統計教程

— モデルによる予測 —

Idiot

2023 年 1 月 8 日

目次

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車の特徴に関して推測できることの種類が決まる。

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。
軽自動車に対してその大きさを予測可能なモデルを使って、トラックの大きさを予測できる。予測できるが、その予測値は実際のトラックの大きさと異なる。モデルが車体長を 3.4m であると予測される。実際のトラックの車体長は 6m よりも大きい。メートル単位でモデルと実際には差が生じる。このように、モデルと実際を調べることで、このモデルではトラックの大きさを推測できないと判定できる。実際には、どれくらいの誤差が生じたときに、モデルが使えないというのかは、予測したいことにより異なる。
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]

1.2.1　統計モデルとデータ

データに統計モデルがよく当てはまるよう指標を定め、その指標を小さくするようにモデルの母数を推定できる。

1.2.2　データへの過剰適合

モデルは改訂することにより、予測の精度をあげることができた。これは、何度も対象を観測することで、モデルと実際の当てはまりを定量的な評価が可能であるから、モデルの作り込みを防ぐことができる。再現性の確保されている現象に対しては、データに当てはまるようにモデルに仮定を足していき、モデルの作り込みを行う。さらに新たなデータとモデルの予測とを定量的な指標を元に評価する。一方で、何度も繰り返し観測可能でない現象を対象にした学問領域において、モデルの作り込みは現在得られているデータを過度によく予測するモデルとなることがある。その結果、構築したモデルが新に得たデータに対して予測精度が落ちてしまうことが多々ある。そのため、データを見た後に、モデルに

[*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 [1]

仮定の追加または変更はしない方が良い。

1.2.3 統計モデルの仮定を自然が満たしているのか

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。その定義を現実の世界にの言葉に変換すること自体が難しい。まず、各変数が独立とは、事象 A,B が同時に起きた確立 $P(A,B)$ がそれぞれが生じる確率の積に等しいということであり、$P(A,B) = P(A)P(B)$ である。そもそも事象生じる頻度が P により決定されているということを考えることができない。それに加えて $P(A,B) = P(A)P(B)$ なども、現実世界の事象に一致する概念がない。

間違っていることを承知の上で、科学的な言葉に変換して、妥当であるかを考察してみる。あえて、得られたデータに相関が全くないと、捉えてみると、現実的には妥当では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、データ間には相関があると考えられる。もし、相関がない実験デザインを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る。コインのトスでは、その裏表の出現確率を二項分布によるものと考えても問題が大きくならない。一方で、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られる。この現象が、サイコロのように乱数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]。

統計モデルを現実の推測に使えないということではない。モデルと現象を比べて予測するためだけにモデルを利用するのであるから、仮定が現実に存在するかはどうでも良い。

■有用な近似が得られるからモデルを使う

Box らは、統計学において正規分布や一次関数で推論することを次のように捉えている [2]。

> Equally, the statistician knows, for example, that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world.

[*2] そう考えてもいいけど、あまり役に立たない

■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。私は、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。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。
実際にモデルに対する認識が研究者によって異なっていると感じている人はいる。学習理論を研究しておられる渡邊 澄夫さんは、情報科学と物理学におけるモデルとして次のような見解を述べている。

> （注意）このことを聞いたとき，どのように感じるかは，人によって ずいぶん違います．情報科学の研究者の人たちは，「目的が違うのだから， 最適なものが違うのは当然であり，まったく不思議ではない」と感じる場合が多いようです． 一方，物理学の研究者の人たちは， 「真の法則が発見できるということと，最良の予測ができることとは，ぴったりと 同じであるべきである」と感じるようです．これは，おそらく，「モデル」という 概念や重みにおいて，情報科学と物理学では大きな隔たりがあることが原因ではないかと 思います．（例題：電子の質量が正確に予言できるのは，量子電磁力学が真の自然法則であるからと 考えられています)．
> 生物学・環境学・経済学に用いられる「モデル」は，上記の意味での情報科学におけるモデルに近いのか， 物理学における理論に近いのか，それとも，その中間に当たるのか，もっと違う種類のものなのか，は，かなり微妙な質問で，一様な回答はないものと思います． 数学者のかたが数理科学の研究に挑もうとされるときには，「モデル」という言葉が表すものが，分野において，場合において，このように様々に異なりうることを認識されておかれるとよろしいでしょう。
> `http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/Bayestheory2.html`

統計や機械学習の分野で有名な Box 氏は、「全てのモデルは間違いである (All models are wrong)」と、次のように説明している。

> For such a model there is no need to ask the question ”Is the model true?”. If ”truth” is to be the ”whole truth” the answer must be ”No”. The only question of interest is ”Is the model illuminating and useful?”.

1.2.4 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。

■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

統計学の言葉をいくつか借りて、本来の意味とは異なった定義で使う。

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。日本人全体の身長に関心があるならば、日本人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]。本書では、モデルから確率変数を生成することを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する[*4]。

1.4.2 誤差・揺らぎ

計測上の手順で生じるデータの差異の平均と各データの差分のことを誤差と呼ぶ。誤差が生じるのは測定者の違いや、計測装置の精度に依存する。
揺らぎとは、ある集団における個体間の差異である。例えば、ある畑で採集された野菜の

[*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 (文献を探すべき)。モデルの仮定を現実が満たすようにすることはできないので、このように考えない方が良い。行き着く先はモデルの仮定を満たすように、検定を繰り返すようになる。もちろん検定ではモデルの仮定を満たしているかを決定することはできない。
[*4] この使い分けは一般的でないし適切ではない。

重量の個体間の差異を揺らぎと呼ぶ。
本書では誤差は、揺らぎよりも十分小さいものとして扱い、揺らぎの性質についてモデルを構築する。

1.4.3　標本、サンプルサイズ、擬似反復、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 得たことになる。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。
擬似反復は、同じ個体においてその特徴を複数回計測し、これを揺らぎとして集団の差異として捉えることである。例えば、17 歳男性の身長について計測することを計画する。サンプルサイズとして、100 個のデータ点から計測することにしたので、10 人から 10 回身長を計測した。結果、100 個の計測データが集まった。このデータでは、通常の 17 歳男性の身長に関する統計モデルと乖離していると結論がつけられやすくなる。
サンプルサイズを標本数と言う流儀の学問もあるようなので注意が必要である[*5]。

1.5　モデルを使った推測

d

[*5] 業界によって様々な慣習があり (https://biolab.sakura.ne.jp/sample-size.html)、業界の慣習に（師匠の言うことに）従った方が余計なトラブルを減らせると考えられる (https://www.jil.go.jp/column/bn/colum005.html)。この言葉くらいは統一して記述したい。本書でも途中で間違った使い方をしてしまうかもしれないが、なるべく間違わないようにしたい

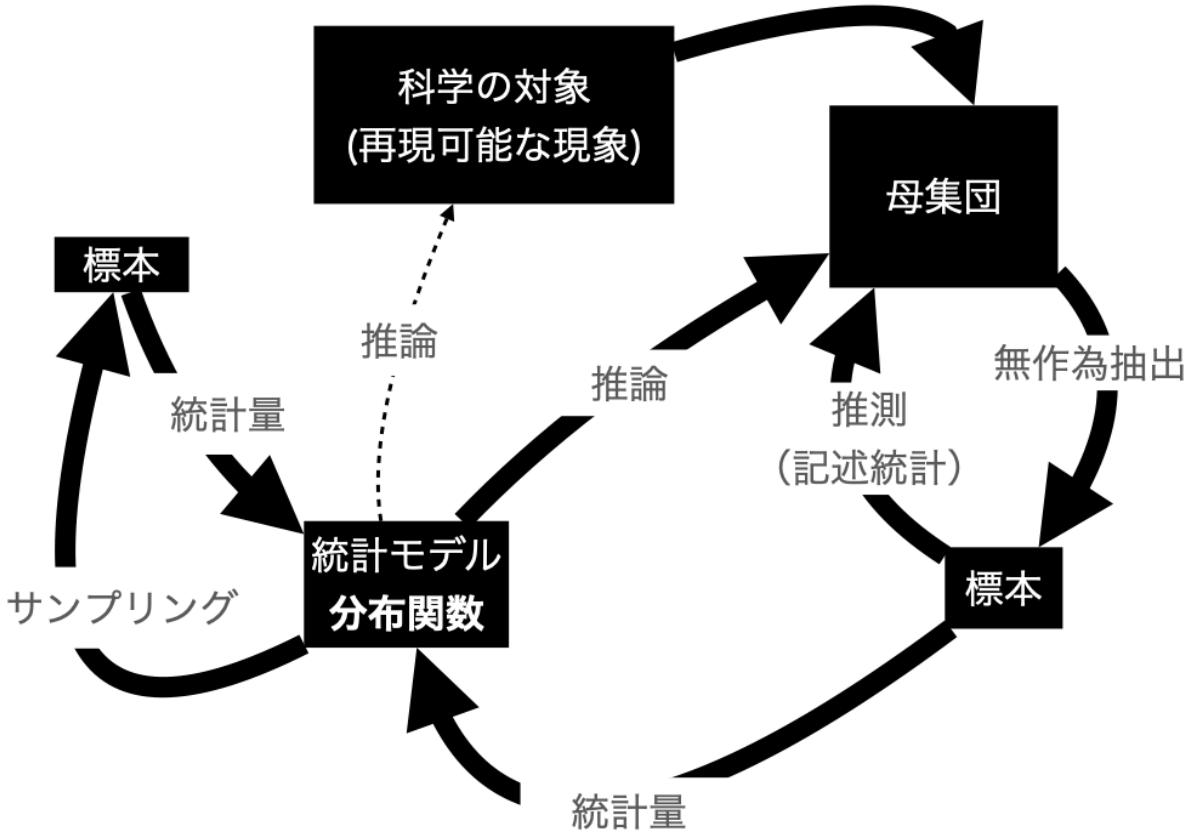


図 1.1　統計モデルによる現象の推測に関する概念図

第 2 章

取り扱うデータの条件

科学的に事象を取り扱うための本書で扱うデータの条件は以下の通りである。これらが無いならば、本書で扱える範囲を超えている。統計学者に相談した方が良い[*1]。

- 再現性 同じような条件であれば、同じような現象が生じるということである。
- 計測誤差 計測により生じた誤差（測定誤差）は揺らぎ（集団内での差）に比べて十分に小さい
- 無作為抽出　なるべく偏りなく母集団からデータを取得する
- 実験デザイン バイアスを小さくなるように計画を行う。
- 予測　データを集めると、データとモデルの予測に関して相違点が明らかになり、モデルの改訂が必要になる。この改訂に終わりはない。

2.1　実験デザイン

わたしが扱える範囲ではないので、他書を読んだ方が良い。今後まとめたい。

2.2　無作為抽出されていない事による過誤

対象を無作為に抽出できていない標本から、統計量を計算し、モデルの母数を推定したとする。このモデルでは、本来設定した母集団に関する予測には誤りが多くなる。例えば、17 歳の日本人男性の身長を母集団に指定したのに、17 歳のバスケ部部員の身長を計測する。その標本を元に、モデルの母数を推定し、母集団に関する推測を行う。すると、その予測は母集団に関して十分なものではなくなる。例えば、平均が大きくなりすぎたり、平均よりも小さな人の割合が予測と異なることが生じる。
やってはいけないとは言い切れないが、偏った集団を計測してしまった場合、その解釈に一工夫が必要になる。

[*1] 最初から統計学者に相談した方が良い

2.3　Questionable Research Practice(QRP)

以下ではやってはいけないことを紹介する。国立研究開発法人　日本医療研究開発機構が出版している研究公正に関するヒヤリ・ハット集の「7 研究データの信頼性、再現性等 」に詳しくまとめられている[*2]。

2.3.1　後付けの母集団かつ $p < \alpha$ を満たす集団

母集団 A を設定し、標本を抽出したものを標本 a とする。標本 a のデータはさまざまな要素から構成されているとする。例えば、ある会社に所属する人の、身長や年収、税金の支払い履歴、ローン残高、労働部署、高校時代の部活などであるとする。この標本から、何らかの属性 A' に当てはまるデータ b を抽出したとする。データ b について特定の統計モデルとの乖離するかを調べ、乖離していることをが判明したとする（乖離を定量的に調べる方法はなんでもいいが、$p < \alpha$ だったと考えても良い)。この結果から、属性 A' に関わると考えられる母集団 A' を再構成する。そこから、母集団 A' を特定のモデルで予測できないと結論づけることはできない。

まず、今集めた標本 a は、母集団 A から集めたものであり、母集団 A' から集めたものではない。よって、母集団 A' から無作為抽出できていない。また、標本 a を無作為抽出したときに付随して得た、母集団 A' の一部の偏った集団のデータである。以上から、母集団 A' に関する無作為抽出とはいえない。図 2.1 には、概念図を示しておいた。

後付けの母集団かつ $p < \alpha$ を満たす集団

$p < \alpha$ であるという標本がデータから発見されたので、標本の特性を持つと思われる母集団を後付けし、その母集団から無作為抽出を行なったことにし、ある統計モデルとデータが乖離していたと言うストーリーを作ったとする。言い換えれば、後付けの母集団ならば、$p < \alpha$ であるという論理を立てたとする。実際には、後付けの母集団でありかつ $p < \alpha$ という集団から作為抽出しているので[*3]、本来の母集団については何もわからない。言い換えれば、母集団に関する拡大解釈が行われたことで、母集団に関しては何もわからないのに、推測を行なったと主張している[*4]。母集団の特徴を知るには、無作為抽出を行い、推測を行う必要がある。

このような母集団に関する拡大解釈を仮説ハッキング ($HARKing$(Hypothesizing After the Results are Known)) といい、この操作により得たデータと仮説について、仮説が元

[*2] https://www.amed.go.jp/content/000064531.pdf

[*3] この場合でも無作為抽出できていると誤解してしまうが、後付けの母集団から無作為抽出できていない！

[*4] 実際調査した母集団は母集団かつ $p < 0.05$ に対して、報告した母集団デカすぎんだろ...

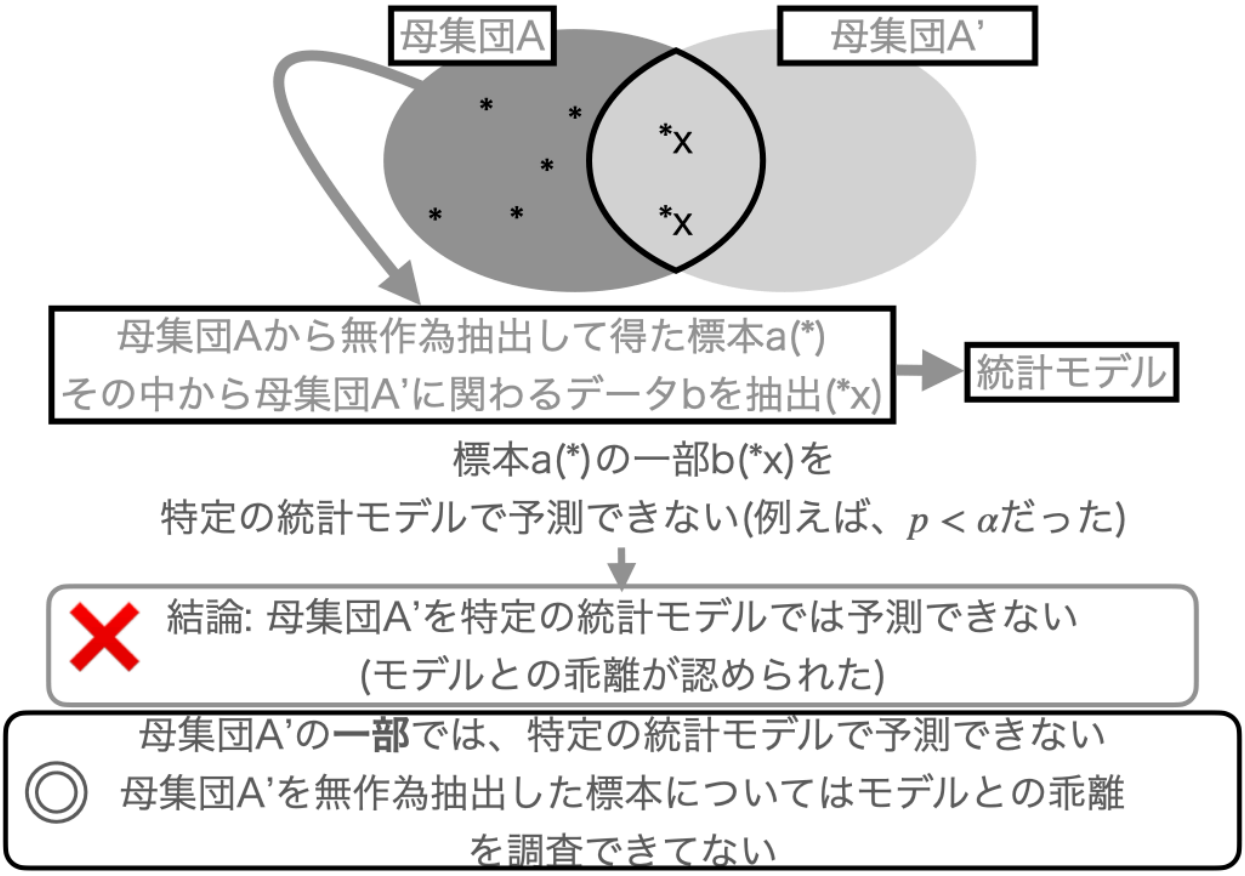


図 2.1 仮説ハッキングの概念図

からあったことにして、報告を行うと、研究不正となる[*5] [*6][*7][*8][*9]。

[*5] HARKing は、再現性の問題という意見もある。https://twitter.com/ykamit/status/1077716200845500416 。この意見に私は同意する。母集団を無作為抽出していないことで、再現できないことが増えると考えられる。

[*6] 多重検定により、p 値が低く推測されることが問題であるというものもある [3, 4]。部分的には同意できるが、私は十分理解できなかった。

[*7] Twitter でのアンケートでは、多くの人が HARking をうまく理解できてないという Twitter でのアンケートもある。https://twitter.com/biomedcircus/status/1088957697368690689

[*8] 探索的なデータ解析においては、帰無仮説の後付けが許されるという主張もある。この意見には同意できない。母集団について拡大解釈をすることは許されない。探索的データ解析により得られるのは、母集団かつ $p < \alpha$ という集団が見つかったということのみ主張できる。これを元に、母集団に関する性質を言及してしまうのはおかしい。

[*9] HARKing については、[5] に詳しくまとめられている

■ HARKing

Yuki Kamitani:
データを操作して p 値をいじる行為を不正と認識している人は多いが、HARKing が不正と思っている人は非常に少ない。私の周辺分野のシニア研究者で理解している人はほぼ皆無（問題を指摘すると一笑に付される）。研究の実践と論文フォーマットの齟齬やフェアプレー精神の問題（？）と理解している人がいた

https://twitter.com/ykamit/status/1077715969827528705

HARking を理解するのは、難しい。無作為抽出したデータから、データを調べた後に、母集団を構成しているのだから、無作為抽出できていると考えてしまいがちになる。

2.3.2　$p < \alpha$ になったら無作為抽出を終える

p 値がある値を下回ったときに、実験を終了するという操作を行なったとする。統計モデルの予測と一致するように、母集団を選択したことになる。この場合、無作為抽出した集団により、設定した母集団に関する性質を調べるという研究目的を達成できない。「母集団かつ設定したモデルにおいて $p < \alpha$ である」集団に関する調査を行なっていることになる。
調査を終えて、この標本についてモデルを使った予測ができないと主張できない。この不正な操作をアステリスクシーキングという。

2.3.3　標本平均が xx になったときに抽出を終える

標本に対して計算できる平均値や分散が理想の（考えているモデル）と一致するまで無作為抽出を繰り返すまたは、一致したときに無作為抽出を終えると、無作為抽出したとは言いきれない。

■計測したデータを報告しない

日本製鉄は 18 日、東日本製鉄所君津地区（千葉県君津市）から有害物質が流出していた問題で、過去の水質測定データに不適切な扱いがあったと発表した。排出基準を超える有害物質が検出されたにもかかわらず、千葉県などに報告していない例があった。有害物質が基準を上回った際、再度測定して基準内に収まる結果を記録していたことも明らかにした。
https://www.nikkei.com/article/DGXZQOUC186070Y2A810C2000000/

アンモニア化合物の漏洩が発生し、着色水の構外への流出が確認され、排水溝から取水したサンプルから、環境規制値を超えるシアンが検出される [a]。その後、シアン除去設備の能力増強などが行われる[b]。さらに、精査していくと、測定データについて不適切な取り扱いがあったことが判明した[c]。ここで、1日のうちに複数回の計測データが存在していたこと、関係機関へ報告していた数値より高い計測データが存在していたことが判明した。

計測データが、予想や基準値よりも大きかったまたは小さかったから、データを削除してはいけない。計測手順を決定し、そして計測したデータは、全て報告しなければならない。

データが恣意的に削除されているかどうかを判定することは非常に難しい。この例でも、データを持たない外部の人間が、不適切な報告が行われていることを判定できなかった。基準値を超えたデータについても記録が残っていたので、報告が適切に行われていないことが明らかになった。

データがなければ、どのような行動を行うだろうか。例えば、保存されたサンプルを再度計測することになる。そのサンプルがなければ、なぜ基準値を超えた値が検出されたのかが徹底的にせいさされることになる。例えば、計測装置の利用手順のミスなどが検証される。ここで異常がなければ、通常のサンプリングが行われ、基準値を超えるデータが取得される頻度が、これまでよりも高いかを調べることになると考えられる。データがなければ、検証のコストが増えてしまうと考えられる。

[a] 東日本製鉄所君津地区における着色水の構外への流出について https://www.nipponsteel.com/common/secure/news/20220624_100.pdf

[b] https://www.nipponsteel.com/common/secure/news/20220706_100.pdf

[c] https://www.nipponsteel.com/common/secure/news/20220818_200.pdf

第 3 章

統計モデル

この章ではついにデータが登場する。データは母集団から無作為抽出によって得られた数値であるとする。データを大文字の $X_1, X_2, \cdots, X_n$ とし、モデルからサンプリングした確率変数を小文字の $x_1, x_2, \cdots, x_n$ とする。統計モデルはデータの出現頻度や統計量などの出現区間などを予測する。まず、その予測可能なことについて列挙する。モデルとデータが異なる場合つまり、データの出現頻度をデータが予測できない場合に生じることについて説明する。

■統計学に数学は必要か

Dr_slump7802:
理論や理屈、式の導出をブラックボックス化し、単に『この実験区なら、このデータならこの検定法、このソフト』みたいな講義になっている大学が多いので、統計学嫌いの学生が増えていく。

https://twitter.com/Drslump7802/status/1610784458655006720

Dr_slump7802:
よく、『統計学に数学の知識は重要でない』と言い切る人がいるが、それは違うと思う。少なくとも分布のグラフや式がどういう関数であるかは理解する必要がある。

https://twitter.com/Drslump7802/status/1610784907328106496

Dr_slump7802:
サンプルデータの条件を把握していることはもちろん前提。

https://twitter.com/Drslump7802/status/1610785188879138816

Dr_slump7802:
敵は「検定法のしくみはわからなくてもいいから，実験結果を判定してくれればいいんだ」と平気で学生に語る農学系教員かな。＞負の教育拡大再生産

https://twitter.com/Drslump7802/status/1610796746355126275

Dr_slump7802:
教員自身はなんとか勉強して使っているけど講義する実力はないし，専門家の非常勤講師を雇う予算もない。だから，数学なしでも成立する学部を目指そうとなっている（苦手だけど勉強するとは大違い）。それが現在の地方国立大学農学部の現状。

https://twitter.com/Drslump7802/status/1610799572766580736

Dr_slump7802:
農学部や生物学科は，もともと数学から逃避した学生比率が他の理系学部より高いので，数理系基礎科目を教えるのは大変労力を要する。しかし，それが面倒なので，そもそも数学を選択にしたカリキュラムの大学も多く，学生の潜在意識どころが，本当に学部教育が「なんちゃって理系」化している。

https://twitter.com/Drslump7802/status/1610800417763659777

数学の勉強が少し必要である。

■数学の勉強方法

教科書 1 冊をペンを使って丸写しすることもある。暗記のためではない。手で書いて考えるために行う。

3.1 正規分布を含んだ統計モデル

次の 3 つを仮定したモデルを正規モデルと呼ぶ。

(1) 独立同分布
(2) その分布は、正規分布
(3) 正規分布の母数 (平均と分散) はそれぞれ μ, σ^2。

この正規モデルを $M(\mu, \sigma^2)$ と書く。σ^2 をある特定の値にしたときのモデルを $M(\mu)$ または $M(\mu; \sigma^2)$ とし、μ を特定の値にしたモデルを $M(\sigma^2)$ または $M(\sigma^2; \mu)$ とする。
母集団から無作為抽出した標本 (データの入った集合) を元にモデルを構築する。正規分布における最尤推定量は、$\mu_{ML} = \bar{X}, \sigma^2_{ML} = \frac{1}{n}\sum(X_i - \bar{X})$ である。最尤推定量を元にした統計モデル $M(\mu_{ML}, \sigma^2_{ML})$ を最尤モデルと呼ぶ。

■最尤モデルが最も良い予測をするかはわからない

赤池は、最尤推定量がデータを推測する上で良い推定量になっているかの根拠にならないことを指摘している**??**。

> R.A.Fisher の研究により、観測データ x が実際に $p(x|a)$ の形の確率分布に従って発生するとき、最尤法が優れた特性を示すことが示された。しかし、応用の場面では、データを生み出す確率的な構造が完全に分かっていることは無いから、Fisher の議論は、最尤法の実上用の根拠を与えない。

本書が扱うデータには、特定の分布形が指定されていないので、最尤モデルが尤もデータに当てはまるモデルを推定することはあり得ない。他の量が良い場合もある。例えば、中央値などを使ってモデルを構築した方がデータに当てはまることがある。

以下では、$M(\mu)$ による予測について説明する。

3.1.1 データが出現しやすい区間

ある決められた確率でデータが出現するとモデルが予測する区間を予測区間という。割合として、よく使われる 95% を設定したものを 95% 予測区間という。正規分布を含んだモデル $M(\mu)$ において、予測区間は比較的簡単に求めることができる。具体的には、正規分布の規格化を行い、標準正規分布に従うように変換を行い、$\frac{x-\mu}{\sigma}$ であるので、予測区間は、

$$\mu - z_{0.05}\sigma < x < \mu + z_{0.05}\sigma$$

である。この範囲に 95% のデータが生じることをモデルが予測する。実際にそのようになるかは不明であり、予測であることを意識した方が良い。

同様に、68% の確率でデータを含むと予測する区間が求められる。

$$\mu - \sigma < x < \mu + \sigma$$

3.1.2 平均値の出やすい区間

次の統計量 Z が標準正規分布 $N(0,1)$ に従うことが、正規分布の再生性によってわかっている。

$$Z(\bar{x}, \mu) = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0,1)$$

ここで $\bar{x}$ は、統計モデル $M(\mu)$ からサンプリングした標本の標本平均値 (データの平均値ではない)、μ, σ は統計モデルで設定した母数平均、母数分散。
$Z(\bar{x}, \mu)$ が、標準正規分布における標準偏差の 2 倍の範囲 ($-2 \sim 2$ の範囲) にあるあるならば、

$$\begin{aligned} & -2 < Z(\bar{x}, \mu) < 2 \\ \rightarrow \quad & -2 < \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} < 2 \\ \rightarrow & \mu - 2\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 2\frac{\sigma}{\sqrt{n}} \end{aligned}$$

である。モデルから決められたサンプルサイズの標本を複数生成し、各標本の平均が $[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}]$ の範囲に入るか判定していくと、その確率は標準正規分布の $[-2, 2]$ の積分値 0.954 である[*1]。この区間のことを 95.4% 信頼区間という。
この積分値の小数点 3 桁以降を切り捨てた数値 0.95 になる範囲は、$[-z_{0.025}, z_{0.025}]$ である。こちらの基準では、

$$\mu - z_{0.025}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{0.025}\frac{\sigma}{\sqrt{n}}$$

である。この区間を 95% 信頼区間と呼ぶ。一般に、確率変数が標準偏差の 2 倍を超えるた値をとる確率は 0.046% である。このことを利用して、統計量についても、標準偏差の 2 倍の範囲の中で見つかることが大抵であることから、$[-2, 2]$ や $[-z_{0.025}, z_{0.025}]$ などの範囲が使われる。

サンプルサイズによる影響

95% 信頼区間の式を見てわかるように、サンプルサイズ n が大きくなれば、$\bar{x}$ が入る範囲は狭くなる。信頼区間がサンプルサイズに依存することを数値的に確認する。図 3.1 は、信頼区間が N に応じて変化する様子を図示した。

[*1] 数学の記法としては間違えているかもしれないがあえ数式で書くと、$\#\{x$ は $M(\mu; \sigma^2)$ からサンプリングされた変数の組; $Z(x) \in [-2, 2]\}$/標本数 $= 0.954$ である。ここで、$\{\}$ は集合であり、$\#$ は、集合の要素の数。$M(\mu)$ からサンプリングした確率変数の組み x について、$P(-2 \leq z(x) \leq 2) = 0.954$ でもある。

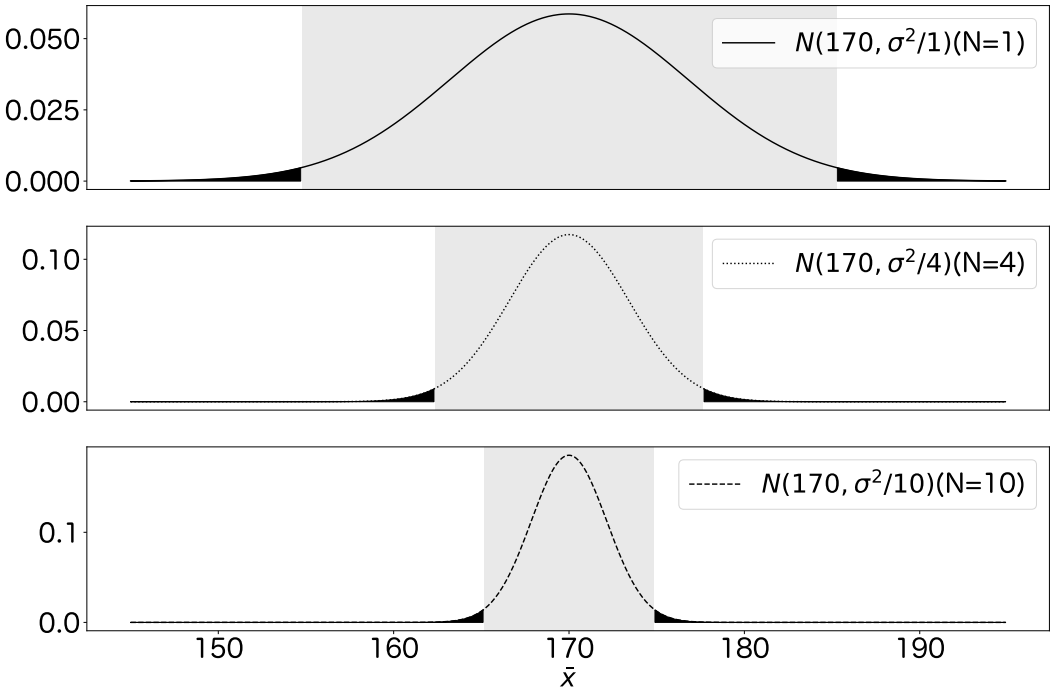


図 3.1 モデル $M(170; \sigma^2 = 5.8)$ における (A)$N = 1$,(B)$N = 4$,(C)$N = 10$ での 95% 信頼区間

信頼区間の中に標本平均が含まれていることは、標本がモデルにり推測可能であることの証拠の一つになる。ただし、予測可能かの判断には、複合的に指標を見る必要がある。

3.2 指数分布を含んだ統計モデル

(1) 独立同分布
(2) その分布は、指数分布 $(\lambda \exp(-\lambda x))$
(3) 指数分布の母数は λ

このモデルを $M_E(\lambda)$ とする。このモデルの 95% 予測区間は、

$$[\frac{1}{\lambda} \log \frac{1}{1-\alpha/2}, \frac{1}{\lambda} \log \frac{\alpha}{2}]$$

である。95% 信頼区間は式 B.3 である。

3.2.1 信頼区間の近似

95% 信頼区間 (式 B.3) を近似的に求める方法がある。中心極限定理を使う。このモデルでは、サンプルの平均および分散は、$E[x] = \frac{1}{\lambda}, Var[x] = \frac{1}{\lambda^2}$ である。このとき、中心極

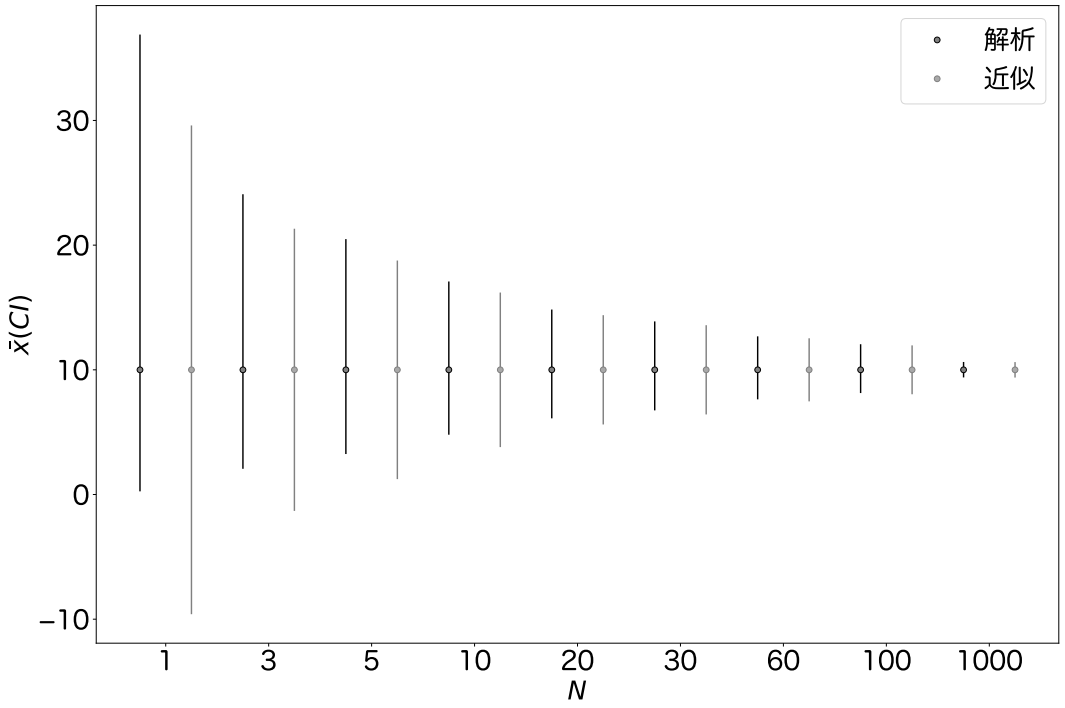


図 3.2　解析的な信頼区間と近似によって求められた信頼区間。$1/\lambda = 10$ とする。横軸にサンプルサイズ。縦軸に、平均値。エラーバーは信頼区間 (CI)。

限定理により、$\bar{x} \sim N(E[x], Var[x]/n)$ である。よって、95% 信頼区間は、

$$\frac{1}{\lambda} - z_{0.05}\frac{1}{\sqrt{n}\lambda} < \bar{x} < \frac{1}{\lambda} + z_{0.05}\frac{1}{\sqrt{n}\lambda}$$

である。

解析的に求めた信頼区間と中心極限定理による近似的な信頼区間を比較する (図 3.2.1)。$\lambda = 10$ としたので、平均は全て 10 である。N が小さいと、解析と近似での信頼区間に差が生じている。近似的な信頼区間は、0 よりも小さな値も出現することを予測している。平均 10 の指数モデルでは、平均が 0 以下になることはない。このように、モデルの想定しない区間も信頼区間に含めている。N が大きくなると、解析と近似での信頼区間の違いは少なくなる。

3.3 モデルとデータの乖離を調べる方法

3.3.1 正規モデルの場合

サンプルサイズ 1 の標本が正規モデル $M(\mu)$ により予測できるのかを考える。$M(\mu)$ であれば、95% の確率で、$\mu - \sigma z_{0.025} \sim \mu - \sigma z_{0.025}$ の間でデータが見つかることを予測する。この中に入っていることが予測可能の目安の一つにはなる。サンプルサイズが小さい場合では、標本分布の形がどの分布に適合するのかを推測しにくので、このモデルが現実を予測していると言い切ることはできない。
標本全体を使えるのならば、

1. 母数平均よりも小さいまたは大きな点が半分程度であることを予測している。
2. 標準偏差の内側言い換えれば、$[\mu - \sigma < x < \mu + \sigma]$ の中にあるデータの数は 68% 程度。
3. $[\mu - 2\sigma < x < \mu + 2\sigma]$ の中にあるデータの数は 95.4% 程度。

などがデータに当てはまるのかを調べる。

3.3.2 母集団の標本が指数分布的に分布していた場合

母集団の分布形と統計モデルに含まれている確率分布関数が著しく異なる場合を考える。母集団分布として、指数分布を仮定しする。これは、自然から指数分布的なデータが得られたときのことを想定している。これを予測するモデルを正規分布の仮定されたモデルとする。
最尤モデルは、$M(\mu_{ML}, \sigma^2_{ML})$ である。ここから、

$$\mu_{ML} - \sigma_{ML} < x < \mu_{ML} + \sigma_{ML}$$

が 68% 予測区間になる。言い換えれば、標準偏差の間に、サンプルの平均が入る確率が 68% であることをモデルが予測している。このことを数値シミュレーションにより確かめる。指数分布からランダムサンプリングを行い、無作為抽出によりサンプルサイズ 10^6 の標本を得たとする。サンプルが上記の区間に入っている割合を計算する。

```
1 N = 10**6
2 sample = expon.rvs(scale=10,size=N)
3 #sample = norm.rvs(loc=0,scale=1,size=N)
4 lambd= np.average(sample)
5 print(np.average(sample),np.std(sample),np.var(sample))
6
7 mu = np.average(sample)
8 s = np.std(sample)
```

```
9
10 a,b = mu-s,mu+s
11 len(sample[np.where( (sample >a) & (sample<b) )])/N
```

この結果、期待していた値 68% よりも著しく大きな割合 86% 程度を得る。これは、モデルでは、正規分布を仮定していたが、実際には指数分布的なデータだったために生じる予測の間違いである。

エラーバー (SD) から読み取れること

正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本を作り、プロットした (図 3.3.2)。それぞれの分布の右側のエラーバーは、68% 予測区間。標本が正規分布であるときには、68% 予測区間の中におよそ 68% のデータが含まれている。一方で、標本が指数分布であるときは、モデルの予測と乖離する。このことは、図から読み解くことが難しい。

エラーバー (SD) だけが描かれた図を見るとデータに対する印象が変わる (図 3.3.2)。図 3.3.2 には、正規分布と指数分布から得られた標本から、最尤推定を行ったモデル $M(\mu_{ML}, \sigma^2_{ML})$ における 68% 予測区間を描画している。データが正規分布的であるならば、データが区間に入っている割合が予測と一致する。一方で、データが指数分布的であるならば、モデルの予測 (エラーバー) から得られることと、実際のデータは乖離する。エラーバーからは、中央からデータが対称に分布しており、その中に、68% のデータが入っていることを我々は読み取ろうとする。データのばらつきを表すために SD を描いた場合、それが正規分布的な標本でない限り、データのばらつきの意味が伝わりにくくなる。実際の論文において、モデルを考えてエラーバーに SD を書いていると断言しがたい。例えば、SD が描かれているのに、正規分布を仮定しない統計モデルにより解析を行うことがある。これは、データの描画においては、正規分布を仮定しているにもかかわらず、仮説検定においては正規分布をもとにした推測をやめていることを意味する。この場合、著者が何を考えて SD を描いたのかを判断することが難しくなる。

測定の精度の良さを示す指標として SD を書くことがある。この場合、ただ単に SD が小さければ良い計測であることを示す意図がある。

3.3.3　正規モデル以外でも使える方法

累積分布によるデータとモデルの比較

標本の累積分布のプロット方法について説明する。標本 $X_1, X_2, \cdots, X_n$ を小さいものの順に並び替えたものを、$X_{r(1)}, X_{r(2)}, \cdots, X_{r(n)}$ とする。ここで、$r(j)$ は、j 番目のデータのインデックスを返す関数である。そして、

$$(X_{r(j)}, j/n) \quad (j = 1, 2, \cdots, n)$$

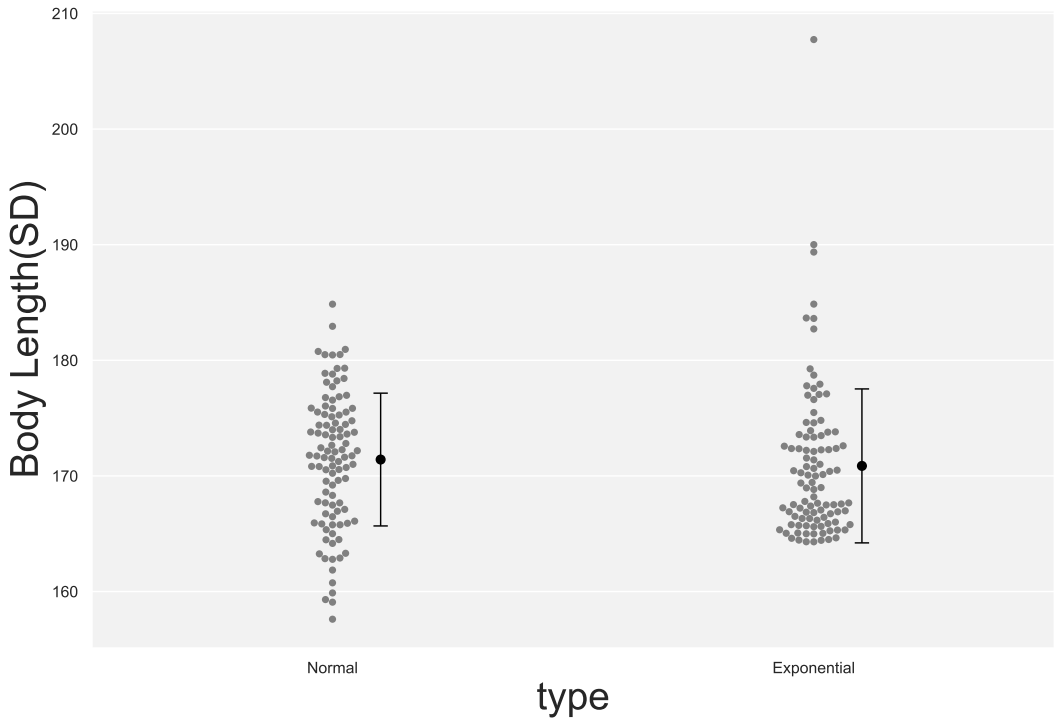


図 3.3 正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本をプロットした。それぞれの右側にあるエラーバーは、正規分布モデルが予測した 68% 予測区間。

をプロットする。言い換えれば、累積分布は、標本を小さい順に並べたものと、順位をサンプルサイズで割ったもののペアをプロットしたものである。
具体的なコードは次のようになる

```
def cummlative_norm(data):
    sorted_data = np.sort(data) # 順番の並び替え X_{r(j)}
    x = np.arange(len(data))/len(data) # データ数分のj/n
    mu_ml,sigma_ml = np.mean(data),np.std(data)
    predict_cdf = norm(mu_ml,sigma_ml).cdf(sorted_data)
    return sorted_data,x,predict_cdf
```

累積分布の傾き

累積分布は、データの密集度が高い範囲において、傾きが大きくなり、密集度の小さい範囲では、傾きが小さくなる (図 3.3.3)。

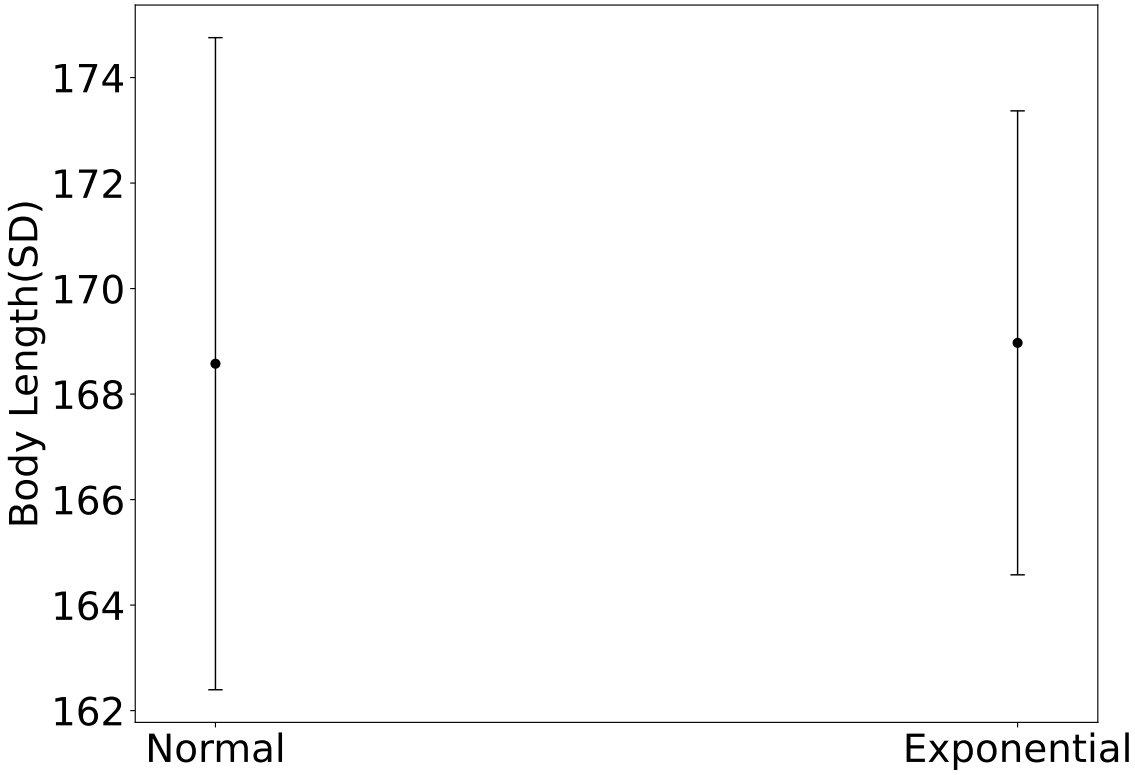


図 3.4　正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本を得た。その標本から推測される 68% 予測区間を描画した。

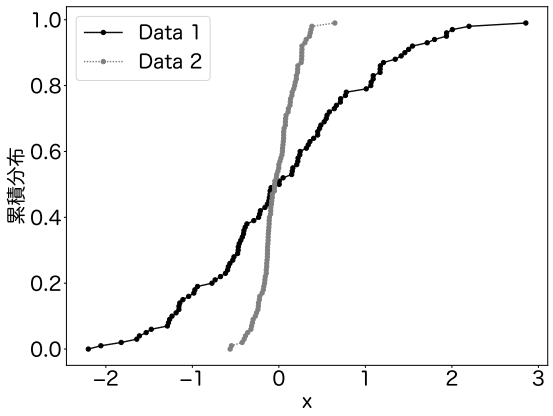


図 3.5　データの累積分布。Data 1 は、正規分布 $N(0, 1)$、Data 2 は正規分布 $N(0, 0.2)$ からサンプリングした。サンプルサイズは 100

データとモデルの比較

図 3.3.4 図 3.3.4 右側に累積分布を描いておいた。データは、(a) 正規分布、(b) 指数分布、(c) ガンマ分布からそれぞれサンプルサイズ 100 の標本である。それぞれに、正規分布の最尤モデルを重ね書きしておいたので、最尤モデルとの乖離具合が把握できる。(a) では、最尤正規モデルとデータが一致している。(b,c) では、最尤正規モデルの曲線上に、データの累積分布の点が乗っていないので、モデルとデータが乖離していることが示唆される。このことから、このようなデータが得られたなら、モデルを再構築したほうが良い。また、サンプルサイズを 30 にした図 3.3.4 では、データが正規分布であっても、正規モデルによって推測することが良いのかはぱっと見では判断しにくく、正規モデルを確信を持って利用しにくくなる。

3.3.4 qq プロットによるデータと正規分布の比較

qq プロットについて説明する。まず上記、$(X_{r(j)}, j/n)$ について、j/n を、$F^{-1}(j/n)$ によって変換する。ここで、累積標準正規分布の逆関数を $F^{-1}(p)$ とする。つまり、

$$(F^{-1}(j/n), X_{r(j)}) \quad (j = 0, 1, \cdots, n)$$

をプロットする。

```
1    def qq_plot(data,ax):
2        sorted_data = np.array(sorted(data))
3        p = np.arange(len(data))/len(data)
4        x_ = norm(0,1).ppf(p)
5        return np.c_[x_,sorted_data]
```

qq プロットを図 3.3.4 図 3.3.4 左側に描いておいた。直線に乗っているデータは、正規モデルの推測が当たりやすいと考えられる。

3.3.5 相対的なモデルのデータへの適合具合

AIC の比較

AIC は、対数尤度に対して、データ由来のパラメータ分、ペナルティを与えたものである。AIC が低いモデルは、相対的に良くデータに当てはまるモデルであり、そのモデルからデータが生成されたことを示唆するものではない。また、AIC の差が 10 あったから良いとか悪いとかではなく、AIC が低いものが相対的に良いモデルと判断されがちになるだけである。AIC が小さいから、良い予測をするということは一般にない。

正規モデルのデータに対する AIC を計算する。母数を最尤推定により決定した最尤モデルのパラメータ数は 2 である[*2]。過去の研究データから、平均 μ を決定し分散については

[*2] データ由来の母数 2 つあるので

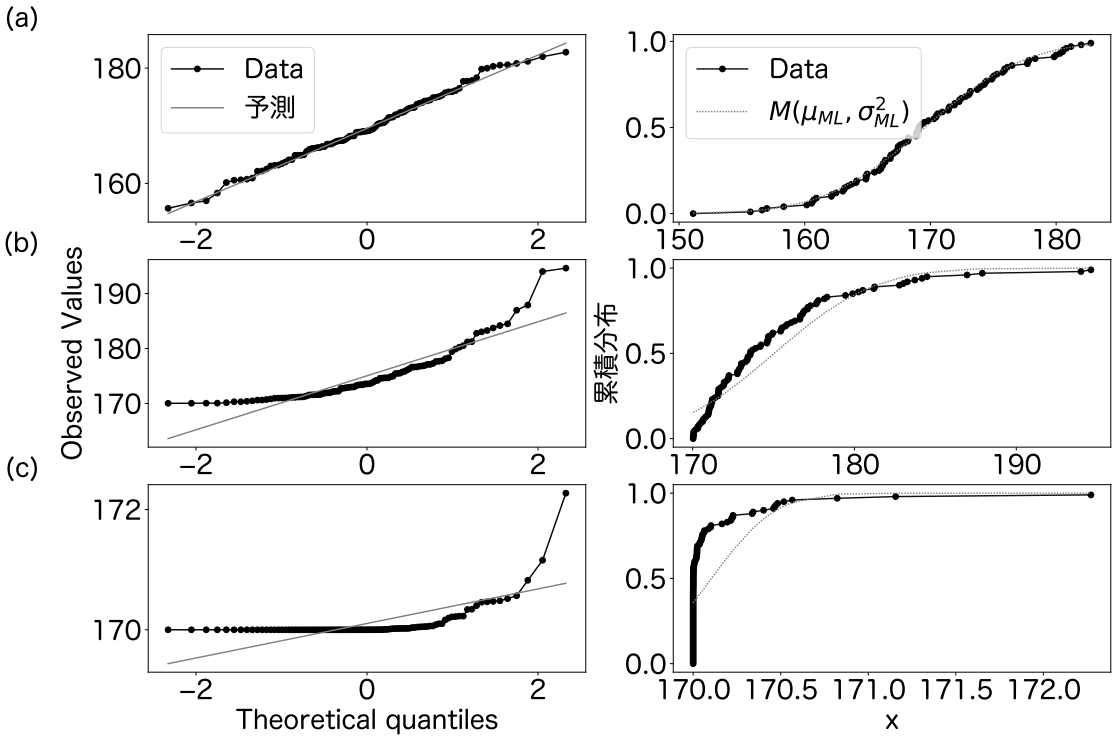


図 3.6　左には qq プロット、右は累積分布と最尤モデルの累積分布。サンプルサイズは 100(a) 正規分布 $N(170, 5.8)$(b) 指数分布 $\lambda = 5.8$(c) ガンマ分布 $s = 0.1$)

最尤推定量により決定したモデル $M(\sigma^2_{ML})$ のパラメータ数は 1 である。

AIC が低いモデルは良い予測をするモデル？

AIC が低いから、良い予測ができるかは不明である。一般に、比較対象のモデルの中で、データへの適合度が相対的に高いモデルである。まず、AIC が低いモデルでも、データの出現を予測しにくい事例を紹介する。

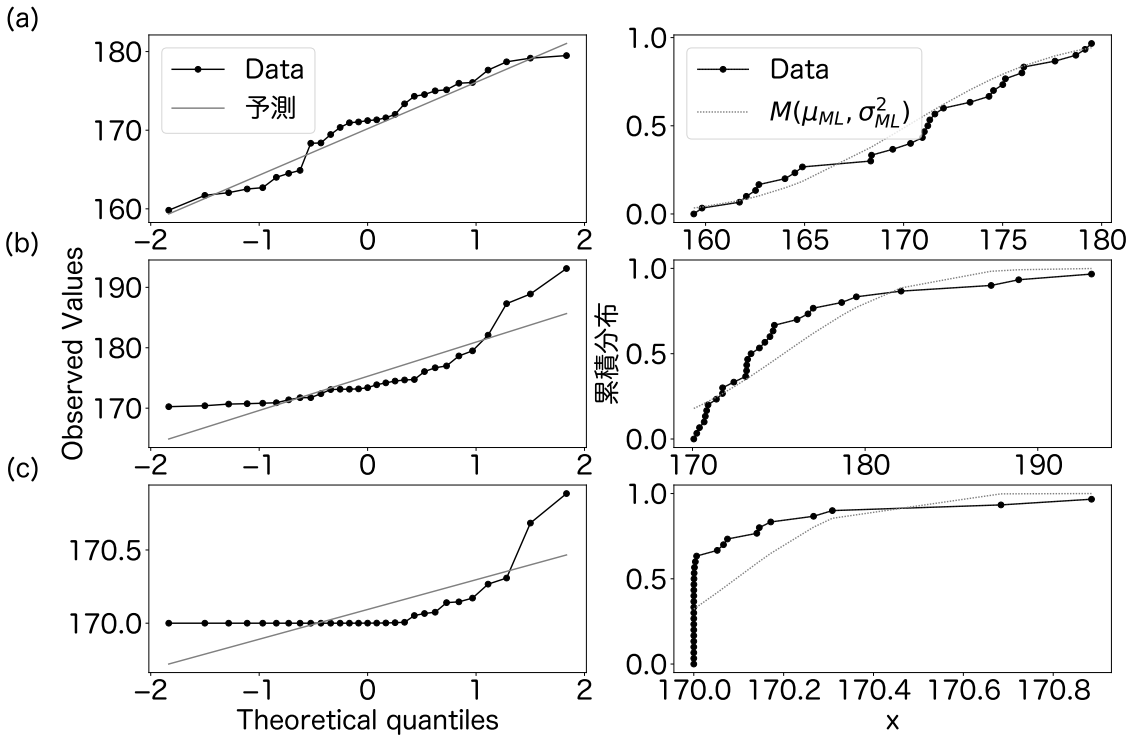


図 3.7　左には qq プロット、右は累積分布と最尤モデルの累積分布。サンプルサイズは 30(a) 正規分布 $N(170, 5.8)$(b) 指数分布 $\lambda = 5.8$(c) ガンマ分布 $s = 0.1$)

第 4 章

モデルにおける統計量の性質

統計モデルからサンプリングを行った標本から、統計量を計算すると、その統計量より偏った値が出現する確率が得られる。この確率が低いとき、標本がモデルからサンプリングされたものではないと一定の割合で判断を下すことにする。このことを利用して、標本をモデルからサンプリングしたと判断するものとそうでないものに仕分けを行う。

4.1 自己標本の批判

統計モデルからサンプリングした標本の統計量が従う確率密度関数が理論的に求められる。正規モデルであれば、

$$Z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0, 1)$$

である。このことを利用すれば、Z 値よりも偏った値のモデルの中で出現する確率が計算できる。例えば、$Z = 0$ であれば、これ以上に偏った値がモデル内で出る確率は 0.5 程度なので、よくある統計量であることがわかる。$Z = 1.96$ であれば、これ以上に偏った値が得られるモデル内での確率は 0.025 程度なので、なかなかのレアさであると判断することがある。ここで注意しなければならないのは、p 値はモデル内における Z 値よりも偏った値の出現する確率であるということである。
言い換えを書いておく。

標本 $\rightarrow$ 統計量 $\rightarrow$ p 値 (p 値の小ささが標本のモデル上での得られにくさの指標になる)

p 値が小さいなら、統計量 Z 値以上に偏った値の得られる確率が低いということであるので、Z 値の元の標本もそもモデルからは得られにくいということを示唆する。つまり、モデルから標本の得られにくさの指標の一つが p 値であるとも言え、p 値が小さいほど、その標本はそのモデルから得られにくい。
モデルが生成したはずの標本であるが、閾値を決めてモデルから生成されたものではないとするのである。モデルが自身から得られた標本を批判するのであるから、自己の標本を批判するのである。これは、モデルを元に、標本がモデルにより予測できるかどうかを考えている。

定義 4.1.1. 統計モデルにおいて、標本の統計量以上に偏った（大きいまたは小さな）値が得られる確率を p 値と呼ぶ。

4.1.1 p 値の計算練習

$Z(\bar{x},\mu) \sim N(0,1)$ により、Z 以上の値が得られる確率も計算できる。つまり、

$$p = \Phi(Z(\bar{x},\mu) > x)$$

を計算させる。モデルからえた標本から計算した統計量を $\bar{x} = 172.4$、サンプルサイズを $n = 10$。モデルとして、正規モデル $\mu = 168, \sigma^2 = 6.8$ とする。このとき、Z 値は $Z(\bar{x},\mu) = 2.04$ であり、これを元に以下のスクリプトを実行すれば、$p = 0.04$ であることがわかる。

```
1 xbar = 172.4
2 mu = 168
3 sigma2 = 6.8**2
4 n=10
5 Z = np.sqrt(n)*(xbar-mu)/np.sqrt(sigma2)
6 print(Z)
7 p=1-norm.cdf(Z,0,1)
8 print(p*2)
```

4.1.2 自己標本の否定確率

あるサンプルサイズの標本をモデルから 100 標本を得たとすると、それぞれの統計量 Z_i をそれぞれ計算できる。全体のうち 95 個の標本についてはモデルから生成されたと判断し、残りの 5 個についてはモデルから生成されていないと判断することにする。これは自己標本の批判を元にすれば可能である。具体的には、正規モデルを利用すれば、その統計量 Z が $N(0,1)$ に従うことがわかっている。Z の値が偏った値になっていれば、その出現頻度は低くなるので、$P(|z| < Z) = 95/100$ となる Z を計算する。この Z は具体的に計算でき、$Z_{0.95} = 1.96$ である。ここから、$|z_i| < Z_{0.95} = 1.96$ となる z_i の個数を数えればおよそ 95 になる。また、$|P(|z_i| > 0.95)| > 1.96$ ならば、$|z_i| > Z_{0.95} = 1.96$ である。式を展開する。

$$\begin{aligned} & |z_i| > Z_{0.95} \\ \rightarrow & \frac{\sqrt{n}|\bar{x}-\mu|}{\sigma} > Z_{0.95} = 1.96 \\ \rightarrow & \mu - \frac{\sigma}{\sqrt{n}} Z_{0.95} << \mu + \frac{\sigma}{\sqrt{n}} Z_{0.95} \end{aligned} \tag{4.1}$$

いくつか言葉を定義しておく。

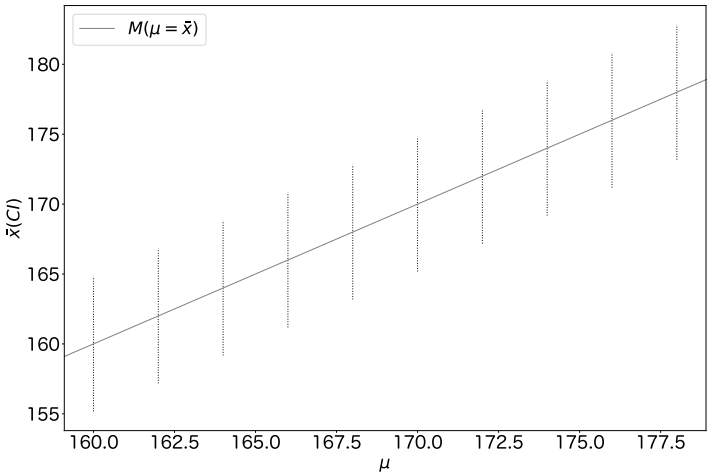


図 4.1 横軸にモデルの母数 μ、縦軸に、モデルが予測する平均値 $\bar{x}$、エラーバーに 95% 信頼区間を描いた。$N = 10, \sigma^2 = 6.8^2$

定義 4.1.2. モデルからサンプリングされた標本のうち、モデルから生成されたものではないと判定する割合を α とし、有意水準と呼ぶ。言い換えれば、α 値は、統計モデルからサンプリングされた値について、これが元の統計モデルからサンプリングなのかどうかを判定する頻度に関する閾値[*1]である。式 (4.1) の範囲を信頼区間といい、これ以外の範囲を棄却域と言う。

本書では、歴史的な習慣にしたがって、$\alpha = 0.05$ を利用して、計算を行う。

4.1.3 μ の変化に応じた信頼区間

信頼区間は、サンプルサイズ n、有意水準 α およびモデルの母数 μ, σ^2 により決まる。ここでは、μ の変化に応じて、信頼区間が変化する様子を確かめる。
図 4.1 には、モデル毎の平均値と信頼区間を描いた。μ の大きさにによらず信頼区間の幅は同じである。各 μ に対して、信頼区間の内側で $\bar{x}$ が 95% の確率で見つかることを統計モデル $M(\mu)$ が推測する。この外側にある $\bar{x}$ になる標本については統計モデルにより推測できないのではないかと疑いがかけられる。

4.2 正規モデルにおける中心間の距離 (効果量)

分散が等しい二つの正規モデル $M_a = M(\mu_a), M_b = M(\mu_b)$ とする。$M_{a'}$ の中心から $M_{b'}$ の中心への距離は、$D = \frac{|\mu_a - \mu_b|}{\sigma}$ となる。D を効果量と呼ぶ。式を変形すれば、

[*1] 閾値（読み：いきち）＝限界値

$D\sigma = |\mu_a - \mu_b|$ であり、中心からの距離が σ 何個分かを D が示す。

4.3　統計量をもとにしたモデル間類似度 (検出力)

母数の異なる二つの統計モデル M_a, M_b について考察する。M_a の信頼区間内の統計量が M_b において出現する確率を検出力という。言い換えれば一方で出現する統計量が他方のモデルにおいて出現する確率である。これは、M_a から M_b へのモデル間の統計量を元にした類似度と言える。

4.3.1　検出力の定義

M_a の棄却できない統計量の範囲 (信頼区間 A) に M_b の統計量が出現する確率を β とする。β を検出力という[*2]。β は、二つの異なるモデルを比較するための指標で、一方のモデルで棄却できない母数がもう一方のモデルで出現する確率である。M_a に対する M_a の検出力 β は、$1-\alpha$ であり、M_a を棄却する閾値を低く設定すると、β は大きな値になる。二つの統計モデルの母数がよく一致するならば、β は $1-\alpha$ に近い値を取り、一致していないならば、β は 0 に近い値を取る。具体的に、α, β を式で書くと、

$$P_a(\mu \in R_a) = \alpha$$
$$P_b(\mu \in A_a) = \beta$$

ここで、R_a, A_a はそれぞれ統計モデル M_a の棄却域、信頼区間、P_a, P_b は、それぞれ統計モデル M_a, M_b における統計量に関する確率密度関数。

4.3.2　正規分布モデルの検出力

具体的に、$P_a(\mu \in R_a), P_b(\mu \in A_a)$ を計算してみる。σ^2 がすでに与えられた正規モデルを $M(\mu)$ とし、$M_a = M(\mu_a), M_b = M(\mu_b)$ とする。M_a または、M_b からサンプリングされた確率変数 $x_1, x_2, \cdots, x_n$ の平均値は、それぞれ $\bar{x}_a \sim N(\mu_a, \sigma/n)$ または $\bar{x}_b \sim N(\mu_b, \sigma/n)$ である。M_a の信頼区間 A_a は、$|\bar{x}_a| < \mu_a + \sigma/\sqrt{n} z_{2.5\%}$ である。このとき、P_a を $N(\mu_a, \sigma)$ の確率密度関数とすると、

$$P_a(\mu \in A_a) = \alpha$$

であるのは定義から明らか。また、P_b を $N(\mu_b, \sigma)$ の確率密度関数とすると、

$$P_b(\mu \in A_a) = \beta$$

である。μ_a と、μ_b が一致していれば、$P_b(\mu \in A_a) = 1-\alpha$ である。μ_b が μ_a から離れていくと、$P_b(\mu \in A_a) = 0$ に近づいていく。

[*2] 検出力を検定力または統計力と呼ぶこともある。
https://id.fnshr.info/2014/12/17/stats-done-wrong-03/

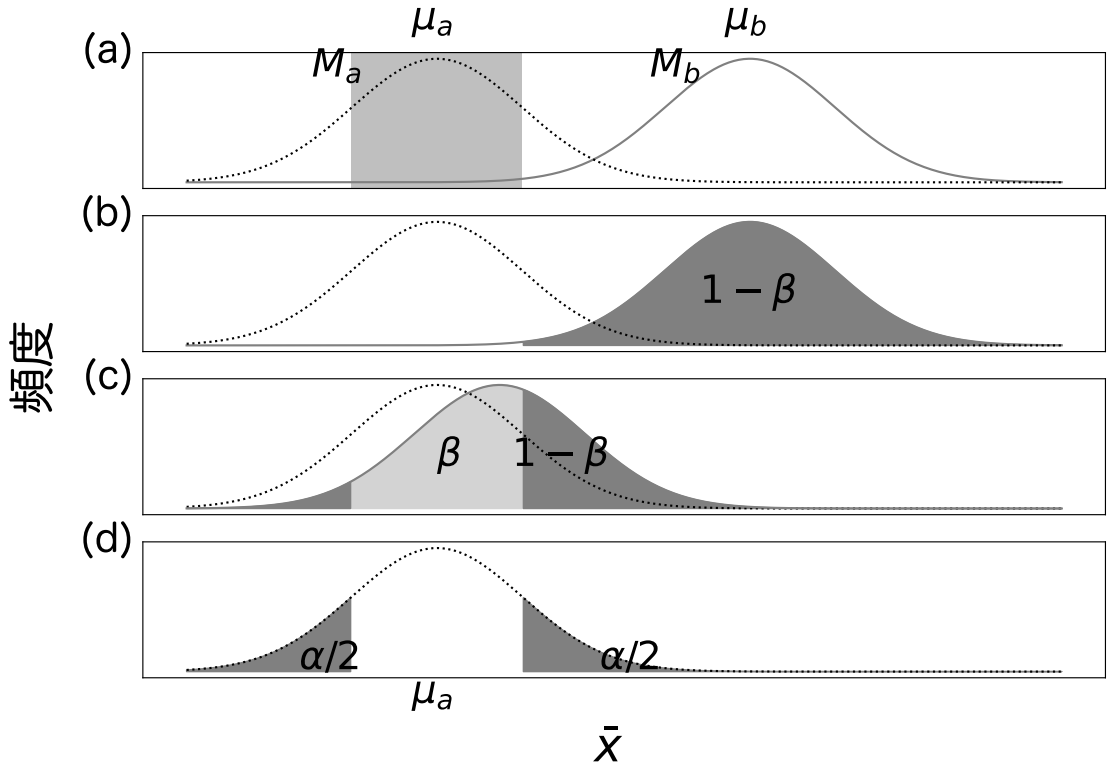


図 4.2　統計モデル M_a, M_b から計算された統計量 $\bar{x}$ の確率分布 P_a, P_b。(a) 灰色の範囲は M_a の信頼区間。(b) 灰色の領域は、$1-\beta$ の領域を示している。β の領域が小さいので、描画できなかった (c)μ_b が μ_a に近いときの β と $1-\beta$ の領域。(d) 灰色の範囲の面積が α を示している。

検出力と α の領域を図示した (図 4.2)。M_a の 95% 信頼区間は、$|\mu| < \mu_a + z_{0.025}\frac{\sigma}{\sqrt{N}}$ である。信頼区間は、図 4.2(a) において灰色で塗った x 軸の範囲である。α は図 4.2(c) の灰色で塗りつぶした領域の面積である。検出力 $1-\beta$ は、M_b における M_a の信頼区間の外側の領域の面積なので、図 4.2(b) の濃い灰色の範囲である。

α を 0 に近づけていくと、信頼区間は徐々に大きくなり、β は大きくなる。α を 1 に近づけていくと、信頼区間は徐々に狭くなり、β は小さくなる。

α、M_a の母数 μ_a、M_b の母数 μ_b を固定したまま、サンプルサイズを変化させ, β の変化を表す (図 4.3)。$\bar{x}$ の確率密度関数 ($N(\mu, \sigma^2/n)$) の分散がサンプルサイズによって変化することは明らかである。このことから、サンプルサイズが大きくなると、信頼区間は徐々に狭くなり、$1-\beta$ は大きくなる。サンプルサイズが小さいときは、$1-\beta$ も小さくなる。

μ_a を固定し、μ_b を変化させたときの検出力 $1-\beta$ を図 4.3.2 に示した。サンプルサイズが大きければ、$1-\beta$ も大きくなることがわかる。

β を定義したことにより、β の数値を決定し、M_a, M_b の違いが β になるために必要なサ

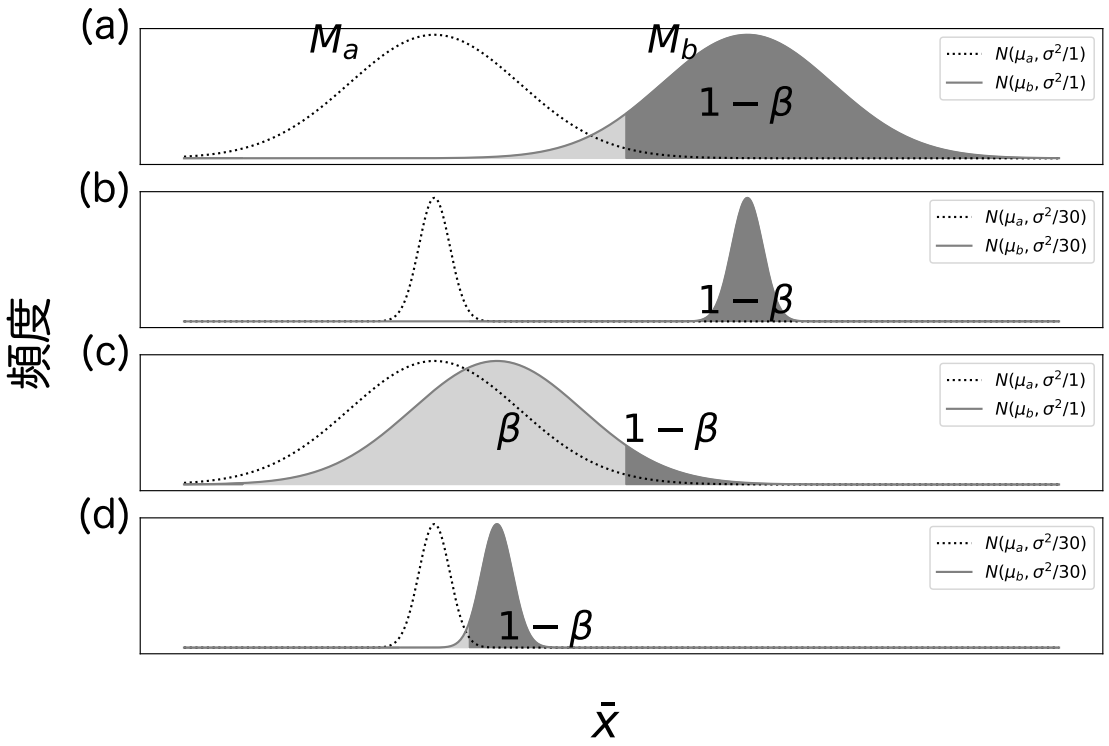


図 4.3　統計モデル M_a, M_b から計算された統計量 $\bar{x}$ の確率分布 P_a, P_b。(a)μ_a, μ_b のサンプルサイズ 1 の平均値がしたがう確率密度関数 $N(\mu_a, \sigma^2/1), N(\mu_a, \sigma^2/1)$。(b)(a) と同じ μ_a, μ_b に対して、サンプルサイズを 30 にした場合の確率密度関数。(c)μ_a, μ_b が (a) よりも近いときの $\bar{x}$ の確率密度関数。(d)(c) と同じ μ_a, μ_b に対してサンプルサイズを 30 にした場合の $\bar{x}$ の確率密度関数。

ンプルのサイズが推測できる。ここでは、μ_a, μ_b が固定されている状況を考える。検出力 $1-\beta$ は 1 に近いほど、M_a, M_b が違うと主張できる。あらかじめ決めたおいた基準の $1-\beta$ を閾値を設定し、それ以上の $1-\beta$ となるサンプルサイズを推測する。サンプルサイズが小さければ、M_a と M_b の違いは曖昧であり、サンプルサイズが大きくなると、はっきりとモデルの違いがわかる。

4.3.3　β の計算

正規モデル M_a, M_b を使って、β を計算してみる。M_a の信頼区間は、

$$-z_{0.025} \leq \frac{\sqrt{n}(\bar{x}-\mu_a)}{\sigma} \leq z_{0.025}$$

より、

$$A_a = \{\mu; \mu_a - \frac{\sigma}{\sqrt{n}} z_{0.025} \leq \mu \leq \mu_a + \frac{\sigma}{\sqrt{n}} z_{0.025}\}$$

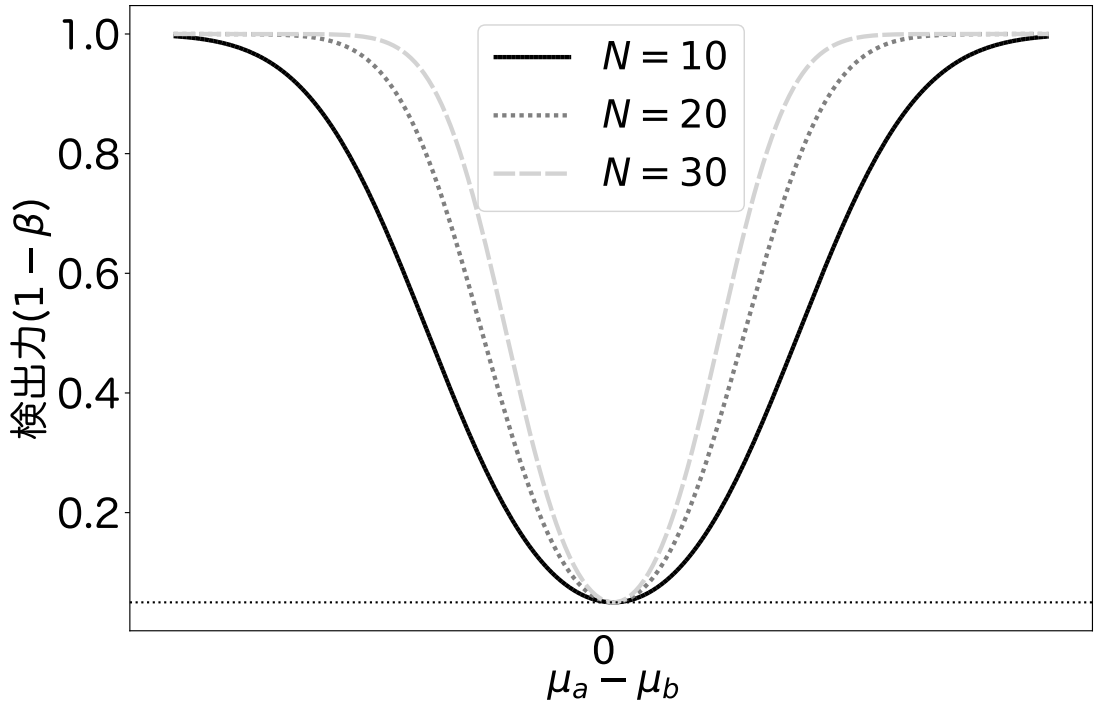


図 4.4　μ_a を変数にしたときの検出力 (検出力関数)。

である。ここで、$a = \mu_a - \frac{\sigma}{\sqrt{n}} z_{0.025}, b = \mu_a + \frac{\sigma}{\sqrt{n}} z_{0.025}$ とおく。棄却域は A_a 以外の μ である。M_b の標本平均 $\bar{x}_b$ は、$N(\mu, \frac{\sigma^2}{n})$ に従うので、A_a の区間で、$N(\mu_b, \frac{\sigma^2}{n})$ の面積を計算すれば良い。ここで、$\frac{\sqrt{n}(\bar{x}_b - \mu_b)}{\sigma} \sim N(0,1)$ である。このことを利用すると、a, b は、$N(\mu_b, \frac{\sigma^2}{n})$ の確率変数だとすると、

$$
\begin{aligned}
A &= \frac{\sqrt{n}(a - \mu_b)}{\sigma} \\
&= \frac{\sqrt{n}(\mu_a - \frac{\sigma}{\sqrt{n} z_{\alpha/2}})}{\sigma} \\
&= -z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b)
\end{aligned}
$$

同様に、

$$
\begin{aligned}
B &= \frac{\sqrt{n}(b - \mu_b)}{\sigma} \\
&= \frac{\sqrt{n}(\mu_a - \frac{\sigma}{\sqrt{n} z_{\alpha/2}})}{\sigma} \\
&= z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b)
\end{aligned}
$$

である。以上より、確率密度関数 $N(0,1)$ において、$-z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b) \leq x \leq z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b)$ の間で積分すれば良い。
$d = \frac{\mu_a - \mu_b}{\sigma}$ とおく。$d = 0.6, n = 9$ とする。このときの β を計算してみる。$N(0,1)$ において、$-z_{\alpha/2} - 0.6\sqrt{n} \leq x \leq z_{\alpha/2} + 0.6\sqrt{n}$ の区間で積分する。

```
A,B = norm.interval(0.95,0.,1)
N = 9
d = 0.6
a,b = A+d*np.sqrt(N),B+d*np.sqrt(N)
print(a,b)
norm.cdf(b,0,1)-norm.cdf(a,0,1)
```

答えは、0.564

サンプルサイズ

d と検出力を指定したときに、M_a, M_b の類似度を検出力以上にするためのサンプルサイズが計算できる。$\beta = 0.1, \fallingdotseq 0.8$ とし、この β を満たすように N を計算した。

```
A,B = norm.interval(0.95,0.,1)
beta = 0.1
d = 0.8
for N in range(10,200,2):
    a,b = A+d*np.sqrt(N),B+d*np.sqrt(N)
    beta_ = norm.cdf(b,0,1)-norm.cdf(a,0,1)
    if beta_ < beta:
        break
print(N)
```

計算を実行すると、18 であることがわかる。

4.3.4　最尤モデルでの β の計算

データを元にしたモデルとモデルの類似度

統計モデル A を $M(\mu = 170)$ とし、統計モデル B を $M(\bar{X})$ とする。ここで、$\bar{X}$ は、無作為抽出によって得られた標本の平均であり、標本の大きさを 100 とする。モデル A,B の間の検出力が計算可能である。$d = \frac{170 - \bar{X}}{6.8}$、$n = 100$ であるので、$\bar{X} = 168$ を得たとすると、

```
A,B = norm.interval(0.95,0,1)
N = 100
d = (170-168)/(6.8)
```

```
4 a,b = A+d*np.sqrt(N),B+d*np.sqrt(N)
5 print(a,b)
6 norm.cdf(b,0,1)-norm.cdf(a,0,1)
```

その検出力は、0.163

4.4 過誤

これまでの議論をまとめる。モデル M_a からサンプリングを行った標本について、モデル M_a に関する標本であるかを判定する。モデルから生成された標本であるが、偏った統計量出会った場合は、モデルから生成されていないと判断する。この頻度を α とした。このように、モデル M_a から生成されたのに、統計量の出現頻度から、このモデルから生成したものではないと言う誤った判断を行う事になる。このことを判断の間違いであると言うことから第 1' の過誤と呼ぶ[*3]。
今度は、モデル M_b からサンプリングを行った標本が、別のモデル M_a からサンプリングされたかを判定する。統計量が信頼区間に入っているかどうかを確認し、入っていなければ、モデル M_a からサンプリングされていないと判定できる。問題が生じるのは、統計量が信頼区間に入っている場合である。これは、実際には、M_a からサンプリングされていないのにもかかわらず間違って、サンプリングされたと判断する事になる。この判断の間違いを第 2' の過誤と呼ぶ。

表 4.1 モデル M_a による自己標本批判

	M_a の信頼区間に標本の統計量が入っていない	M_a の信頼区間に標本の統計量が入っている
モデル M_a の標本	モデル M_a の標本ではないと判定 (第 1' の過誤)	モデル M_a の標本と判定
モデル M_b の標本	モデル M_a の標本ではないと判定	モデル M_a の標本であると判断 (第 2' の過誤)

■過誤はデータとモデルを比較したときに生じる判断ミス

データとモデルを比べたときに、誤ってモデルが間違いと判定することを第一の過誤と教科書において紹介していることがある。誤ってモデルが間違いと判断するとはどのようなことなのかの定義がないので、この定義の意図がわからない。
本書では、モデルからサンプリングした標本とモデルを比較したときに生じる間違

[*3] Neyman-Pearson とは異なる過誤を定義したので、1' および 2' とした。仮説検定において、Neyman-Pearson と Fisher を混ぜ合わせて過誤を定義することが現在の主流である。こちらの定義では、さまざまな誤解が生じている [6]

いとして過誤を定義した。標本は無作為抽出によって得られたものではない。

■正解と回答の違い

あるデータ群に対してそのデータの特徴を元に、Yes または No とアノテーションをつける。データからその Yes または No を予測する手順を開発する。その手順によって得た回答と、正解（真の値）の一致と不一致は以下の通りになる (表 4.2)。回答と一致したら、True、一致しないなら False。Yes と予測したら Positive、No と予測したら Negative とする。回答が Yes な問題に、Yes と答えることは (手順が正しい予測を行なった)、True Positive といい、No と答える (手順が間違えた予測を行なった) ことは False Negative という。回答が No な問題に、Yes と答えることを、False Positive、また、No と答えることを True negative という。
モデル M_a の標本に Yes を対応ずけ、モデル M_b の標本に No を対応付ける。標本を元に、Yes または No を判定する手順をモデル M_a を元にした統計検定を利用する。この問題において回答が FP となったものを第 1' の過誤であり、FN となったものが第 2' の過誤である。

表 4.2　正解と回答の違い

	負例 (真の値)	正例 (真の値)
正例 (予測値)	偽陽性 (FP)	真陽性 (TP)
	予測が外れた	予測が当たった
負例 (予測値)	真陰性 (TN)	偽陰性 (FN)
	予測が当たった	予測が外れた

4.5　自己否定の過推定

統計モデルの中で、統計モデルを統計量により検査するときに、モデル自身を絶対にダメなモデルと判断しすぎてしまうことを自己否定の過推定と言う。この過誤は 2 つの要因に分解でき、[*4]、不適切な統計量を使用することで、棄却域と統計量の違いにより生じる α_1、そして、検定を繰り返して生じる α_2 である $(0 < \alpha_2 \leq 1)$。$\alpha_2 = \alpha$ となっていれば、有意水準 α の検定ができる。α_1 は、統計モデルと、その統計量の関数になっており、言い換えれば、統計量が統計モデルの中で設計通りの振る舞いをしているかを測る指標である。正規モデルを使い、統計量 T を使った場合、$\alpha_1 \approx 0$ であるが、指数モデルを使い、統計量 T を使った場合、α_1 が指定した α よりも多くなる。これを見ていく。α_2 は、$\alpha \times 2$

[*4] α_2 は α_1 に関係するので実際には、分解できない。気持ちとしては、α_2 は、α_1 を変数に持つ関数である $\alpha = \alpha_2(\alpha_1)$。

以上になる場合、軽視されることはないが、α_1 が同程度の隔たりになる場合においては無視され、α_1 は α_2 よりも軽視されがちであることも説明する。

4.5.1 どんな統計モデルでも T 統計量で調べよう (α_1)

統計モデルの分布の仮定が正規分布以外の場合においても、T 統計量を使ってモデル自身を検証できるのかを調べる。次の統計モデル $M_E(\lambda)$ を構築する。

1. $X_1, X_2, \cdots, X_n$ は i.i.d
2. 指数分布
3. λ

母数 $\lambda = 1$ とした統計モデルを $M_E(1)$ とする。$M(1)$ からランダムサンプリングした確率変数 $x_1, x_2, \cdots, x_n$ から次の統計量を計算する。

$$T = \frac{\bar{X} - 1}{\sqrt{\frac{\sigma^2}{n}}}$$

ここで、$T \sim t(n-1)$ とする。T 値が $t(n-1)$ の棄却域に入っている頻度を数値計算により計算する。具体的に、平均 1 の指数分布または、平均 1、標準偏差 1 の正規分布からサンプルを得て標本を作る。その標本を 100000 回取得する。このとき、T 値を計算し、T 値いじょの値が得られる確率 p を計算する。その p が $p < 0.05$ となる割合を計算する。以上をサンプルサイズを変化させてシミュレーションを行なった。平均 1、標準偏差 1 の正規分布の場合、T 値は $t(n)$ 分布に従ので、$p < 0.05$ となる頻度も、5% 程度になることが期待される。一方で、平均 1 の指数分布の場合、T は $t(n-1)$ 分布に従うとはいえない。このことから、$p < 0.05$ となる頻度は計算してみなければわからない。
シミュレーションの結果、正規分布から標本を得た場合、$p < 0.05$ になる割合は、サンプルサイズに依存せず、5% 程度であり、期待通りである。一方で、指数分布から標本を得た場合、$p < 0.05$ になる割合はサンプルサイズに応じて変化しており、また、どのサンプルサイズでも $p < 0.05$ となる割合は 5% より多い。
このことから、指数モデルの α_1 は、$\alpha_1 > 0.05$ であることがわかり、統計量を正しく選ばなかったことで、自己否定の過誤が期待した 0.05 よりも大きくなっていることがわかる。

■サンプルサイズが xx 以上あるから t 検定

サンプルサイズがある値以上あるので、中心極限定理により、t 検定が利用できるというものもある[a]。このロジックが読み込めなかったので、その謎を明らかにすべく我々はアマゾンの奥地へ向かった。
データが指数分布的であるときに、t 検定を使うときに生じる問題は上でみた通りであり、$p < 0.05$ となる標本の割合が多くなっているので、間違った推測をする可能性が高くなる。他の分布関数でもおそらく同じような現象が現れる。このことか

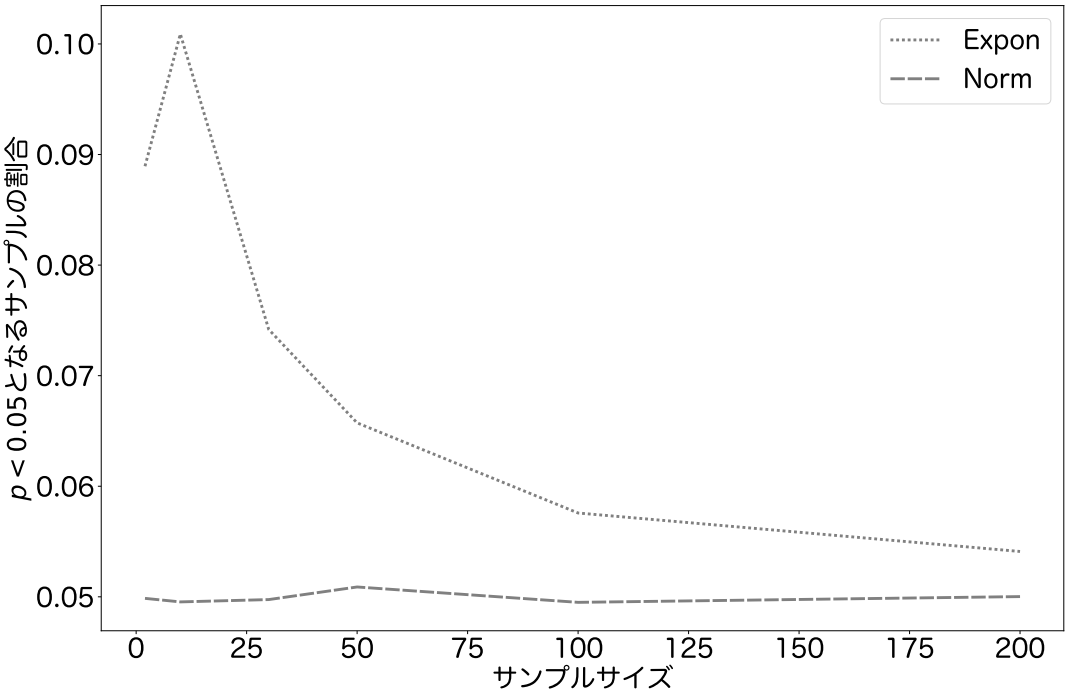


図 4.5　正規分布または指数ぶんぷから得た標本の T 値から計算した p 値で、$p < 0.05$ 以下になる割合

ら、我々は「t 検定が利用可能である」は正確ではなく、「t 検定を使うことができるが、間違った推測である確率が高くなる」ということだと推察した。
業界によっては、サンプルサイズが xx 以上であれば、過誤を無視して良いというふうに言われることもある。実際には、設計したモデルと

[a] http://id.ndl.go.jp/bib/024660739

4.6　検定を繰り返し使おう (α_2)

ここまでは、一つの標本に対して、統計モデル $M(\mu)$ により推測できるかを考えていた。ここでは、$\sigma^2 = 10^2$ とした正規モデル $M(\mu)$ によって複数の標本について推測できるかを仮説検定を指標にし考える。標本が 3 個あるとする。このとき、それぞれの標本の統計量 T が信頼区間に入っている確率は、$(1-\alpha)$ である。全ての標本の統計量 T が信頼区間に入っている確率は、その積 $(1-\alpha) \times (1-\alpha) \times (1-\alpha) = (1-\alpha)^3$ であり、この確率で統計モデルは棄却されない。一方で、棄却される確率は、$1-(1-\alpha)^3$ である。表 4.3 は、標本数に応じた α_2 である。標本数が大きくなるについれて、α_2 が大きくなることが

表 4.3 標本数に応じた α_2

標本数	$\alpha = 0.05$	$\alpha = 0.01$
1	0.05	0.01
2	0.0975	0.0199
3	0.142	0.0297
4	0.185	0.0394

わかる。
α_1 がレベル α の検定になっていない場合、α_2 はさらに有意水準 α から隔たりの多い数値になる。

4.7 類似度の過誤

統計モデルの間の類似度を検出力といった。統計モデルに対して、不適切な統計量を与えたとき、検出力を歪める。これを類似度の過誤といい、その確率を β' で表す。直接またはシミュレーションを行い β を計算することがおそらくできそうだが、面倒なので行わない。

4.8 データとモデルの比較

ここで、いくつかのことを定義しておく。

定義 4.8.1. 統計モデルと標本を比較して、モデルが母集団のことを予測できないとさまざまな指標をもとに判断するとき、統計モデルを却下すると宣言する。

ここで、母集団から無作為抽出した標本 (モデルから生成された標本ではない) を正規モデルにより、予測できるかを考える。上記の議論と同様に、標本から、統計モデルにあった統計量を計算し、統計量よりも偏った値が出現する確率 (p 値) を計算する。p 値が小さければ、モデルにより予測できないと考え、値が 1 に近いほど、もしかしたらモデルで予測できるのかもしれないと考える[*5]。標本を元に、モデルにより予測ができないかを考えている。
以上のことは、托卵行動に例えることができる。モズは、カッコウに対して卵を託す托卵を行い、カッコウは、モズの卵とは気が付かず、そのまま育てる。ここで言い換えたいのは、カッコウは統計モデルであり、卵は標本そして、モズは科学者である。統計モデルは、モデルからのサンプリングされた標本を巣穴に置いている。卵の情報を要約した統計量が、モデル由来であることをモデルはその統計量の出現頻度を推測できる。出現頻度が p 値である。モデルの巣に自然から無作為抽出した標本を科学者が置く。その標本の統計

[*5] p 値だけで判断してはいけない

図 4.6　統計量を使ったモデルとデータの比較に関する概念図

量の出現頻度をモデルは推測できる。得られた推測から、標本がモデルの卵であることを判定するのは科学者である。この手順だけでは科学者はモデル鳥と標本卵を比較しているだけであり、標本卵を構成しているデータそのものとモデル鳥を比較していないということに注意しなければならない。

■偶然の差が生じたかを確かめたい

「偶然の差が生じたかを確かめたい」や「こんなことが起こる確率は 5% くらい」という言葉を統計学の教科書で見たことがある。これらは、本書での説明とは異なる前提をもとに議論を進めており、本書と解釈の互換性はない。

科学では、実験で得られたデータは、同様の実験を行った場合、同様のものが得られるということが前提になっている。このことを現象に再現性があると言う。再現性のないデータを現状の統計学で扱うことや、現実の現象が得られる確率を議論することは困難である。

本書の前提を元にすれば、「こんなこと（これ以上に偏った統計量値）が（モデル内

で）起こる確率は 5% くらい」ということを省略して「こんなことが起こる確率は 5% くらい」と言うことはできる。また、現実において起こりやすいのかどうかについては議論できない。

4.8.1　p 値を使った判断に関する注意

p 値を元に統計モデルとデータの不一致を考えるとき、p 値はモデルとデータの乖離を示す指標の一つであると言うことを意識しなければならない。このことを忘れてしまい、次の間違った判断を行うことがある。

1. p 値が 0 に近いならば、統計モデルによりデータを予測できないと判断する
2. p 値が 1 に近いならば、統計モデルによりデータを予測できると判断する

それぞれのデータがどのようなものなかのかを確認してみる。

p 値が 0 に近い → 統計モデルによりデータを予測できないと判断
p 値が 1 に近い → 統計モデルによりデータを予測できると判断

■ P 値が小さければ、モデルの仮定のうち少なくとも一つが間違い

> P 値が小さければ、データと帰無仮説の矛盾している程度が大きいので、P 値が小さければ帰無仮説は棄却するんだと統計の教科書には書かれています。実はそうではなくって、今お話ししたように小さい P 値が何を意味するかというと、たくさんある統計モデルの仮定のうちどれか一つが間違っているあるいは、複数のものが間違っている。決して帰無仮説だけが間違いの対象ではなくって、先程のように、小さい P 値が選択的に報告してあれば、結果としては誤った結果になります。・・・・[a]

P 値が小さければ、モデルの仮定のうち少なくとも一つが誤っているというものがある。私はこの意見に賛成できない。
モデルの中で標本の統計量以上の値の出現確率を計算したものが P 値である。P 値によって、仮定の間違いを主張できるような値ではない。ある母数を持つモデルによりデータの平均値を予測できなかったことを示唆すのが P 値である (パラメトリックモデルの範囲であれば)。正規分布や独立同分布ではないことを示唆することはおそらくない。
ただ、モデルとデータの比較を行なった後、データが目的にあっているのかを調べなければならない。

[a] 京都大学大学院医学研究科 聴講コース 臨床研究者のための生物統計学「仮説検定と P 値の誤解」

佐藤 俊哉 医学研究科教授 https://www.youtube.com/watch?v=vz9cZnB1d1c

■モデルの仮定を満たせるのか

最初の原則。最初に述べられている原則ですが、P 値はデータと特定の統計モデルが矛盾する程度を示す指標の一つであるというふうに書かれています。ここでですね、統計モデルは何かって言うと、統計モデルは必ず一連の仮定のもとで構成されています。どんな仮定かと言いますと、統計の教科書をみますと、「データが正規分布している」とか、「平均値が等しい」などが統計モデルに必要な仮定とされているのですが、まず、一番大切なことは、データを撮るときに、先程の試験のように、薬剤のランダム割り当てが行われているとか、対象者を剪定するときにランダムサンプリングがなされているか、こういったことも統計モデルの仮定に含まれています。それから当然、研究計画がきちんと守られているかも統計モデルが必要とする前提の一つです。例えば、先程の臨床試験で言えば、結果の解釈も変わってきます。最後まで対象者が追跡できているのか。追跡不能とからつだくがあったとすると、統計モデルの後世に影響を与えます。もちろん解析方法も妥当な結果を与える解析方法でなければいけない。こういったことを満たしていなければ、統計モデルの仮定を満たしているとは言えない。[a]

この意見は統計モデルに関する仮定と実験計画の二つの要素が混じっている。実験計画を統計モデルの仮定を満たすように設計するという意見だと考えられる。この意見に賛成しない。

まず、統計モデルの仮定が自然において対応するものが、本書においてはない。また、「平均値が等しい」という仮定であるが、ある平均値をもつ統計モデルとデータを比べるさいに、データの平均値が異なる場合においても、統計モデルを使ってそのデータの出現頻度などを推定することが可能である。このことは、モデルの仮定をデータが満たさなければならないことを示唆していない。

次に、実験計画については、科学者がみたい効果を見るために設定しているのもである。ランダムサンプリングしているのは、対象に偏りがないようにし、さまざまな対象である特徴の変化を与え、その集団内での変化を計測するために行う。対象の選定に偏りがあった場合、本当に推測したかったことが推測できない。例えば、成人以上を対象にした試験なのに、60 歳だけしかからサンプリングできなかったなら、成人に対しての言及はできない。また、偏りのあるデータを偏りを前提としていない統計モデルにより解釈するのはこんなんである。この困難さを回避するためにも実験デザインを守った無作為抽出であった方が良い。

いずれも本書の方針とは異なる。

[a] 京都大学大学院医学研究科 聴講コース 臨床研究者のための生物統計学「仮説検定と P 値の誤解」
佐藤 俊哉 医学研究科教授 https://www.youtube.com/watch?v=vz9cZnB1d1c

第 5 章

尤度比を使ったモデルとデータの比較

モデル M において得られるデータ元に、母数を最尤推定する。新たに作られた最尤モデル上での尤度と元のモデル M での尤度の比がある分布に従うことがわかっている。このことを利用して、もともモデル M でデータを予測してもいいのかを考察する。
前の章で統計検定をモデル鳥によって説明した。あるモデル鳥が生んだ標本卵に関する統計量のばらつきの特徴と、研究者が持ってきたデータ卵を比較し、そのモデル鳥が産んだと判定していいのかを考える方法と説明した。ここでも、どのモデルが生んだ標本卵に関わる予測なのかを考えなくてはならない。あるフルモデルにおける予想分布をまずは考察する。

5.1 概要

定義 5.1.1. 母数の数が異なる統計モデル間の尤度の比を次のように定義する。

$$Dev(D, M_1, M_2) = -2\log \frac{M_1\text{における } D \text{ の尤度}}{M_2\text{における } D \text{ の尤度}}$$

ここで、M_1 は、$M(\theta_1, \theta_2)$,M_2 は $M((\theta_{1'}, \theta_2))$ であり、θ_1 と θ_2 のベクトルの要素数の和は、$\theta_{1'}, \theta_2$ のベクトルの要素数の和に等しい。D は標本とする。

M_1 における標本 D から、最尤モデル $M_1(\hat{\theta_1})$ を構築したとき、$Dev(D, M_1, M_1(\hat{\theta_1}))$ は、

$$Dev(D, M_1, M_1(\hat{\theta_1})) \sim \chi_p^2$$

であることがわかる。ここで、p は、θ_1 の要素数。例えば、正規モデル $M(170, 5.8^2)$ から標本 D を生成する。その μ に対する最尤モデル $M(\hat{\mu})$ は、次のようになる。

$$Dev(D, M(170, 5.8), M(\hat{\mu})) \sim \chi_1^2$$

データ由来の母数の個数が 1 なので、自由度 1 の χ^2 分布に従う。

あるデータ D' に対して、最尤推定を行なったモデル $\hat{M}$ について、次のことがわかる。

$$Dev(D, \hat{M}, \hat{\hat{M}}) = -2\log \frac{\hat{M}\text{における } D \text{ の尤度}}{\hat{\hat{M}} \text{ における } D \text{ の尤度}} \sim \chi_p^2$$

ここで D は、$\hat{M}$ の標本であり、$\hat{\hat{M}}$ は、標本 D を使って $\hat{M}$ のいくつかの母数の最尤推定を行なったモデル、p は $\hat{\hat{M}}$ において最尤推定を行なった母数の個数。

5.1.1　データと当てはめモデル $\hat{M}$ の比較

データ D を当てはめていないモデル M に対して、D を予測できるかを確かめる。パラメータの個数が少ないモデル M' についての尤度比は次で計算できる。

$$Dev(D, M, M')$$

M によって、D について十分な予測ができるならば、パラメータを任意の個数減らしたモデル M' との尤度比は、$Dev(D, M, M')$ が自由度は元のパラメータ数-自由度 $+1$ となる χ^2 分布に従う。このことから、$Dev(D, M, M')$ が比較的小さな値であれば (p 値は比較的大きくなっている)、モデル M によって予測可能であることの根拠の一つになりうる。

5.1.2　データと当てはめモデル $\hat{M}$ の比較

当てはめたモデル $\hat{M}$ とデータの比較。次を計算する。D' を観測データとする。

$$Dev(D', \hat{M}, \hat{\hat{M}})$$

ここで、M の標本 D' の最尤推定モデルを $\hat{M}$ とし、$\hat{\hat{M}}$ は $\hat{M}$ の p 個の母数に対して最尤推定を行なったモデルである。$Dev(D, \hat{M}, \hat{\hat{M}})$ の分布 (χ_p^2) の中で、$Dev(D', \hat{M}, \hat{\hat{M}})$ が珍しい値を取っていたなら、$\hat{M}$ から考えられる尤度比の変動の中では、比較的大きな変動が起きていることが示唆される。このことから、$\hat{M}$ で標本を予想しない方が良いのではないだろうかと判断する。
注意しなければいけないのは、$\hat{\hat{M}}$ の方が良いとは言えてないことである。当てはまりの良さは、尤度の大小関係を見れば良い[*1]。

> **■尤度比検定で $p < 0.05$ だったので M_1 より M_2 がより適合的だ**
>
> 尤度比検定で $p < 0.05$ だったので M_1 より M_2 がより適合的だという判断はしないほうが良い。尤度比検定において p 値が小さいということは、M_1 における尤度比の予測値の中で、比較的大きな尤度の変化が実験データでは生じていることを示したことになる。これは、M_1 の中での変動と比較しているだけである。相対的に M_2 の方がより適合的であることを示唆していない。

[*1] 次は何かがおかしい「尤度の大小関係が有意であることを確かめるのが尤度比検定である。」。

より適合的であることを示す量は、尤度である。対数尤度が小さい方がデータに対して適合しているという判断ができる。

5.2 尤度比検定

母数の個数が k 個のモデル $M(\theta)$ とする (θ は k 次元ベクトル)。モデル $M(\theta)$ からサンプリングしたサンプルサイズ n の標本 $x = (x_1, x_2, \cdots, x_n)$ を得たとする。この標本 X から θ のうち r 個の母数に関する最尤推定量を $\bar{\theta}$ 得たとする。$\bar{\theta}$ のうち $k-r$ 個はモデル由来の母数であり、r 個は標本から推定した母数である。このことから、$\bar{\theta}$ は自由度 r の母数のベクトルと言う。
もとのモデル $M(\theta)$ における標本 X に対する尤度は、$L(\theta, x)$ とする。また、最尤モデル $M(\bar{\theta})$ での尤度は、$L(\bar{\theta}, x)$ とする。このとき、これら尤度の比がカイ二乗分布分布に従うことがわかっている[*2]。つまり、

$$-2\log\lambda(X) \sim \chi^2_{k-r}$$

ただし、

$$\lambda(X) = \frac{L(\theta, x)}{L(\bar{\theta}, x)}$$

である。

■滅多に観察されない逸脱度

> 有意確率が小さければ (通常は 5% 以下)[a]、2 つのモデルの「逸脱度の差」は滅多に観察されないほど大きな値であると判断する。

これは本書とは異なる方針の科学における指針である[b]。
本書では、ある統計モデルが予測した統計量と比較して大きな統計量が得られたからといって、現実的に滅多に観察されないとは解釈しない。
本書が扱いたい科学において、一般化線型モデルを使えば、現実での起こりやすさが検証できるということはおそらくない。

> 2 つのモデルの「逸脱度の差」が大きいことから、すなわち、要因を覗くことでモデルの当てはまりが十分に悪くなることから、その要因は有意な要因であると判断する。

これも本書とは異なる分野を研究しているのだと思われる。
尤度比の差の統計量を実データの尤度比の差と比べてわかるのは、フルモデルで予測または当てはめしない方が良さそうということである。より当てはまりのモデルかどうかは尤度比を比べればよい。

[*2] ただしいくらかの条件がある

[a] おそらく p 値が小さければ
[b] この話は後でもう一度考えて見た方がいい気がする。できないはずであるが、できるとする論文が多い。なぜなんだろう TODO。

5.3　正規モデルにおける尤度比検定

σ_0^2 を設定した正規モデル $M(\mu_0;\sigma_0^2)$ について考察する。この正規モデルからサンプリングを行なった標本 X とする。標本から得た最尤正規モデルを $M(\bar{x};\sigma_0^2)$ とする。それぞれのモデル内での標本 X の尤度を $L(\mu_0,X), L(\bar{x},X)$ とする。具体的な数式は、

$$L(\mu_0,X) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\sum(x_i-\mu_0)}{2\sigma^2}) \tag{5.1}$$

$$L(\bar{X},X) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\sum(x_i-\bar{X})}{2\sigma^2}) \tag{5.2}$$

$$\tag{5.3}$$

これらから $\lambda(X)$ を計算すると、

$$-2\log\lambda(X) = -2(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2) \tag{5.4}$$

$$= \frac{n}{\sigma_0^2}(\bar{x}-\mu_0)^2 \sim \chi_1^2 \tag{5.5}$$

$$\tag{5.6}$$

である。

数値実験

モデルと同じ確率密度関数からサンプリングを行い、尤度比検定を行なってみる。
数値実験を行なってみる。具体的に、正規分布 $N(170,5.8^2)$ からサンプリングした標本 1000 個を集める。標本から平均値を求め、これを最尤推定量とする (xbar)。この最尤モデル $M(\mu;\sigma^2=5.8^2)$ における標本の尤度を計算する (loglike2)。同様に、モデル $M(170;\sigma^2=5.8^2)$ における標本の尤度を計算する (loglike)。以上から尤度比を計算し、それが χ_1^2 分布と一致することを確かめる。以下がコードである。

```
1 norm_ = norm(170,5.8)
2 data_ = norm.rvs(170,5.8,size=(1000,10))
3 xbar = np.average(data_,axis=1)
4 loglike_ = np.prod(norm_.pdf(data_),axis=1)
5 #loglike2_ = np.prod(norm(xbar,5.8).pdf(data_),axis=0)
6 #print(np.prod(norm(xbar,5.8).pdf(data_),axis=1),xbar)
7
8 loglike2_ = []
```

```
for item in data_:
    #print(item.shape)
    a = norm(np.average(item),5.8).pdf(item)
    loglike2_.append(np.prod(a))

y = -2*np.log(loglike_/loglike2_)
x = sorted(y)
y_ = np.arange(len(y))/len(x)
plt.plot(x,y_)
plt.plot(x,chi2.cdf(x,df=1))
plt.show()
```

$N(170, 5.8^2)$ と $N(175, 5.8^2)$ と言う 2 種類の密度関数からサンプリングを行いそれぞれ結果を図 5.1(a) および (b) に示す。図 5.1(a) は、モデルとデータの分布が一致していることから、累積分布が χ_1^2 の累積分布にかなり近いことがわかる。図 5.1(b) は、モデルとデータが一致していない状況での結果を示している。尤度比の多くが右に移動しており、標本の多くが χ_1^2 において珍しいと判定されやすくなっている。

5.3.1 データとモデルの乖離を検証する

モデル上において、その標本を元にした最尤モデルにおける尤度比が χ_1^2 に従うことを示した。このことを元に、データをモデルによって予測可能かを調べる。手順は、

1. 標本を x とする。
2. モデル M における最尤推定量を計算する。
3. モデル M および最尤モデル M_{MLE} における標本 x に対する尤度を計算する
4. 尤度比および $-2\log\lambda(x)$ を計算し、χ_1^2 において珍しい値なのかを検証する。

実際に、正規モデルにおいてこの手順をなぞってみる。$M(\mu;\sigma^2)$ における最尤モデルは、$M(\bar{x};\sigma^2)$ である。それぞれのモデルにおける尤度を計算し、$-2\log\lambda x$ を計算すればよい。

5.4 複雑なモデルでの尤度比検定

次のモデル $M(\beta_1, \beta_2)$ を考える。

1. x_i は定数。
2. y_i は以下に示す分布 $p(y_i;\lambda_i)$ に従う。
3. $\lambda_i = \exp(\beta_1 + \beta_2 x_i)$
4. $y_i \sim p(y_i;\lambda_i) = \frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}$

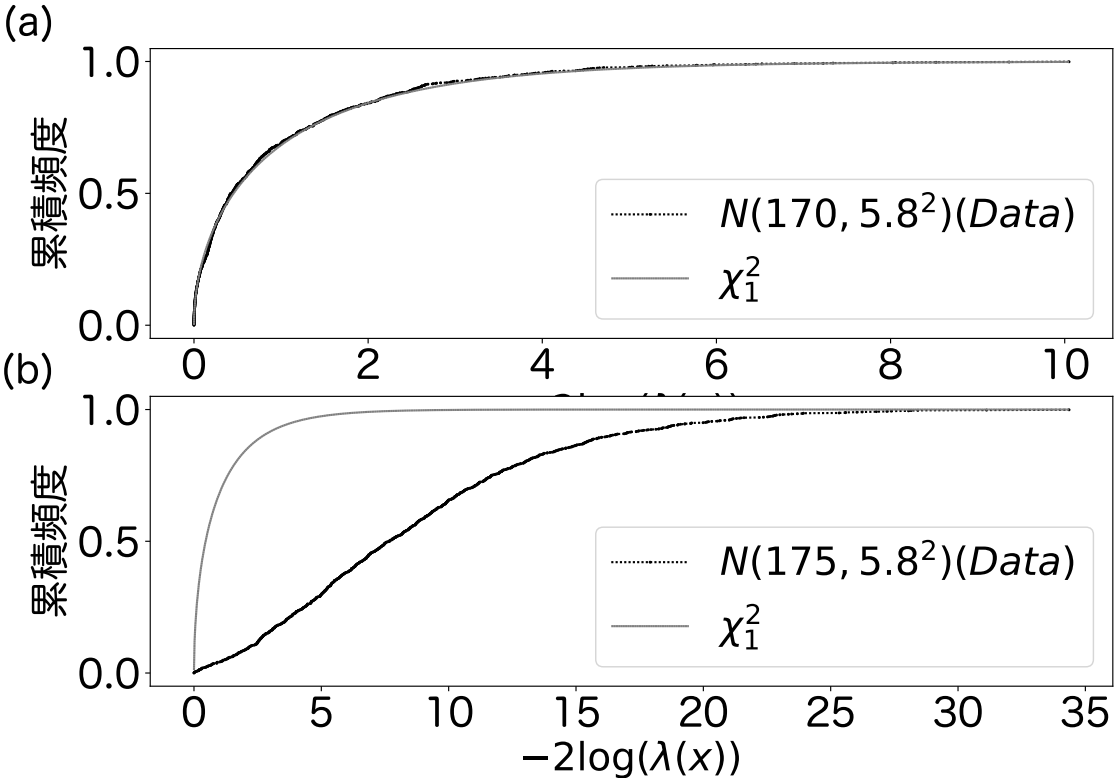


図 5.1　対数尤度比の累積頻度。モデルは正規モデル $M(170; \sigma^2 = 5.8^2)$。(a) 標本を $N(170, 5.8^2)$ からサンプリングした結果。(b) 標本を $N(175, 5.8^2)$ からサンプリングした標本。

無作為抽出した標本 x における 2 つの最尤モデルを考える。最初のモデルは、$\beta_2 = 0$ とした上で、β_2 に関する最尤推定を行なったモデル $M_1 = M(\hat{\beta}_1, \beta_2 = 0)$ である。このモデルでは、x_i に応じて、λ_i が変化しないので、λ が常に一定のモデルになる。言い換えれば、y が母数 $\lambda = \exp(\beta_1)$ のポアソン分布となるモデルである。次のモデルは、β_1, β_2 の両方について最尤推定を行なったモデル $M_2 = M(\hat{\beta}_1, \hat{\beta}_2)$ である。このモデルにおいて、(x_i, y_i) はペアになっており、x_i に応じて y_i が揺らぎを持って決まる。
ここで、M_1 における尤度比が χ_1^2 に従うことを確かめる。手順は以下の通りである。

1. M_1 においてサンプリングを行い、(x_i, y_i) からなる標本 X を得る。x_i は、既存の標本 x のものを使う。(x_i, y_i) に関してバラバラになった標本が得られる。
2. M_1 における標本 X の尤度 L_1 を計算する。
3. M_2 における標本 X の尤度 L_2 を計算する。
4. $-2\log\frac{L_1}{L_2}$ を計算する。以上を繰り返す。

以上を行うと、χ_1^2 に従うことがわかる。図 5.2a,b に結果を載せている。コードを書いておく。

```
df  = pd.read_csv("https://raw.githubusercontent.com/tushuhei
    /statisticalDataModeling/master/data3a.csv")

def get_dd(d):
    d['y_rnd'] = np.random.poisson(np.mean(d.y),len(d.y))
    model1 = smf.glm(formula='y_rnd~1',data=d,
    family=sm.families.Poisson())
    model2 = smf.glm(formula = 'y_rnd~x',data=d,family=sm.
        families.Poisson())
    #print(fit1.summary())
    fit1 = model1.fit()
    fit2 = model2.fit()
    return fit1.deviance - fit2.deviance

l = []
for i in range(1000):
    l.append(get_dd(df))


x = sorted(l)
y = np.arange(len(l))/len(l)
plt.plot(x,y)
plt.plot(x, chi2.cdf(x,df = 1))
plt.show()
```

データと、最尤モデル M_1 との比較は同様に、

1. 標本 x の尤度 L_1 を M_1 上で計算する。
2. 標本 x の尤度 L_2 を M_2 上で計算する。
3. $-2\log\frac{L_1}{L_2}$ を計算する。

最尤モデル M_1 においてデータ x が予測できないなら、$-2\log\frac{L_1}{L_2}$ が大きな値を取る。最尤モデル M_1 からサンプリングされた標本の尤度と、最尤モデル M_2 での尤度を比較すると、χ_1^2 に従う。なぜならば、ここにおける最尤モデル M_2 のパラメータ β_2 はほとんど 0 と変わりなく、小さな値をとるので、M_1 と違いが少ない。標本が M_1 からサンプリングされていないなら、モデル 2 での最尤推定の結果、β_2 も 0 から離れてしまい、尤度比も大きくなるはずである。M_1 で標本 x を予測しない方がよくないことを示す証拠の一つになる。ただし、M_2 が良い予測モデルであるのかは不明である。

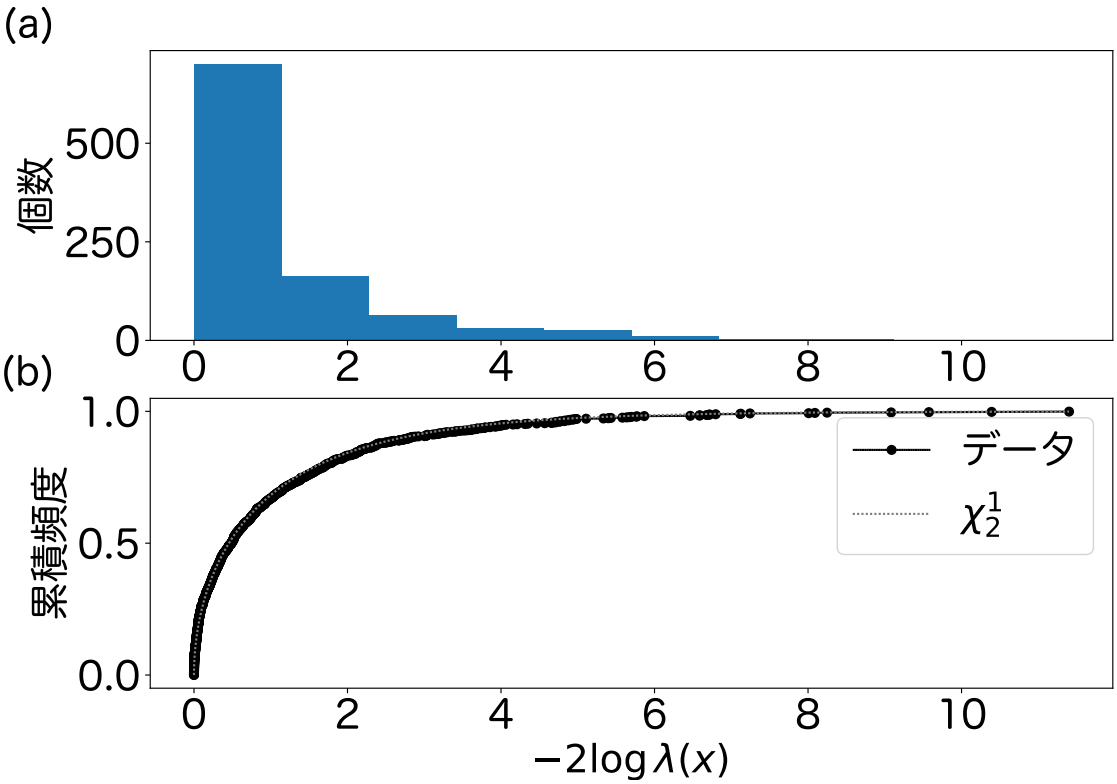


図 5.2　M_1 における対数尤度比の累積頻度。(a) ヒストグラム (b) 累積分布

M_1 からサンプリングした標本で、M_1 および M_2 を推定したモデルの尤度比は χ^2_1 に従う。実際の標本を元に、M_1, M_2 を推定し、そのモデルの尤度比は、M_1 による予測ができるならば、χ^2_1 程度だと考えられる。実際にこの例では、尤度比が大きくなり、χ^2_1 においては珍しい値になったので、M_1 により予測しない方が良さそうと言う根拠の一つになりうる。M_2 の $\hat{\beta}_2$ のパラメータがどうなっているかなども気にした方が良さそう。

第 6 章

身長を予測する統計モデル

6.1　正規分布を組み入れた統計モデル

日本人の 17 歳男性の身長を予測する統計モデルを構築する。この統計モデルは次の 1-3 から構成される。

(1) 独立同分布
(2) その分布は、正規分布
(3) 正規分布の母数 (平均と分散) はそれぞれ $\mu, \sigma^2 = 5.7^2$。

μ を変数としたこの統計モデルを $M(\mu)$ とする。およその平均値は日本にいれば母集団の分布をなんとなく知っているので、$\mu = 171.1$cm であると推測できる。母集団のばらつき具合を意識することが少ないので、分散の値を決定することは難しい。今回は、カンで 5.7^2 としました[*1]。

■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。

[*1] 統計データを覗き見した。分散を経験で推定できる人は少ないはずです。標準偏差の二倍の範囲に大体 90% の人が入っているので、言い換えれば、大体 160cm くらいの人は見ることが少なくなってくることから、分散は、大体 $5^2 \sim 6^2$ 位であることは推測できます。

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。本書では、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、現象について推測を行う。

6.2　統計モデルによる推測

$\mu = 171.1$ としたときの統計モデル $M(171.1)$ を使って、身長に関する推測を行う。

6.2.1　○○ cm 以下、◇◇ cm 以上の人の割合

まず、母集団に $180cm$ 以下、$180cm$ 以上の人の割合を推測する。正規分布関数を使い、$P(x > 180)$ を計算する。

```
1 norm.cdf(180,171.1,5.7)
2 1-norm.cdf(180,171.1,5.7)
```

結果、　$P(x < 180) = 0.940$ より、$P(x > 180) = 0.059$ ということが分かります。このことから、母集団から 100 人無作為抽出を行うと内 $5 - 6$ 人程度は $180cm$ 以上であることが推測できる。

もう一つ、$160cm$ 以下の人割合を推測する。

```
1 norm.cdf(160,171.1,5.7)
2 1-norm.cdf(160,171.1,5.7)
```

結果、$P(x < 160) = 0.059$、$P(x > 160) = 0.940$ と推測できる。

$P(x < 160)$ と $P(x > 180)$ が極めて近い値でるのは、利用した正規分布は、母平均 $\mu = 171.1$ を中心に、対称に分布する関数なので、171.1 からおよそ $10cm$ 離れた $160cm$ 以下の人と $180cm$ 以上の人ではおよそ同じくらいの割合でいると推定される。

6.2.2 擬似的に無作為を行うサンプリング

10 人分のデータをサンプリングしてみると、以下の数値が得られる。10 人を母集団から無作為抽出すると、およそこのようなデータが得られることがあると推測できる。

```
1 168.575192 164.5988088 162.7027275 163.9689649 169.8187076
    174.8851702 172.767133 165.0665034 175.7370453 163.0385381
```

6.3 統計モデルとデータの比較 1

統計モデル $M(171.1)$ による推測と実データを比較し、モデルがデータを推測できていることを確認する。17 歳男性の身長を無作為抽出して標本を得るには時間とお金がかかるので、公開されているデータ[*2][*3]を使う。このデータは文部科学大臣があらかじめ指定した 1410 校の高校に在籍する生徒を対象にした標本である。

■ $170cm$ を少し超えた人が多いのは、不正 (無作為抽出の手順に異常) があったから？

> 「生物学上、グラフは曲線になっていなければならないが、169cm の部分はへこんでいる。これは先生や生徒による四捨五入で生まれるサバ読みの結果。身長が 170cm なのか 169cm なのかで気持ち的に違ってきますからね」と話すと、食料自給率や犯罪発生件数とは異なる微笑ましいサバ読みのトリックに、出演者一同、笑みを浮かべていた。[a]

このように、データが統計モデルに一致しないことから、データに不正な操作が加わっているという推測がされることがある。議論となっている身長のデータを観察してみる。図 6.1 上を見ると、確かに、170 を超えたあたりの度数は、169 の度数よりも多い。また、$170cm$ 以下のデータは統計モデルの度数よりも低く、$170cm$ 以上のデータは統計モデルの度数よりも大きい。一方で、図 6.1 下の累積相対度数を見ると、度数と同様の変異は少ないように見える。このようなデータと統計モデルの相違の原因は、不正な計測により生じたと断言できるのだろうか。

データとモデルの相違が生じる原因が、不正な計測だけではないことを確認する。具体的には、データを統計モデルからサンプリングし、そのデータが統計モデルと一致するかを観察してみる (図 6.2)。図を見るとわかるように、サンプリングを行った場合、$168cm$ 付近で、度数が曲線よりも上にくる部分がある。また、$170cm$ より小さいところでは、統計モデルよりもデータの度数が上にあり、$170cm$ より大きな

[*2] https://www.e-stat.go.jp/dbview?sid=0003107092
[*3] https://www.e-stat.go.jp/dbview?sid=0003037791

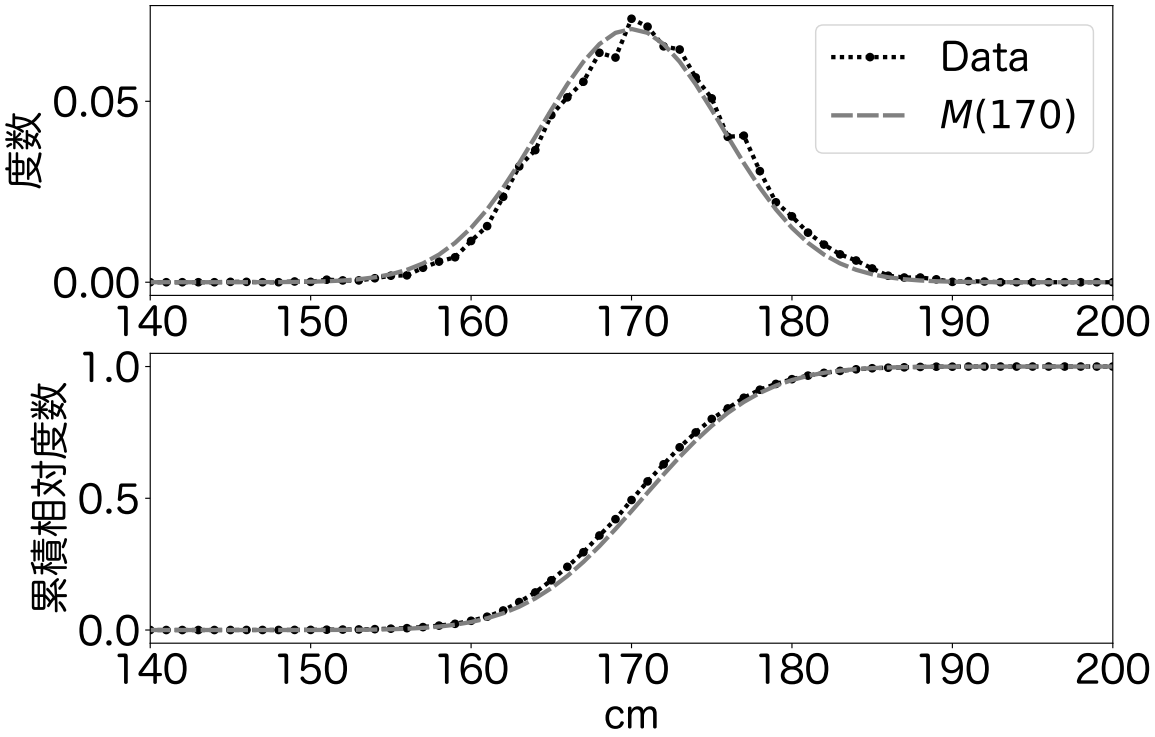


図 6.1　17 歳の男性から無作為抽出したデータ。上は、データと統計モデル $M(170)$ の度数。下は、データと統計モデル $M(170)$ の累積相対度数

ところでは、統計モデルより、データの度数が下にある。このように、統計モデルによりサンプリングし、統計モデルとサンプリングデータを比較した場合でも、ズレが生じる。これは、不正なモデルの予測とデータの間のズレが計測以外から生じることを示唆している。

不正を見つけるには、次の経験が必要である。恣意的な操作を一切介入させない、かつ、無作為にデータを取得する条件のもと、得られたデータ、と同じ計測方法・同じ生徒において、教員が計測したデータこの二つのデータが一致しないならば不正な操作が加わったことが疑える。

データ解析をするには、常に、データを収集する手順が守られていないことを疑うことをするべきである。例えば、髪の毛や靴などを履いる人がそうではない人と同じように計測をされると、平均値が大きくなる。身長の低い生徒に対してその傾向が高ければデータには歪みが生じやすくなる。計測を行なった先生方の疲れなども考慮すれば、データ収集の手順の誤りにより、データが偏ることもある。

データの収集には多大な労力がかかっている。誰かがどこかで腰を痛めながら高校

生の身長を測る仕事をしていることは心に留めておくべきで、不正があったと主張するのは、彼らの仕事を低く評価しすぎではないだろうか。おそらく先生たちは、正確に計測できるように正確に手順を満たすように計測しているはずである。不正を疑うならば、それなりに確証できる証拠を提示すべきである。具体的には、自分が手順を守って計測したデータと、先生が測ったときのデータにおいて、それらの間の差を示すべきである。
もう一つこの論者と私とで異なる点は、生物学データのグラフが曲線になるべきという点である。私は、推論のために統計モデルを利用しているので、統計モデルとデータが一致しない場合でも、推測に利用できると考え、統計モデルを利用する。一方で、この論者は、統計モデルとデータが一致すべきと考えている。言い換えれば、データが統計モデルに従うことを前提にする立場と、データを推論するために統計モデルを仮定すると言う立場がある。

[a] 国民を欺く“統計のウソ”　知らないと怖い“統計トリック”を専門家が解説 https://times.abema.tv/articles/-/5640846 2022/04/30 確認

■軽いパンばかり買わされる

ある国では、ある時期、パンを作るための道具、手順、材料が政府からパン屋に配布され、パン屋がパンを作ることになっていた。パンを焼くための型は、完成時に $1000g$ になるように設計されており、手順を厳密に守り作ったパンは確かにおよそ $1000g$ になっていた。どの季節に作っても手順を守りさえすれば、$1000g$ になったのだ。この材料、道具をパン屋が利用し、手順にそってパンを作れば、やはりパンはおよそ $1000g$ になるはずである。
その国では、小麦の値段が高騰しており、支給された小麦をそのまま売った方が儲かるという状況になっていた。そんなとき、パンが $1000g$ よりも軽いと感じた数学者が、数ヶ月にわたりパンの重量を計測していった。その結果、パンの重量は平均で $950g$ となっており、本来の $1000g$ よりも、軽いことがわかった。
このとき、パン屋が不正をしていると主張できる。手順を踏めば平均で $1000g$ になるパンが平均およそ 950 になったのは、パン屋が手順通りにパンを作っていないことを疑える。手順を守って作れば $1000g$ になるという経験（データ）があるから疑うことができる。

6.3.1　サンプルサイズが大きい場合

データと統計モデルを比較する。$180cm$ 以上の割合は、0.0642 であり、モデル $M(171.1)$ の推測値 $P(x > 180) = 0.059$ と数値が近い。また $160cm$ 以下の割合は、0.023 程度であり、統計モデルの推測値 $P(x < 160) = 0.025$ であり、やはり数値が近い。

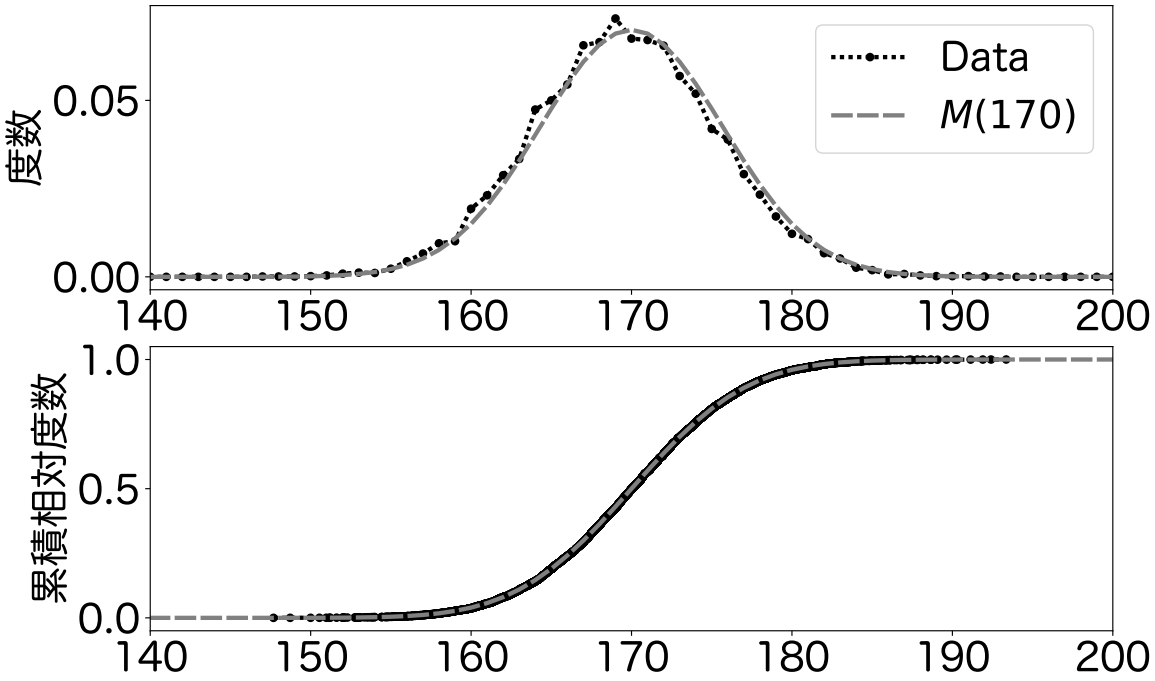


図 6.2　上: 正規分布を含む統計モデル $M(170)$ によりサンプリングされた Data の頻度と、統計モデルの頻度。下: 上と同じデータ・統計モデルの累積相対頻度

ここまでは、$M(171.1)$ を用いて、母集団を推測した。統計モデル $M(170)$ の代わりに $M(168)$ により推測を行うとデータとの一致具合を確かめる。$180cm$ 以上の人を推測すると $M(168)$ では $P(x > 180) = 0.03$ であり、統計モデル $M(171.1)$ の推測 $P(x > 180) = 0.059$ よりもさらに実際の計測値 0.0642 と乖離している。これは、$M(168)$ では、ピークが平均値の 168 に移動するので、$180cm$ を超える割合がさらに低くなるので、実際の数値から離れる。

一方で、160 以下の人では、$M(168)$ では、$P(x < 160) = 0.08$ 程であり、$M(171.1)$ の推測値 $P(x < 160) = 0.025$ よりも、実際の数値 0.023 から離れている. これも、$M(168)$ では、ピークが 170 よりも小さな値になるので、$160cm$ より小さい人の割合が大きくなるので、予測と実際のデータの不一致度が大きくなる (表 6.1 にまとめておいた)。このように、統計モデルの母数に応じて、現実の予測精度が変化する。

この統計モデルの予測の良さが分かったのは、無作為抽出を繰り返して、サンプルサイズを大きくしたときのデータの分布を得ていることによって、そのデータとモデルとを比較をすることで、$M(171.1)$ が $M(168)$ より良い統計モデルであることを判別できた。

では、データが十分でない場合においても、推測とデータの一致を基準にして、より良い統計モデルを選ぶことはできるのでしょうか？

表 6.1 統計モデルとデータの比較

統計モデル	$P(x < 160)$	$P(x > 180)$
データ	0.023	0.0642
M(171.1)	0.025	0.059
M(168)	0.08	0.03

6.3.2 サンプルサイズが小さい場合

母集団のことをほとんど知らない場合において、統計モデルとデータの比較はできるが、これを元に統計モデルが良いことを検討できない。サンプルサイズ 10 の標本が二つ得られたとします（実際には、コンピュータを使って正規分布からサンプリングした。このデータは母集団から無作為抽出したと考える)。標本は、次の通り。

```
1 sample1 = [162.56944902, 178.42128764, 171.15286336,
      172.2581195 , 160.21499345, 175.35072013, 173.17952774,
      173.73301156, 179.52758126, 178.35924221]
```

表 6.2 統計モデルと小さいサンプルサイズの標本

統計モデル	$P(x < 160)$	$P(x > 180)$	$\bar{X}$
標本 1	0	0	172.8
M(171.1)	0.025	0.059	171.1
M(168)	0.08	0.03	168

$180cm$ 以上の人は、0 人、$160cm$ 以下の人も 0 人、どちらの統計モデルでも推測と一致しているかを推測できない [表 6.2]。標本平均 $\bar{X} = 172.8$ であり、$M(170)$ の母数 170 が $M(168)$ の母数平均 $168cm$ でどちらも同じ程度の差である。サンプルサイズが小さいときには、統計モデルの予測とデータを比較できないことがあるので、予測精度の良いモデルがどれかを決定できないことがある。

6.4 統計モデルとデータの比較 2

6.4.1 モデルの平均を含む信頼区間の個数

実際に、$M(\mu = 170)$ を使って、サンプルサイズを 10 とし、標本を 100 個作ってみると、その標本平均の分布は、図 6.3B である。それぞれの標本に対して、最尤モデル $M(\bar{x}_i)$ を作り、信頼区間を描いたものが図 6.3A である。図 6.3A の 170cm のところにある縦の線は、元の統計モデル $M(\mu = 170)$ の母数平均である。元の統計モデルの母数 170cm を跨いでいる信頼区間の個数はこの図では 96 個ある。コンピュータシミュレーションをする

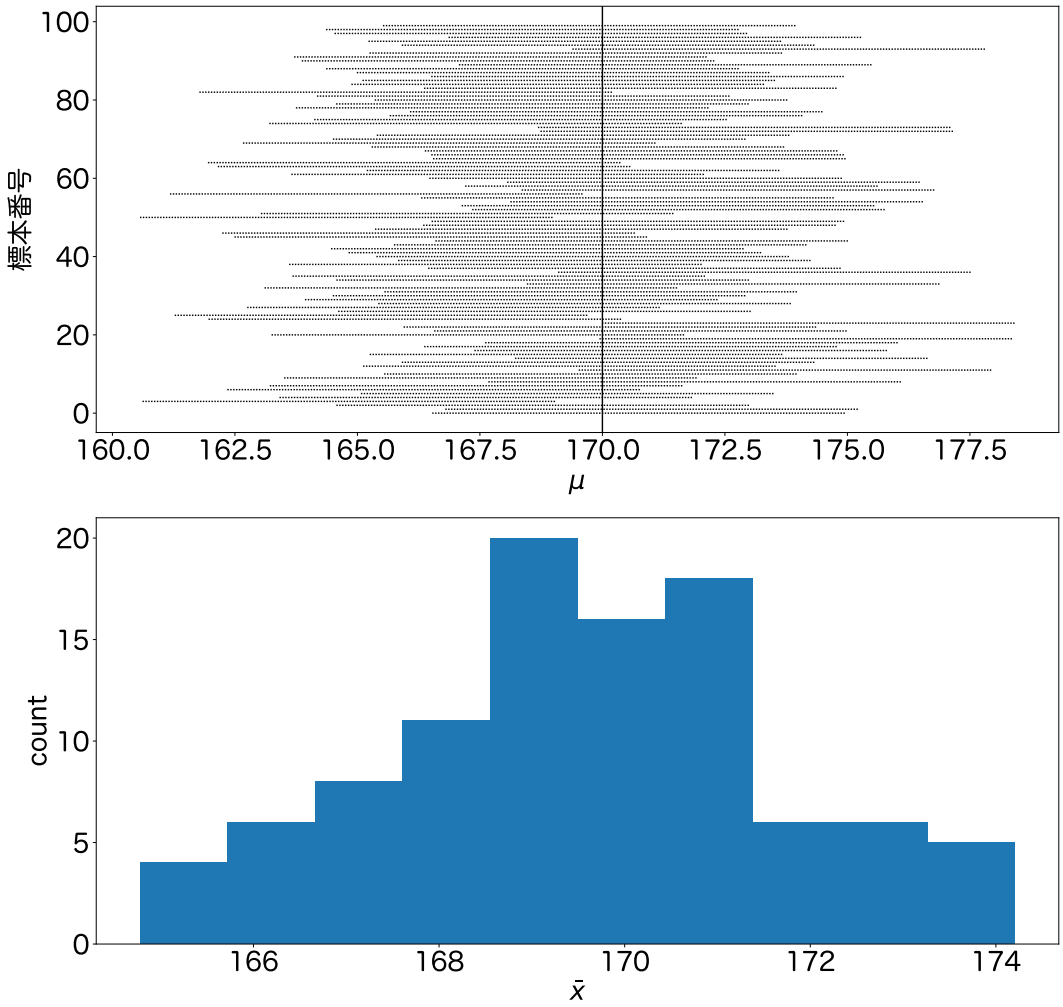


図 6.3　(A) モデルから標本を得て、その標本から信頼区間を計算し、表示したもの。
(B) 標本平均の分布

と、μ を跨いでいる信頼区間の個数はおよそ 95 個である。このことは、信頼区間の定義から明らかである。

■信頼区間は、データをたくさん取ったときに (サンプルサイズが同じ標本をたくさん集めたときに)、その範囲に真値が 95% の確率で含まれるの区間のこと

信頼区間は、データをたくさん取ったときに、その範囲に真値が入る 95% の確率で含まれるの区間のこと[a]。このように解説されることがある。データを元に、統計モデルの母数を決定したときに、信頼区間が得られる。さらに計測を行い標本を作ると、標本の標本平均がこの信頼区間の間に含まれる確率が 95% であることを主張していると考えられる。
一般に、母集団が統計モデルにより、よく推測できるならば、無作為抽出の標本平均が 95% くらいの確率で信頼区間に含まれる。そうではないならば、95% 信頼区間にモデルの母数が含まれる確率は 95% とは異なる値をとることがある。モデルが推測に適さないことは科学においてはよくあり、たった数回の試験を元に構築したモデルにおいて、この解釈を適用するのはやめておいた方が良い。

[a] https://www.slideshare.net/simizu706/ss-123679555

6.5 統計検定量によるモデルの評価

これまでは、統計モデル $M(\mu)$ における信頼区間・棄却域の計算を行った。今回は、p 値を計算する。無作為抽出により得られたデータ $\bar{x}$ がこれ以上偏る確率は、$\phi(z > \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma})$ である。
棄却されるモデルが観測されたデータの平均値 $\bar{x}$ に応じて変化することを視覚的に確認しておく。図はさまざまな $\bar{x}$ を得たときにその信頼区間を描いたものである。この信頼区間の範囲内にある μ であれば、統計モデル $M(\mu)$ は棄却されない。例えば、$\bar{x} = 170$ であれば、$M(170)$ は棄却されない。一方で、$\bar{x} = 165$ あたりであれば、その棄却域は $\mu = 170$ を含まないので、$M(170)$ は棄却される。

6.5.1 データの統計検定量と統計モデルの評価

実際の標本のサンプル $X_1, X_2, \cdots, X_{10}$ について、その標本平均を $\bar{X}$ とする。$M(\mu = 171)$ において、$\bar{X}$ 以上の値が得られる確率を計算する。$\phi(z)$ を標準正規分布とすると、$\phi(z > \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma})$ を計算する。具体的な数値として、$\bar{X} = 172$ モデルの母数を $\mu = 171$ なら、$p = \phi(z > \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}) = 0.289$ であり、$\bar{X} = 169$、モデルの母数を $\mu = 171$ の場合、$p = \phi(z > \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}) = 0.866$ である。統計モデル $M(\mu = 171)$ において、これらの標本平均は、そこまで珍しいものではない。言い換えれば、このモデルにより、母集団について予測ができるかもしれないことを示唆している。

```
1  xbar = 172
2  mu=171
3  sigma = 5.7
```

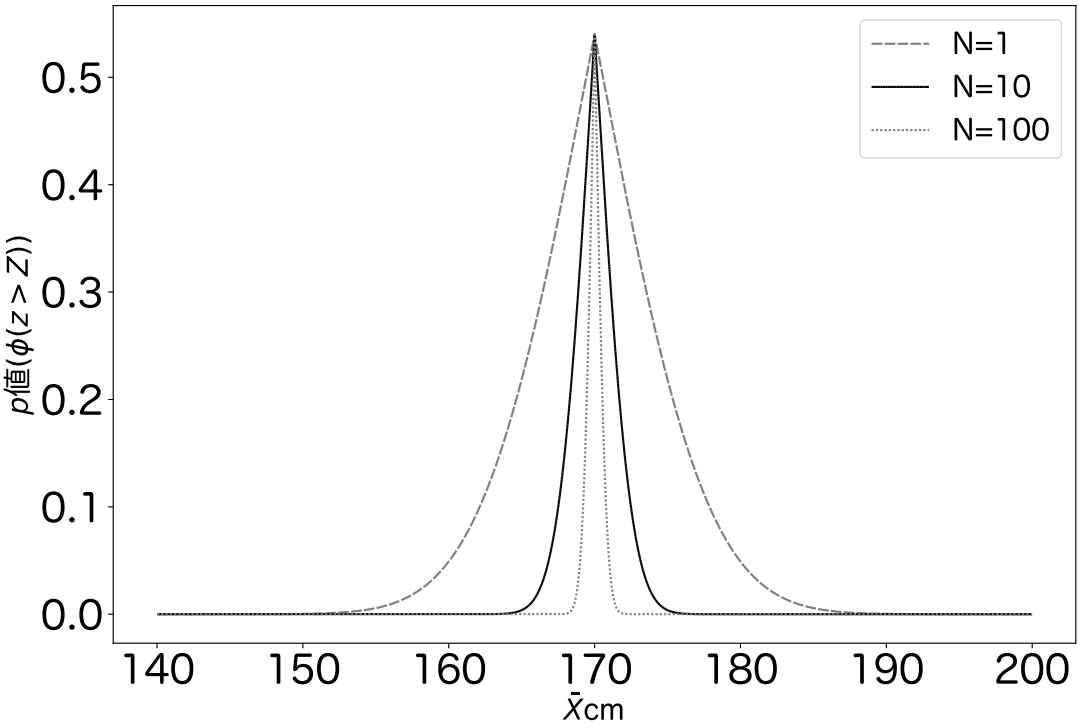


図 6.4　各標本平均 $\bar{X}$ が $M(\mu = 171)$ において得られる確率。$N = 1, 10, 100$ の場合。

```
4    N=10
5    c = np.sqrt(N)*(xbar-mu)/sigma
6    1-norm.cdf(c,0,1)
```

6.5.2　標本平均と p 値

$M(171)$ の上で、各 $\bar{X}$ に対して $p = \phi(z > Z(\mu))$ を計算する。これを図示したのが図 6.4 である。標本平均 $171cm$ をピークに左右対象に p 値が減少している。モデルの母数 μ と $\bar{X}$ が近ければ、p 値が大きく、離れるほど p 値が小さい。言い換えると、得られたデータが統計モデルによって推測できそうであれば、p 値が小さく、離れるほど p 値が小さくなる。このことから、p 値が一つの目安になることが示唆される。

第 7 章

誤差論

誤差論は、計測に対する信頼性を定量的に扱う方法論である。計測に利用する計測機は、同じ対象について計測を行うと、毎回異なる値を示す。複数回の計測のうちその平均値を真の値といい、平均値と計測値の差を誤差という。その誤差は以下の誤差の公理を満たすものとする。

1. 誤差には中心がある。中心を対象に同等にデータが生じる。
2. 絶対値の小さい誤差の方が大きな誤差よりも現れる頻度が高い
3. ある程度以上の大きな誤差は生じない

7.1　標準偏差か標準誤差か

計測のばらつきが小さな計測機はより良い計測機であることが言える。このことを示すために、SD をグラフの中に書くことがある。同一の対象を複数回計測したとき、その平均値はより正確な値へと収束する。

7.1.1　標準偏差

正規分布を含んだ統計モデルを仮定し、そのモデルの上で、予想されるサンプルがおよそ 68% の確率で出現する範囲は、母数分散 σ^2 より以下の範囲になります。

$$[\mu - \sigma, \mu + \sigma].$$

モデルの母数分散は不明な場合、母集団から無作為抽出を行なって集計した標本の偏差 s を計算します。

$$s = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}.$$

このことから、モデルが予測するサンプルが 68% の確率で出現する範囲は

$$[\mu - s, \mu + s].$$

これを図示したものが、図です。言い換えれば、これは、68% 予測区間である。

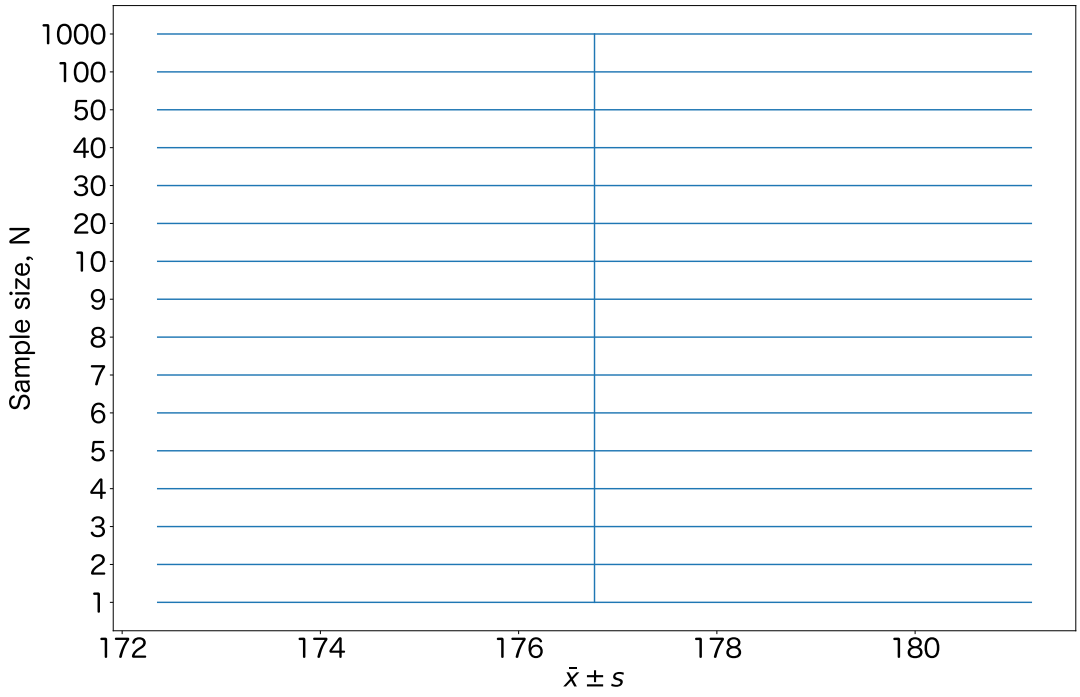


図 7.1　サンプルサイズに応じた標準誤差の広がり

7.1.2　標準誤差

標準誤差 SE は、標準偏差 s をサンプルサイズ N の平方根で割ったものである。

$$SE = \frac{s}{\sqrt{N}}$$

$\bar{x} \sim N(\mu, \sigma^2/\sqrt{(N)})$ であるので、モデルの上で $\bar{x}$ が以下の範囲に出現する確率は、およそ 68.2% である。

$$[\bar{\mu} - SE, \bar{\mu} + SE]$$

よって、μ の値は一般にわからないので、標本平均 $\bar{X}$ を用いて、

$$[\bar{X} - SE, \bar{X} + SE]$$

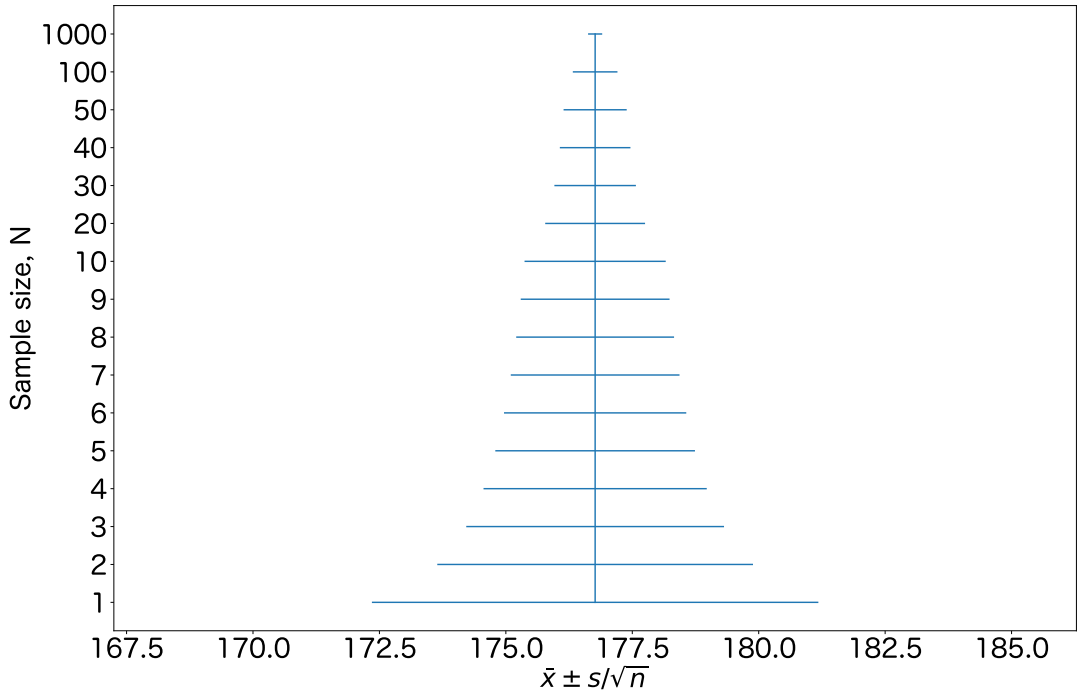


図 7.2　サンプルサイズに応じた標準偏差の広がり

である。統計モデル $M(\bar{x}, s^2)$ からサンプリングした $\bar{X}$ がこの範囲に得られる確率が 68.2% である[*1] [*2]。言い換えれば、SE は、68% 信頼区間である。

[*1] σ が変曲点であるから使ったと思われるが、なぜ、68.2% または、$\bar{X} \pm SE$ の範囲を使ったのかはわからなかった。誤差論の立場では、以下の記述が見つかった [7]。
したがって、おおよそ 70% の確率で誤差の絶対値は σ より小さいことがわかるので、これを測定値の信頼度の目安として用いる。

[*2] 標準偏差に ± を付けるな！：医療論文に多い？ `https://biolab.sakura.ne.jp/mean-sd.html`。まだ読めていない。$\pm SE$ という表記はよろしくないらしい。

第 8 章

統計モデル 2

ここでは、二つの確率変数から得られた確率変数についてその性質を議論する。

8.1 正規分布二つを含んだ統計モデル

次の 3 つを仮定したモデルを正規 2 モデルと呼ぶ。

(1) x_i および、y_i はそれぞれ独立同分布
(2) その分布は、正規分布
(3) 正規分布の母数はそれぞれ $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ とする。

この正規 2 モデルを $M(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ と書く。σ_1, σ_2 を特定の値に設定したモデルを $M(\mu_1, \mu_2)$ または、$M(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2)$ とし、μ_1, μ_2 を特定の値に設定したモデルを $M(\sigma_1^2, \sigma_2^2)$ または $M(\sigma_1^2, \sigma_2^2; \mu_1, \mu_2)$ とする。データから最尤推定を行なった母数を持つモデル $M_{ML} = M(\mu_{1,ML}, \mu_{2,ML}, \sigma_{1,ML}^2, \sigma_{2,ML}^2)$ を最尤正規 2 モデルという。

8.2 分散について事前知識のある場合

分散が先行研究において明らかに成っているとき良い予測を行えるモデル $M(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2)$ について考える。最初に、$\sigma_1 = \sigma_2$ の場合、次に $\sigma_1 \neq \sigma_2$ それぞれにおける信頼区間および検出力を考える。

8.2.1 信頼区間

統計モデル $M(\mu_1, \mu_2; \sigma^2, \sigma^2)$ により、$x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_m$ とサンプリングされたとする。次の統計量を定義する、

$$Z = ((\bar{x} - \mu_1) - (\bar{y} - \mu_2)) \frac{\sqrt{mn}}{\sigma\sqrt{m+n}}$$

ただし、$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ である。Z は、$Z \sim N(0,1)$ となることがわかっている。

信頼区間は、Z の大きさ $|Z|$ によって決まる。$\alpha = 0.05$ とすると、

$$
\begin{aligned}
&|Z| < z_{0.025} \\
&\rightarrow |(\bar{x} - \mu_1) - (\bar{y} - \mu_2)| \frac{\sqrt{mn}}{\sigma\sqrt{m+n}} < z_{0.025} \\
&\rightarrow |(\bar{x} - \mu_1) - (\bar{y} - \mu_2)| < z_{0.025} \frac{\sigma\sqrt{m+n}}{\sqrt{mn}}
\end{aligned}
$$

式を展開すると、

$$
(\mu_1 - \mu_2) - z_{0.025}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \bar{X} - \bar{Y} \leq (\mu_1 - \mu_2) + z_{0.025}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.
$$

を得る。統計モデル $M(\mu_1, \mu_2)$ によるサンプリングによって得られた平均値の差の大きさは、右辺よりも小さくなることが、95% くらいの確率でモデル内でおこる。実際に、何度も統計モデルからサンプリングを行なってみると、95% くらいの確率でこの等式が成り立っている。計算機で試してみる。

```
1  mu1 = 10
2  mu2 = 10
3  sigma = 5
4  m=10
5  n=20
6
7  norm1 = norm(mu1,sigma)
8  norm2 = norm(mu2,sigma)
9
10 sample1 = norm1.rvs(size=(m,1000))
11 sample2 = norm2.rvs(size=(n,1000))
12
13 xbar1 = np.average(sample1,axis=0)
14 xbar2 = np.average(sample2,axis=0)
15 A = np.abs(xbar1-xbar2) < norm(0,1).interval(1-0.05)[1]*sigma
       *np.sqrt(m+n)/np.sqrt(m*n)
16 len(np.where(A==True)[0])
```

およそ 950 程度の標本で、不等式が成立していることが確かめられる。
計算機で計算するには、次のコードが使える。

```
1  def tTest(X,Y,sigma):
2      x_bar,y_bar = np.average(X),np.average(Y)
3      M,N = len(X),len(Y)
```

```
4       Z = (x_bar-y_bar)*np.sqrt(M*N)/(sigma*np.sqrt(M+N))
5       p=norm.cdf(Z,0,1)
6       return 1-p
7   tTest(X,Y,5.7)
```

```
1   def rejectRange(X,Y,sigma):
2       M,N = len(X),len(Y)
3       Z = sigma*(np.sqrt(M+N))/np.sqrt(M*N)
4       za,zb= norm.interval(0.95,0,1)
5       return Z*za,Z*zb
6   rejectRange(X,Y,5.7)
```

Z の不等式を変形していくと、次がわかる

$$(\bar{X}-\bar{Y})-z_{0.025}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2 \leq (\bar{X}-\bar{Y})+z_{0.025}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

この式の意味は、何だっけ？？？ TODO

8.2.2 検出力

$M(\mu,\mu;\sigma^2,\sigma^2)$ における統計量 Z が、もう一つの統計モデル $M(\mu_1,\mu_2)$ においての出現頻度を計算する。これは 1 標本のモデルと同様に検出力という。
$M(\mu,\mu)$ において、

$$\frac{\bar{x}-\bar{y}}{U} \sim N(0,1)$$

である。ここで、$U=\sigma\frac{\sqrt{m+n}}{\sqrt{mn}}$ である。また、$M(\mu_1,\mu_2)$ において、

$$\frac{\bar{x}-\bar{y}}{U} \sim N(\frac{\mu_1-\mu_2}{U},1)$$

である。上式の信頼区間 $-z_{\alpha/2} \sim z_{\alpha/2}$ が下式において出現するのは、確率分布が平行移動しているので、その区間を $[A,B]$ とすると、ぞれぞれ

$$A=-z_{\alpha/2}+\frac{\mu_1-\mu_2}{U}$$
$$B=z_{\alpha/2}+\frac{\mu_1-\mu_2}{U}$$

である。この区間に統計量が出現する頻度は、

$$\beta=\Phi(B)-\Phi(A)$$

により計算できる。ここで、$\Phi(x)$ は、標準正規分布の累積分布関数である。

8.2.3　σ が異なるモデルでの検出力

信頼区間は、

$$-z_{\alpha/2} \leq \frac{\bar{x}-\bar{y}}{U} \leq z_{\alpha/2}$$

ここで、$U = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ である。また、$M(\mu,\mu)$ における統計量を Z とすると、$Z = \frac{\bar{x}-\bar{y}}{U} \sim N(0,1)$ でり、また、モデル $M(\mu_1,\mu_2)$ において $\frac{(\bar{x}-\mu_1)-(\bar{y}-\mu_2)}{U} \sim N(\mu_1-\mu_2,1)$ であることから、

$$\begin{aligned} A &= \frac{(a-(\mu_1-\mu_2))}{U} \\ &= (\mu_1-\mu_2)/U - z_{\alpha/2} \end{aligned}$$

同様に、

$$\begin{aligned} B &= \frac{(b-(\mu_1-\mu_2))}{U} \\ &= (\mu_1-\mu_2)/U + z_{\alpha/2} \end{aligned}$$

よって、

$$\beta = \Phi(B) - \Phi(A)$$

である。

8.3　母分散の事前知識がないときの統計モデル

正規モデル $M(\mu_1,\mu_2,\sigma^2,\sigma^2)$ について考える。

8.3.1　信頼区間

t_{m+n-2} を自由度 $m+n-2$ の t 分布上の上側 100α 点とする。言い換えると、t 分布の確率密度関数を p^t とすると、$p^t(T > t_{m+n-2,\alpha}) = \alpha$ となる点 $t_{m+n-2,\alpha}$ である。
このとき、正規モデル $M(\mu_1,\mu_2,\sigma^2,\sigma^2)$ からサンプリングを行った $x_1,x_2,\cdots,x_n,y_1,y_2,\cdots,y_m$ について、

$$|t_0| = \frac{|\bar{x}-\bar{y}|}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

が成り立つ。ただし、

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^n (x_i-\bar{x}) + \sum_{i=1}^m (y_i-\bar{y})}{n+m-2} \\ \bar{x} &= \frac{1}{n}\sum_{i=1}^n x_i \\ \bar{y} &= \frac{1}{m}\sum_{i=1}^n y_i \end{aligned}$$

である。以上より、信頼区間は、

$$|t_0| \leq t_{m+n-2,\alpha/2}$$

である。

8.3.2 検出力

検出力の計算には、σ が事前研究により明らかでなければならない。ここでは σ が特定できているモデルにおける検出力を調べる。統計モデル $M(\mu_1, \mu_2)$ において、次の統計量が非心 t' 分布に従うことがわかっている。

$$t_0 \sim t'(n+m-2, \lambda)$$

$\lambda = \sqrt{\frac{nm}{n+m}}\Delta$、$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$ であり、$t'(n+m-2, \lambda)$ を自由度 $n+m-2$、非心パラメータ λ の非心 t 分布と言う。モデル $M(\mu, \mu)$ での信頼区間は、$|t_0| < t_{n+m-2,\alpha/2}$ だったので、検出力は、$P^{t'}$ を非心 t' 分布の確率密度関数だとすると、

$$\begin{aligned} 1-\beta &= 1 - P^{t'}(|t| \leq t(n+m-2, \alpha/2)) \\ &= P^{t'}(t \leq -t(n+m-2, \alpha/2)) + P^{t'}(t \geq t(n+m-2, \alpha/2)) \end{aligned}$$

である。ここで、確率密度関数に関する近似式

$$P^{t'}(t' \leq w) \approx \Phi\left(\frac{w(1-\frac{1}{4\phi}) - \lambda}{\sqrt{1+\frac{w^2}{2\phi}}}\right)$$

が成り立つ [[8]][*1]。ただし、Φ は、標準正規分布の累積分布関数であり、$\phi = n+m-2$ である。

[8] より、例題を解いてみる。$\alpha = 0.05, n = 10, m = 8, \mu_1 = 5.6, \mu_2 = 5.0, \sigma = 1.0, n+m-2 = 16, \lambda = \sqrt{n \times m/(n+m)} \times \frac{\mu_1 - \mu_2}{\sigma} = 1.265$ とする。検出力は、

$$\begin{aligned} 1-\beta &= P^{t'}(t \leq -t(16, 0.05)) + P^{t'}(t \geq t(16, 0.05)) \\ &= P^{t'}(t \leq -2.12) + P^{t'}(t \geq 2.12) \\ &= P^{t'}(t \leq -2.12) + (1 - P^{t'}(t \leq 2.12)) \\ &\approx \Phi(\frac{-2.12(1-1/(4\times 16)) - 1.265}{1+(-2.12)^2/(2\times 16)}) + 1 - \Phi(\frac{2.12(1-1/(4\times 16)) - 1.265}{\sqrt{1+2.12^2/(2\times 16)}}) \\ &= \Phi(-3.139) + 1 - \Phi(0.770) \\ &= 0.222 \end{aligned}$$

数値計算

数値計算でも確かめてみる。手順は、次の通りである。モデル $\mathrm{A}M(5.6, 5.6; \sigma^2 = 1.0^2)$ からサンプリングを $N = 10^4$ 回行い、統計量 $t0$ を計算する。その 95% 信頼区間 $[A, B]$

[*1] 私は証明を読んでいない。いつか読む。

を求める。モデル $\mathrm{B}M(5.6, 5.0; \sigma^2 = 1.0^2)$ からサンプリングを N 回行い、統計量 $t1$ を計算する。$t1$ の中で、信頼区間 $[A, B]$ の外側にあるものが検出力 $1 - \beta$ である。

```
1  n=10
2  m=8
3  mu1=5.6
4  mu2=5.0
5  sigma = 1.0
6  phi = n+m-2
7  N = 10000
8
9  sample1 = norm(mu1,sigma).rvs(size = (n,N))
10 sample2 = norm(mu1,sigma).rvs(size = (m,N))
11 xbar = np.average(sample1,axis=0)
12 ybar = np.average(sample2,axis=0)
13 S2 = (np.sum((sample1-xbar)**2,axis=0)+np.sum((sample2-ybar
       )**2,axis=0))/float(n+m-2)
14 S = np.sqrt(S2)
15 t0 = (xbar-ybar)/(S*np.sqrt(1/n+1/m))
16 A,B = np.quantile(t0,q=[0.025,0.95+0.025])
17
18 sample1 = norm(mu1,sigma).rvs(size = (n,N))
19 sample2 = norm(mu2,sigma).rvs(size = (m,N))
20 xbar = np.average(sample1,axis=0)
21 ybar = np.average(sample2,axis=0)
22 S2 = (np.sum((sample1-xbar)**2,axis=0)+np.sum((sample2-ybar
       )**2,axis=0))/float(n+m-2)
23 S = np.sqrt(S2)
24 t1 = ((xbar-ybar))/(S*np.sqrt(1/n+1/m))
25
26 print(len(np.where((t1 < A ) | (t1 >B))[0])/N)
```

0.22 に近い値が得られる。

第 9 章

モデルを使った研究の進め方

9.1　指針

いくつか言葉を定義する。

定義 9.1.1. 標本に対してデータを見て、適合したまたは適合するように構築したモデルを、適合モデルと言う。標本を見る前に、既存の研究で構築してあったモデルを、予測モデルと言う[*1]。適合モデルには標本のデータや統計量を含んであり、予測モデルには、これまで得られた標本の統計量などが設定してあるが、現在注目している標本のデータは含まれていない。

図 9.1 には、統計モデルを使った研究の概念図を示した。なんらかの生物学的問いに対して、それを観察するための実験デザインを構築する。このとき、既存の研究成果を確認し、予測モデルを構築する。
次に、現象からデータを取得し、その予測モデルで標本を予測できるかを検証する。データの取得後、予測モデルをデータに合わせるように変更してはいけない。データに合わせたモデルで標本の乖離を調べた場合、それは適合モデルの適合具合を示したことになる。研究者らが普段使っている仮説検定は、ある一つのモデルが適合しないことを示そうとした途中の結果である。p 値以外にも様々な指標を使って予測モデルにより予測が可能であるかを示さなければならない。
予測モデルの予測と標本が乖離していようがいまいが、標本に適合するモデルを探索する。言い換えれば、どのような分布関数が適合するのかやその分布形での母数やパラメータを推定する。このモデルはデータに最もよく適合したモデルであるだろうから、適合モデル 1 と呼ぶ。現段階では、適合モデルが予測に適しているかは不明である[*2]。
構築した適合モデル 1 と生物学的な問いを元に新たな実験計画を設計し、実験を行う。ここでも先ほどと同様に、既存の研究成果を元に予測モデルを立て、予測可能かを検討して

[*1] 予測モデルと定義するのは良くないかもしれない。すでに定義された言葉であるので。
[*2] 標本を分割して、データの適合に使うものと予測に使うデータにしておけば、予測可能な程度を測ることもできる

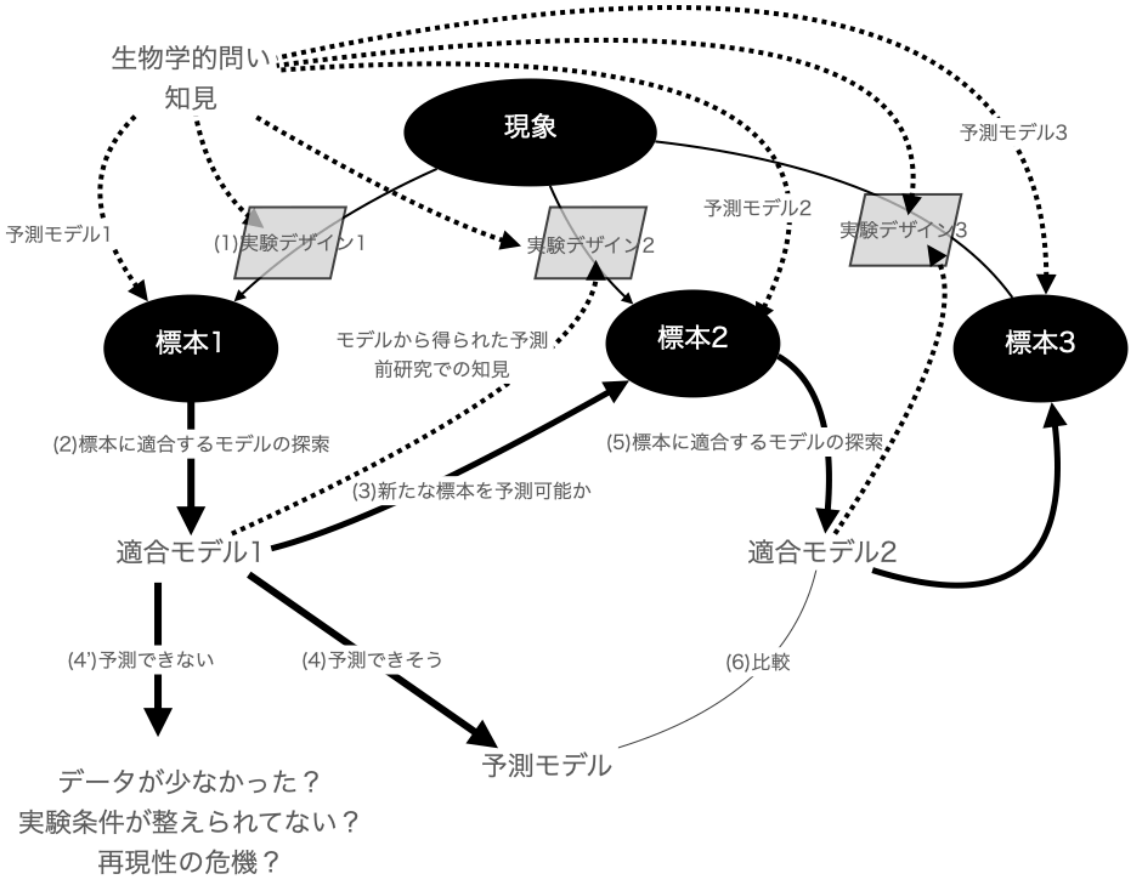


図 9.1　統計モデルを使った研究のフロー図。

も良い。ここでの予測モデルは適合モデル 1 と一致しているか極めて類似したモデルである。もしも標本 2 が適合モデルで対象にしていた特徴量を計測しているならば、適合モデル 1 の予測性能を測る。予測できるならば、モデルの予測可能性が示される。予測モデルにより標本 2 を予測できないならば、その理由は次のものが想定できる。

1. データが少数だったため、良い適合モデルを選べなかった（分布形の推定までできなかった等）
2. 同様の実験条件を整えることができなかった (影響を受ける要因があったことを発見できた)
3. 再現できない実験だった（実験デザインの見直し・再現できない研究の積み重ねを阻止）

標本 2 に対する適応モデルを構築する。予測モデルと適応モデル 2 の差異を調べる。以上のプロセスを繰り返すことで、生物学の予測可能な領域を増やしていく。
予測モデルにより予測可能かを検証することそして、標本に対して適合したモデルを探索

することこの繰り返しにより研究が進む。

■検定はデータを見てから・見ないで行うべき

(1) データを見ないで、構築済みのモデルとの乖離を調べるという方針がある。これはすでにわかっていることから、どのような現象が生じるのかをまとめ、モデルを構築する。このモデル構築は、データを見る前の、実験を計画する時点で構築できる。このモデルと得られたデータとの乖離を検定により調べることで、既存の知識を元にした予測モデルの予測性能を検証している。予測モデルとデータの乖離が発見できたのなら、適合モデルを調べることで、新しいモデルの方が良いのかを議論できる[a]。ただし、データにどのような値が含まれているのかは一度見ることになる。
(2) データを見てから、モデルを構築して検定を行うという方針。こちらは、データを見ながら適合するモデルを構築する。言い換えれば、適合モデルを構築するという方針である。こちらの方針でも、$p < \alpha$ であればよしという解析を行なっている。
(2) を行なったのに、(1) の手順でモデルを構築して $p < 0.000$ を得たという報告はしてはいけない。逆も同様にしてはいけない。嘘はつかない方が良い。

[a] 一般に、ここまでやらない。やった方が良い

■適合モデルを探しても、データとモデルの乖離 ($p < \alpha$) を報告するだけ

$p < \alpha$ を見つけるとなんだかよし！と言いたくなる。

9.1.1 どれが科学的成果だろうか

1 つ目の試験での成果は次のことになる。

1. 標本があるモデルに適合しなかった (p 値・モデルの予測とデータが一致しないことを示す証拠)
2. 標本に適合する適合モデル（モデルの分布形・母数）

2 つ目の試験における成果は次のことである。

1. 標本 2 を適合モデル 1 が予測可能か。適合モデル 1 が予測にも使える。研究結果の予測可能性を確認。
2. 適合モデル 1 が標本 2 を予測できなかった理由の探索
3. 標本 2 の適合したモデル。適合モデル 2 と、適合モデル 1 の差異。

これら以外にも、モデルから予想される生物学的な情報も科学的成果である。例えば、正規分布でそれぞれ予想できそうな群 A,B があったとして、それらのモデルの平均値間の距離から何が言えるのかを考えることが求められる。

■生物学者は $p < \alpha$ に興味がある

生物学における論文の多くは適合モデルを見つけようとしたのか、予測モデルで予測ができることを評価したかったのかという違いをはっきりと明記しない。一般化線形モデルを使っていれば、適合モデルを探索したかったのだろうとあたりをつけることができる。2 群の t 検定を行なっているなら、これまでは正規モデルが適合モデルであり、このモデルとデータを比較したのだろう。
正規分布的ではないデータが得られたなら、これまでの適合モデルと異なる結果が得られたという点は報告するべきである。このことを報告せずに p 値を報告する。生物学はデータの特徴に関する情報を捨てている。

9.1.2　適合しすぎな適合モデルは良いモデル？

1. 汎化性能 あるデータに対して適合させたモデルを元に、それ以外のデータに対する予測性能。
2. 過学習したモデル　モデルがデータに対して当てはまりが良すぎることで、汎化性能が著しく低下した状態のモデル。過度に適合したモデルは未知のデータに対する予測性能が落ちることが予想される。
3. モデルの表現力 モデルがデータに適合できる度合い。表現力が高いほど過学習しやすいモデルになる。

9.2　ダメモデルを羅列する研究例

図 9.2 に従来の研究フローの概念図を示した。標本に当てはまらないモデルを構築し、標本とモデルが乖離していることを示す。これには、$p < \alpha$ であることが使われる。α には 0.05 が多くの場合、採用される。このことに根拠はない。標本の分布形に関わる情報は消失し、標本に当てはまるモデルの探索を行わない。情報を捨てることによって、各研究が独立して実行される。
このような研究は行わない方が良い。

9.3　2 群に対する研究

標本が二つの群、A 群 B 群となっており、これらの違いを定量的に求めると言う問題がある。例えば、ある生物のオス、メスにおいて、体長が異なることを知りたいと言う問題である。これまでは 2 群の t 検定などを使い、$p < 0.05$ であれば、違うという判定を行っていた[*3]。これは、正規 2 モデル $M(\mu, \mu, \sigma^2, \sigma^2)$ における統計量の予測と乖離していることを示しているにすぎない。

[*3] これだけだと何がどう違うのかは言及できてない

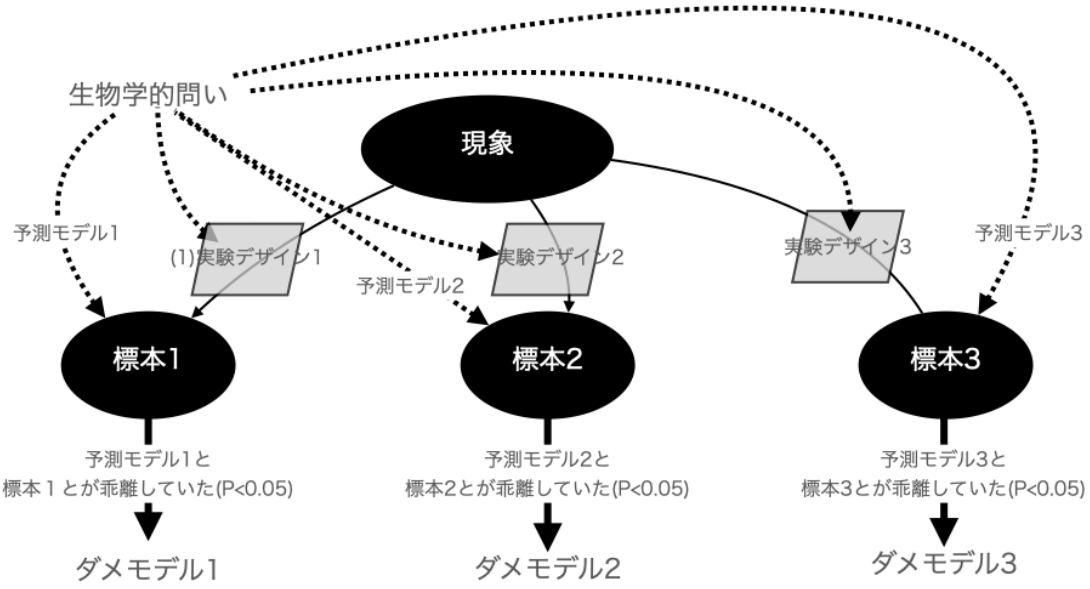


図 9.2　統計モデルを使った従来研究のフロー図。

母数が同じモデルでは統計量が適合していようがいまいが、次の問いは、標本に適合したモデルはどのようなもかである。そこで、適合する統計モデルとその母数を探索する。ここでは、対数尤度や AIC などを使い、単に相対的に当てはまりがそれなりに良いモデルを探す。最適なものが将来的に良い予測をするわけではないことに注意しなければならない。

次に、モデルの性質を調べる。母数の異なる二つの正規モデル M_a, M_b が適合したならば、そのモデルの差異を調べる。統計量の出目の意味での類似度を示す検出力や、中心からの距離を分散で規格化した効果量などを求める。例えば、効果量 d が 0.001 程度であれば、M_a, M_b の中心は極めて近く、モデル M_a でどちらの群のデータも当てはまりが良いと判定できる。この計算により、二つのモデルが異なる予測をするのかを明らかにする。以上により、異なるモデルが標本に適合していることが示せる。このことから、二つの群は異なる性質のモデルで予測した方がいいという示唆が得られる。

次の試験において、生物学者はその生物が標高によって体長が異なるのではないかと言う問いを立てる。さらに、オス・メスでその違いが顕著であるのではないかなどと考え、調

査方法を構築する。新たに得た標本において、オスメスそれぞれが M_a, M_b により標本を予測できているのかを調べる。こうすることでモデルの予測能力と、一つ前の試験における研究の予測可能性を確認できる。予測できないならば、なぜ予測できなかったのかを問い直すことになる。
さらにこの標本に対する適合モデルなどを調べる。標高データをモデルに組み込むことでより良い予測ができると考え、標本に適合するモデルを探索する。線形回帰モデルや一般化線形モデルなどが候補になる。

9.3.1　アヤメ (iris) に関する推論

公開されているアヤメのデータを使って、研究の進め方について検討する。このアヤメのデータでは 3 種 (setosa,versicolor,virginica) のがく片の幅、がく片の長さ、花弁の幅、花弁の長さのデータが記録されている。データサイズは、150 で、種によって 50 ずつ記録されている。Python のライブラリ sklearn から簡単にデータを呼び出せる。

```
1 from sklearn import datasets
2 iris = datasets.load_iris()
```

9.3.2　アヤメのがく弁の幅を予測するモデル

ここでは、アヤメという植物が見つかったときどのようにモデルを構築するかを考える。アヤメについてその種が 3 種類の分類が行われる前で、1 種類であると考える。アヤメという植物を発見し、無作為に 150 個体を採集、がく弁の幅を計測したとする。
我々は、がく弁の幅を予測するモデルを構築したい。この目的を達成できるかはわからないが、一手目に行うのは、データに適合するモデルを探索することである。データをみると、ある点を対称に同じくらいの数のデータがあることがわかる。このことから、正規モデルが候補にあげられる。データから平均と分散を求めると、$\bar{X} = 3.05, \sigma = 0.434$ であった。このことから、最尤正規モデルを構築する $M_a = M(3.05, \sigma^2 = 0.434^2)$ である。最尤正規モデルがデータに適合しているかをみる。このモデルの予想では、μ より大きいまたは小さいデータの個数は半数程度である[*4]。また、$\mu - \sigma \sim \mu + \sigma$ の中にあるデータは 68% 程度である。表 9.1 がデータが予測にあっているかを示している。どちらの指標も予想に合っている。また、$> \mu$ と $\mu <$ となるデータの個数の比率も 1 に近い 1.24 であった。これはモデルがデータに適合していることを示している。

表 9.1　aa

	$< \mu$	$> \mu$	Data Size	$< \mu$ Rate	$> \mu$ Rate	$< \mu / > \mu$	$\mu - \sigma \sim \mu + \sigma$
All	83	67	150	0.55	0.45	1.24	0.673

[*4] 中央値と平均値が十分近ければ割合を調べなくてもいいかも

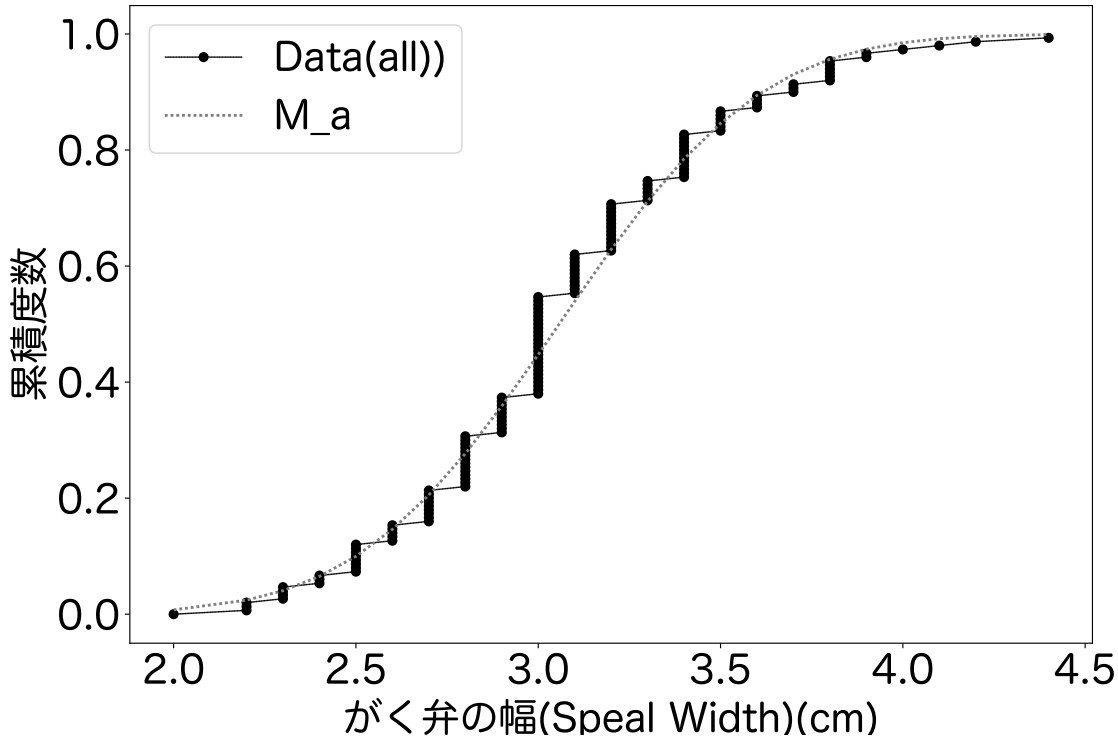


図 9.3　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M_a

図 9.3 は、データの累積度数とモデルの累積度数を示している。データとモデルの累積度数に関する予測がそれなりに一致していることがわかる。このこともモデルがデータに適合していることを示唆している。

9.3.3　アヤメの分類の細分化

アヤメの分類を細分することになり、virginica とそれ以外とすることになった。これら二つのグループにおいて、がく弁の幅はこれまでに作ったモデル M_a により予測することができるだろうか。新たにデータを取得し、モデルの予測とデータを比べることでモデル M_a の予測性能を測ることができる。今回はデータを得るのが難しいので、もう一度同じデータを使い、モデルのデータに対する予測性能を測る[*5]。表 9.2 がモデル M_a の予測に対する実際のデータの性質を示している。virginica とそれ以外はモデル M_a によって十分予測できていない。モデル M_a の平均母数 μ より小さなものと大きなものの比が 1 より離れた値を取っている。また、データの 68% が見つかるという予測をする区間には、68% とはかけ離れた割合のデータが存在する。図 9.4 には、モデル M_a とデータの累積

[*5] モデル構築に使ったデータを再び使うので、モデル M_a の予測の良さを測れていない

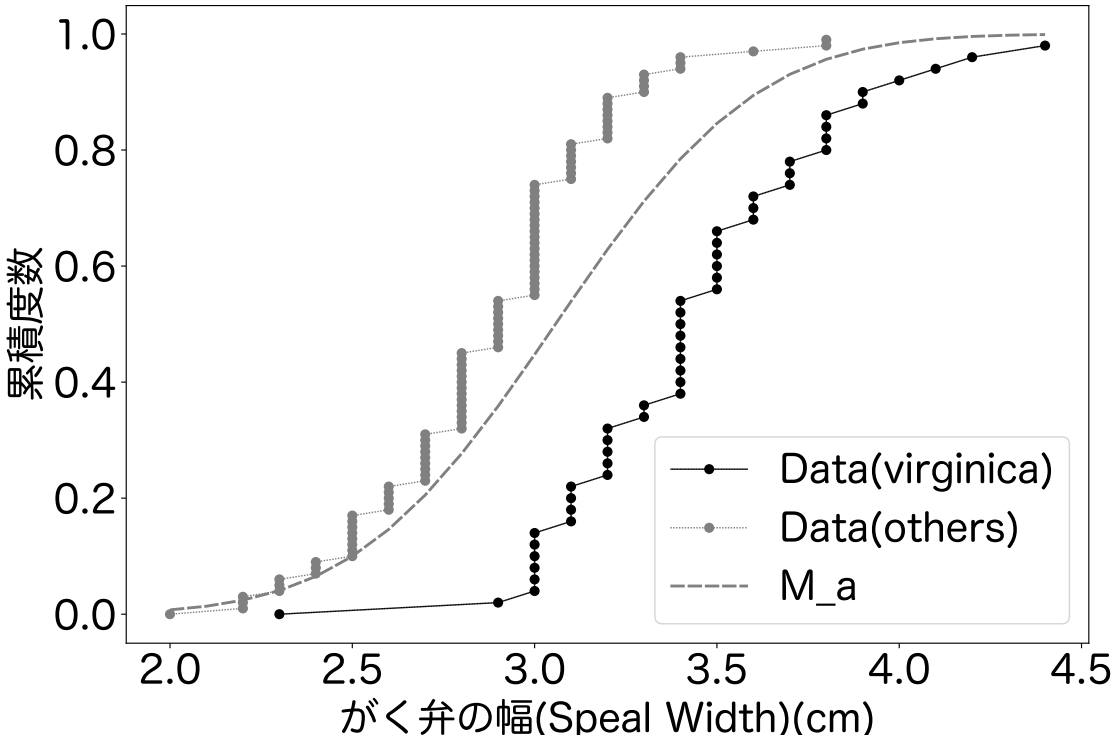


図 9.4　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M_a

度数を表示している。モデル M_a の累積度数の上にデータの点がないことからも、モデルとデータが乖離していることが示唆される。

表 9.2　アヤメ (virginica とそれ以外) のがく弁の幅に関するデータの割合。モデル M_a の平均値を μ としたとき、μ より小さなデータの個数と割合 ($< \mu$、$< \mu$ Rate)。μ より大きなデータの個数 ($> \mu$) と割合 ($> \mu$Rate)。$< \mu$ と $> \mu$ の割合。$\mu-\sigma \sim \mu+\sigma$ の中にあるデータの個数 (68%) と割合 (68%Rate)。

	$< \mu$	$> \mu$	$< \mu$ Rate	$> \mu$ Rate	$< \mu / > \mu$	68%	68%Rate	Data Size
virginica	8	42	0.16	0.84	0.19	27	0.54	50
others	75	25	0.75	0.25	3.00	74	0.74	100

モデル $M_a = M(\mu = 3.05, \sigma^2 = 0.434^2)$ における統計検定量も利用する。次のことがわかっている。

$$Z = \frac{\sqrt{N}(\mu - \bar{x})}{\sigma} \sim N(0, 1)$$

統計検定量 Z を計算した結果が表 9.3 である。Z の絶対値は、2 より大きくモデルとデータが乖離していることがわかる。
これらのことから、モデルの改訂をした方が良いことが示唆される。

表 9.3　統計検定量 Z

	$\bar{x}$	σ	Z
virginica	3.43	0.38	-6.03
others	2.87	0.33	4.27
M_a	3.05	0.434	-

表 9.4　新たなモデル M_v, M_o による予測とデータの適合具合

	$<\mu$	$>\mu$	$<\mu$Rate	$>\mu$Rate	$<\mu/>\mu$	68%Rate	Sample Size
0	28	22	0.56	0.44	1.27	0.72	50
1	46	54	0.46	0.54	0.85	0.72	100

以上のことは論文においては、統計統計量より偏った値が得られる確率（p 値）または $p < 0.05$ が報告される。すでに議論した通り、p 値だけでモデルとデータの乖離を検証すると、モデルの予測性能が過度に低いと判定されることがある。さまざまな指標を元にモデルの予測性能を測るべきである。

9.3.4　新たなモデルの構築

virginica と others に適合するモデルをそれぞれ構築する。累積分布はどちらも正規分布的になっている。M_a を構築するときと同じようにそれぞれの平均と分散を求め (表 9.3 の通りである)、データが平均に対して対称に分布していること、$\mu - \sigma \sim \mu + \sigma$ の間にあるデータが 68% 程度であるかを確かめる。
表 9.4 には、新たなモデル M_v および M_o の予測とデータの適合具合を示している。どの指標もモデルの予想と一致しており、モデルがデータと適合していることを示唆している。
図 9.5 はモデルの累積度数とデータの適合具合を示している。それぞれのモデルがそれぞれデータをよく予測していることがわかる

9.3.5　更なる生物学的な種の細分化

アヤメの others についても細分化することになった setosa,versicolor である。これらについて予測モデル M_v は良い予測をするかを調べ、予測できないと判断したのなら、モデルを構築し直す。
これまでのアヤメに関する議論はなかったことにして、まっさらな状況でモデルを使う方法について考える。状況設定として、これまで同一だと分類されていたアヤメを 2 種類に分けるということになった。この 2 種類のがく弁の幅ついてこれまでと同じモデルを使って予測できるだろうか。この 2 種類を予測するモデルとして、正規 2 モデル $M_2 = M(\mu, \mu, \sigma^2, \sigma^2)$ とし、M_2 によりがく弁の幅が予測できるかを考える。同じアヤメ

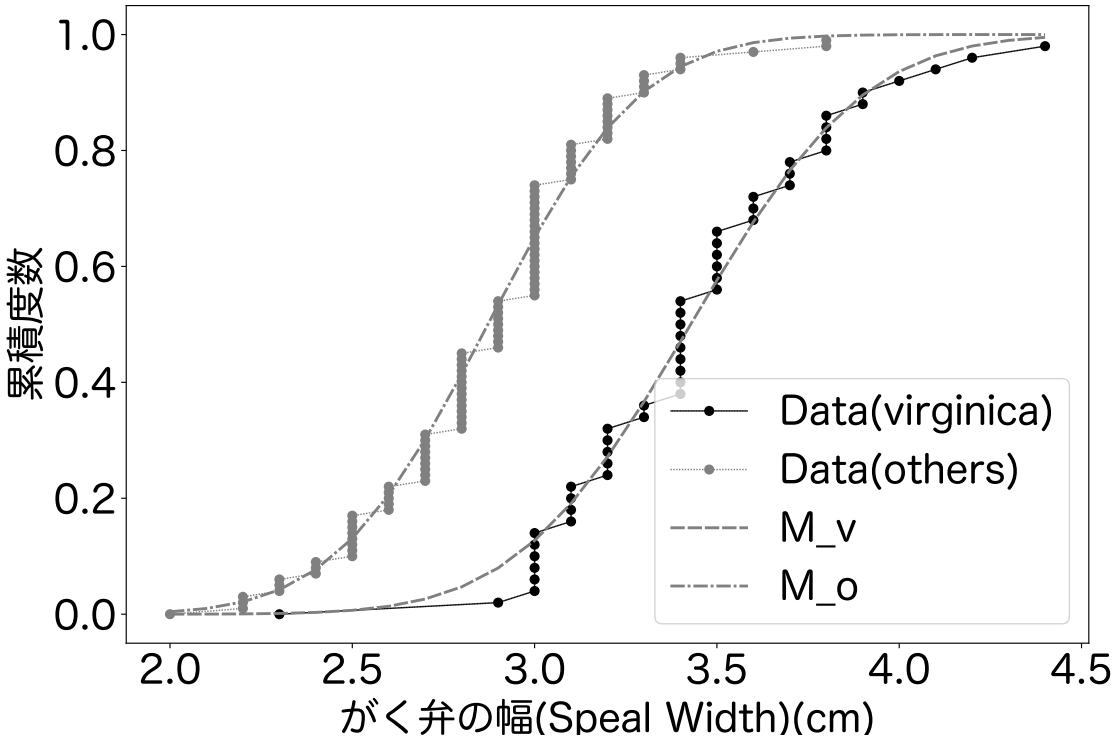


図 9.5　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M_v, M_o

という分類に属するのだから、がく弁の幅も同じ程度だろうと考え、μ が同一のモデルを選んだ[*6]。また、正規分布を仮定したのも、これまでアヤメのがく弁は正規モデルで予測していたからである[*7]。

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

ここで、$s^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$ である。パラメータについては、表 9.5 にある通りである。これを元に計算すると、$t = 9.45$ 程度である。このことから、モデル M_2 では予測しにくいと考えられる。

[*6] 頻度論を元にした生物学の教科書では、母数を同一にするのは、そうでないことを示すため (背理法) と説明されることがある。本書の方針とは異なる。本書では、これまで同一だと思われていたアヤメの分類が増えたので、それまではがく弁についてもある統計モデル M で予測されていたと言う前提があったとする。実際にデータと適合モデルが既存研究において提案されてたとする。母数が同じと考えられていたモデルで予測可能であるかを調査するために、モデルの統計量に関する予測を利用する。この論証は背理法ではない。

[*7] と言う設定の上で推論を進めていく。実際には、既存研究でどのように扱われていたかを調べる必要がある。もしも指数分布で推測していたなら、分布形や統計検定量を変える。既存の研究成果で予測ができるのかを調べるためである。既存の予測が当たらなかったら、新たなモデルを提案する。既存のモデルが予

表 9.5　setosa、versicolor のがく弁の幅のデータの統計量

	Ave.	Sigma	N
setosa	3.43	0.38	50
versicolor	2.77	0.31	50
setosa and versicolor	3.10	0.48	100

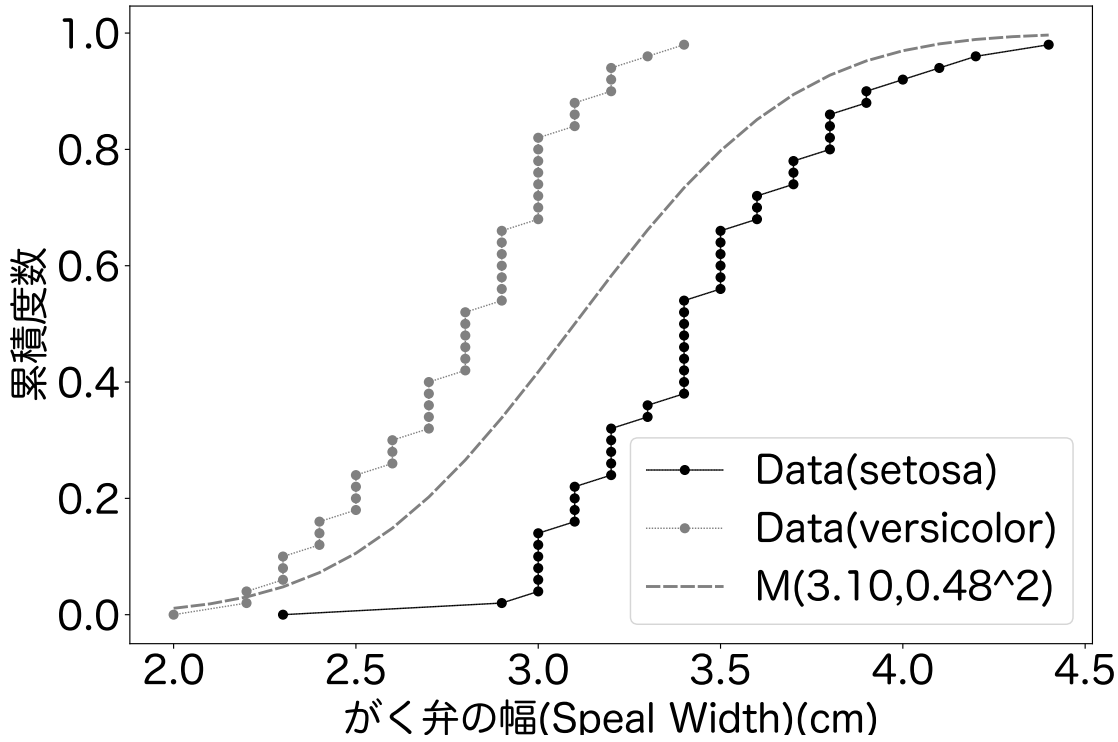


図 9.6　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M

最尤モデル $M = M(3.10, 0.48^2)$ のデータへの適合具合を見る。累積度数はモデルと離れていることがわかる (図 9.6)。以上から、M_2 ではデータと適合していないことがわかる。二つのモデルを構築し、それぞれが versicolor と setosa のデータに適合するかを調べる。面倒なので勝手にやってくれ。

種が細分化されたならば、モデルも更新すべき？

種は細分化されたが予測モデルは同一のものを使ってもいいという判断を下すこともある。言い換えれば、生物学的な特徴を元に種を細分化したが、別の特徴は同じモデルで予測できるということもある。

測には適さないことや新しいモデルの提案は科学的な成果である。報告するべき事柄である。

練習問題: ペンギンの身長

3 つの種のペンギンの身長・体重・くちばしの長さ・性別・観察年度・観察した島などの観測データが公開されている (`https://github.com/mcnakhaee/palmerpenguins`)。このデータを使って、種によって身長を異なる統計モデルを使って推測した方が良いのかを考察せよ。また、性によって異なるモデルを使った方が良いのかを考察せよ。

9.4　ダニの個体数

grouseticks はライチョウのひなの頭についているダニの個体数をスコットランド (北緯 57 度 7 分、西経 3 度 19 分) で調べたデータである。Python でも呼び出すことができる。

```
1  import statsmodels.api as sm
2  data = sm.datasets.get_rdataset("grouseticks", "lme4").
       data
3  ticks = data['TICKS'].values
```

■カウントデータだからポアソン分布

カウントデータならポアソン回帰で![a]

- もしこの観測データ (縦軸) がカウントデータだったら?

まずい点: 等分散ではないに直線回帰?

まずい点: モデルによる予測は「負の個体密度」?

カウントデータだからポアソン分布を仮定しよう。その理由は、正規分布などであれば、負の個体数が出てくるので、現象をよく予測できていないと言うことが挙げられている。ここでいうカウントデータはある一定時間に起きた事象の回数のことではなく、種子の個数のことである。

このことは本書では推奨しない。予測が現象を反映していないことは、大きな問題ではない。例えば、身長の分布の推定に正規分布を使った。正規分布の推定では負の身長は、非常に低い確率で出現する。これは物理的にあり得ない。では、この仮定がダメなのかと言えばそんなことはない。あり得ない部分は無視して、予測したいことが予測できれば問題にならない。

また、大学入試の共通テストの得点分布はおよそ正規分布で推測可能な形になっている。このことからテストの得点は正規分布すると考えてはいけない。何も考えずにテストを作って、ある集団に解かせてみると、その分布は正規分布とは程遠い。授業の得点を予測するのに正規分布は仮定しないほうがいいことがわかる。状況に応じて予測できそうなモデルを構築する必要があることを示唆している。

カウントデータなのだからポアソン分布を仮定することは本書では勧めない。データに適合しないならば、より多くの適合しそうなモデルを探索してみることを勧める[b]。

[a] P.19 https://kuboweb.github.io/-kubo/stat/2011/y/skubostat2011y.pdf
[b] 分散も平均も変化するデータなので、正規分布を仮定したモデルだと計算が破綻するのかもしれない。このことを私は調べてない。

付録 A

数理統計学

データの出現頻度を近似する式である確率密度関数、累積分布関数について説明し、様々な形の確率密度関数について説明する。さらに、特定の分布に従う確率変数が、その分布関数から生成された確率変数であることを確かめる方法について説明する。最後に、モデルの確率変数への当てはまりの良さの相対的な指標である尤度を導入し、尤度を最大にする母数を推定する方法を説明する。さらに、モデルのパラメータの数に対するペナルティを導入した指標の AIC を導入する。

A.0.1 確率密度関数

A.0.2 累積分布関数

aaa

A.0.3 相補累積分布関数

1 から累積度数を引いたものは、相補累積分布関数と呼ばれ、ある値よりも大きな値をえる確率示し、数式では、

$$1 - F(x) = f(X > x) \tag{A.1}$$

$$= \int_x^\infty f(z)dz. \tag{A.2}$$

図 A.1(c) に図示した。累積分布関数と相補累積分布関数のどちらかを表示するかは、分野によって異なる[*1]。

[*1] 生物学の分野などでは、より大きな値を得る確率を重視することがあるので、累積分布関数よりも、相補累積分布関数が好まれることがあるように私は感じている。

A.1 確率変数

A.1.1 確率変数がある分布関数に従う

確率変数 x が、ある分布関数に従うとは、

A.2 正規分布

正規分布の確率密度関数は、

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{A.3}$$

ここで、μ, σ^2 は、正規分布のパラメータで、それぞれ母数平均、母数分散です。母数平均は最も出現頻度の高い数値を表しており、この値を中心にし、対象に分布が広がります。言い換えれば、$\mu - a$ と、$\mu + a$ の出る確率は同程度になります。母数分散は、数値のまとまり具合を示します。σ が大きくなるほど、μ の近くの数値が出現する頻度は小さくなり、より離れた場所での出現頻度を高くします。正規分布関数に確率変数が従うことを $X \sim N(\mu, \sigma^2)$ とかく。
正規分布においてその母数を $\mu = 0, \sigma = 1$ とするとき、標準正規分布といい、$N(0,1)$ で表す。確率変数 Z が標準正規分布に従うとき、その確率密度関数は

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) \tag{A.4}$$

であり、図 A.1(a) である。標準正規分布の累積分布関数は、

$$\Phi(x) = p(X < x; 0, 1) \tag{A.5}$$

$$= \int_{-\infty}^{x} \phi(z)dz \tag{A.6}$$

$$= \frac{1}{2}(1 + \mathrm{erf}\frac{\mathrm{x} - \mu}{\sqrt{2\sigma^2}}) \tag{A.7}$$

であり、図 A.1(b) である。
相補累積分布関数は、

$$1 - \Phi(x) = p(X > x; 0, 1) \tag{A.8}$$

$$= \int_{x}^{\infty} \phi(z)dz. \tag{A.9}$$

A.2.1 正規分布に従う確率変数の出現しやすさ 1

標準正規関数に従う確率変数が 95% の確率で見つかる範囲を求めてみます。標準正規関数は、0 を中心にして、対称な関数なので、正負の値が同じ程度の確率で見つかります。

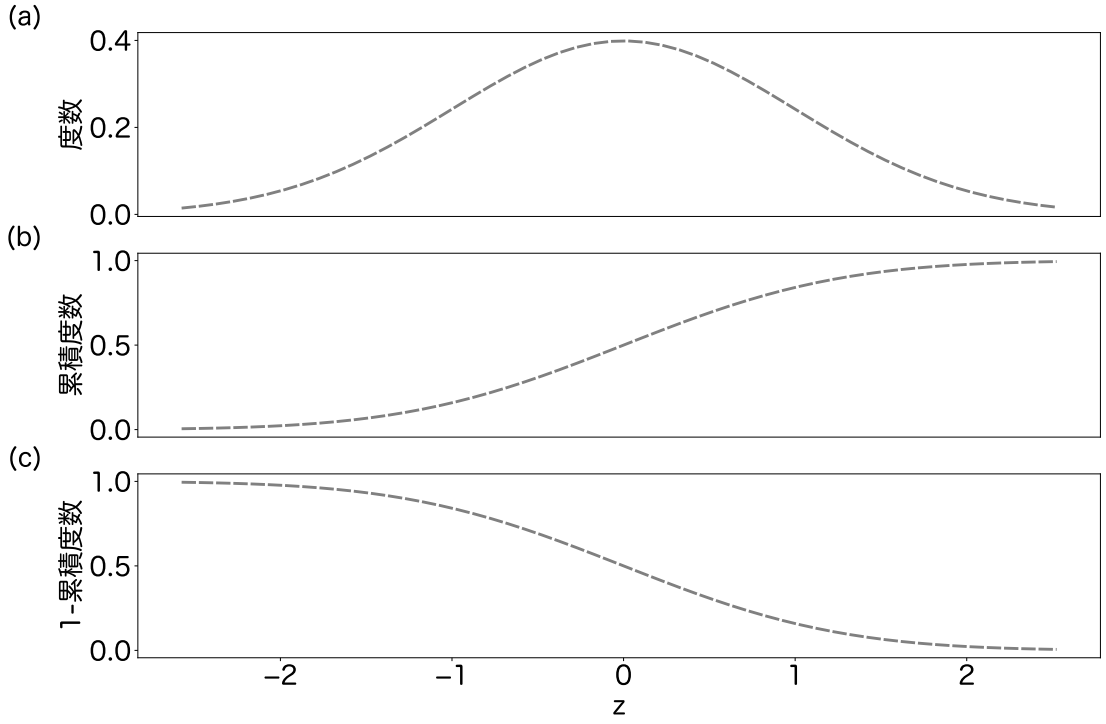


図 A.1 標準正規分布 (a) 確率密度関数 (b) 累積度数分布 (c)1-累積度数分布

言い換えれば、$0 \sim a$ までの積分値と、$-a \sim 0$ までの積分値が同じになります。そこで、次の積分を考えて、その最小値となる値を見つけてみます。

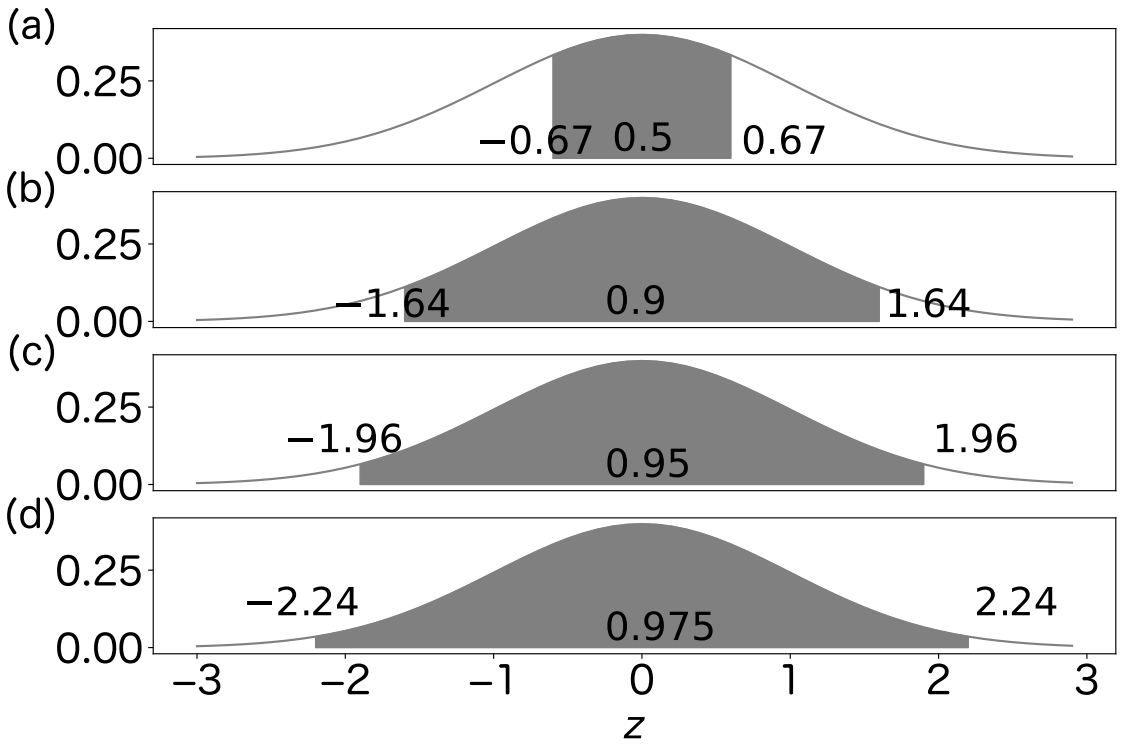
$$\int_{-a}^{a} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})dz = 0.95 \tag{A.10}$$

```
1 b,a = norm.interval(0.95,0,1) # 積分値が0.95になる範囲を計算
2 print(norm.cdf(b, loc=0, scale=1)−norm.cdf(a, loc=0, scale
      =1)) # 0.95になるかを確認
3 print(b,a) # その範囲を表示
```

$0 < \alpha < 1$ に対して、$\Phi(z_\alpha) = 1 - \alpha$ となる z_α を上側 100% 点という。$z_{0.05} = 1.64, z_{0.025} = 1.96$ の値は後でよく使う。

より、一般的には、$\alpha(0 \leq \alpha \leq 0)$ を指定すると、その半分 $\alpha/2$ となる積分範囲の末端を a_1 とします。数式で書くと、

$$\int_{-\infty}^{a_1} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})dx = \frac{\alpha}{2}. \tag{A.11}$$



同様に、右側の範囲の末端を a_2 とします。数式で書くと、

$$\int_{a_2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})dx = \frac{\alpha}{2}.$$

これを書き換えると、次と同値です。

$$\int_{-\infty}^{a_2} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})dx = 1 - \frac{\alpha}{2}.$$

標準正規分布 $z \sim N(0,1)$ において 95% の確率で確率変数が見つかる範囲を調べることはできましたが、正規分布 $x \sim N(\mu, \sigma^2)$ においてでは、どの範囲になるのでしょう。次の定理を使えば簡単に計算ができます。

定理 A.2.1. 確率変数 x が、$x \sim N(\mu, \sigma^2)$ であるならば、$\frac{x-\mu}{\sigma} \sim N(0,1)$ である。

定理 A.2.2. $\alpha(0 \leq \alpha \leq 1)$ に対して、$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) = \alpha$ を満たすとき、$\int_{-\infty}^{\mu+\sigma z} \frac{1}{\sqrt{2\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma}) = \alpha$ である。同様に、$\int_{z}^{-\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) = 1 - \alpha$ を満たす z について、$\int_{\mu+\sigma z}^{\infty} \frac{1}{\sqrt{2\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma}) = 1 - \alpha$ である。

言い換えれば、標準正規分布の軸上の点 z を、$[-\infty, z]$ の範囲での積分値を保ったまま、正規分布 $N(\mu, \sigma^2)$ 上の点に変換するには、$\frac{x-\mu}{\sigma} = z$ を x について解けば良いことになります。

この定理により、以下をとけば、値が 95% の確率で得られる範囲がわかります。

$$\frac{x-\mu}{\sigma} = z_{0.025}$$
$$\rightarrow x = \mu + \sigma z_{0.025}$$

また、

$$\frac{x-\mu}{\sigma} = -z_{0.025}$$
$$\rightarrow x = \mu - \sigma z_{0.025}$$

以上により、$x \sim N(\mu, \sigma^2)$ が 95% の確率で見つかる範囲は、$[\mu - \sigma z_{0.025}, \mu + \sigma z_{0.025}]$ であることがわかります。同様に 90% の確率で見つかる範囲は、$[\mu - \sigma z_{0.05}, \mu + \sigma z_{0.05}]$ です。

A.2.2 より大きな値をとる確率

x を標準正規分布の確率変数とし、$(x \sim N(0,1))$ また、$x \leq 0$ であるとします。。x 以上の大きな値を取る確率は、$P(X > x) = 1 - \Phi(x)$ で計算できます。同様に、$x < 0$ であるときは、より小さな値を取る値が、$P(x < X) = \Phi(x)$ で同様に計算できます。図 A.2 には、x に対して、より異なった値を取る確率を書いています。
x の大きさ $|x|$ よりも大きな値を取る確率は、以上の二つの和で次のようにかけます。

$$P(|x| > z) = 1 - \Phi(|x|) + \Phi(-|x|) \tag{A.12}$$

式を見ると正の数で x より大きな値を取る確率と、負の数で x より小さな値を取る確率の和になっていることが確認できます。$P(|x| > z)$ はより極端な値を取る確率などと言う方もされます。
計算してみます。$x = 1.64$ であれば、$\Phi(1.64) = 0.95$ より、それ以上に大きな値を得る確率は、$P(X > 1.64) = 0.05$ です。また、$x = -1.64$ であれば、$\Phi(-1.64) = 0.05$ です。よって、$|x| = |1.64|$ よりも大きな値を得る確率は $P(|1.64| > X) = 0.1$ です。

A.2.3 $N(0,1)$ での珍しい値は、$N(0,2)$ では珍しくない？

以上の議論により、$N(0,1)$ において、$z = 1.64$ 以上の値が出る確率はおよそ 5% である。では、$N(0,2)$ において $z = 1.64$ が出る確率はいくつだろうか。$N(0,2)$ において、$z = 1.64 \times 2$ 以上に大きな値が出る確率は、およそ 5% である。このことから、$N(0,2)$ において $z = 1.64$ 以上の値が出る確率は、5% より大きいことがわかる。具体的に、計算をしてみると、その確率は 0.206 程度であることがわかる。

```
1 1-norm.cdf(1.64,0,2)
```

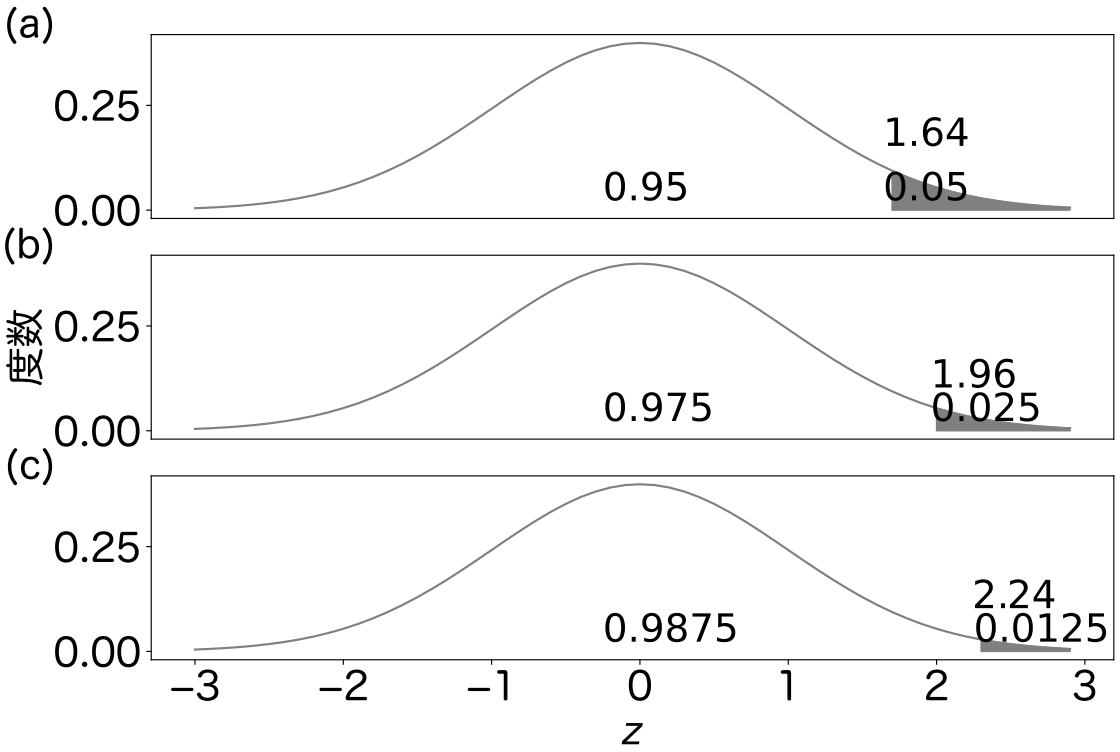


図 A.2 標準正規分布におけるより大きな値 (より偏った値) を取る確率。(a)$z = 1.64$ より大きな値を取る確率は 0.05。(b)$z = 1.96$ より大きな値を取る確率は 0.025。(c)$z = 2.24$ よりも大きな値を取る確率は 0.0125

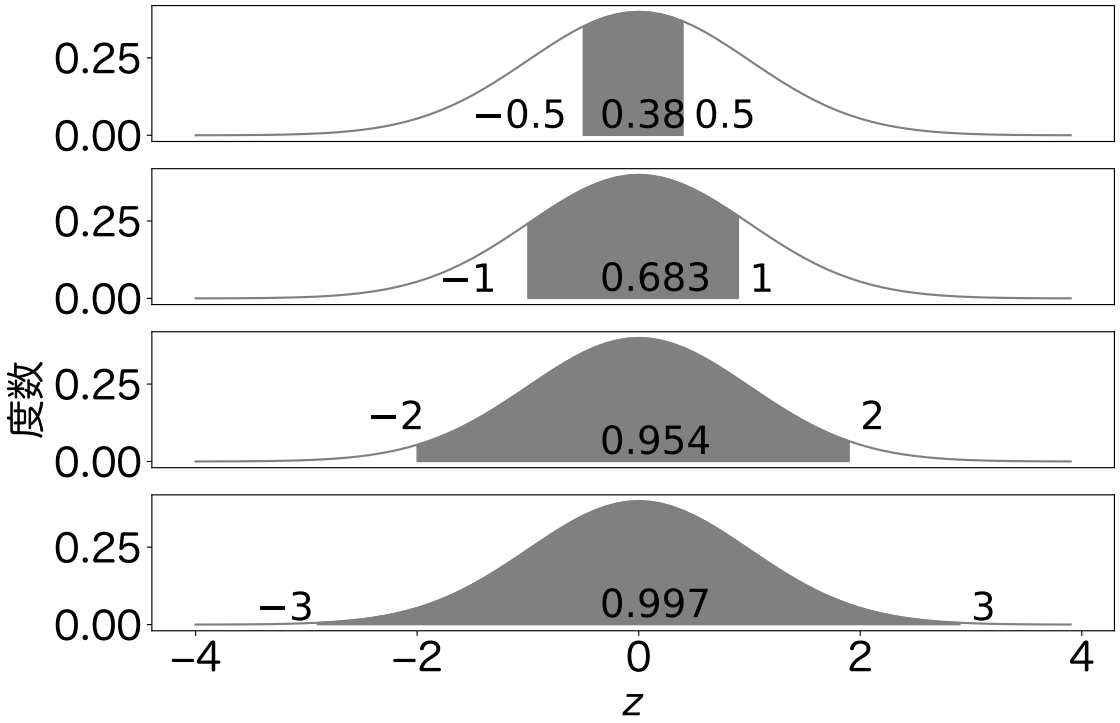
A.2.4 $N(1.96, 1)$ で出てくる値は、$N(0, 1)$ において珍しい？

$N(1.96, 1)$ において、1.96 以上の値が出る確率は、50% です。明らかに、よく出る値であることがわかります。一方で、$N(0, 1)$ においては、1.96 以上の値が出る確率は、2.5% くらいなので、珍しい値になります。このように、確率分布の母数が変化すると、珍しい値も変化します。

A.2.5 正規分布に従う確率変数の出現しやすさ 2

確率変数のしやすさを表す基準として、σ を基準にして、定数 a 倍の範囲 $[\mu - a\sigma, \mu + a\sigma]$ を使う方法もあります。標準正規分布では、分散が 1 なので、その 0.5 倍、1 倍、2 倍、3 倍の範囲はそれぞれ $[-0.5, 0.5]$,$[-1, 1]$,$[-2, 2]$,$[-3, 3]$ になります。この範囲に入る確率は、それぞれ 0.38,0.683,0.954,0.997 です。それぞれの範囲と確率は、図 A.2.5 に図示しました。

σ の定数倍の範囲に値が見つかる確率は、σ の大きさに依存しないことが証明できます。



言い換えれば、$[-0.5\sigma, 0.5\sigma], [-\sigma, \sigma], [-2\sigma, 2\sigma], [-3\sigma, 3\sigma]$ の範囲に値がある確率は、上記と同じで、それぞれおよそ 0.38,0.683,0.954,0.997 になります。

表 A.1　σ を基準にした値の出やすさ

出現確率	$N(0,1)$	$N(\mu, \sigma^2)$
0.38	[-0.5,0.5]	$[\mu - 0.5\sigma, \mu + 0.5\sigma]$
0.683	[-1,1]	$[\mu - \sigma, \mu + \sigma]$
0.954	[-2,2]	$[\mu - 2\sigma, \mu + 2\sigma]$
0.996	[-3,3]	$[\mu - 3\sigma, \mu + 3\sigma]$

A.3　指数分布

確率変数 X が指数分布に従うことを $X \sim Exp(\lambda)$ と書く。指数分布の確率密度関数は、

$$f(x) = \lambda \exp(-\lambda x).$$

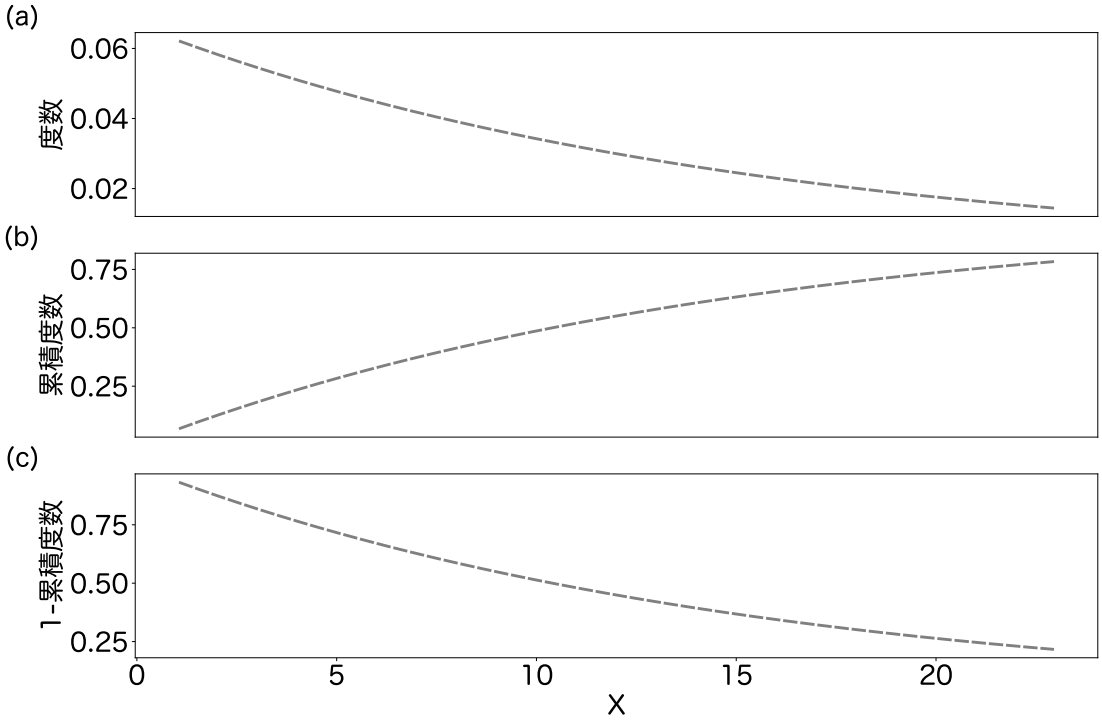


図 A.3 指数分布 $\lambda = 1/15$(a) 確率密度関数 (b) 累積度数分布 (c) 相補累積度数分布

ここで、λ は、$\lambda > 0$ であり、指数分布の母数である。期待値は $E[X] = \frac{1}{\lambda}$ で、分散は、$V[X] = \frac{1}{\lambda^2}$ である。累積分布関数は、

$$F(x) = 1 - \exp(-\lambda x).$$

正規分布は、母数平均を中心として、左右対称に分布していた。言い換えれば、$\phi(\mu + x) = \phi(\mu - x)$ である。一方で、指数分布は、左右非対称に分布が広がり、小さな値は大きな値よりも出現確率が高いので、$f(E[X] + a) \neq f(E[X] - a)$ である。また、正規分布では、母数平均と母数分散がそれぞれ独立なので、それぞれの特徴を独立に動かすことで、期待値や分散が独立に変化する。指数分布では、母数が一つであり、母数を変化させると、期待値と分散は同時に変化する。

A.3.1　指数分布に従う確率変数の出現しやすさ

指数分布の確率密度関数を区間 $[a,b]$ で積分したときに、$\alpha(0 \leq \alpha \leq 1)$ になる $[a,b]$ を求めます。条件として、

$$\int_0^a \lambda \exp(-\lambda x)dx = \alpha/2$$
$$\int_0^b \lambda \exp(-\lambda x)dx = 1-\alpha/2$$

を満たすとする。a について、とくと、

$$\int_0^a \lambda \exp(-\lambda x)dx = \alpha/2$$
$$1-\exp(-\lambda a) = \frac{\alpha}{2}$$
$$\rightarrow a = \frac{1}{\lambda}\log\frac{1}{1-\alpha/2}$$

b については、同様に、

$$b = \frac{1}{\lambda}\log\frac{\alpha}{2}$$

以上より、この積分の条件で、$100(1-\alpha)\%$ の確率で値を得る範囲は、$[\frac{1}{\lambda}\log\frac{1}{1-\alpha/2}, \frac{1}{\lambda}\log\frac{\alpha}{2}]$ である。図 A.4 は、指数分布により、サンプルサイズ 1000 の標本を 100 回作って、各標本においてデータが区間 $[\frac{1}{\lambda}\log\frac{1}{1-\alpha/2}, \frac{1}{\lambda}\log\frac{\alpha}{2}]$ に入った割合をシミュレーションし、そのヒストグラムを表示している。確かに、95% くらいの割合でその区間にデータが入っている。

A.4　カイ二乗分布

確率変数 X がカイ二乗分布に従うことを $X \sim \chi_k^2$ と書く。ここで、k はカイ二乗分布の母数で、自由度を示し、自然数を取る。確率密度関数は、

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}\exp\left(-\frac{x}{2}\right).$$

ここで、$\Gamma(k/2)$ はガンマ関数を表す[*2]。累積分布間数は、

$$F(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)}.$$

ここで、$\gamma(k/2, x/2)$ は、不完全ガンマ関数である[*3]。この関数も左右非対称である。

[*2] $\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)dt$ である。

[*3] $\gamma(a,x) = \int_0^x t^{a-1}\exp^{-t}dt$ である。ガンマ関数も、不完全ガンマ関数も計算できなくても問題はない。コンピュータを使えばすぐに計算してくれる。

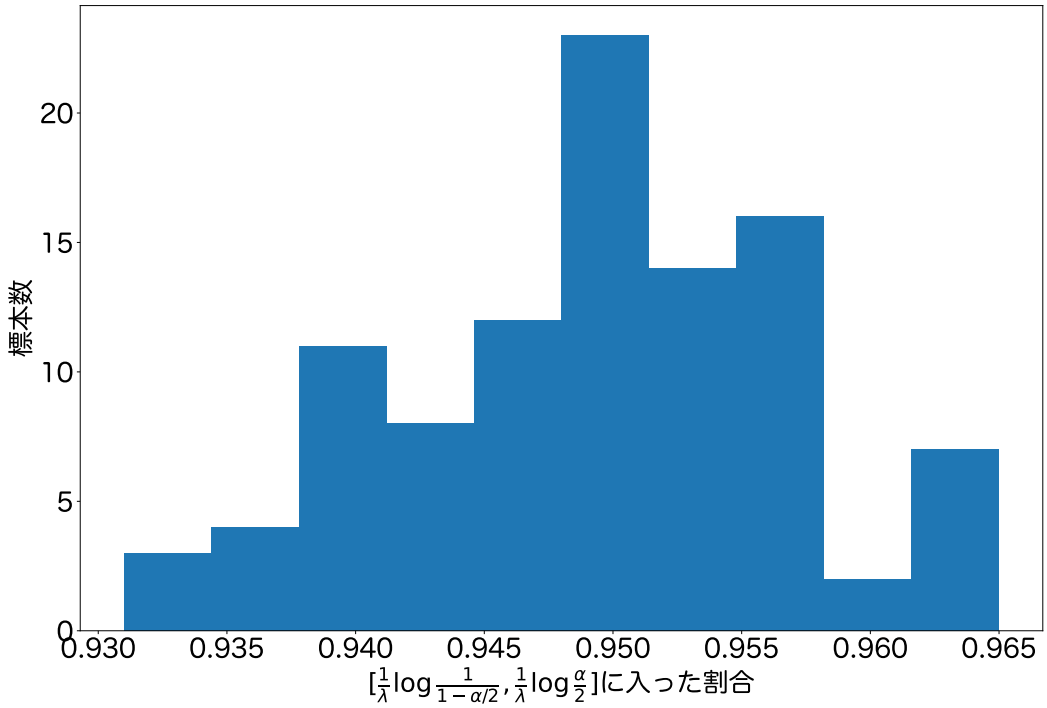


図 A.4 指数分布 $\lambda = 1/10$ からサンプルサイズ 1000 の標本を 100 回シミュレーションし、各標本においてデータが区間 $[\frac{1}{\lambda}\log\frac{1}{1-\alpha/2}, \frac{1}{\lambda}\log\frac{\alpha}{2}]$ に入った割合を計算した。そのヒストグラム。

A.4.1 カイ二乗分布に従う確率変数の出現しやすさ

カイ二乗分布の確率密度関数を区間 $[a, b]$ で積分したときに、$\alpha(0 \leq \alpha \leq 1)$ になる $[a, b]$ を求めます。条件として、

$$\int_0^a \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\left(-\frac{x}{2}\right) dx = F(a) - F(0) = \alpha/2$$
$$\int_0^b \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\left(-\frac{x}{2}\right) dx = F(b) - F(0) = 1 - \alpha/2$$

を満たすとする。代数的に a, b について解くことが難しいので、数値的に計算してみた結果を載せておく (表 A.2)。この a, b をそれぞれ $\chi_k^2(\alpha), \chi_k^2(1-\alpha)$ と書くことがある。

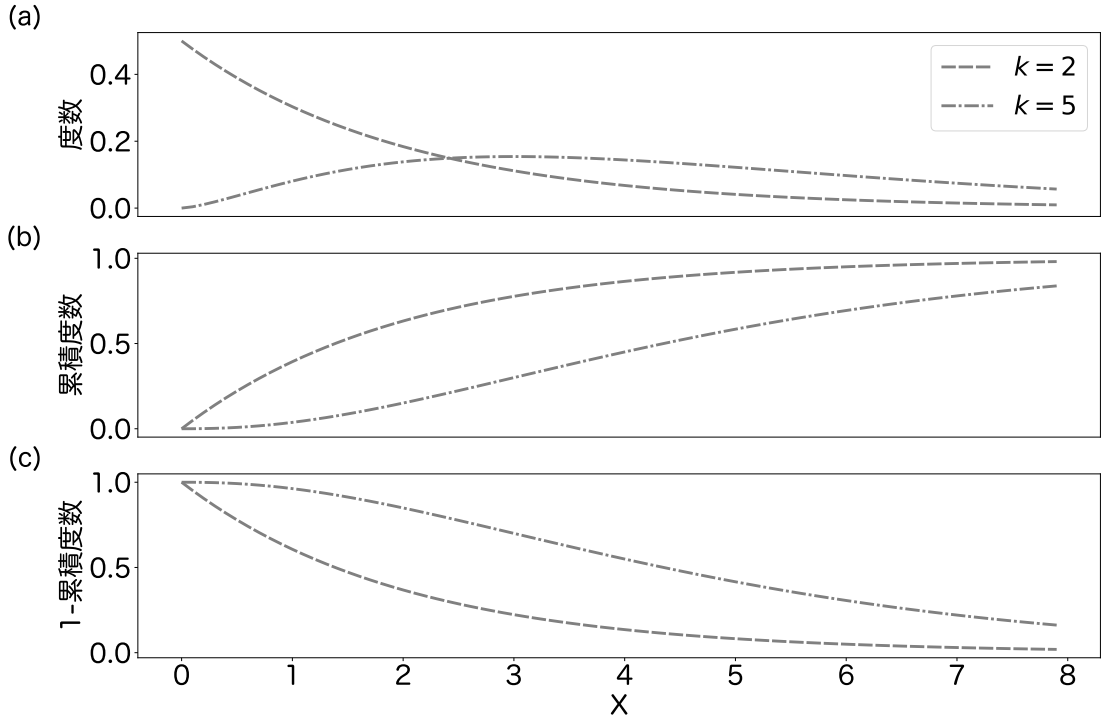


図 A.5　カイ二乗分布

表 A.2　$\alpha = 0.05$

k	a	b
1	0.0009	5.02
3	0.215	9.3484
5	0.831	12.832

A.5　t 分布

確率変数 T が t 分布に従うとき、$T \sim t(\nu)$ と表記する。確率密度関数は、

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + t^2/\nu)^{-(\nu+1)/2}.$$

ここで、ν は、0 より大きな実数である。この関数を見ただけでは、すぐには判別するのは難しいかもしれないが、$f(t)$ には t が関係する部分は $(1 + t^2/\nu)$ だけである。二乗の項があるので、偶数関数であることがわかり、0 を中心にした対称な関数 $f(t) = f(-t)$ であることがわかる。累積分布関数は著者には難しすぎるので、記述しない。wikipedia などで調べれば正しそうな数式が書かれている。

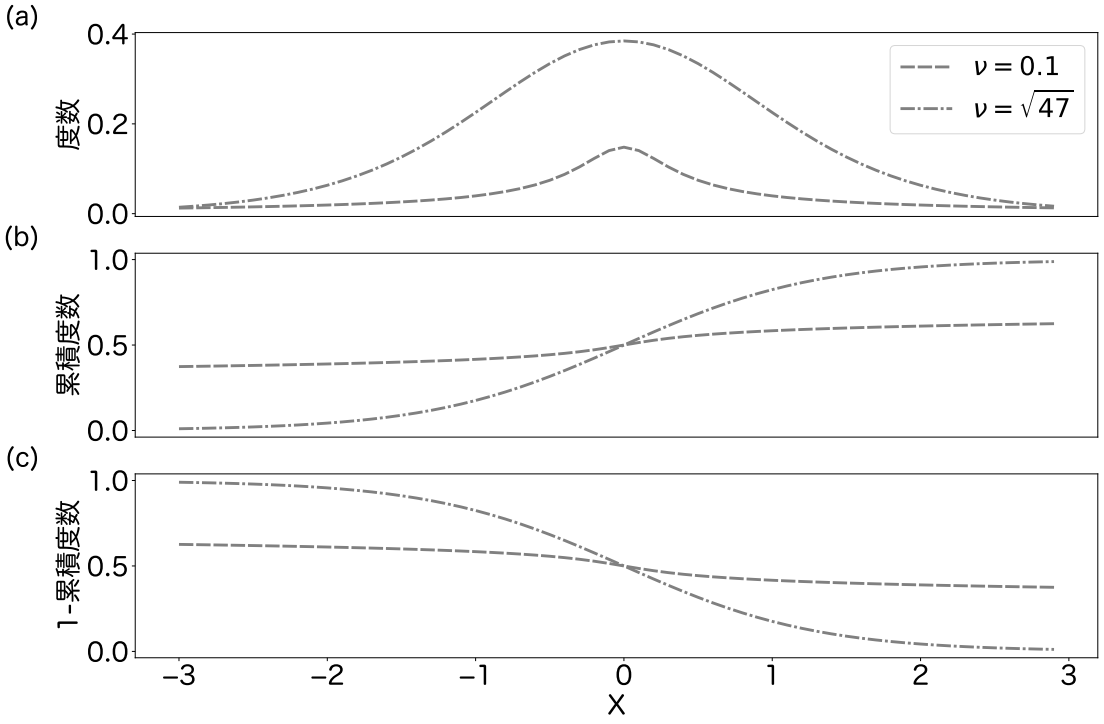


図 A.6 t 分布

A.5.1 t 分布における珍しい値

t 分布における $|T|$ 以上の値が得られる確率が α 程度になる $|T|$ のリスト。例えば、$n = 10$ の t 分布において $|T| = 1.81$ 以上の値が得られる確率は、0.1 程度である。

表 A.3 t 分布における $|T|$ 以上の値が得られる確率が α 程度になる $|T|$ のリスト

n	p=0.1	$p = 0.05$	$p = 0.025$
1	6.31	12.70	25.45
5	2.01	2.57	3.16
10	1.81	2.22	2.63

A.6 統計分布の関係

同一の確率分布からサンプリングされた複数の確率変数 $X_1, X_2, \cdots, X_n$ を得たとき、それを要約した要約統計量がどのような分布関数に従うのかを考察する。

A.6.1 正規分布の再生性

$X \sim N(\mu_1, \sigma_1^2), Y \sim (\mu_2, \sigma_1^2)$ とするとき、$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$ より、$a = \frac{1}{2}, b = \frac{1}{2}$。すると、$\frac{X}{2} + \frac{Y}{2} \sim N(\frac{\mu_1+\mu_2}{2}, \frac{\sigma_1^2}{2^2} + \frac{\sigma_2^2}{2^2})$ である。$\mu_1 = \mu_2, \sigma_1 = \sigma_2$ とすると、$\frac{X+Y}{2} \sim N(\mu_1, \frac{\sigma_1^2}{2})$ が成り立つ。このことを利用すると、$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ とすると、$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$ である。よって $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$。また、$\bar{x}$ の出現しやすい区間は、

$$-z_{0.025} < \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < z_{0.025}$$

である。式を変形すると、

$$\mu - z_{0.025}\frac{\sigma^2}{n} < \bar{x} < \mu + z_{0.025}\frac{\sigma^2}{n}$$

がわかる。以上をまとめておく。

定理 A.6.1. $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ とすると、$\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ ただし、$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$。また、$\bar{X}$ の出現しやすい区間は、$\mu - z_{0.025}\frac{\sigma^2}{n} < \bar{x} < \mu + z_{0.025}\frac{\sigma^2}{n}$ である。

定理 A.6.2. $X_1, X_2, \cdots, X_{n_1} \sim N(\mu_1, \sigma_1^2), Y_1, Y_2, \cdots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ ただし、$\mu_1 \neq \mu_2, \sigma_1 \neq \sigma_2$ とする。正規分布の再生性により、$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$ である。次が成り立つ。$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ であり、

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

A.6.2 指数分布の再生性

指数分布 $Exp(\lambda)$ と、ガンマ分布 $Ga(1, \frac{1}{\lambda})$ は、同一の密度分布関数であり、それは $f(x) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda})$ である。ガンマ分布には、分布の再生性があり、$X \sim Ga(a_1, b), Y \sim Ga(a_2, b)$ であるなら、$X + Y \sim Ga(a_1 + a_2, b)$ である。このことを、n 個の確率変数 $X_1, X_2, \cdots X_n \sim Exp(\lambda)(= Ga(1, \frac{1}{\lambda}))$ に適用すると、$X_1 + X_2 + \cdots + X_n \sim Ga(n, \frac{1}{\lambda})$ である。以上によって、$n\bar{X} \sim Ga(n, \frac{1}{\lambda})$ ただし、$\bar{X} = X_1 + X_2 + \cdots + X_n$ である。再生性については、確率母関数を利用することで証明できる。

定理 A.6.3. $X_1, X_2, \cdots, X_n \sim Ga(1, \frac{1}{\lambda})$ ならば、$n\bar{X} \sim Ga(n, \frac{1}{\lambda})$

証明. $Ga(1, \frac{1}{\lambda})$ の確率母関数は、$M_X(t) = (1 - \frac{1}{\lambda}t)^{-1}$ である。確率変数 $X_1 + X_2 + \cdots +$

X_n の確率母関数は

$$M_{n\bar{X}} = M_{X_1+X_2+\cdots+X_n} = M_{X_1}M_{X_2}\cdots M_{X_n} \tag{A.13}$$

$$= (1-\frac{1}{\lambda}t)^{-1}(1-\frac{1}{\lambda}t)^{-1}\cdots(1-\frac{1}{\lambda}t)^{-1} \tag{A.14}$$

$$= (1-\frac{1}{\lambda}t)^{-n} \tag{A.15}$$

以上より、$n\bar{x} \sim Ga(n, \frac{1}{\lambda})$ である。 □

A.6.3 正規分布と t 分布の関係

$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ とする。統計量 T を、

$$T = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{\sqrt{n}}}}.$$

ここで、$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$、$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ である。この統計量 T は、$t(n-1)$ 分布に従うことが知られている。統計量 T の中に母数 σ が入っていないので、σ わからないときでも、T を計算すれば、それが $t(n-1)$ に従うことがわかる。
2 つの正規分布 $X_1, X_2, \cdots, X_{n_1} \sim N(\mu_1, \sigma_1^2), Y_1, Y_2, \cdots, Y_{n_2} \sim N(\mu_2, \sigma_1^2)$ とする。このとき、

$$T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{U^2}{n_1}+\frac{U^2}{n_2}}}$$

は、n_1+n_2-2 の t 分布に従う。ここで、U は、

$$U^2 = \frac{(n-1)U_1^2+(n_2-1)U_2^2}{n_1-1+n_2-1}$$

であり、U_1, U_2 は、不偏分散

$$U_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})$$

$$U_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar{Y})$$

である。

A.7 尤度・対数尤度・AIC

定義 A.7.1. 確率変数の組み $(x_1, x_2, x_3, \cdots, x_n)$ が、ある同時確率密度関数 $P(X_1, \cdots, X_n|\theta)$ から得られたとする。ここで、θ は密度関数 $P(X)$ の母数。このとき、

θ を変数として考えるとき、次を尤度関数という[*4] [*5]。

$$L(\theta) = P(X_1, \cdots, X_n|\theta)$$

ここで、$x_1, x_2, \cdots, x_n$ が独立であるならば、同時確率密度関数は、X_i の密度関数の積に等しいので、尤度関数は次の形に書き換えられる。

$$L(\theta) = P(X_1|\theta)P(X_2|\theta)\cdots P(X_n|\theta)$$

尤度関数に対数をつけたものを、対数尤度関数という。

$$l(\theta) = \sum_{i=0}^{n} \log f(x_i|\theta)$$

$N(0,1)$ において、確率変数 $X^1 = (x_1, x_2, x_3) = (0, 0, 0)$ を得たとする。$N(0,1)$ において 0 の出現確率は $P(0) = 0.398$ である。このことから、尤度はその積で計算でき、$L(0) = 0.398^3 = 0.063$ である。また、別の確率変数の組 $X^2 = (x_1, x_2, x_3) = (1.96, 1.96, 1.96)$ を得たとすると、$N(0,1)$ における 1.96 の出現確率は、$P(1.96) = 0.058$ より、尤度は、$L(0) = 0.05^3 = 0.0001$ である。このことは、確率変数 X^1 は X^2 よりも得られやすいことを示唆する。もしもこの X^1, X^2 が、$N(1.96, 1)$ において得られた場合は、尤度はそれぞれ、$0.0001, 0.063$ となり、尤度の大小関係が逆転する。
具体的に、標準正規分布から 100 個の確率変数をサンプリングし、正規分布 $N(\theta, 1)$ の確率密度関数における対数尤度関数を計算し、尤度関数の変化を図示した (図 A.7)。これを見ると、上に凸な 2 次関数のように見える。実際に、対数尤度関数を展開してみると、$l(\theta)$ が θ に関する 2 次関数になっていることがわかる。

$$l(\theta) = \sum_{i=0}^{100} \log f(x_i|\theta) \tag{A.16}$$

$$= \sum_{i=0}^{100} \log \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i-\theta)^2}{2}\right) \tag{A.17}$$

$$= -\frac{100}{2}\log(2\pi) + \sum_{i=0}^{100} \frac{(x_i-\theta)^2}{2} \tag{A.18}$$

この式より、2 次関数であることは明らかである。

[*4] wikipedia にて尤度を調べると、尤もらしさの指標と出る。この言い換えは適切であるとは言えない。尤度は確率密度関数の積で、密度関数の母数を変数にした関数である。数学における定義を、現実に当てはまる言葉に言い換えできない。尤度という言葉にほぼ意味がない。犬度と言って、尤度のことを指しても良い。`https://ja.wikipedia.org/wiki/`尤度関数.

[*5] 数学において定義された言葉は必ず一意に定まる。定義を見れば尤度が何かが書いてあるのだから、尤度とは何かという問いは意味がない。一方で生物学者はしばしば定義を言い換えたがる。同じ言葉を異なる使い方で用いて議論することがある。議論している人の間で全く定義が異なることもありえる。

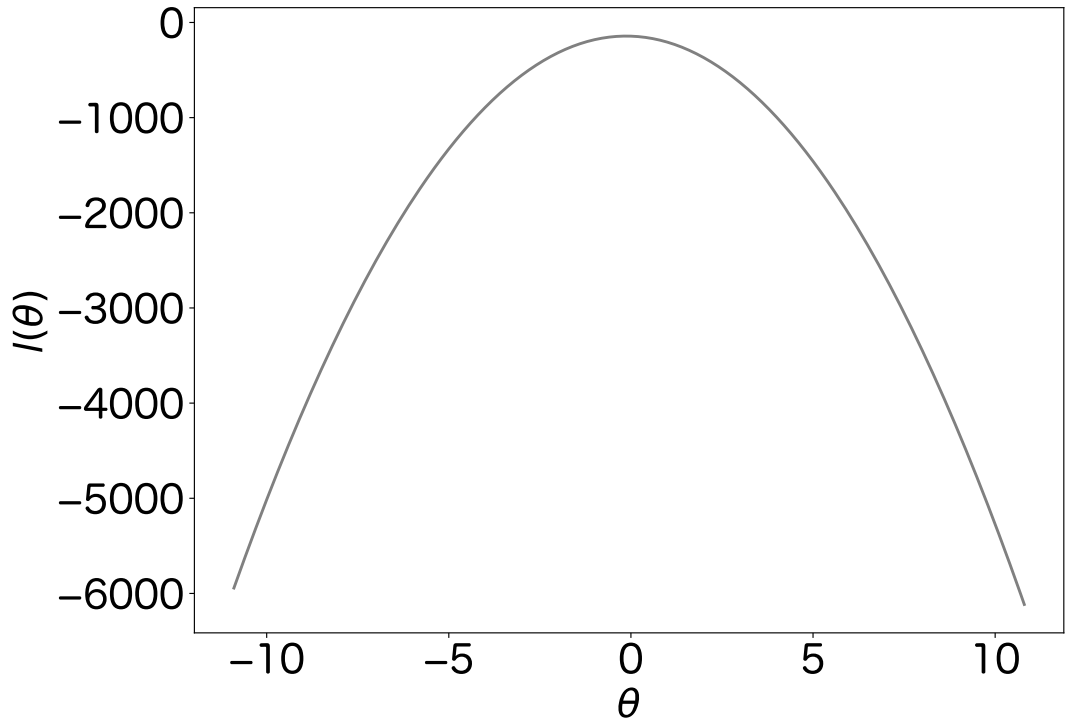


図 A.7 $N(\theta, 1)$ における対数尤度関数。確率変数は、$N(0, 1)$ からサンプリングした。

A.7.1 最尤推定

定義 A.7.2. 尤度関数 $l(\theta)$ を最大にする θ を最尤推定量という。

正規分布における最尤推定量を計算してみる。正規分布は、母数を二つ持つので、尤度関数も 2 変数関数である。まず、対数尤度関数は、

$$l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$l(\mu, \sigma^2)$ を μ で微分する。

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{\sum_i (x_i - \mu)}{\sigma^2}$$

$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = 0$ とおいて、μ について解くと、

$$\mu_{ML} = \frac{\sum_i x_i}{n}$$

これが最尤推定量となる[*6]。

[*6] 最尤が maximum likelihood なので頭文字を取った ML を μ の足に書いて μ_{ML} とした

同様に σ^2 に関する微分を行う。

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (x_i - \mu)$$

これが 0 と等しいとき、σ^2 について解く。

$$\sigma^2_{ML} = \frac{\sum_i (x_i - \mu)^2}{n}$$

最尤推定量は、分布関数によって異なるので、計算してみるとよい。

A.7.2 AIC(an information criterion)

確率分布から対数尤度を求め、対数尤度の低い確率分布は、その中で相対的にデータに対して当てはまりの良い確率分布であると考えることができる。最尤推定量を使った確率分布関数は、データを使って分布関数を決定しているので、データを使わずに求めた分布よりも、データに対して良い分布関数になりがちである。そこで、対数尤度に対して罰則項を加えた AIC を使って、データに対する当てはまりの良さを計算することがある。

$$AIC = -2\log f(x|\theta) + 2k$$

ここで、k はデータによって決まったパラメータの個数である。

A.7.3 マトリョウシカになったモデル

複数のパラメータにより決定されるモデル $M(\alpha, \beta, \gamma, \omega)$ がある。このパラメータの中からモデルに影響を与えないように値を個体したモデル $M(\alpha, \beta, \gamma)$ や $M(\alpha, \beta, \omega)$ や $M(\alpha)$ などが構成できることがある。このようにパラメータの個数が少なくなったモデルを元のモデルからネストされたモデルや入れ子になったモデルと呼ぶことがある。また、元のモデル $M(\alpha, \beta, \gamma, \omega)$ を「フルモデル」と呼ぶ。
モデル $M(\alpha, \beta, \gamma)$ をフルモデルとしたとき、$M(\alpha, \beta)$ はネストされたモデルでるが、$M(\alpha, \omega)$ はネストされたモデルではない。

付録 B

数理統計の補足

B.1 正規分布の検定 1

また、母数分散 σ について次が成り立つ。

定理 B.1.1. $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ について、次が成り立つ。

$$Y = (n-1)(\frac{S_x}{\sigma})^2 \sim \chi^2_{n-1}$$

ここで、$S_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2, \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ である。

B.2 指数分布を含むモデル

定理 B.2.1. $x_1, x_2, \cdots, x_n \sim Ex(\lambda)$ とする。$x_1 + x_2 + \cdots + x_n \sim \gamma(n, \lambda)$ である

n を自然数とし、ガンマ分布 $Ga(\frac{n}{2}, 2)$ を特に、カイ 2 乗分布といい、χ^2_n で表す。

定理 B.2.2. n を自然数とする。$G \sim \Gamma(\frac{n}{2}, \beta), Y_n \sim \chi^2_n$ とすると、$P(G \leq w) = P(Y_n \leq 2\beta w)$

証明. $w > 0$ に対して、

$$\begin{aligned}
P(G \leq w) &= \int_0^w \frac{\beta^{\frac{n}{2}}}{\Gamma(n/2)} x^{n/2-1} \exp(-\beta x) dx \\
&= \int_0^{2\beta w} \frac{\beta^{\frac{n}{2}}}{\Gamma(n/2)} \left(\frac{t}{2\beta}\right)^{n/2-1} \exp(-\beta t/2\beta) \frac{dt}{2\beta} (x = t/(2\beta)) \\
&= \int_0^{2\beta w} \frac{1}{2^{n/2}\Gamma(n/2)} t^{n/2-1} \exp(-t/2) dt \\
&= P(Y_n \leq 2\beta w)
\end{aligned}$$

□

以上より $n\bar{x} \sim \Gamma(n, \lambda)$ である。このとき、λ の信頼区間を求める。λ の下限は、

$$P(G \leq n\bar{x}) = \frac{\alpha}{2} \tag{B.1}$$

を満たし、λ の上限は、

$$P(G \leq n\bar{x}) = 1 - \frac{\alpha}{2} \tag{B.2}$$

を満たす。下限の式を変形していく。

$$\begin{aligned} \alpha/2 &= P(G \leq n\bar{x}) \\ &= P(Y_{2n} \leq 2n\lambda_l \bar{x}) \\ &\to 2n\lambda\bar{x} = \chi^2_{2n}(1-\alpha/2) \\ &\to \lambda = \frac{\chi^2_{2n}(1-\alpha/2)}{2n\bar{x}} \end{aligned}$$

上限についても同様に、

$$\begin{aligned} 1 - \frac{\alpha}{2} &= P(G \leq n\bar{x}) \\ &= P(Y_{2n} \leq 2n\lambda\bar{x}) \\ &\to 2n\lambda\bar{x} = \chi^2_{2n}(\alpha/2) \\ &\to \lambda = \frac{\chi^2_{2n}(\alpha/2)}{2n\bar{x}} \end{aligned}$$

以上によって、$\frac{1}{\lambda}$ の信頼区間は、

$$\frac{2n\bar{x}}{\chi^2_{2n}(\alpha/2)} \leq \frac{1}{\lambda} \leq \frac{2n\bar{x}}{\chi^2_{2n}(1-\alpha/2)} \tag{B.3}$$

B.2.1 ２標本・指数分布

$X_1, X_2, \cdots, X_n \sim i.i.dExp(\theta_1), Y1, Y_2, \cdots, Y_n \sim i.i.dExp(\theta_2)$ とする。帰無仮説 H_0 を、$H_0 : \theta_1 = \theta_2$ とし、対立仮説 H_1 を、$H_1 : \theta_1 \neq \theta_2$ とする。帰無仮説のもとで、尤度関数 L_{H_0} は、

$$L_{H_0} = \theta^{-n_1-n_2} \exp\{-\theta^{-1}T\} \tag{B.4}$$

ただし、$T = \sum_{i=0}^{n} X_i + \sum_{i=0}^{n} Y_i$ である。$\frac{\partial H_0}{\partial \theta} = 0$ となる θ は、

$$\frac{\partial H_0}{\partial \theta} = \{-(n_1+n_2) + \theta^{-1}T\}\theta^{-n_1-n_2-1} \exp(-\theta^{-1}T). \tag{B.5}$$

より、$\theta_0 = \frac{T}{n_1+n_2}$ である。θ_0 を L_{H_0} に代入すると、

$$L_{H_0} = \theta_0^{-n_1-n_2} \exp(-n_1-n_2). \tag{B.6}$$

同様に、対立仮説のもとで、尤度関数 L_{H_1} は、

$$L_{H_1} = \theta_1^{-n_1} \exp(-\frac{n_1}{\theta_1}\bar{x})\theta_2^{-n_2} \exp(-\frac{n_2}{\theta_2}\bar{y}) \tag{B.7}$$

$\frac{\partial H_1}{\partial \theta} = 0$ となる θ_1 を計算する。

$$\frac{\partial H_1}{\partial \theta_1} = \{-n_1\theta_1^{-n_1-1}\exp(-\frac{n_1}{\theta_1}\bar{x}) + n_1\bar{x}\theta_1^{-n_1-2}\exp(-\frac{n_1}{\theta_1}\bar{x})\}\theta_2^{-n_2}\exp(-\frac{n_2}{\theta_2}\bar{y}). \quad \text{(B.8)}$$

$\frac{\partial H_1}{\partial \theta_1} = 0$ より、$(-n_1 + n_1\bar{x}\theta_1^{-1})\theta_1^{-n_1-1} = 0$ より、$\hat{\theta}_1 = \bar{x}$ である。同様に、$\hat{\theta}_2 = \bar{y}$。以上によって、L_{H_1} は、

$$L_{H_1}(\hat{\theta}_1, \hat{\theta}_2) = (\hat{\theta}_1)^{-n_1}\exp(-n_1)(\hat{\theta}_2)^{-n_2}\exp(-n_2) \quad \text{(B.9)}$$

である。
尤度比は、

$$\Lambda = \frac{L_{H_1}}{L_{H_0}} = \frac{(\hat{\theta}_1)^{-n_1}(\hat{\theta}_2)^{-n_2}\exp(-n_1-n_2)}{\theta_0^{-n_1-n_2}\exp(-n_2-n_2)} \quad \text{(B.10)}$$

$$= \left(\frac{\theta_0}{\hat{\theta}_0}\right)^{n_1}\left(\frac{\theta_0}{\hat{\theta}_1}\right)^{n_2} \quad \text{(B.11)}$$

尤度比検定より、$-2\log\Lambda \sim \chi_1^2$ である。
$Ga(\alpha, \beta)$ について、以下が成り立つ

$$kX \sim Ga(\alpha, \beta/k) \quad \text{(B.12)}$$

$$\frac{1}{k}X \sim Ga(\alpha, k\beta) \quad \text{(B.13)}$$

$$\chi_{2n}^2 = Ga(n, 2) \quad \text{(B.14)}$$

以上を使うと、

$$n\bar{X} \sim Ga(n, \frac{1}{\lambda})$$
$$\frac{n}{2}\bar{X} \sim (\frac{n}{2}, \frac{1}{\lambda})$$
$$\frac{n}{2\lambda}\bar{X} \sim Ga(\frac{n}{2}, 2) = \chi_n^2$$

また、ガンマ分布とベータ分布の関係より、$X_1 \sim Ga(\alpha_1, \beta), X_2 \sim Ga(\alpha_2, \beta)$ ならば、$\frac{X_1}{X_1+X_2} \sim Beta(\alpha_1, \alpha_2)$ である。以上より、$Z = \frac{n\bar{X}}{n\bar{X}+m\bar{Y}} \sim Beta(n, m)$ である。このことから、棄却域 $(z_1 \leq Z \leq z_2)$ を求めることができる。具体的には、

$$\int_0^{z_1}\frac{1}{Bn,m}z^{n-1}(1-z)^{m-1}dz = \alpha/2, \int_{z_2}^{\infty}\frac{1}{Bn,m}z^{n-1}(1-z)^{m-1}dz = \alpha/2. \quad \text{(B.15)}$$

この解 z_1, z_2 を計算すれば良い。

B.2.2 中心極限定理

定理 B.2.3 (中心極限定理)**.** 期待値 μ と分散 σ^2 を持つ独立分布に従う確率変数列 $X_1, X_2, \cdots$ に対し、$S_n = \sum_{k=1}^n X_k$ とおくと、S_n は、期待値 0、分散 1 の正規分布に分布収束する。

統計学のユーザーの中には、次のことが成立すると考えている[*1]。

仮説 B.2.1 (中心極限仮説)**.** 一般のデータについて、データのサンプルサイズを大きくすると、その平均 $\bar{x}$ は、$\bar{x} \sim N(\mu, \sigma^2/n)$。

反例がすぐに出てくるので、このような仮説は一般には成り立たない。例えば、データがコーシー分布から生成されている場合、成り立たない。
少なくとも一人は、次のように考えていた。

仮説 B.2.2 (中心極限仮説 2)**.** サンプルサイズを大きくすると、正規分布に近づく。

これも間違いである。中心極限定理の前提のもと、標本平均を集めると正規分布に近づくことはありえる。

[*1] 中心極限仮説が成り立つと考えている人は多い。というよりも、中心極限仮説が成り立つことにして、科学的な議論を進めるという方針を取っているのだと思われる。
http://www.ner.takushoku-u.ac.jp/masano/class_material/waseda/keiryo/R10_inference.html#3_%E4%B8%AD%E5%BF%83%E6%A5%B5%E9%99%90%E5%AE%9A%E7%90%86 .
https://yukiyanai.github.io/stat2/clt.html.

付録 C

仮説検定の 3 つの枠組み

現在利用されている仮説検定の枠組みを紹介する。それぞれの枠組みは前提が異なっており、それぞれの問題で良い推測を与える。適切な枠組みを選ばなければ間違った解釈をしてしまう事になる。我々の行なっている科学では、F 型の枠組みを使う。以降の章では、F 型の問題について考える。

■有意性検定・仮説検定

Fisher は、帰無仮説を設定し、帰無仮説とデータを比較検討する方法を構築した。これを、有意性検定という。これに対して、Neyman-Pearson らは、帰無仮説に加えて、対立仮説を設定し、データを元に帰無仮説を棄却するかの判断を有意水準により行う意思決定の枠組みを構築した。これを、仮説検定と呼ぶことがある [6]。現代の科学の多くは、Fisher と Neyman-Pearson の両方を組み合わせ、帰無仮説・対立仮説を設定し、帰無仮説とデータの乖離を p 値によって調べ、棄却するかを検討する。ここでの p 値は、Neyman-Pearson の解釈から、20 回に 1 回程度のことを有意と呼ぶことにしている。この流派をハイブリッド仮説検定法 [9] と呼ぶことにする。

C.1　F 型

F 型では、既存の研究結果からモデルを構築し、そのモデルによりある集団の性質が予測できるかを調べる。その後、計測結果を元に、その集団を予測できそうなモデルを構築する。次の実験では、構築したモデルを利用し、予測が可能かを調べていく。または、同じ集団についてさまざまな計測を行い、予測性能をあげるようにモデルを構築する。

C.1.1　解決できる問題

C.2　NP 型

NP 型では、すでに前提になっていることから著しく外れたことが起きたことを検出するための方法である。言い換えるなら、まず、何度も計測を行い、モデルが事象をよく予測できるようにモデルを構築する。そして、母集団から無作為抽出し、標本を得る。最後に、標本の統計検定量がモデルの予測を著しく外れているならば、これまで計測していた現象が得られので、前提となる計測または母集団が変化したと疑う。

C.2.1　解決できる問題

NP 型で扱う問題をいくつか挙げる。

> ある調味料の製造ラインでは, 砂糖の含有量 (g) は, 原料の不均一や製造ラインの狂いなどから変動するが, 標準偏差は常に一定で $\sigma = 3$ の正規分布に従っているとしてよい. 各製品の砂糖の含有量が $\mu = 60$ になるように調整してラインを稼働させて, しばらくしてから, 25 個の製品を抜取検査したところ, 砂糖の含有量の平均値は $\bar{x} = 61.63$ であった. その時点で製造ラインは $\mu = 60$ を保持していると言えるだろうか.

正規モデル $M(60, 3^2)$ によって予測ができるという前提条件を満たしている。この前提を元に、無作為抽出した 25 個がそのモデルにより予測できるのかを調べる。標本全体とモデルの累積分布などを比較する方法もあるが、ここでは、検定によって調べてみる。このモデルでは、

$$Z = \sqrt{n}\frac{\mu - \bar{x}}{\sigma} \sim N(0, 1)$$

である。変数を入れれば、$Z = 2.72$ となる。$\Phi(Z) < \Phi(1/20 = 0.05)$ であり、モデル内で、20 回に 1 回よりも少ない頻度で観測されないようなことが現実で起きている。または、$\Phi(Z) < \Phi(2\sigma = 2)$ であり、$2\sigma = 2$(標準正規分布の中で) よりも珍しいことが起きているので、モデルでの予測ができないことが起きている。偶発的に生じた可能性も捨てられないが、製造過程に不具合が生じているのではないかと推測される。

C.2.2 解釈

■第一の過誤・第二の過誤・統計モデルが正しい

Neyman-Pearson 流の統計学においては、α, β を次のように定義する。帰無仮説が正しいとき、誤ってそのモデルを棄却してしまう間違いを第一の過誤といい、この確率を α とする。また、対立仮説が正しいのに、帰無モデルを採択する間違いを第二の過誤といい、この確率を β とする。

C.3 ハイブリッド型

C.3.1 解決できる問題

付録 D

仮説検定の実際

実際に利用されている仮説検定について説明する。ここではあえて間違っているとされる方法も含めている。
正しい仮説検定について説明してある論文は多い。それらの中でも、数学は同じだが、手順には差異がある。

D.1　仮説検定における前提

仮説検定とは、仮説を採用するかを決定する方法である。数多くの研究で、科学的な検証がなされたことを示すために、仮説検定が利用されている。
その手順は、データの出現頻度を予測するという方法論を十分に利用しきれていない[*1]。
仮説検定では、モデルの代わりに仮説を使う。モデルの母数に関する仮説のみを帰無仮説と定義し、帰無仮説で指定した母数ではないを対立仮説と呼ぶ。帰無仮説の元、標本の統計量以上に偏った値が得られる確率 (p 値) を計算する。p 値が 0.05 よりも小さいならば、対立仮説を採択し、$p > \alpha$ ならば判断を保留する。
仮説検定の枠組みでは、データが前提を満たさなければならないと考えられていることが多い[*2]。例えば、データは独立同一の分布関数から得られている[*3]。また、正規分布を仮定しているのであれば、データの分散と帰無仮説の分散が等しいなどである。そのため、仮説検定を行う前に、いくつかの仮説検定を行い、これらの前提を確かめる。正規分布の仮定は、*Shapiro* 検定を使う。その後、正規分布であれば、等分散検定などを行う。これらの前段階の検定では、p 値が設定した α よりも小さければ、対立仮説を採択し、$p > \alpha$

[*1] 予測を一切考えないので、モデルという考えが存在しないように見える
[*2] 仮説検定を使う研究者にとって、モデルを使った予測であるということは意識されない。モデルの仮定ではなく、仮説検定をするための前提のことである
[*3] これを確かめる方法はあるのだろうか。言い換えるなら、現象が数学的分布関数により生じていることを確かめる必要がある

であれば、帰無仮説を採用する[*4][*5][*6]。さらに、最終的な仮説検定においては、$p < \alpha$ ならば帰無仮説を棄却し、$p > \alpha$ ならば、判断を保留する[*7][*8][*9]。

■ p 値の解釈

p 値は分野によって多様な解釈がなされることがある [9, 10]。
よい解釈として以下の 6 つの原則が示されている [9]

1. p 値はデータと特定のモデルが矛盾する程度を示す指標の一つである。
2. p 値は調べている仮説が正しい確率や、データが偶然飲みで得られた確率を図るものではない。
3. 科学的な結論や、ビジネス、政策における決定は p 値がある値を越えたかどうかのみに基づくべきではない。
4. 適正な推測のためには、全てを報告する透明性が必要である
5. p 値や統計的優位性は、効果の大きさや結果の重要性を意味しない。
6. p 値は、それだけでは統計モデルや仮説に関するエビデンスの、よい指標とはならない。

■ p 値への誤解

誤解とされる解釈はも引用しておく [11][a]。以下の解釈は、統計ユーザーの流派によらず間違いであるとされることが多い[b]。

1. $p = 0.05$ ならば、帰無仮説が真である確率は 5% しかない。
2. $p \geq 0.05$ のような有意でない結果は、グループ間に差がないことを意味する
3. 統計的に有意な発見は客観的に重要である
4. p 値が 0.05 より大きい研究と小さい研究は矛盾する
5. p 値が同じ研究は帰無仮説に対して同等の証拠を提供する。
6. $p = 0.05$ は、帰無仮説のもとで 5% しか起こり得ないデータを観察したことを意味する
7. $p = 0.05$ と $p \leq 0.05$ は同じことである。

[*4] 検定により対立仮説や帰無仮説を採択することはできないが、仮説検定においてはできるという立場をとる。

[*5] 検定ではモデルを決定できない。仮説検定においてはそれができるということにして、仮説の論証がなされている。

[*6] 採択すると言い切ったが、前段階の検定においては、採択または棄却と判断してるといっておいた方が現状にあっている。

[*7] 仮説検定の手順は分野によって少しずつ異なるので、指導教員に手順を聞くことを勧める。留年したくないなら、魔術を信仰した方が良い。やれと言われたことをやらなければ論文は通らない。

[*8] 複数回の検定を行うので、多重検定の問題もあり、想定された α 水準を満たされないことが指摘されている。

[*9] 仮説検定が廃止されたとしても、過去の研究においては仮説検定が使われており、それら過去の研究を理解する必要がある。この理由から仮説検定を理解しなければならない

8. p 値は不等式の形で書かれるものである (例えば、$p = 0.015$ のときは $p \leq 0.02$ とする)。
9. $p = 0.05$ は、帰無仮説を棄却したとしたら、第一種の誤りの確率が 5% しかないことを示す。
10. 有意水準 $p = 0.05$ のもとで、第一種の誤りの確率は 5% になる。
11. ある方向を向いた結果やその方向の結果があり得ない差異を気に留めないのであれば、片側の p 値を用いるべきである。
12. 科学に関する結果や処方の方針は p 値が有意であるかどうかに基づくべきである。

[a] 原典は [12] である。孫引き引用である
[b] これらの誤解を採用している科学者もいないとは言えない。教科書でも誤解を広めていることがある

D.2　仮説検定の手順

仮説検定の手順を確認します。仮説検定では、仮説の前提が正しいことを決定する必要がある。これは、特定の分布関数にデータが従っていることを前提にし、前提が正しいならば、帰結も正しいと考えており、正しいデータを使わなければ仮説を検証できないと考えているからである[*10]。ゆえに、データと想定した仮説の前提を満たしていることを注意深く検証しながら、仮説検定を利用することが求められている[*11]。データが前提を満たすように、得られたデータを仮説の前提と一致するように、数学的な変換を施すこともある[*12]。

1. 仮説検定が使える前提が何かを確認する。前提は以下のようになることが多い。
    - 確率変数は独立同一分布に従う
    - 分布関数 (正規分布など)
2. 有意水準 α を設定する（さまざまな業界で 0.05 が設定される)。
3. 母集団から無作為抽出を行い、標本を得る
4. 標本が仮説の前提を満たしていることを確認する。標本分布と仮説の前提の分布関数がある程度一致していることを調べる。正規分布を前提にしているなら、正規分布の検定を行う。

[*10] 実際には、前提は検証できないのだが、仮説検定においては、できると決定されている
[*11] 科学的仮説検定では、現象を予測するためにモデルを使ったので、モデルの仮説をデータが満たさなくても良い
[*12] 標本分布の形が失われる

5. 標本から統計量を計算し、その値以上に大きな値をとる確率を計算する (p 値)。
6. p 値が α 以下であれば、帰無仮説を棄却し、対立仮説を採択する。
7. p が α 以上であれば、判断を保留する（最終検定前の検定では採択する)。

■過誤の概念に対する懸念

第一の過誤・第二の過誤に関する批判として [13] がある。[14] において引用されていた部分を引用しておく。

> 過誤の概念は非現実的である。根本的な問題は、我々が真実を知らないことである。現実の臨床試験では、我々は実験から学び、真実を知りたいと願うのであって、真実がすでに知られており、我々の観察を判断するのに利用できる、というようなものではない。現在利用できる情報だけに基づく決定は、それ以上の情報が利用できるときには間違っていたことがわかることもあり得る。それ以上の情報が得られないとき、決定を行なった元になる情報でその決定の評価を行うことは理論的に不可能である。一つの試験では、試験さそのものから得られる情報が、利用できる唯一の情報である。利用できる情報の調査と競合する利害の注意深いバランスを考慮した後でのみ、仮説の棄却や採択の判断が行われる。その後の試験の情報が利用できるようになるまでは、現在の判断が正しいか誤りかを判断する情報は存在しない。従って、一つの試験にとっては、過誤の考え方は全く意味を持たない。

D.3　モデルの設定

仮説検定とモデルに対応付ができるように、仮説の設定をモデルに言い換えて説明する。帰無仮説 $\mu = \mu_0$ ををを含む統計モデル $M(\mu_0)$ を帰無モデル (M_{H_0})、対立仮説 $\mu \neq \mu_0$ を含む統計モデル $M(\mu \neq \mu_0)$ を対立モデル (M_{H_1}) と呼ぶ。一般に、統計モデルの否定したい母数 μ_0 を帰無仮説と言い、その母数ではないという $\mu \neq \mu_0$ を対立かせつと言う。具体的には、データがある特定の母数 μ をもつ統計モデルの信頼区間に含まれるか否かによって、統計モデルが棄却されるかを調べる。

- i.i.d
- 数学関数
- 統計モデルの母数を μ とし、$\mu = \mu_0$

一番最後の仮説が帰無仮説と言う。対立仮説を含む統計モデル M_1 は、M_0 と同様の仮説 (1),(2) から構成されまるが、仮説 3 は統計モデル M_0 と M_1 で異なる。

- i.i.d

- 数学関数
- 統計モデルの母数を μ とし、$\mu \neq \mu_0$

一番最後の仮説が対立仮説である。M_1 の最後の仮説は、M_0 の最後の仮説の否定系である。
二つの統計モデルを作って、M_0 で計算される信頼区間に、データから得られる統計量が入らないなら、M_0 は棄却される。逆に、統計量が信頼区間に入るなら、M_0 が採択されることはない。分野によっては採択することがありうる。このように、否定したい仮説を設定し、少なくとも帰無仮説を含む統計モデルがだめだったと判断する。

参考文献

[1] 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.

[2] George EP Box. Science and statistics. *Journal of the American Statistical Association*, Vol. 71, No. 356, pp. 791–799, 1976.

[3] 池田功毅, 平石界. 心理学における再現可能性危機：問題の構造と解決策. 心理学評論, Vol. 59, No. 1, pp. 3–14, 2016.

[4] 中村大輝, 原田勇希, 久坂哲也, 雲財寛, 松浦拓也. 理科教育学における再現性の危機とその原因. 理科教育学研究, Vol. 62, No. 1, pp. 3–22, 2021.

[5] Norbert L Kerr. Harking: Hypothesizing after the results are known. *Personality and social psychology review*, Vol. 2, No. 3, pp. 196–217, 1998.

[6] 祐作大久保, 健大會場. p 値とは何だったのか： Fisher の有意性検定と neyman-pearson の仮説検定を超えるために. 生物科学 = Biological science, Vol. 70, No. 4, pp. 238–251, 04 2019.

[7] 光田暁弘. 誤差の取り扱い. 2005.

[8] サンプルサイズの決め方. 統計ライブラリー. 朝倉書店, 2003.

[9] 土居淳子. 帰納的推論ツールとしての統計的仮説検定―有意性検定論争と統計改革―. 京都光華女子大学人間関係学会 年報人間関係学, No. 13, 2010.

[10] 医療統計解析使いこなし実践ガイド:. 羊土社, 2020.

[11] アレックス・ラインハート〔著〕西原史暁〔訳〕. ダメな統計学. Sage Publications Sage CA: Los Angeles, CA, 2014.

[12] Steven Goodman. A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, Vol. 45, No. 3, pp. 135–140, 2008. Interpretation of Quantitative Research.

[13] M.X. Norleans. 臨床試験のための統計的方法. サイエンティスト社, 2004.

[14] 毒性試験に用いる統計解析法の動向 2010:. 薬事日報社, 2010.