# 科学統計教程

Idiot

2022 年 7 月 8 日

## 0.1 前書き

本書の特徴
注意点

1. 統計検定のみで帰無仮説を含むモデルを採択または棄却する
2. $N = 30$ であれば中心極限定理よりある特定の検定が使える
3. データの出現頻度が正規分布によりよく近似できる場合のみを考える。または、標本分布が正規分布であることを前提とする。

これらの魔術を使わずに統計学を使う方法および、推測可能なことについて考える。

1. 仮説検定で推測可能な事象 (type I,II error)
2. モデルと現象の乖離により生じる大きな間違えを含む推定
3. モデルの母数が現象を捉えていないことにより生じる事象

孫引き引用をした箇所は孫引きしたと書いておいた。今後読んで、引用に修正することもある。
一般的な数理統計学の教科書 [1, 2, 3, 4, 5, 6, 7, 8, 9]、統計モデルについては [10]。生物学者が統計学を使うときの視点は [11] に詳しいが、中心極限定理の説明が十分ではないと感じる。

# 第 1 章

# 科学的推論

科学におけるモデルおよび、数理統計学におけるモデルについて説明し、その違いを明らかにする。

## 1.1　モデル

モデル (模型) とは、現実を表していると思わせるような、作られたものであり、次の特徴を備えています。

1. モデルは本物の特徴の一部を推測可能。本物との乖離の程度も推測できる
2. (1) を行うために、複数の仮定により構築される。また、それらの仮定は、現状の知識では明らかではないまたは、現実的には成立していないことがある。
3. モデルは間違った推測をする。

例えば、車のプラモデルはモデルの一つです。本物の特徴の一つである大きさを推測可能にするため、スケール（例えば、1/24 など）を決めて作られている。ドアや車体の幅を計測し、スケール倍すれば、本物の大きさを推測できる。普段長さを測れない場所であっても、手のひらに収まるプラモデルであれば、どの部分でも推測が可能になる。言い換えれば、本物の車がなくても、スケールを維持した車のプラモデルを持っていれば、簡単に大きさに関する推測が可能になる。
本物の車を持って来れば、本物の様子を推測することが可能であるので、本物の車は、車自身のモデルということができるが、車を車自身のモデルとすると、それまであった利便性が損なわれる。おおよその車体の長さが知りたいのに、わざわざ長い測りが必要になることや、手に持って観察することもできない。このように、細部まで推測可能にするというのは、デメリットになることがあり、モデルとして利用することはない。
細部まで推測可能なモデルは使うことは稀であり、車のモデルとして、大きさの尺度を保っていない直方体のブロックを使うことがある。このモデルでも推測できることがある。３台の同じ車を縦列駐車するのに必要な長さなどは、直方体三つ分と推測が可能である。モデルの作り込みの程度によって推測できることは様々である。

真球を車のモデルし、車の大きさに関する推測を行うと、現実の大きさと推測は大きく乖離することが考えられる。モデルが本物の推測に使えないということに判断を下すには、本物のデータとモデルの出す推測を複数の指標から比較し考察することになる。
モデルは本物ではないが、推測に役にたつ物として利用する。モデルと本物が極めて一致するように感じられることもあると思うが、モデルは本物ではない。

## 1.2　統計モデル

統計モデルについて説明し、モデルを使って現実を推測することを概念図を用いて説明する。まず、統計モデルは、数理統計の知識を使いモデルを構築され、現実を推測するために用いられる。簡単な統計モデルを例に挙げると、次のような仮説から構築される。

1. (仮定 1) 確率変数が同一の分布から独立に得られる (i.i.d)
2. (仮定 2) その分布関数は、$f(x)$ と書ける。
3. (仮定 3) 分布関数の母数に関する仮説[*1]

統計モデルにより推定したい対象またはデータが、統計モデルの仮定から外れていることは多々ある。まず仮定 1、独立性と同一の分布という仮定は、数学的厳密な定義がある。科学におけるその意味を考えることで、その仮定が現実の世界において当てはまらないことがわかる。まず、各変数が独立とは、得られたデータに相関が全くないこと考えられ、それは科学においては当てはまっていないことの方が多い。例えば、人の身長を計測器により繰り返し観測すると、その計測器や扱う人の癖がデータに含まれ、それはデータの傾向を決定する因子となり、相関があると考えられる。もし、相関がない計測状況が設定できたとしても、人の身長はその背景にある社会や遺伝的な繋がりが因子となっており、相関が無いと言い切ることは難しい。同一の分布とは、同一の数学的規則に自然が支配されていることを仮定していると考えられ、サイコロやコインのトスではそのように考えても矛盾しない。人の身長は、母父の大きさや成長過程における栄養の量などの因子によって成長すると考えることは科学的ではあるが、サイコロをふって決定されていると考えるのは妥当とは言い切れない[*2]。仮定が科学において妥当とは言い切れないモデルを使って推測を行うことになる。
以下では、統計モデルを $M(\boldsymbol{a})$ とし、ここで $\boldsymbol{a}$ は、仮定 3 の統計モデルの母数であり、母数が複数あることも考慮し、ベクトルで表記しておく。

**■学問間に生じているモデルに関する認識の違い**

モデルが本物であるか否かは、学問領域によって認識が異なっている。生物物理学の視点では、モデルは現実を推測するための偽物のことだと考えていることの方が

[*1] 三番目の仮説のみを統計モデルと主張する流派もある [12]
[*2] そう考えてもいい

多い。モデルが自分の知りたいことをうまく予測してくれさえいればいいという立場である。一方で、数学では、モデルを現実と捉える傾向がある。モデルにより世界が支配されていると考えているのである。例えば、ある数学者は、流体モデルに解が安定的に存在するかがわからないから飛行機に乗りたくないと思っていると言う雰囲気がある。現代の統計学はどちらかと言うと数学者が作った枠組みを統計ユーザーが受け入れてしまったため、ユーザーたちは、数学者のように世界を捉えようとしているように見える。

**■なぜ正規分布を仮定できるのか**

数理統計学の本には、正規分布を前提にして書かれていることが多々あることから、科学において統計を利用するには、その前提が満たされる必要があるという考えがある。私も以前はそのように考えており、同様の考えにハマってしまう人は少なくない。

**Katsushi Kagaya:**
学生のころ先生とデータについて議論していて（生物学分野です）「そもそもなぜ正規分布が仮定できるのか…」とおっしゃって二人でしばらく固まったことを思い出します。実現可能性の考え方から学ぶのが良いのかなと思います

https://twitter.com/katzkagaya/status/1209656621523058691

学問の世界において、分布関数に関する仮定が可能な理由についての認識は様々である。数学においては、仮定をして結論を導くことはよくある。数学から離れた科学の領域では、仮定することに対して妥当性や客観的であること要求していることもある。この考えに反して、私は、恣意的に考えたモデルを使って推測をしてみるという考えに基づいて、統計モデルを構築し、自然について推測を行うと考えている。

### 1.2.1　数理モデルの機能

数理モデルには予測・サンプリングという機能がある。それぞれ説明していく。

**■予測**　次に説明するサンプリングを使うことで出現しやすい場所を数値的に計算することが必要となるモデルもある。

**■サンプリング**　サンプリングは、モデルを使ってデータを生成する方法である。モデルが説明したいデータの出現頻度をよく予測できるなら、モデルが生成したデータは実際に得られるデータと似たものになる。

## 1.3　数理統計学におけるモデル

数理統計学は、モデルが生成した有限個の確率変数からモデルの母数を推測する方法論を提供している。

## 1.4　統計学の用語

### 1.4.1　母集団、無作為抽出、サンプリング

母集団は興味のある対象全体の集団のことである。例えば、17 歳男性の身長に関心があるならば、17 歳男性の全員の集合が母集団である。無作為抽出とは、偏りなく母集団からデータを取得することである。無作為抽出することで、都合の良い結果が集まらないようにしている[*3]。サンプリングは無作為抽出の英訳である。本書では、モデルからの無作為抽出のことをあえてサンプリングとカタカナで記述し、現実の作業である無作為抽出と区別をする。この使い分けは一般的でない。

### 1.4.2　標本、サンプルサイズ、標本数

**定義 1.4.1.** 母集団から無作為抽出して得た標本に含まれるデータの個数をサンプルサイズ（標本の大きさ）といい、その数を $T$ や $n$ で表す。同じ実験を繰り返して行ない、複数の標本を作ると、その標本の個数を標本数という。モデルからサンプリングした場合も、その確率変数の集まりを標本という。モデルの標本において、標本の大きさが大きいものを大標本、小さいものを小標本と言う。

例えば、無作為抽出しデータを 20 個得る実験を 30 回繰り返した場合、サンプルサイズ 20 の標本を 30 得たことになります。言い換えれば、標本数 30 で、サンプルサイズは 20 であると言う。
サンプルサイズを標本数と言う流儀の統計学もあるようなので注意が必要である。[*4]

## 1.5　モデルを使った推測

[*3] 無作為抽出しなければならないのはモデルの仮定 1 を満たすためだという主張を見かけたことがある。モデルの仮定と現実を対応づけることはほとんど無理なので、そのように考えなくても良い。文献を探すべき

[*4] 業界によって様々な慣習があるので、業界の慣習に（師匠の言うことに）従った方がいいようにも思う (`https://www.jil.go.jp/column/bn/colum005.html`)。方言により定義が異なることがある文献がまとめられている (`https://biolab.sakura.ne.jp/sample-size.html`)。

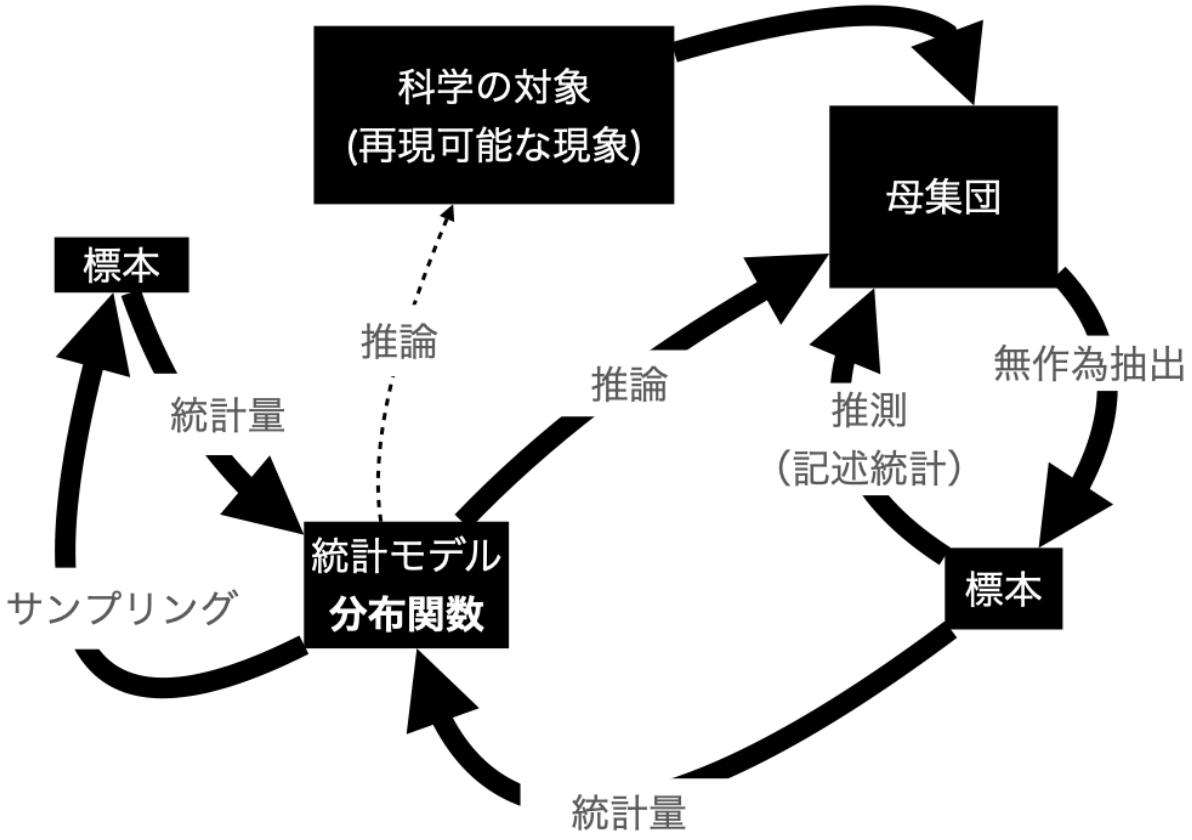


図 1.1　標準正規分布 (a) 確率密度関数 (b) 累積度数分布 (c)1-累積度数分布

# 参考文献


[1] 白石高章. 統計科学の基礎: データと確率の結びつきがよくわかる数理. 日本評論社, 2012.

[2] 宮岡 悦良野田 一雄. 入門・演習　数理統計. 共立出版, 1990/05/01.

[3] 確率・統計入門. 岩波書店, 1973.

[4] 数理統計学: データ解析の方法. 東洋経済新報社, 1963.

[5] 統計的機械学習: 生成モデルに基づくパターン認識. Tokyo tech be-text. オーム社, 2009.

[6] 確率と統計: 情報学への架橋. コロナ社, 2005.

[7] 統計学. 東京大学出版会, 2016.

[8] 現代数理統計学の基礎. 共立講座数学の魅力. 共立出版, 2017.

[9] 現代数理統計学. 学術図書出版社, 2020.

[10] データ解析のための統計モデリング入門: 一般化線形モデル・階層ベイズモデル・MCMC. 確率と情報の科学 / 甘利俊一, 麻生英樹, 伊庭幸人編. 岩波書店, 2012.

[11] 統計思考の世界: 曼荼羅で読み解くデータ解析の基礎. 技術評論社, 2018.

[12] 塩見正衛. 仮説検定と¡i¿p¡/i¿値問題：草地学・農学における統計的手法の正しい利用のために. 日本草地学会誌, Vol. 66, No. 4, pp. 209–215, 2021.