

KYOTO UNIVERSITY

統計的モデリング基礎①

～概要・導入～

鹿島久嗣
(情報学科 計算機科学コース)

DEPARTMENT OF INTELLIGENCE SCIENCE
AND TECHNOLOGY

第一回では、今学期の講義の(暫定的な)進め方と、講義内容全体の導入を行います。

今学期の講義について

まず、今学期の暫定的な講義実施法について説明します。

本講義の学習の進め方： PandA上で学習を進めてください

- 現在、物理的な講義を実施できず、この状況は少なくとも5月の連休明けまでは続く見込みです
- 各自、PandA上で学習を進めてください
 - PandAのページ：<https://bit.ly/2wg7vwU>
 - 連絡事項・課題提出等はこちらから行います
 - 資料はWebに置きます：<https://bit.ly/2V6ebWY>
 - 特定の教科書は使用しません

現在、物理的な講義は実施できない状況にあり、これは少なくとも5月の連休明けまでは続くと思われます。

当面のところ、本講義ではPandAを利用して学習を進めますので、各自、必ず定期的にPandAの本講義のページを確認してください。

連絡事項や、課題の提出等は基本的にここで行います。

また、PandAからもリンクを貼りますが、講義資料はすべてWeb上に置きますので、これを使って学習を進めてください。

本講義では特定の教科書を使用することはしませんが各話題において適宜参考図書等は紹介します。

成績評価：

PandA上での課題と中間・期末試験またはその代替による

- 通常では中間試験と期末試験で成績をつける
 - (情報学科 1 回生の「アルゴリズムとデータ構造入門」と同じ感じ)
- 今学期は状況に合わせて対応するが、いまのところ：
 - PandA上での各回の理解を確認するための課題
 - 中間試験・期末試験の一部あるいはすべてがレポート課題等に代わる可能性あり

気になる成績評価ですが、通常だと、中間試験と期末試験の成績(のみ)によって成績を評価しています。

今学期の状況では、試験の一部あるいはすべてが実施できない可能性も高いため、その場合にはレポート課題等に置き換えることになります。

また、PandA上での学習が本格的に開始されたあとは、毎回の学習確認のための(比較的簡単な)課題が出ますので、これらも成績評価に加わります。

いずれにしても、PandA上等での連絡を行いますので、必ず定期的に確認してください。

導入

ここからは講義の導入に移ります。

本講義の目的： 統計的モデル化の基礎を身につける

- 我々は、研究や業務で出会う様々な種類のデータから適切な判断を下したい（自動的なシステムあるいは、人間の意思決定をサポート）場面にしばしば遭遇する
 - 例：実験データ、社会調査データ、検査・診断データ、売り上げデータ、行動データ、Webサイトのログ等々
- そのために、観測されたデータに基づいて、不確実な現象の特性を捉え、将来の観測値の確率分布を推定し、予測や制御に資する統計的モデル化の基礎を学習する
 - 現在注目を浴びている機械学習（≡人工知能）の基礎でもある

本講義の目的を確認します。

皆さんは今後、研究室で研究を実施したり、その後社会に出て仕事をするようになったときに、様々な種類のデータに出会うことになります。それらのデータから研究上・業務上の適切な判断を下したい場面に出会うでしょう。

データの種類にはさまざまありますが、たとえば、研究上の実験データ、社会調査データ、検査・診断データ、売り上げデータ、行動データ、Webサイトのログ等々、世の中にはたくさんの種類のデータがあります。これらのデータをうまく利活用して、様々な決定を自動的に行う知的システムを構築したり、あるいは人間の意思決定に役立つ知見を得ることが求められます。

以上を踏まえたくえで、本講義の目標は、上記の目標を達成するための基本的な知識を身につけること、つまり、これらの観測されたデータに基づいて不確実な現象の特性を捉え、将来の観測値の確率分布を推定し、予測や制御に資する統計的モデル化の基礎を学習することです。現在、人工知能や機械学習が注目を浴びていますが、上記の知識は、人工知能・機械学習を学ぶ上でも大変役に立ちます（というよりむしろ、そのものといっても過言ではないかもしれません...）

統計的モデルが世の中で使われている例： 顧客の購買行動の予測に基づく推薦

■ Webショッピングサイトでの商品推薦の例を考える：

– 誰に何を薦めると買ってくれるだろうか？ 下記はタコ焼き機を買った人に推薦される商品



■ 消費者の購買行動を予測し、購入しそうなものを推薦する

– 過去の購買履歴をもとに、ある商品を買ってくれるかどうか予測

- これまでに購入した商品のリストから、将来ある商品を購入する確率を推定する

– 最も購買可能性が高いものから提示すればよさそう

統計的モデリングが世の中で使われている例をひとつ見てみます。

みなさんもよくWebショッピングサイトで買い物などをしますが、サイトを眺めていると、様々な商品をおすすめされるのを見ることがあると思います。

そのサイトをよく使っていると、自分の興味ある商品をかなり精度よく薦めてくれることがあると思います。

こういった機能は、しばしば「推薦システム」と呼ばれ、ショッピングだけでなく、様々な情報提供サービス上で見ることができます。

彼らは、私たちがそのサイトで行った過去に購入した商品の履歴をもとに、他の商品を買う可能性を推定し、その中で購入可能性が高い商品を提示しています。

このようなことは一見不可能にも思えますが、非常にたくさんの方がこのサイトを利用したデータが蓄積されてくると、どういう商品を買ったどういう人が、別のどういう商品を買いやすいかといった傾向がみえてくるのです。

本講義のトピック： データ解析の基礎的項目

1. 回帰モデル：線形回帰モデルと最小二乗法による推定など
2. モデル推定：最尤推定、事後確率最大化等のモデル推定の枠組み
3. モデル選択：情報量基準、交差確認等に基づくモデルの選択
4. 質的変数の予測モデル：ロジスティック回帰モデルなど
5. 様々なデータに対する確率モデル：時系列、テキスト、...
6. ベイズ推定：ベイズ統計の枠組みに基づく統計モデル推定
7. 因果推論：相関関係と因果関係の違い、因果関係の推定法

本講義で触れるトピックはおおむね上記の内容になります。
多少の追加や順番の前後があると思います。

データとはなにか： たとえば表形式データ

■ 項目と値の組で構成される

(各行が1つの企業、業種や会社規模などで表されている)

全学ライセンスあり
(医・薬あたりではデフォクトらしい...)

JMPサンプルデータ



The screenshot shows the JMP Pro interface with a data table titled 'Companies'. The table has 10 columns: 'タイプ' (Type), '会社規模' (Company Size), '売上(\$M)' (Sales), '利益(\$M)' (Profit), '従業員数' (Number of Employees), '従業員一人あたりの利益' (Profit per Employee), '資産' (Assets), and '利益/売上 (単位:%)' (Profit Margin). The rows represent different companies, such as 'Computer' and 'Pharmaceutical', with varying sizes like 'small', 'medium', and 'big'. The table is annotated with callouts: '項目' (Item) pointing to the column headers and '値' (Value) pointing to the data cells.

タイプ	会社規模	売上(\$M)	利益(\$M)	従業員数	従業員一人あたりの利益	資産	利益/売上 (単位:%)	
1	Computer	small	855.1	31.0	7523	4120.70	615.2	3.63
2	Pharmaceutical	big	5453.5	859.8	40929	21007.11	4851.6	15.77
3	Computer	small	2153.7	153.0	8200	18658.54	2233.7	7.10
4	Pharmaceutical	big	6747.0	1102.2	50816	21690.02	5681.5	16.34
5	Computer	small	5284.0	454.0	12068	37620.15	2743.9	8.59
6	Pharmaceutical	big	9422.0	747.0	54100	13807.76	8497.0	7.93
7	Computer	small	2876.1	333.3	9500	35084.21	2090.4	11.59
8	Computer	small	709.3	41.4	5000	8280.00	468.1	5.84
9	Computer	small	2952.1	-680.4	18000	-37800.0	1860.7	-23.05
10	Computer	small	784.7	89.0	4708	18903.99	955.8	11.34
11	Computer	small	1324.3	-119.7	13740	-8711.79	1040.2	-9.04
12	Pharmaceutical	medium	4175.6	939.5	28200	33315.60	5848.0	22.50
13	Computer	big	11899.0	829.0	95000	8726.32	10075.0	6.97
14	Computer	small	873.6	79.5	8200	9695.12	808.0	9.10
15	Pharmaceutical	big	9844.0	1082.0	83100	13020.46	7919.0	10.99
16	Pharmaceutical	small	969.2	227.4	3418	66530.13	784.0	23.46
17	Pharmaceutical	medium	6698.4	1495.4	34400	43470.93	6756.7	22.32
18	Computer	big	5956.0	412.0	56000	7357.14	4500.0	6.92
19	Pharmaceutical	big	5002.7	681.1	42100	16178.15	8224.8	11.64

具体的なデータがどのようなものかを見てみます。

データといって一番わかりやすいのは、たとえばExcelのような表計算ソフトウェアで扱うような、表形式のものが挙げられます。

ここで挙げたのはJMPと呼ばれる統計ソフトウェアに含まれているサンプルデータですが、各行がひとつの企業を表しています。

表は項目(業種や売り上げ、会社規模など)と、実際の値(ある会社の売り上げが100\$Mだとか)から構成されています。

ちなみに、京都大学はJMPの全学ライセンスを契約しているので、みなさんも使うことができますが、情報系だとあまりなじみがないかもしれせん。

医学・薬学などの分野では非常にポピュラーのようですが...

データをもとにやりたいことの例： 予測や因果関係の抽出

- 前述のデータを利用してやりたいこととして、例えば：
 - 予測：会社の売上げから利益を予測したい
 - モデル推定・選択：予測の式をデータからどのように得るか
 - 因果推論：従業員を減らすと、従業員ひとりあたり利益は伸びるかなどが考えられるだろう
- さらに進んで、以下のようなことも考えられるかもしれない：
 - ベイズ推定：データが少ないときにどうするか？
 - 様々なデータ：会社説明のテキストがあったらどうするか？

たとえば、前頁のデータをつかってどのようなことができそうでしょうか。

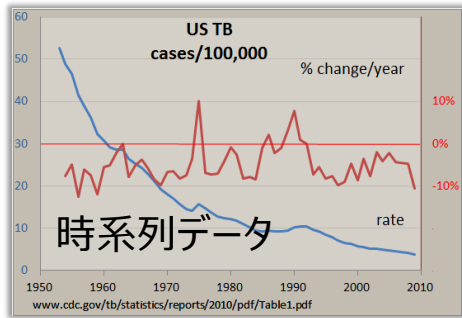
たとえば、会社の売上げをもとに、その利益を予測したりできるかもしれません。そのためには、データの背後に潜む関係性としてどのような仮定を設けるか（「モデル」と呼ばれます）、それをどのようにデータから発見するかを考える必要があります。

あるいは、経営的な判断を迫られたときに、従業員を増やしたり減らしたりすることで、利益がどう変わっていくかを知りたいかもしれません。

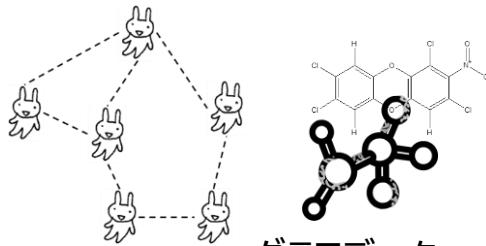
前頁のデータのほかに、たとえばテキスト情報（会社の説明など）があったらどうなるでしょうか？ あるいは、データがとても少ない場合はどうなるでしょうか？ といったことにも興味があるかもしれません。

表形式以外のさまざまなデータ： 時系列、テキスト、グラフなど...

- 時系列
- テキスト
- グラフ



https://en.wikipedia.org/wiki/Time_series#/media/File:Tuberculosis_incidence_US_1953-2009.png



前述の表形式のデータの他にも、世の中には様々な形式のデータがあります。

たとえば、時間とともに推移する株価や、温度センサーが記録するデータなどは、時系列データと呼ばれます。

Wikipediaの記事やTwitterのつぶやきなどは、テキストデータです。

あるいは、SNS上の人々のつながりや、化合物中の原子間の共有結合など、モノ・コトの間の関係を表したものは「グラフ」データと呼ばれます。

統計的モデル化の目的： 「部分」から「全体」を知ること

- すべての場合（母集団）を網羅的に観測できることは少ない

- 「記述統計」と「推測統計」

- 記述統計：全数調査を前提とする

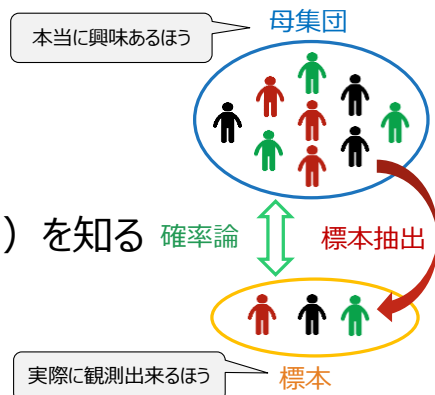
- 推測統計：標本調査を前提とする

- 部分（標本）から全体（母集団）を知る

- 過去から未来を予測する

- 母集団と標本は「確率論」でつながる

- 母集団は対象となる集合の要素すべて、あるいは、何らかの確率分布に従っていて、標本はそこから確率的に取り出されたと考える



我々はこれからデータをもとに統計モデル化を行うわけですが、その主な目的は「部分から全体を窺い知ること」であるといえます。

たとえば、人が商品を買うかどうかを知りたいと思ったとき、究極的には、地球上すべての人間（さらにいえば今後生まれてくるすべての人間）について、それを知りたいわけですが、このように我々が真に興味をもっている集団のことを母集団と呼びます。

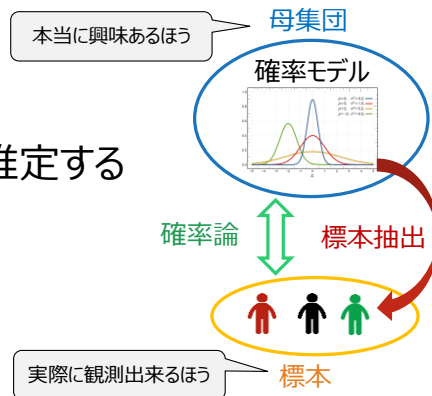
母集団を直接観測できれば言うことなしですが、実際には、コスト的な問題、あるいは原理的に不可能などの理由で全数調査を実施することは困難です。

そこで、母集団の一部をとりだした「標本」をもとに、全体に関して間接的に推測を行うこととなります。

標本は通常、母集団から（できるかぎり一様に）ランダムに抽出されます。これは確率的なプロセスですので、母集団と標本は確率論によってつながっているといえます。

確率モデルとは何か： データとデータの「間」をつなぐもの

- 全数調査のかわりに、部分（限られたデータ）から全体を知るためには、データとデータの間を補間する必要がある
- そのためにはデータの分布に関する仮定が必要になる
 - 仮定 = 確率モデル
- データから確率モデルを推定する
 - より具体的には、モデルパラメータを推定する
- モデルの利用法：
 - モデルを用いて全体の性質を知る
 - 未来のデータについて予測を行う



13

KYOTO UNIVERSITY

母集団を、その一部である標本(データ)から知るということは、何の仮定もなくこれを行うことはできません。全てを観測することはできないのですから。

無限個のデータを持ちうる母集団を、有限のデータしかもたない標本から知るためには、データとデータの間を何らかの方法で補間してやる必要があります。

そのためには、データとデータの間がどのような形になっているかという仮定が必要です。

母集団のデータが全体として、つまり分布としてどのような形をしているかについて我々が設ける仮定を、確率モデルと呼びます。

通常、確率モデルは、ある種の確率分布(たとえば、単純なものだと後述する正規分布など)を仮定します。

これはデータの分析者が、データの性質に関する知識や、これまでの経験、あるいは単純に数学的な扱いやすさなどの様々な理由で決定することが多いです。

確率モデルのタイプを決定したら、あとは、データをもとに、その確率モデルのより具体的な形を決定することになります。

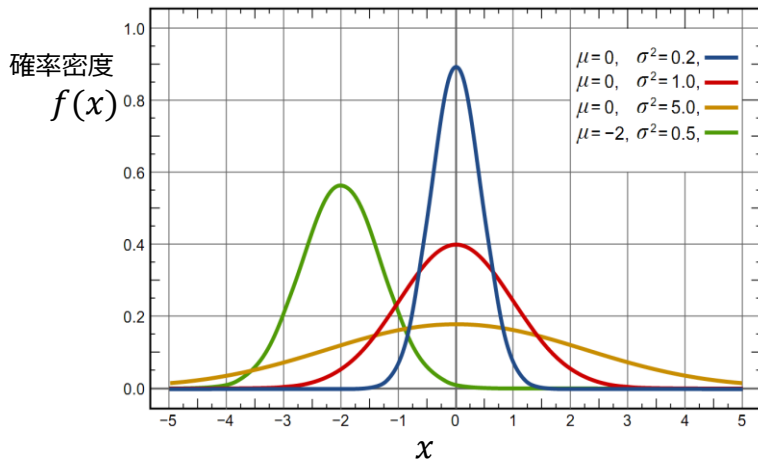
多くの確率モデルには、パラメータと呼ばれる、いくつか「調整ねじ」があり、データに合うようにこれを調整することが、いわゆる統計的な推定にあたります。その具体的な方法については、追々学んでいきます。

ひとたびパラメータを含めてモデルが決定されると、これを調べることで、母集団についての性質を伺い知ることができたり、あるいは、未知のデータに対する予測を行うことができます。

たとえば、前述の、商品の購買行動をモデル化したとすると、標本に含まれない人の購買行動を予測することができます。こういった予測は、応用上、非常に有用です。

代表的な確率モデル： 正規分布

- 量的な確率変数に関する最も基本的な確率分布の一つ
- データは平均値 μ を中心にバラつき度合 σ で散らばる



正規分布の確率密度関数

$$f(x) = N(x|\mu, \sigma^2) \\ = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ただし以下を満たす

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

さきほどチラッと出てきましたが、確率モデルのなかでも、もっとも標準的なものが正規分布と呼ばれるものです。

正規分布は、量的な(連続値をとる)確率変数についてのモデルであり、一次元の場合、2つのパラメータとして平均 μ 、分散(あるいはその平方根である標準偏差 σ)をもつ、釣鐘状の分布です。

平均と分散を変えると、中心の位置や、その周辺でのデータのばらつき(広がり)が変わるのがわかります。

確率モデルとは： データの生成過程

- 母集団は対象となる集合の要素すべて、あるいは、何らかの確率分布に従っていて、標本はそこから確率的に取り出されたと考える
- モデルはデータの生成器として理解できる
 - ボタンを押すとデータが出てくる機械（のようなもの）
- サイコロのモデル：出目 X の確率 $P(X = i) = \frac{1}{6}$
- ある行動をとるかどうかのモデル：
ある人のとる行動 X が a である確率 $P(X = a) = 0.8$
- 多くの場合、個々のデータは同じ分布に従い、独立に生成されると仮定する（= i.i.d: identically & independently distributed）

確率モデルは、データが生成される過程として考えることもできます。

さきほど我々は、我々が本当に興味のある対象である母集団は、対象の要素すべてからなる集合、あるいは、なんらかの確率分布によって表されるしました。

確率分布に従う（確率分布で表される）ということは、標本（データ）がそこから確率的に取り出された＝作り出されたと考えることもできます。

データが「作り出される」のイメージがやや分かりにくいかもしれませんが、例えば、この世の全てを司る神が、世の中のあらゆるもの（たとえば人間）を作りだそうとしている状況を想像してみましょう。

神様といえど、世の中の全ての人をひとりひとり考えて作るのは非常に面倒ですので、そこで「確率モデル」という機械をつくりました。この機械は、ボタンを押すと、性格や体格などがある一定の範囲内でランダムに決定され、人がひとり生み出されます。ボタンを何度も押せば、ひとりひとり違うけれども、それでも全員人間らしい性質をもった人が、たくさん作り出されることとなります。

たとえば、サイコロのモデルは、それぞれの出目がどの確率で出るかを指定する離散分布で表すことが考えられます。理想的なサイコロでは、これは6分の1づつになります。サイコロをふるたびに、この離散分布によって、出目が決定され、我々がそれを観測することになります。

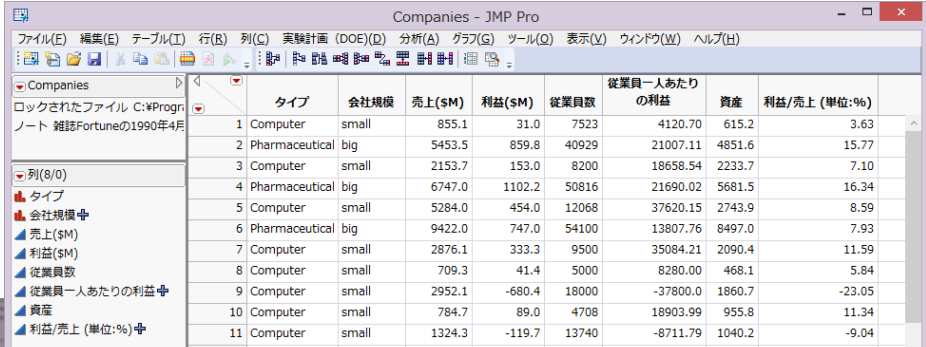
多くの場合、我々は個々のデータは、お互いに独立に、同じ分布に従って生み出されていると仮定します。これは厳密には成り立たないこともありますが、扱いやすさから、多少のことには目をつぶって使用することが多いです。

初等的なデータ分析

ここからは初等的なデータ分析について入っていきます。

基本的なデータの種類： 質的データと量的データ

- 統計データには質的データと量的データがある
 1. 質的データ：
男/女、好き/普通/嫌いなどの記号を値にとるデータ
 2. 量的データ：
温度や身長など数値を値にとるデータ (連続尺度)



The screenshot shows a JMP Pro window titled 'Companies - JMP Pro'. The main data table is displayed with the following columns: タイプ (Type), 会社規模 (Company Size), 売上(\$M) (Sales), 利益(\$M) (Profit), 従業員数 (Number of Employees), 従業員一人あたりの利益 (Profit per Employee), 資産 (Assets), and 利益/売上 (単位:%) (Profit Margin). The data rows are numbered 1 through 11, representing different companies.

	タイプ	会社規模	売上(\$M)	利益(\$M)	従業員数	従業員一人あたりの利益	資産	利益/売上 (単位:%)
1	Computer	small	855.1	31.0	7523	4120.70	615.2	3.63
2	Pharmaceutical	big	5453.5	859.8	40929	21007.11	4851.6	15.77
3	Computer	small	2153.7	153.0	8200	18658.54	2233.7	7.10
4	Pharmaceutical	big	6747.0	1102.2	50816	21690.02	5681.5	16.34
5	Computer	small	5284.0	454.0	12068	37620.15	2743.9	8.59
6	Pharmaceutical	big	9422.0	747.0	54100	13807.76	8497.0	7.93
7	Computer	small	2876.1	333.3	9500	35084.21	2090.4	11.59
8	Computer	small	709.3	41.4	5000	8280.00	468.1	5.84
9	Computer	small	2952.1	-680.4	18000	-37800.0	1860.7	-23.05
10	Computer	small	784.7	89.0	4708	18903.99	955.8	11.34
11	Computer	small	1324.3	-119.7	13740	-8711.79	1040.2	-9.04

さきほど見たように、表形式データは、我々が扱うもっとも典型的なデータです。
さきほど見た企業データでも、業種タイプや、売上など様々な項目がありました。
この項目には大きく分けて2タイプ:質的データと量的データがあります。
質的データとは記号を値にとるデータ、一方で量的データとは数値を値に取るデータです。
ただし、両者の切り分けは、ややあいまい

質的データと量的データの分類： さまざまな尺度

- 質的データ：記号を値としてとるデータ
 - 名義尺度：値が単なるラベルとして扱われる
(例：「男」「女」)
 - 順序尺度：順序に意味がある
(例：「好き」>「普通」>「嫌い」)
- 量的データ：数値を値としてとるデータ (連続尺度)
 - 間隔尺度：数の間隔に意味がある (例：温度)
 - 比例尺度：数の比にも意味がある (例：身長)
 - 原点に意味があるともいえる

質的データ・量的データのそれぞれもさらに細分されます。

質的データのほうは、値にはラベル以上の意味をもたない名義尺度と、順序がある順序尺度に分けられます。

一方、量的データのほうは、間隔に意味のある間隔尺度、比のほうに意味がある比例尺度に分けられます。

量的データの例： 体重データ

- 100名分の体重データ（1次元）：このままだとわかりにくい

No.	体重	No.	体重	No.	体重	No.	体重	No.	体重
1	48	21	52	41	52	61	55	81	54
2	48	22	50	42	57	62	54	82	55
3	40	23	55	43	56	63	55	83	52
4	52	24	53	44	50	64	52	84	49
5	60	25	49	45	49	65	50	85	51
6	55	26	56	46	52	66	50	86	55
7	52	27	52	47	51	67	48	87	50
8	55	28	56	48	45	68	52	88	51
9	53	29	50	49	46	69	52	89	45
10	50	30	52	50	50	70	50	90	56
11	53	31	50	51	49	71	55	91	53
12	62	32	55	52	50	72	50	92	50
13	48	33	50	53	53	73	56	93	53
14	55	34	56	54	58	74	54	94	55
15	45	35	66	55	52	75	48	95	55
16	48	36	49	56	48	76	54	96	51
17	50	37	55	57	65	77	50	97	48
18	50	38	58	58	56	78	49	98	52
19	50	39	48	59	50	79	52	99	63
20	48	40	58	60	60	80	52	100	68

では、具体的に量的データの分析例を見てみましょう。

これは100名分の体重データです。項目としては体重しかない1次元のデータになります。

こうして100名分を並べてみて、なにか思うところがあるかという、ちょっとこれではわかりにくいですね…。

つまり、母集団の形がこれだけからは見えてこないということです。

もう少し整理して、見やすい形にもっていきたいところです…。

量子化： 量的データを理解しやすくするための量子化

- 生データのままでデータを理解するのは困難
- 量子化：データがとりうる値の範囲を、あらかじめ定めた区間（階級）に分け、観測される数値の入る階級によって集計を行う
 - 観測される数値が実数（連続値）の場合には、厳密な値は表現できないので必ず量子化を行う
 - CDに録音されている音響信号も16 [bits]で量子化、各時刻の振幅は0～65535の整数で表現
 - 例：体重の場合
 - 観測する最小単位を1kgとし最小単位より小さい端数を丸める
 - あるいは、5kgずつの区間に分け、それぞれの区間で集計する

生の数値データを並べただけではわかりにくいので、整理しましょう。

ここで行うのが量子化と呼ばれる作業です。

なんだか量子力学的なスゴイ何かをやるのかと思いきや、要するに、

0以上10未満

10以上20未満

...

といったように区間を切って、それぞれの区間にデータが何個ずつ入るかを調べるだけです...

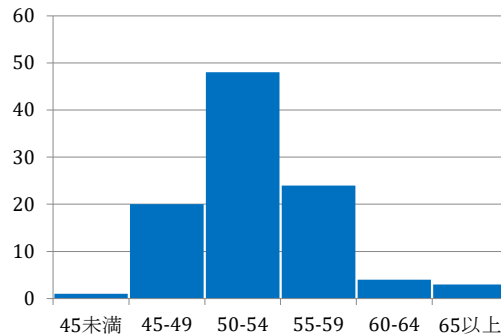
量的データの集計： 度数分布表とヒストグラム

- ヒストグラムでデータ分布を視覚化
 - 度数分布表：各階級の度数をカウント
 - ヒストグラム：度数分布のグラフ表現

度数分布表（階級幅5kg）

階級	度数
45未満	1
45～49	20
50～54	48
55～59	24
60～64	4
65以上	3

ヒストグラム

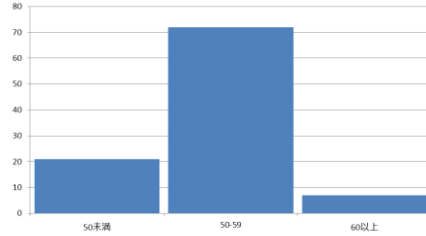
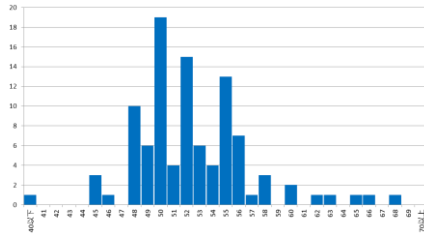


量子化したデータを表にまとめたものが度数分布表です。
これでどの範囲にどのくらいのデータがあるのか大分わかりやすい。
おおまかにですが、母集団の形が見えてきました。

これを視覚的に表したのがヒストグラムです。
分布の形が一目瞭然で非常にわかりやすいです。

ヒストグラムと階級幅の関係： ヒストグラムでは幅の決め方で見た目が大きく変わる

- 階級幅1の場合と10の場合でヒストグラムの形が変わる



- スタージェス (Sturges) の方法： $K = \log_2 N + 1$

- データが100個： $\log_2 100 + 1 = 7.643856 \rightarrow 8$ 階級ぐらい
- データが50個： $\log_2 50 + 1 = 6.643856 \rightarrow 7$ 階級ぐらい
- データが25個： $\log_2 25 + 1 = 5.643856 \rightarrow 6$ 階級ぐらい

ヒストグラムは大変見易い表現ですが、実は量子化の幅(区間の切り方)によって形が大きく変わるといふ欠点があります。

左の図は量子化の幅(階級幅)を非常に細かくした場合、かなりガタガタしています。

一方、右の図のように量子化幅を大きくしすぎると、これもちょっと大雑把すぎますね。

階級幅の決め方には色々なやり方がありますが、ひとつ簡単な経験的方法として、スタージェスの方法と呼ばれる決め方があります。

ここでNはデータの数、Kは区間の数です。データの数を決めると、区間の数が決まります。

そのほかの集計：

度数・累積度数・相対度数・累積相対度数

- データ： $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ をいくつかの階級： $I_1, I_2, I_3, \dots, I_K$ に分割する

- 度数： $f_1, f_2, f_3, \dots, f_K$

– $x_i \in I_k$ を満たす i の個数

– 累積度数： $F_k = \sum_{i=1}^k f_k$

– 相対度数： $\frac{f_k}{N}$

– 相対累積度数： $\frac{F_k}{N}$

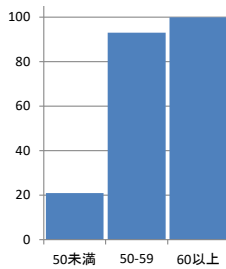
階級	度数	累積度数	相対度数	累積相対度数
45未満	1	1	1%	1%
45-49	20	21	20%	21%
50-54	48	69	48%	69%
55-59	24	93	24%	93%
60-64	4	97	4%	97%
65以上	3	100	3%	100%

度数分布表の他にも、各区間に入るデータ数を直接見るのではなく、累積で見る累積度数という見方もあります(そのメリットは次頁で)。

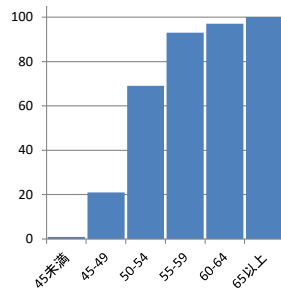
また、全体のデータ数で正規化した、相対度数・相対累積度数なども使われます。

累積度数と階級幅の関係： 累積度数は階級幅にそれほど左右されない

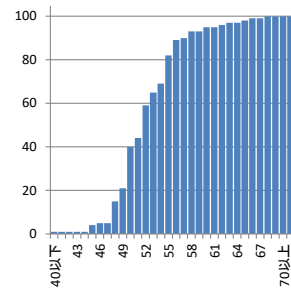
- 累積度数は階級幅にそれほど左右されない
 - むしろ階級幅が小さいほうが分布の様子がよくわかるくらい…



階級幅10kg



階級幅5kg



階級幅1kg

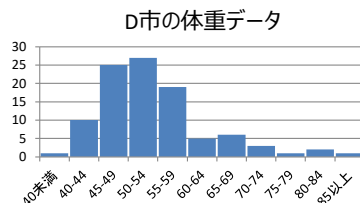
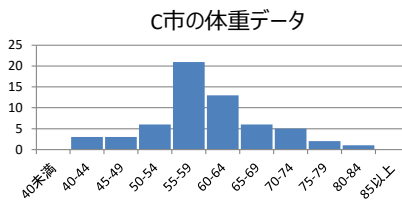
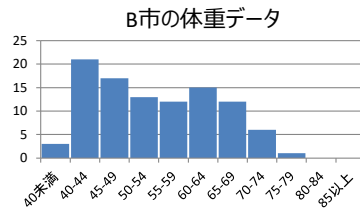
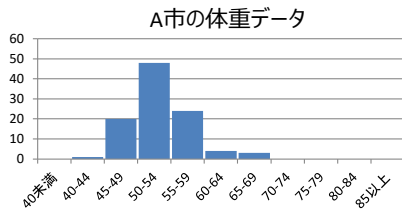
累積度数を考えるひとつのメリットは、これが階級幅に左右されにくいという点です。

上記の図は、階級幅を変えながら、累積度数を図示したのですが、度数分布の場合と異なり、大きな質の差が見られません。

むしろ、階級幅を小さくしていくと、より形がはっきりしていくようにも見えます。

複数種類のデータを比較したい場合： ヒストグラムの形を表す指標がほしい

- ヒストグラムから分布の形状はよくわかるが、一覧性には欠ける
- ヒストグラムの特徴を表す少数の指標で代表したい



ヒストグラムを見ると、データの分布が視覚的にとらえられることがわかりました。
データを入手したら、まずヒストグラムをみるというのがデータ分析の第一歩といえます。

さて、さきほどの体重データの例をさらに進めて、今度は4つの市でそれぞれ集められた体重データが手に入ったとします。

これまでの考え方でいくと、まずは4つのデータそれぞれにヒストグラムを書いてこれらと比較することで、それぞれの市の傾向が見えてくるのが期待できます。

では、もっと市の数が増えたら?? 数個のヒストグラムはまだ一覧性がありますが、これが増えていくと、ヒストグラムでもわかりにくいという状況になるでしょう。

つまり、より多くの種類のデータをみるには、ヒストグラムではまだ代表性に欠けると。もっともっとデータを端的に表す指標が欲しいものです。

データの代表値： 標本平均・中央値

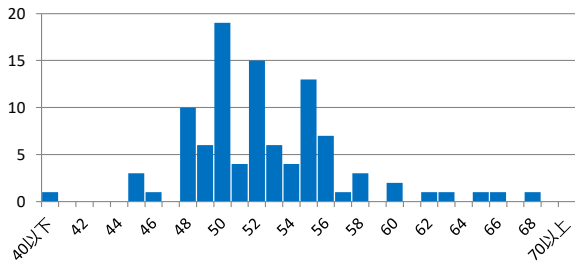
■ データ $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ の特徴を表す数値

– 標本平均： $\bar{x} = \frac{1}{N}(x^{(1)} + x^{(2)} + \dots + x^{(n)})$

- $\operatorname{argmin}_x f(x) = (x^{(1)} - x)^2 + (x^{(2)} - x)^2 + \dots + (x^{(n)} - x)^2$

– 中央値 (median)：大きいほうからだいたい $\frac{n}{2}$ 番目の値

- 外れ値の影響を受けにくい



そこで出てくるのが、標本平均です。

ヒストグラムでは分布の「形」をみましたが、これをもっとシンプルに、データの「真ん中」だけ見ようというものです。

標本平均と同じく、データの「真ん中」をみるのが中央値です。

中央値は、標本平均に比べるとやや計算が面倒ですが、外れ値(たとえば、入力ミスで体重5000kgや、-50kgなどのデータ)がデータに紛れていた場合に、その影響を受けにくいといった利点もあります。

中央値は、厳密にはデータの総数 N が奇数個なら小さい方から $\frac{N+1}{2}$ 番目の値 偶数個なら小さい方から $\frac{N}{2}$ 番目の値と $\frac{N}{2} + 1$ 番目の値の平均

データ分布の代表値： 分散・四分位点・箱ひげ図

- 平均だけでは不十分な場合もある

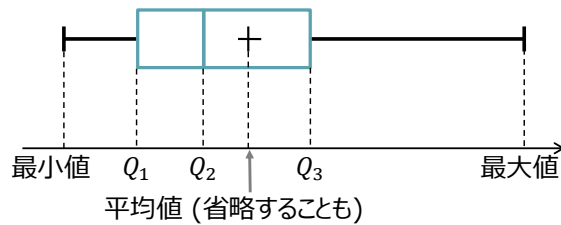
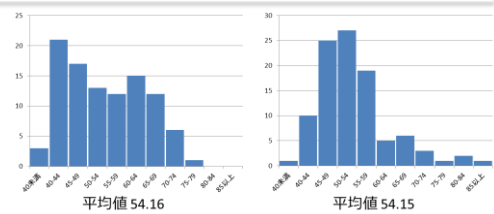
- 分布の形も知りたい

– データのばらつき：分散

– 4分位点：整列したデータを四等分する位置にある値

- Q_1 ：25%点、 Q_2 ：50%点（中央値）、 Q_3 ：75%点、

- 箱ひげ図による可視化



27

KYOTO UNIVERSITY

標本平均・中央値のように分布の「真ん中」に関する情報だけだと、一気に代表値がひとつになってしまうので、分布の「形」についての情報は失われてしまいます。

そこで、もう少し情報を加えるとすると、おそらく「真ん中」のまわりはどう分布しているかという情報が役に立ちそうです。

これを表すのが分散です。

あるいは、中央値が上(下)から50%のところを示すのに対して、25%、75%などの点もみよというものが、四分位点です。

これらをまとめて、箱ひげ図として視覚化することもあります。

不偏分散： データのばらつきをあらわす

- 不偏分散 $\hat{\sigma}^2$: データのばらつきを表す

$$-\hat{\sigma}^2 = \frac{(x^{(1)} - \bar{x})^2 + (x^{(2)} - \bar{x})^2 + \dots + (x^{(n)} - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{x})^2$$

– 平均と分散でデータを捉える = 背後に正規分布を仮定

- ばらつきを表す類似の指標：

– 変動係数CV (coefficient of variation) $\frac{\hat{\sigma}^2}{\bar{x}}$

- 相対標準偏差 (relative standard deviation: RSD) と呼ばれる
- 平均値が異なる二つの集団のばらつきを比較するのに用いる

– 偏差値 T_i : $x^{(i)}$ を平均値50・標準偏差10となるようにスケールした値

さきほど、データのばらつきを表す指標として分散を挙げましたが、実際にデータから計算するのは不偏分散と呼ばれる指標がよく使用されます。

$n-1$ で割っているところが気になるかもしれませんが、いずれまた説明することになると思います

不偏分散のスケールを調整したのも、たまに使われます。

練習問題： ストリームデータの平均・分散の計算

- ストリームデータ：時々刻々到着するデータ
 - 時刻 t においてデータ $x^{(t)}$ が観測される
 - 例：センサーデータ
- これまでに観測されたデータの平均・分散を、各時刻で $O(1)$ で保持したい
 - 定義に従って素朴に計算すると $O(t)$

標本平均・不偏分散を実際に計算する問題を考えてみてください。
時々刻々データが到着するような状況は応用上もよく現れる重要な問題です。

まとめ：

統計的モデル化の導入と量的データの初等的分析

- 観測されたデータを理解し、予測をおこなうためには、データの背後でデータを生み出す確率モデルを考える
- モデルをデータから推定する必要がある
- データには量的データ、質的データがある
- 量的データの初等的分析には、ヒストグラム等を用いて可視化したり、平均・分散などの指標でとらえる
- 次回以降：2変数の関係の分析（相関・回帰）

