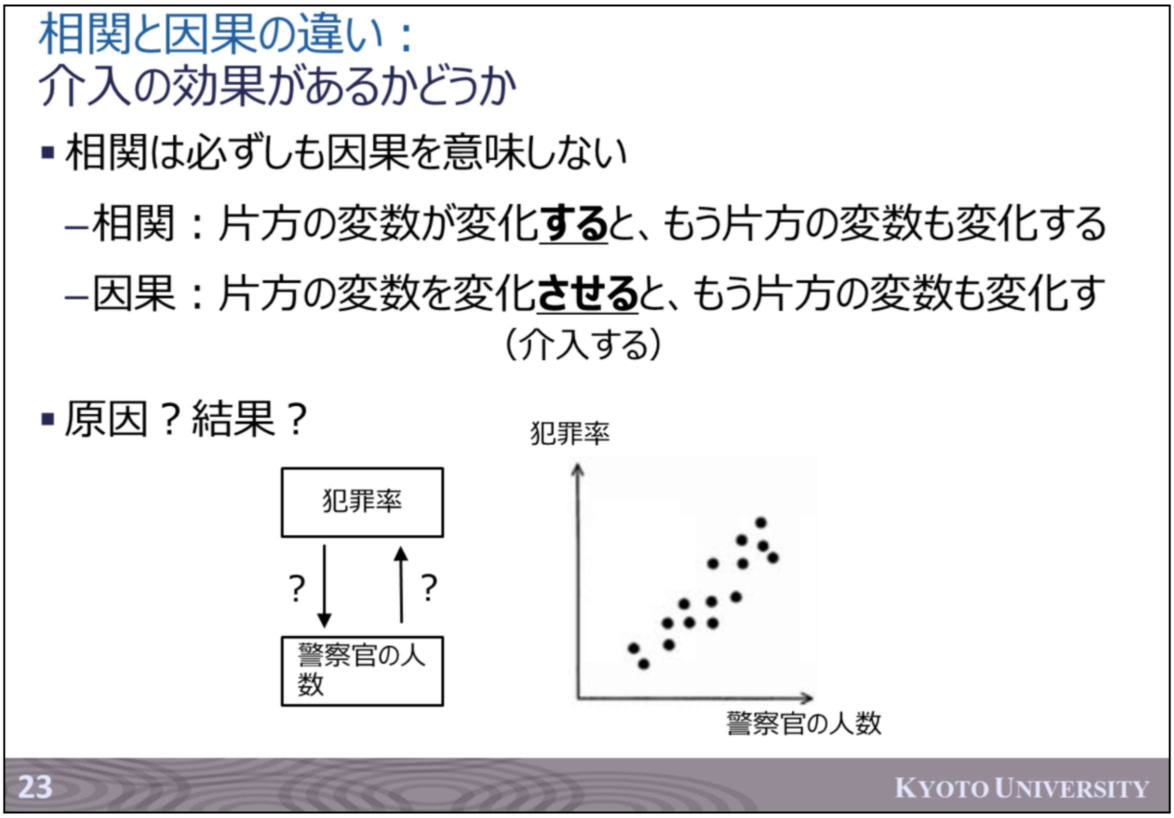
相関関係があるからといって因果関係があるとは限らない場合の、よくある原因として、2つの変数の背後に共通の原因が存在し、その原因によって2つの変数が操られ、結果として相関を生じる場合があります。

このようなケースを擬似相関と呼ぶことがあります（「擬似相関」というと、相関がないように聞こえますが、因果関係はないが、相関関係はあるという意味です）

有名な例として：

・アイスクリームの売り上げとプールでの溺死事故数の間の正の相関（背後に猛暑という共通の原因があるが、両者には因果関係はない）

・赤ん坊の出生数とコウノトリの数の間の正の相関（おそらく両方が9か月前の天候と関係していると考えられている）



さきほど述べたように、因果関係と相関関係の違いは、片方の変数を操作することで、もう片方の変数を操作できるかどうかということになります。

因果関係については、本講義の後半で、また詳しく扱いますが、世の中には、自らの主張を正当化するために、相関関係をもとに因果関係を主張する人たちがいます。

いまのところは「見た目の相関にだまされないようにしよう」ということを覚えておきましょう
。