空想統計教程

— 機械学習からはじめる科学研究のための統計学 —

Idiot

2023 年 12 月 10 日

目次

第 1 章　空想統計教程 8
1.1　生物統計学の問題点 8
1.2　生物統計学と本書の違い 9
1.2.1　ASA の p 値に関する解釈 9
1.2.2　仮説検定 10
1.2.3　仮説検定とモデル選択 10
1.2.4　頻度論 10

第 2 章　モデル 14
2.1　モデル 14
2.2　統計モデル 15
2.2.1　統計モデルとデータ 15
2.2.2　データへの過剰適合 15
2.2.3　統計モデルの仮定をデータが満たしているのか 16
2.2.4　統計モデルの機能 19
2.3　統計学の用語 19
2.3.1　母集団、無作為抽出、サンプリング 19
2.3.2　誤差・ばらつき 20
2.3.3　標本、サンプルサイズ、擬似反復、標本数 20
2.3.4　確率 21
2.4　機械学習の用語 21
2.4.1　モデルを使った推測 23
2.4.2　モデルの種類 23
2.4.3　評価指標 23
2.5　疑似乱数 25

2.5.1 乱数生成法 25

第 3 章 取り扱うデータの条件 26
3.1 実験デザイン 27
3.2 無作為抽出されていない事による過誤 27
3.2.1 仮説検証型にしなければいけない 28
3.2.2 $p < \alpha$ になったら無作為抽出を終える 29
3.2.3 統計量が xx になったときに抽出を終える 29
3.3 Questionable Research Practice(QRP) 29
3.3.1 HARKing 29
3.4 Garbage in, garbage out 35
3.4.1 実験手順を守っていないデータ 36
3.4.2 再現できなかったデータ 36
3.4.3 外れ値 36

第 4 章 統計モデル 38
4.1 正規分布を含んだ統計モデル 40
4.1.1 データが出現しやすい区間 41
4.1.2 平均値の出やすい区間 41
4.2 二つの正規モデルの比較 43
4.2.1 中心間の距離は差の絶対値 44
4.2.2 ばらつきの差異は標準偏差の比 44
4.3 正規モデルにおける中心間の距離 (効果量) 44
4.3.1 分散が等しい 2 つのモデルの効果量 44
4.3.2 最尤モデルの中心間距離 44
4.3.3 分散が等しいとするとおかしな例 45
4.4 指数分布を含んだ統計モデル 49
4.4.1 信頼区間の近似 49
4.5 対数正規分布を含んだ統計モデル 49
4.6 モデルとデータの乖離を調べる 51
4.6.1 チェビシャフの不等式 51
4.6.2 正規モデルの場合 52
4.6.3 母集団の標本が指数分布的に分布していた場合 52

4.7 累積分布によるデータとモデルの比較 56
4.7.1 累積分布の傾き 56
4.7.2 データとモデルの比較 56
4.8 qq プロットによるデータと正規分布の比較 57
4.9 AIC(相対的なモデルのデータへの適合具合) 59

第 5 章 モデルにおける統計量の性質 61
5.1 自己標本の批判 61
5.1.1 p 値の計算練習 62
5.1.2 自己標本の否定確率 62
5.1.3 母数平均の変化に応じた信頼区間 63
5.2 統計量をもとにしたモデル間類似度 (検出力) 63
5.2.1 検出力の定義 64
5.2.2 正規分布モデルの検出力 65
5.2.3 β の代数計算 66
5.2.4 最尤モデルでの β の計算 69
5.3 過誤のまとめ 69
5.3.1 サンプルサイズの設定 70
5.4 自己否定の誤推定 71
5.4.1 有意水準 α で検定が行えている 72
5.4.2 どんな統計モデルでも T 統計量で調べよう 72
5.4.3 検定を繰り返し使おう 75
5.4.4 最小の p 値を採用しよう 78
5.4.5 サンプル追加による p 値の変化 80
5.4.6 いつかは有意になる 82
5.4.7 いつかは有意にならない 84
5.4.8 まとめ 85
5.5 データとモデルの比較 86
5.5.1 p 値を使った判断に関する注意 88
5.5.2 有意水準 α で検定できない例 90
5.5.3 p 値を使うことが常に最適な判断材料 91
5.5.4 いつかは有意になる 91
5.5.5 モデルの性能 92

5.5.6 何度も検定しよう 94
5.6 サンプルサイズを決める 96
5.6.1 サンプルサイズを決めることができない 97
5.6.2 あたかも論理的にサンプルサイズを決める 97
5.6.3 実験後の検出力 98
5.7 まとめ 99

第 6 章 モデルにおける尤度比の性質 100
6.1 概要 100
6.1.1 尤度比の従う分布の数値計算 101
6.1.2 正規モデルにおける尤度比検定の計算 102
6.1.3 データとモデルの乖離を検証する 103
6.2 データと当てはめモデル $\hat{M}$ の比較 104
6.2.1 注意点 104
6.2.2 まぜると危険 105
6.3 複雑なモデルでの尤度比検定 106
6.3.1 データとの比較 109
6.4 One-way ANOVA 110

第 7 章 身長を予測する統計モデル 114
7.1 正規分布を組み入れた統計モデル 114
7.2 統計モデルによる推測 115
7.2.1 ○○ cm 以下、◇◇ cm 以上の人の割合 115
7.2.2 擬似的に無作為抽出を行う 116
7.2.3 母数によって変化する予測 116
7.3 統計モデルとデータの比較 1 119
7.3.1 極端な値を使って調べる 119
7.4 統計モデルとデータの比較 2 121
7.4.1 モデルの平均を含む信頼区間の個数 121
7.5 検定統計量によるモデルの評価 122
7.5.1 データの検定統計量と統計モデルの評価 122
7.5.2 標本平均と p 値 124

第 8 章 モデルを使った研究の進め方 126

8.1 指針 . . . 126
8.1.1 どれが科学的成果だろうか . . . 129
8.2 ダメモデルを羅列する研究例 . . . 130
8.3 2 群に対する研究 . . . 131
8.3.1 アヤメ (iris) に関する推論 . . . 132
8.3.2 アヤメのがく弁の幅を予測するモデル . . . 132
8.3.3 アヤメの分類の細分化 . . . 133
8.3.4 新たなモデルの構築 . . . 135
8.3.5 更なる生物学的な種の細分化 . . . 136
8.4 ダニの個体数 . . . 139

第 9 章 親子の身長の研究 141
9.1 独立ではない変数を持つモデル . . . 141
9.1.1 $r^2 \leq 1$ の証明 . . . 142
9.2 2 変量正規分布 . . . 142
9.2.1 確率密度関数 . . . 143
9.2.2 決定係数 R^2 と相関係数 ρ^2 の関係 . . . 146
9.2.3 独立な 2 変数モデルの回帰平均二乗誤差 . . . 147
9.2.4 相関係数が 0.8 のとき . . . 148
9.3 誤差モデル I . . . 148
9.4 誤差モデル II . . . 149
9.5 観測点を直線により予測する . . . 149
9.5.1 まとめ . . . 155

第 10 章 親子の身長の関係 157

付録 A 数理統計学 158
A.1 基本的な統計量 . . . 158
A.1.1 平均分散 . . . 158
A.1.2 逐次更新 . . . 159
A.1.3 標本の追加による平均値分散の更新 . . . 159
A.2 確率変数 . . . 160
A.2.1 確率変数がある分布関数に従う . . . 160
A.2.2 平均・分散 . . . 161

A.3 正規分布 162
A.3.1 正規分布に従う確率変数の出現しやすさ 1 162
A.3.2 より大きな値をとる確率 165
A.3.3 $N(0,1)$ での珍しい値は、$N(0,2)$ では珍しくない？ 166
A.3.4 $N(1.96,1)$ で出てくる値は、$N(0,1)$ において珍しい？ 167
A.3.5 正規分布に従う確率変数の出現しやすさ 2 167
A.4 指数分布 167
A.4.1 指数分布に従う確率変数の出現しやすさ 169
A.4.2 指数分布に従う確率変数の予測区間 170
A.5 カイ二乗分布 172
A.5.1 カイ二乗分布に従う確率変数の出現しやすさ 172
A.6 t 分布 173
A.6.1 t 分布における珍しい値 174
A.7 統計分布の関係 175
A.7.1 正規分布の再生性 175
A.7.2 指数分布の再生性 175
A.8 尤度・対数尤度・AIC 176
A.8.1 最尤推定 177
A.8.2 AIC(an information criterion) 179
A.9 マトリョウシカになったモデル 179
A.9.1 尤度比 180
A.9.2 データから推定したモデルの尤度比 181
A.9.3 尤度比検定 181
A.9.4 中心極限定理 182

付録 B 統計モデル 2 183
B.1 正規分布二つを含んだ統計モデル 183
B.2 分散について事前知識のある場合 183
B.2.1 信頼区間 184
B.2.2 検出力 185
B.2.3 σ が異なるモデルでの検出力 188
B.3 母分散の事前知識がないときの統計モデル 188
B.3.1 信頼区間 188

B.3.2 検出力 189
B.4 自己否定の誤推定 191
B.4.1 検定を繰返し使おう 191

付録 C 仮説検定の実践 192
C.1 仮説検定における手順 192

付録 D 検定とモデル 194
D.1 正規モデル 194
D.2 正規 2 モデル 195
D.3 独立性の検定 195
D.3.1 検定 195
D.3.2 モデル 196
D.4 独立性の検定 197
D.4.1 検定 197
D.4.2 モデル 197
D.4.3 計算例 198
D.5 指数分布を含むモデル 199
D.6 指数 2 モデル 200

参考文献 202

第 1 章

空想統計教程

本書は統計学を使って科学的な推論・予測を行うプロセスを説明したものである。ただし、このプロセスを通して研究を行っている人は少なく、認められていない方法である。ゆえに本書が説明する全ては、空想である。なぜこのような空想が必要であるのかを本章で説明する。

1.1. 生物統計学の問題点

一般的な生物統計学の書籍には、論文を査読プロセスに耐えるための方法論が記載されてる。そのため、生物学の各分野に特化したものが多く、それぞれ独自の特徴を持っている。例えば、以下が問題がある。

1. 繰り返し検定を行っていき、最終的な検定方法を決定する。いわゆる検定フロー
2. p 値のさまざまな解釈間違い
3. 統計モデルとデータの比較で言えそうにないことまで言う。
4. $p < 0.05$ であるので対立仮説を採択する。
5. 何のために行っているのか不明な解析

これらの特色は統計を理解している科学者から否定的に批判されており、生物学者たちもよく理解していると考えられる。しかし、なぜこれらの特色が使い続けられているのだろうか？

一般的に、生物統計学の書籍を執筆するのは、多くの論文を学術誌に投稿し、掲載された研究者たちである。掲載には 2 人程度の査読者が割り当てられ、論文を査読され、そこで使われている統計がその学術において妥当であるかどうかが調査される。この査読プロセ

スに耐えられるように統計を使うことが必要であり、その経験を元にして一般的な生物統計学の書籍が執筆されているのだと考えられる。
ただし、このような目的が論文の掲載になっているため、科学的に何をすべきかを考えることは二の次になってしまっているのではないだろうか。
生物統計学の本では、t 検定を行う前に正規性の検定を行うことを求めているものが多い。正規性の検定を行っても、サンプルが正規分布から抽出されたかどうかを確認できないから、生物統計以外では推奨されていない方法である。
もし正規性の検定を行わないことが査読プロセスで要求されているなら、正規性の検定を拒否した著者の論文は出版されにくくなると予想される。Publish or Perish の世界において、論文が出版されなければ、その研究者のキャリアは短命となることが想像できる。一方で、業界内で認められた統計解析を行い、その結果に基づいて論文を出版した研究者は、研究者としてのキャリアを築いていくことができる。このような研究者は、査読プロセスで得られた知見を元に、その業界内での適切な解析方法を学び、その後、生物統計学の本を執筆することなる。言い換えれば、推奨されない方法を学んだ科学者は生き残りやすく、そうではないものは生き残りにくいというバイアスが働きやすくなる。

1.2. 生物統計学と本書の違い

以下には生物統計学の書籍と異なる点を示す。

1.2.1. ASA の p 値に関する解釈

p 値に関する解釈は、アメリカ統計協会 (ASA) が 2016 年に発表した声明に従う [1]。以下に引用しておく。

> P-values can indicate how incompatible the data are with a specified statistical model.

日本計量生物学会が公開した翻訳では、

> P 値はデータと特定の統計モデルが矛盾する程度をしめす指標のひとつである。

この翻訳された文章では incompitable を矛盾としている。本書では適合と翻訳する。日本語の翻訳を書き換えると、p 値は、データと特定のモデルの適合を示す指標の一つである。データとモデルの適合とはどういう意味なのかを後で定義する。

ASA が p 値をどのように使ってほしいと考えているかについて関係者ではない私は理解していない。ASA の発表を私の都合に合わせて使っているにすぎない。また、ASA の解釈を採用している生物統計学の書籍は非常に少数である。

1.2.2. 仮説検定

生物学的問いからある 2 数の生物の計測可能な特徴量に関する予測を立てる。その予測とは以下のようになる。

1. OO 条件と XX 条件では という特性に $\cdots$ という差がある
2. A 群と B 群の特徴量の平均値が異なる

現在の生物統計学においては、これらの問いを有意差検定を用いて解決を試みている[*1]。具体的には、母数が同一のモデルを仮定し、そのモデルにデータが適合するかを検証する。適合していなければ、母数が異なると判定をくだしている。このような扱いは本書では推奨しない。本書ではこのような方法によってわかることは、検定統計量について適合しなさそうなモデルが明らかになった。これよりも多くのことを、有意差検定であきらかにはできない。
本書では、これらの問いについて A 群と B 群に共通する特徴について、その特徴を測定した標本が適合するモデルをそれぞれの群に対して特定する。さらに、それらのモデルの性質の違いを明らかにする。ここまでが本書で扱う内容である。生物学ではさらに、それらの適合モデルの違いが生物学的にどのような影響をもたらすのかについて考察する必要がある。この点は、本書で扱っていない。

1.2.3. 仮説検定とモデル選択

p 値を使ったモデル選択は、仮説検定と数学的に同じである。あるモデルにおける統計量の統計的性質を用いてモデル内でのデータを元にした統計量以上の値が出現する確率が p であり、仮説検定とモデル選択は同じ作業を行なっている。

1.2.4. 頻度論

定義 1.2.1. 頻度主義とは次のこととする *[?]*。

[*1] 失敗している

ある計算手続を決めたとき、その確率的性質を導き、観測データに対して計算手続きを行って得た出力に、導いた確率的性質をそのまま適用する。

頻度主義において、データがある分布関数に従っていることを前提とし、その母数を推定することが初等的な問題設定となる。扱う問題として、次の様なものがある。

大量に生産された製品の中から無作意抽出された製品に関するデータについて、そのデータの平均値と母標準偏差があたえられているとする。このとき、このデータの平均値を用いて信頼区間 95% で母平均を推測することなどに統計的な推測の考えは活用できる。[*2]

これまでの計測および解析により分布形が明らかになっていることを前提として、少数のデータからその分布形を推定することに利用されている。
一方で、生物学においては、分布形の前提が常に確定しているとは言い切れないので、分布形推定からおこなうことになる。さらに、頻度主義で行われている確率的性質を観測データにそのまま適用し、解釈するということは、生物学においてそのまま適用するしばしば不適切な解釈となる。
前提の見直しは頻度論の教化書では扱うことが少ない[*3]。本書では、頻度論統計学の知見を元に、頻度論が対象にしていない範囲で推論を行う方法を説明する。具体的には、正規分布で予測できることがわからない対象に対してモデルを構築する考え方と推定方法について説明する。さらに、頻度主義で明かになっているモデルの性質を用いて、そのモデルのデータへの適応の妥当性について議論する。そのとき、議論かのうとなる事柄について整理していく。

■「数理科学を使えば統計の ” 主義 ” を争う必要ない」

1. 「“ 数理的な方法 ” を使っても、主義の争いが解決しない」ということを示唆する事実が存在する
2. 頻度主義とベイズ主義の論争を「どちらの方法が正しいか」という争いとして捉えると論争の全体像を見誤る

a

[*2] https://www.stat.go.jp/teacher/dl/pdf/c3index/guideline/high/math.pdf
[*3] 生物統計学の教化書では、正規性の検定を行うことになっていが、検定をしてもそのデータが正規分布に従うとは言えない。

[a] 「数理科学を使えば統計の ”主義 ”を争う必要ない」という主張について検討する https://ameblo.jp/yusaku-ohkubo/entry-12588890730.html

■生物統計学では言えないことも言える

統計学を使っても主張できないことを生物統計学では主張していることが多々ある。この理由を探そうとしても大抵徒労に終わる。生物統計の解析手順いわゆる検定フローをやってるだけである。ある統計処理をしても判定できないことを判定できることにして処理を進めていることが多い。

■生物学部の卒論発表会に登場する統計学者

生物学部の卒論発表会に、統計学を学んだ先生方が登場することがある。彼等は、卒論生の発表をきき、検定フローに従って作業を行っていないばあ、厳しく叱咤していく。例えば、この検定を実施する前に、検定の仮定について検定しているのか？していないなら、この結果を信じることができないなど。生物学の先生は、統計学の先生に御礼を言い、そのおかげもあり、次の年も統計学の先生が学生たちを叱咤しに卒論発表会にくる。統計学の先生が行っていることは本当に良い行為であろうか？彼らの言うことを聞いていれば、我々にとって知りたいことがより理解できるのだろうか。
この統計学の先生は、日本以外にも出没しているように思える。

> A lot can go wrong with statistical inference, and this is one reason that beginners are so anxious about it. When the framework is to choose a pre-made test from a flowchart, then the anxiety can mount as one worries about choosing the “ correct ” test. Statisticians, for their part, can derive pleasure from scolding scientists, making the psychological battle worse.

[a]

[a] Statistical Rethinking 2nd Edition Chapter.1 P.4 より

Sato Shuntaro ｜佐藤俊太朗:

研究計画書を見ていると、「統計解析は実施しない」という記載もちょくちょく見る。検定や回帰をしないから。集計（要約）も統計解析です。

https://twitter.com/Shuntarooo3/status/1726817830690271516

第 2 章

モデル

モデルについて説明する。

2.1. モデル

モデル (模型) は次の特徴を備えている。

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. 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である。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。車体の幅などを計測し、スケール倍すれば、実物の大きさを推測できる。手のひらに収まるような小さなプラモデルでも、スケールが同じであれば、どの部分でも実物と同様に推測することができる。つまり、本物の車がなくても、スケールが合ったプラモデルがあれば、簡単に大きさを推測することができる。言い換えれば、実測した数値をモデルに組み込むことで、推定の精度を上げることができる。
車がもっとも正確なモデルと言えるが、実際に車を用いて様子を推測することは、手間がかかりますし、モデルの利便性を損なう。また、細部まで推測可能にするためには、より高精度の測定器具が必要になるため、モデルとして利用する際のデメリットになることがあります。
モデルは、推測したい内容に応じて構築されるため、完全に現実を再現する必要はありま

せん。例えば、同じ車を 3 台縦列に駐車するために必要な長さは、単純な直方体 3 つ分で推測できます。そのため、車のモデルには大きさの尺度を保っていない直方体のブロックを使用することがあります。
モデルは対象としていなかったものについても予測できる。例えば、軽自動車の大きさを予測するモデルを使ってトラックの大きさを予測することができるが、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 と予測した場合、実際のトラックの車体長は 6m よりも大きいこ。つまり、メートル単位でモデルと実際との差が生じている。
このように、モデルと実際を比較することで、このモデルではトラックの大きさを推測できないことが判明することがある。ただし、実際にどの程度の誤差が生じた場合に、そのモデルが使えないと判断されるかは、予測したい内容によって異なる。

2.2. 統計モデル

まず、統計モデルは数理統計の知識を用いて構築され、現実を推測するために使用される。例えば、以下のような仮説から構築される簡単な統計モデルがあります。

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、$f(x)$ で表される
3. (仮定 3) 分布関数の母数に関する仮説[*1]

2.2.1. 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程度を評価する指標を定め、その指標を最小化するようにモデルの母数を推定することができる。

2.2.2. 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

[*1] なお、中には三番目の仮説のみを統計モデルと主張する学派もある [2]

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、モデルに仮定の追加または変更はしない方が良い。

2.2.3. 統計モデルの仮定をデータ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、独立同一分布という仮定は、数学的厳密な定義がある。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。この定義と現実の世界に対応関係がない。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 $P(A,B) = P(A)P(B)$ 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、得られたデータの間に相関関係が全くないと、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。また、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][*3]。

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にモデルを利用するのであるから、仮定と現実との間に対応関係があるかまた、仮定を満しているということはどうでも良い。

[*2] そう考えたとしても大きな誤解にはならないのだろうが、役に立つことはないはずである

[*3] 例えば、Statistical Rethinking 2nd Edition Chapter.4 P.84 なども確認するべき

■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 [3]。

> 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.

■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

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?".

この考えかたは、最近の書籍でも紹介されている[a]。

[a] 例えば、Statistical Rethinking 2nd Edition Chapter.2 P.32 など

2.2.4. 統計モデルの機能

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。

2.3. 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。

2.3.1. 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、場所や社会的属性に対して偏りなく母集団からデータを取得することである。例えば、日本在住の 17 歳男性の身長を母集団としたとき、東京の板橋区に住む人のみを集めるのではなく、全都道府県に住む人を無作意に選択するということである。この処理により、住所に依存しないデータが収集可能になる。無作為抽出することで、都合の良い結果や偏った特性を持つ集団によるデータにならないようにしている[*4]。
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[*5]。

[*4] 無作為抽出が必要とされるのは、モデルの確率変数が独立同一分布であるという仮定を満たすためだという主張がある (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、本書ではこのように考えないことにする

[*5] この使い分けは一般的でないし適切ではない。

2.3.2. 誤差・ばらつき

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。
ばらつきとは、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重量の個体間の差異をばらつきと呼ぶ。
本書では観測により生じている誤差は、ばらつきよりも十分小さいものとして扱い、ばらつきの性質についてモデルを構築する。

■誤差を集団の特徴とし、変異（ばらつき）ととらえた

統計学の利用目的の一つは誤差論であり、これはある一つの計測対象を複数回計測した際により妥当な値を推定することである。この考え方を、ある社会集団の中における個体間の計測値の変異をあつかうため、社会学に導入したのはケトレーである。ケトレーは変異を誤差と捉え、平均値を平均人としてあつかった。この考えかたはばらつきに特に注目していない。一方、ダーウィンの従弟、フランシス・ゴルトン (1822-1911)) は、 ある集団における個体間の計測値の違いを、集団の特徴としてととらえた。 集団中での計測値のばらつきかたに意味を持たせた[a]。

[a] 現在では一般的な生物統計学において、平均値を重視することや誤差論における真値と平均値の意味を同一にとらえる。ゴルトンが重要だと捉えた変異の特徴を、現在では扱わなくなったのには何かしらの理由がありそうだ。TODO

2.3.3. 標本、サンプルサイズ、擬似反復、標本数

定義 2.3.1. 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、これを集団の変異として捉える

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[*6]。

2.3.4. 確率

確率をどのように定義するのかは昔から盛んに議論されている。Laplace らは、確率を次の様に定義した。起こりやすさに差異の認められない全ての場合の数に対する、期待していた事象の場合の数の比率。頻度主義において確率は次の様に定義される。対象の無限回の試行において、それぞれの事象が出現した割合。ベイズ主義において確率は「信念の度合い」と定義する。
Kolmogorov は確率を測度として定義した。本書ではこの定義を採用し、確率その物に関する疑問についてはこれ以上考えないことにする。

サンプルサイズの増加

本書では、サンプルサイズを大きくしていけば、モデルの予測と一致するとは考えない。

2.4. 機械学習の用語

■内挿／外挿　内挿とは、学習したモデルの仮定を用いて、学習データの近傍において、予測値を出力すること。外挿とは、学習したモデルの仮定を用いて、学習データから十分外れた点において、予測値を出力すること。

■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。

■学習したモデル　データをモデルに与え、データにモデルを当てはめたものを学習したモデルという。

[*6] 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、間違わないようにしたい。

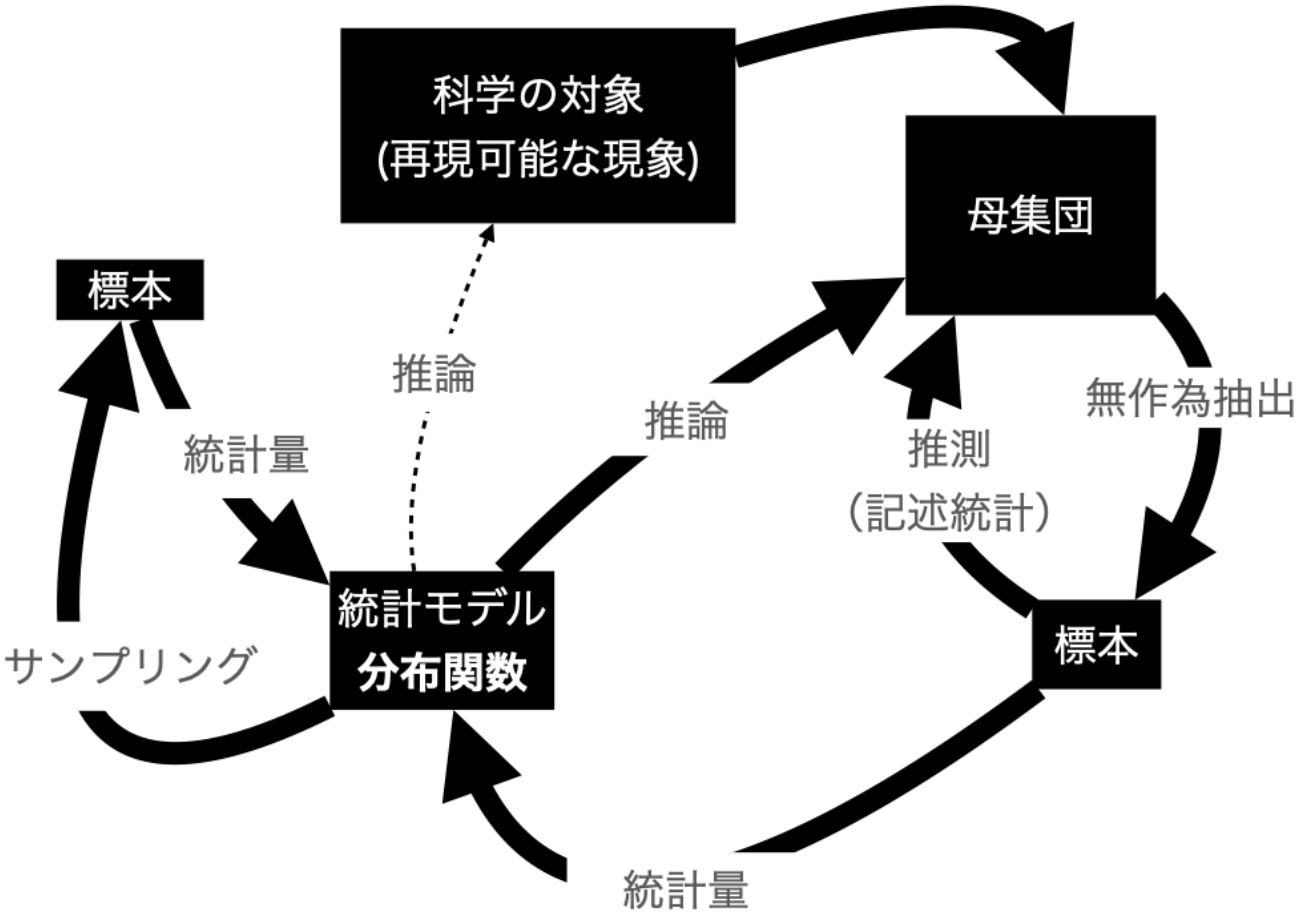


図 2.1　統計モデルによる現象の推測に関する概念図

■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。

■モデルの表現力　モデルがデータに適合できる度合い。パラメータや母数の個数のことを指していることが多い。表現力が高いほど学習用に与えたデータそのものを予測しやすくなることがある。

2.4.1. モデルを使った推測

2.4.2. モデルの種類

よく扱われるモデルを説明する。ただし、論文などではここでの説明通りにはなっていないことが多い。
データや数値を全く取り入れていないモデルをヌルモデルという。例えば、統計的仮説検定で出て来る統計モデルのいくつかは、データを取り入れない状態でもその統計的性質が判っており、その性質を使い、推論を行う。
ベースラインモデルは、既存のデータの一部およびモデルを使い、構築されたモデルのことである。次のモデルをベースラインモデルと呼ぶことがある。

1. データの平均値またはデータの一部を抜き出した何らかの定数。例えば 100 という値がデータに多く含まれていれば、100 を予測値とするモデルをベースラインモデルとする。
2. 過去の研究での最新のモデル。モデルの性能を比較するためにベースラインモデルとすることがある。

2.4.3. 評価指標

予測値の予測の良さを数量にして表したものを評価指標という。図で、モデルがそれなりによい予測を行っていることを確認した後に、それを定量的に評価するために用いる。評価指標だけでは、そのモデルが良いかはわからない。

平均二乗誤差

観測値と予測値の差の二乗に関してその平均値を RMSE という。

$$RMSE = \frac{1}{\text{サンプルサイズ}} \sum (\text{観測値} - \text{予測値})^2$$

平均的に観測値と予測値がどれくらいズレているのかを示す数値であり、RMSE が小ければ、それなりによい予測をしているだろうと判断ができる。

R^2

回帰分析では、平均値をもとにしたベースラインモデルと回帰モデルとの比較をおこなうため、R^2 という指標が登場する。具体的には、次の形で回帰モデルの予測の良さを示す物である。

$$R^2 = 1 - \frac{\sum(y_i - f(x_i))^2}{\sum(y_i - \bar{y})^2} \tag{2.1}$$

ことばでは、次の様になる。

$$R^2 = 1 - \frac{\sum(\text{観測値} - \text{予測値})^2}{\sum(\text{観測値} - \text{観測値の平均})^2} \tag{2.2}$$

ここで、(x_i, y_i) はデータ、$f(x_i)$ は回帰モデルの予測、$\bar{y}$ は、y_i の平均値である。分子は、回帰モデルとデータとの二乗誤差、分子は平近値を基にしたベースラインモデルとデータとの二乗誤差である。二つのモデルの二乗誤差を比較したものが R^2 になっている。モデルの上では、R^2 は $0 \sim 1$ の値を取る。

データを入れた場合、R^2 が 1 に近ければ、回帰モデルの予測誤差が十分小さいことを示し、R^2 が小くなると (負の値になることもある)、回帰モデルの予測誤差がベースラインモデルの予測誤差よりも大きい。R^2 が小さな値をとっているということは、回帰モデルよりも平近値を基にしたベースラインモデルの方が、平均二乗誤差の面で性能がよいことを示唆する。

例えば、$R^2 = 0.25$ であれば、

$$\begin{aligned} &\frac{\sum(y_i - f(x_i))^2}{\sum(y_i - \bar{y})^2} = 1 - R^2 \\ \rightarrow &\sum(y_i - f(x_i))^2 = (1 - R^2)\sum(y_i - \bar{y})^2 \\ \rightarrow &\sum(y_i - f(x_i))^2 = 0.75\sum(y_i - \bar{y})^2 \quad (R^2 = 0.25 \text{ を代入}) \end{aligned}$$

と計算できる。これを読み替えると、構築したモデル $f(x_i)$ による予測と実際の差の二乗和、ベースラインモデルの誤差の二乗和を 2 割 5 分したものに等しいことがわかる[*7]。構築したモデルによる予測がベースラインモデルと比較して、相対的に良い予測をしていることを数値で示すことができる。

[*7] 2 乗和誤差 2 割 5 分引きのモデル

2.5. 疑似乱数

乱数（でたらめな数）とは、人間の意図が加わらず、作為的でなく、規則性がなく、再現性もない数値列のことである。決定論を元に計算が行われているコンピュータ上では理想的な乱数列数を生成することできない。そこで、乱数の代わり擬似乱数を用いて、でたらめをまねた数値を生成することができる。

2.5.1. 乱数生成法

■逆関数法　逆関数法は、ある確率密度関数 $f(x)$ に従う疑似乱数を生成する方法の一つである。$f(x)$ の累積分布関数を $F(x)$ とする。$F(x)$ は、値域を $[0,1]$ にとる関数である。このことから、$0 \sim 1$ の一様乱数 u を生成し、累積分布関数の逆函数 $x = F^{-1}(u)$ を計算する。すると、x は、確率密度関数 $f(x)$ に従う。

■Box-Muller 法　x_1, y_1 を区間 $[0,1]$ に分布する一様乱数とする。次の、U, V は $U, V, i.i.d \sim N(0,1)$ である。

$$U = \sqrt{-2\ln x_1}\cos(2\pi y_1)$$
$$V = \sqrt{-2\ln x_1}\sin(2\pi y_1)$$

このことから、平均 μ 分散 σ^2 に従う乱数を得るには、以下のように変換すればよい。

$$U \times \sigma + \mu$$
$$V \times \sigma + \mu$$

第 3 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。

- 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。

これらが無いならば、本書で扱える範囲を超えている。統計学者を専攻した科学者に相談した方が良い[*1]。

■何を学ばなければならないか

[3] から引用しておく。

> We may ask of Fisher
> Was he an applied statistician?
> Was he a mathematical statistician?
> Was he a data analyst?
> Was he a designer of investigations?

[*1] 最初から相談した方が良い

> It is surely because he was all of these that he was much more than the sum of the parts. He provides an example we can seek to follow.

3.1. 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。

■あとでまとめたい TODO

'Pseudoreplication' problem - 接着したコドラートは何が悪いのか `http://www.mus-nh.city.osaka.jp/iso/argo/nl02/nl02-21-32.html`

3.2. 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。

Sub Group の解析を一般論にする

母集団 A を設定し、その標本を 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。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

図 3.1　Sub Group の解析結果から一般論を展開する

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。
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[*2]、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[*3]。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。

3.2.1. 仮説検証型にしなければいけない

論文として投稿するときには、データから仮説を生成し、その仮説を検証したというように体裁をととのえるように要求されることが多々ある。

■仮説検証型／仮説探索型

> **Satoshi Tanaka (@sato51643335):**
> 論点が盛りだくさんでこの問題に詳しくないと理解しづらいかもしれません。検証的研究における HARKing は不正ですが、探索的研究では問題ではないので、誤解が起こらないと良いのですが。
>
> https://twitter.com/sato51643335/status/1645911659213643776

仮説検証型または仮説探索型だとしても、計画し、実行したことそして、得られたデータを改竄／隠蔽なく報告することが必要である。

[*2] 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！
[*3] 実際調査した母集団は母集団かつ $p < \alpha$ に対して、報告した母集団デカすぎんだろ...

3.2.2. $p < \alpha$ 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。

3.2.3. 統計量が 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。

3.3. 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[*4]。

3.3.1. HARKing

仮説を元に実験を計画し、計測を行い、解析を行う。その仮説以外の仮説を検証するために実験を行ったと報告することを、仮説ハッキング ($HARKing$(Hypothesizing After

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

the Results are Known)) という[*5] [*6][*7][*8][*9]。これは、研究不正である。
HARking が起こる過程を説明する。データを解析したあとに、その解析結果をもとに仮説を構築する。さらに、あたかも元からその新たな仮説を検証することを目的とし研究をおこなったかのように偽り報告をおこなう。もちろん、仮説がなかったかのように改竄するのも研究不正になる。
データを解析した後に、仮説を作ると、データに合せた仮説になりやすく、そのデータにのみ適合的な仮説になる。機械に特定のデータを読み込ませ、そのデータしか予測できなくなることを「モデルが過学習した」と言う。ここでいっているデータへの過剰適合した仮説は、人間がデータを過学習した結果、そのデータのみを予測する仮説をつくることであると言い替えることができる。

■ 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

[*5] HARKing は、再現性の問題という意見もある。https://twitter.com/ykamit/status/1077716200845500416 。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 [4, 5]。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、[6] に詳しくまとめられている

なる。

■学術分野によってことなる HARKing の認識

Takefumi Nakazawa:
生態学では、データを取ってから研究の筋書き（イントロ）を考える事がある。研究の動機が対象への興味だったり、結果が予想と違う事が多かったりするからだろう。研究発表には新規性が必要なので、どうやって大きな新規性を謳うかのか、どういうイントロを構成するか、特に学生さんは悩むだろう。

https://twitter.com/Take_Nakazawa/status/1635760385436557312

Takefumi Nakazawa:
新規性を謳うには、自分の材料や技術に捉われず、他の多くの研究事例の中で、自分の研究をできるだけ一般化する必要がある。攻めれば大風呂敷と言われ、守れば新規性が弱いと言われ、勉強不測のために新規性がないと言われることもある。この辺の「スキル」が上手な人は早く論文を書ける。

https://twitter.com/Take_Nakazawa/status/1635762588788326400

Takefumi Nakazawa:
研究の材料と技術は研究室で提供され、体系的に習得しやすいが、筋書きを作る技術には、他の研究に関する知識に加えて、ある種のセンスが必要なのではないかと思う。知識は学べば得られるが、新規性を謳うには既存の知識を統合した上で、欠けている大きな穴を見つけなければいけないからだ。

https://twitter.com/Take_Nakazawa/status/1635764423599198208

Takefumi Nakazawa:
生態学会で沢山のポスター発表を見て、こんなデータでこんな筋書きを作るのかと感じることが多々ある。きっと当初はそんな筋書きはなかったのだろう。始める前より面白い筋書きになって、寧ろ面白いと感じたかもしれない。データを取るだけではなく、これが生態学なのかって勉強した人もいるかも。

https://twitter.com/Take_Nakazawa/status/1635765545932042242

Takefumi Nakazawa:
でも、本心で言えば、みんながみんな大風呂敷を広げて生態学の偉大な理論を見つけたかのように新規性を高らかに謳っている学会は、少し距離を置いてみてみると、異様な学会に見えたりもする。外部の人にとっては、何が重要な研究テーマ間の花、生態学は何を目指しているのか、わからないのではないか。

https://twitter.com/Take_Nakazawa/status/1635766641157087233

Takefumi Nakazawa:
自然界は複雑で多様だから、そんなことは仕方がないし、それがむしろ面白いとは思う。時々見られる大風呂敷に感動することもある。でも、新規性の呪縛に捉われて、見ている現象や系の面白さが台無しになっているような発表もあったりして、心苦しく思うこともある。

https://twitter.com/Take_Nakazawa/status/1635768265611022336

Takefumi Nakazawa:
研究発表の場だから、発表者は聴衆のために一般化された筋書きを用意しているだけかもしれない。センスのある筋書きはそれはそれで面白いけど、私が一番面白いと感じるのは、他人のために無理やりに一般化された筋書きではなく、本人が面白いと思って取ったデータをその思いのまま伝えた発表なのである

https://twitter.com/Take_Nakazawa/status/1635769045046919169

Ohkubo Yusaku:
（ガチのフィールド研究室だったので「データからの筋書き」が生態学を駆動する大きな力であることは認めた上で、）幅広い学術領域で HARKing の問題が指摘されてる以上もう少し現状のプラクティスを見直すか、「データからの筋書き」を擁護する真面目な理屈を考える必要があるように思える。

https://twitter.com/Ohkubo2021/status/1637804153039917059

Takefumi Nakazawa:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0200303

https://twitter.com/Take_Nakazawa/status/1637832643567050754

ある学術分野では HARKing が行われている。「やってはいけないのになぜやってるんですか？」ときかれても困る。やってるのである。

Ken McAlinn:
これは (広義の) 研究不正だと思うんだけど、そういう認識がないんだろう。それ自体は仕方ないんだけど、悪いインセンティブが働きすぎで是正はかなり難しいように思う。引用 https://twitter.com/Take_Nakazawa/status/1635760385436557312

https://twitter.com/kenmcalinn/status/1637816619605827584

ballman:
これ書かれてる先生自身はその問題を認識さてるというのは置いといて、自分の友人でも生態学やってる人いますが統制された実験がそもそも可能かというところもあるので、分野的には探索探索で割り切ったらええんちゃうかなとは思います

https://twitter.com/katsuymd/status/1637819533355548672

Ken McAlinn:
あまり研究不正として認識はしてないと思うんですよね。まぁ探索なら探索でいいんですけどね。

https://twitter.com/kenmcalinn/status/1637822015175262208

HARKing は研究不正。実験を行った後で、仮説を作成し、その仮説を検証するために実験を行ったと述べれば不正。もとの仮説を改竄しても研究不正となる。

■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

3.4. Garbage in, garbage out

ある集団の統計的性質をモデルを使って調べたい。特定のモデルにおいて発生頻度が著しく低い集団については、そのモデルで出現頻度に関して十分な予測ができない。言い換えれば、モデルにおいてメジャーな集団における性質を調べるためには、モデルにおいて発生頻度の低い集団を省くことで、メジャーな集団の性質を理解しやすくなる。そこで、データの中からいくつかを取り除くことが必要になる。

ゴミデータを統計処理したならば、ゴミのような推論しか得られない。言い換えれば、意味のある結果を返しにくくなったり、解釈が難しくなることがある。このことを Garbage in, garbage out と呼ぶ。
特定の統計量を予想していた値に近づけるためにデータを取り除ってはいけない。我々の行いたいことは、注目したい現象の統計的特徴を調べることである。
成功した実験結果と失敗した実験結果とで計測した数値が似ているにていたとする。この場合、すべてを混ぜて解析をおこなうと、統計量が過大または過小評価される[*10]。

3.4.1. 実験手順を守っていないデータ

実験手順を守っていないデータを何ら断りも入れずに、手順を守ったデータと同一視し、解析してはいけない。

3.4.2. 再現できなかったデータ

ある手順に沿って実験を行ったとしても常に同一の結果が得られないことがある。例えば、次のようなことが考えられる。

- 動物に休眠を誘発する薬剤を投与したにもかかわらず、休眠しなかった
- ある遺伝子が発現するように手順を踏んだにもかかわらず発現しなかった

いずれもその後、ある量を計測し、それを解析するという手順だとする。同一視できない現象のデータを混ぜて解析することは辞めておいたほうがいい。
解析対象とした結果が生じた数と実験回数も報告の対象になる。報告しなければ、データの隠蔽と判断されかねない。

3.4.3. 外れ値

外れ値とは、実験手順通りに処理を行い得られたデータではあるが、事前の想定を外れた値のことである。次のデータを外れ値とすることがある

- 研究者が想定していた計測値から著しく離れた値をとったサンプル
- 既存研究で想定したモデルではめったに出現することがないと考えられるサンプル

[*10] どちらかといえば、解析担当者が、なんとなくサンプルサイズは大きいほうがいいという気分があり、すべてを同一視して解析をしたいという気分になり、ゴミを含めたまま解析しがち。

なんらか想定が研究計画にないならば、サンプル内において注目する集団を恣意的に決定し、それ以外を外れ値として処理することになる。どのような基準を用いてサンプルを外したのかや実験手順を元に行われた全サンプルの数も、報告の対象になる。外した理由や存在を報告しないならば、データの隠蔽行為である。
統計解析を行った後で、仮説に合わないと判断した少数のデータを除いて、隠蔽してはいけない。具体的には、在る統計量が期待した値に達しなかった[*11]等の理由から、いくつかのデータを除いて、等計量を期待通りの値にすることが想定される。

■欠損値を平均値で埋める

> **@M123Takahashi:**
> 予測モデルの精度を上げるためであっても，欠測値を平均値で代入処理することは一切お勧めしません．添付の図のとおり，予測値 yhat2 は実測値 y1 に対して一列になってしまうので，MCAR であっても，平均値代入法で予測モデルの精度は下がります．
>
> https://twitter.com/M123Takahashi/status/1658036202325573633

> **@M123Takahashi:**
> 予測精度を上げるときの欠測値処理をするときは，平均値の代入は常に悪いので，このツイートの内容も残念ながら的外れです．決して精度の上がらない方法を用いることは，倫理的に正当化される場面はないと思います．(2/2)
>
> https://twitter.com/M123Takahashi/status/1658036205194461186

本書が扱っている範囲では、代表値でデータを埋めないほうがよい。

[*11] $p > 0.05$ だった。

第 4 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。

4.1. 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) $x_1, x_2, \cdots, x_n, i.i.d. \sim F$
(2) その分布 F 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。以下では、正規モデル $M(\mu)$ による予測について説明する。

■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、最尤推定量が母数を推測する上で良い推定量であるとは限らないことを指摘している [7]。

> 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。

本書で扱うデータは、分布形が指定されていないため、最尤推定法で得た母数を取り入れたモデルが予測に適しているとは限らない。例えば、中央値などの他の推定量を母数に入れたモデルの方が、良い予測が可能となることがある。

4.1.1. 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。

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

4.1.2. 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、次の様に式変形できる。

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

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、その頻度は 0.954 である[*1]。これは、標準正規分布の $[-2, 2]$ の積分値と当然一致する。この区間を 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。この範囲では、

$$\begin{aligned} & -z_{0.025} < \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} < z_{0.025} \\ \rightarrow & \mu - z_{0.025}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{0.025}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

である。この区間を 95% 信頼区間と呼ぶ。
正規モデルにおいて、確率変数が標準偏差の 2 倍を超える値をとる確率は 4.6% である。ある確率変数以上の数値がみつかる確率を 4.6% を切上げた 5.0% を採用し、統計量についても、5% となる範囲として $[-z_{0.025}, z_{0.025}]$ がよく使われる[*2]。

サンプルサイズによる信頼区間への影響

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 4.1 は、信頼区間が N に応じて変化する様子を図示した。
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、複合的に指標を検討する必要がある。

[*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、$\#\{x \text{ は } M(\mu; \sigma^2) \text{ からサンプリングされた変数の組}; Z(x) \in [-2, 2]\}/\text{標本数} = 0.954$ 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。

[*2] 正規モデルの場合、標準偏差の 2 倍の区間に確率変数が 95.4% ほど見付かるが、あとで説明するように、他のモデルではそのようにならないこともある。

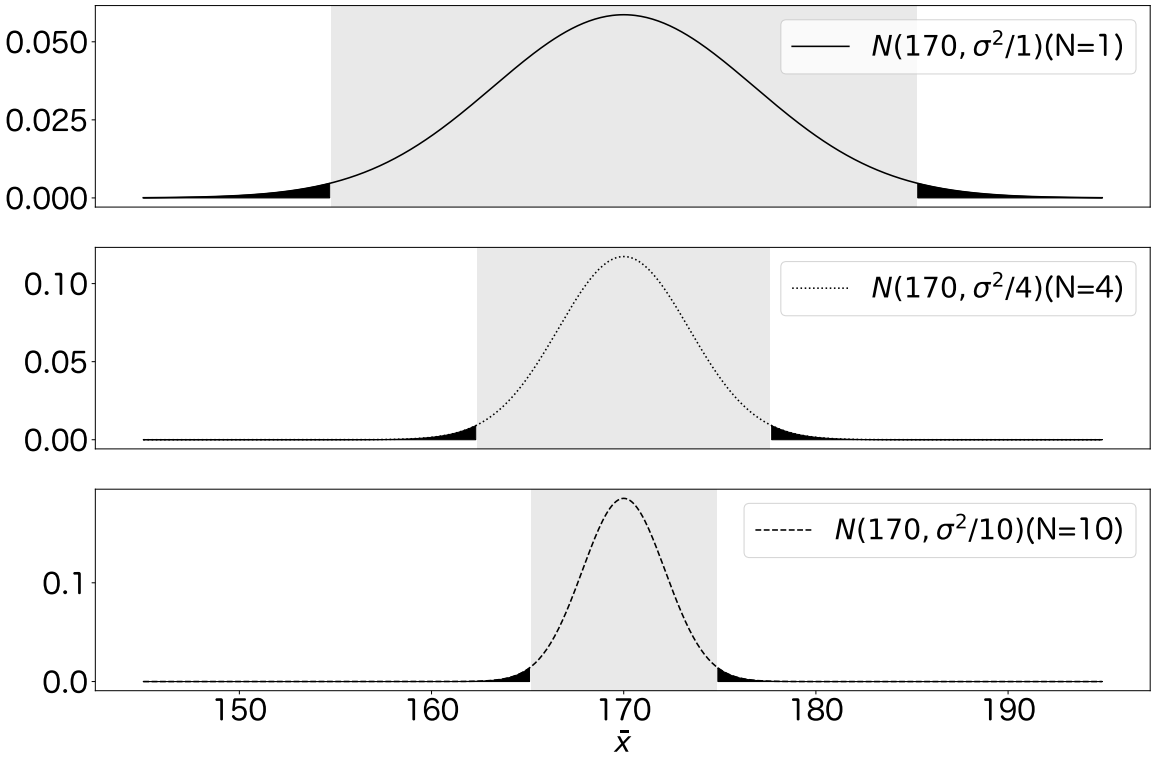


図 4.1　モデル $M(170; \sigma^2 = 5.8)$ における (A)$N = 1$,(B)$N = 4$,(C)$N = 10$ での 95% 信頼区間

4.2.　二つの正規モデルの比較

2 つの正規モデルを、それぞれ次のとおり定義する。

$$M_A = M_N(\mu_a, \sigma_a^2)$$
$$M_B = M_N(\mu_b, \sigma_b^2)$$

この二つのモデルを比較する。

4.2.1. 中心間の距離は差の絶対値

中心間距離は、モデルの中心 μ_a から μ_b までの距離として、その差の絶対値で定義する。

$$|\mu_a - \mu_b|$$

である。

4.2.2. ばらつきの差異は標準偏差の比

正規モデルにおけるデータのばらつき方の違いは、モデル間の標準偏差の比によって定量的に評価する[*3]。

$$\frac{\sigma_a}{\sigma_b}$$

標準偏差の比が ~ 1 であるならば、だいたい同じような標準偏差であることを示し、1 より十分大きいならば、モデル M_b のばらつきは、モデル M_A のばらつきにくらべてかなり大きいことがわかる。

4.3. 正規モデルにおける中心間の距離 (効果量)

ここでは、分散が等しく、平均が異る二つの正規モデルについて、その間の距離を考える。

4.3.1. 分散が等しい２つのモデルの効果量

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、$D\sigma = |\mu_a - \mu_b|$ であり、２つのモデルの中心からの距離が標準偏差 σ の D 個分であることを示す[*4]。

4.3.2. 最尤モデルの中心間距離

二つの母集団 A,B からそれぞれサンプリングを行った標本 $X = (X_1, \cdots, X_n), Y = (Y_1, \cdots, Y_m)$ について、次の最尤正規モデルを考える。それぞれの集団にたいして、モ

[*3] 問題:ばらつきを標準偏差の差で定義しないのはなぜでしょう。

[*4] 効果量の記述をみたら、二つのモデルの確率密度関数があたまに浮かび、それらのピークの間が $D\sigma$ くらい離れている図があたまに浮んでくるだろうか

デルとして、$M_A(\bar{X}, \sigma_{AB}^2), M_B(\bar{Y}, \sigma_{AB}^2)$ とする。ただし、ここで $\bar{X}, \bar{Y}$ はそれぞれ標本 X, Y の平均値であり、σ_{AB} は、二つの標本からプール (重み付き平均) した標準偏差である。具体的には、

$$\begin{aligned}\sigma_{AB}^2 &= \frac{n\sigma_A^2 + m\sigma_B^2}{n+m} \\ &= \frac{n \times \frac{1}{n}(\sum_i^n (X_i - \bar{X})^2)}{n+m} + \frac{m \times \frac{1}{m}(\sum_i^m (Y_i - \bar{Y})^2)}{n+m} \\ &= \frac{(\sum_i^n (X_i - \bar{X})^2)}{n+m} + \frac{(\sum_i^m (Y_i - \bar{Y})^2)}{n+m}\end{aligned}$$

である。最尤モデル M_A, M_B の中心からの距離を、標準偏差で割った量を D とする。D を具体的には、

$$D = \frac{|\bar{X} - \bar{Y}|}{\sigma_{AB}}$$

である[*5]。いつでも効果量を報告することは推奨できない。このことを調べてみる。

4.3.3. 分散が等しいとするとおかしな例

分散の異る標本を得た場合、効果量の意味がとりにくいことを示す。
ふたつの条件 A, B で何らかの実験を行ったとして[*6]、それぞれで表 4.1 のような統計量が得られた。それぞれの条件で、正規モデルで推測してみても良さそうだと判断したとする[*7]。
分散の違いを示すため、分散比を計算すると、分散の比は、$\frac{19.63}{10.31} \sim 2$ である。分散が同じとしてモデルを作れば、それぞれの標本を予測しにくくなると考えられる[*8]。この判断を行えば、効果量によりモデル間の距離を測るのも難しいと判断ができる。あえてここでは、同じでもええかと実験者が考えた場合、モデルとデータが乖離してしまうことを示す。
プールした分散は、$\sigma_{AB} = 16.37$ である。プールした分散を用いたモデルは、それぞれ

$$M_A(0.78, 16.3^2)$$
$$M_B(10.27, 16.3^2)$$

[*5] ここで考えた D は一般には、Cohen's D と呼ばれている量と一致する。ただし、モデルを考えずに定義するのがならわしである。

[*6] 実際には、正規分布 $N(0, 20^2), N(10, 10^2)$ からサンプルサイズ 100 の標本を得た。

[*7] ここで複数の疑問が生じる。平均はプールしないのは理由をおおよそ理解できるが、分散をプールしたほうがいい理由は自明ではない。データが二つの標本分散が同じ程度であることを示し、それらを統一したモデルで詳細をしらべたいという意思の部分がある。

[*8] 分散がいっぽうの 2 倍程度違うので、同じ分散としてモデルを構築し、データを予測できるとは考えにくい。

表 4.1　条件 A, B で得られた標本の統計量

	条件 A	条件 B
サンプルサイズ	1000	1000
平均	0.78	10.27
分散	19.63	10.31
最小値	-61.08	-21.53
25%	-12.00	3.53
50%	0.83	10.31
75%	14.08	17.30
最大値	79.17	44.33

である。最尤モデルはそれぞれ

$$M_A^{\mathrm{ML}}(0.78, 19.63^2) \tag{4.1}$$
$$M_B^{\mathrm{ML}}(10.27, 10.31^2) \tag{4.2}$$

である。それぞれのモデルがデータをよく予測するかを確認する。表 4.2 は、それぞれのモデルにより推定した、1 標準偏差,2 標準偏差の区間に期待される量のデータが含まれているかを調べた結果である。それぞれの実験において、最尤モデルは、モデルが期待する区間に対応する量のデータが含まれている。それぞれ、694 個と 671 個であり、モデルの予測と十分近いといってもいいだろう。一方で、プールされた分散を用いたモデルは明かに、予測と乖離する。これは、二つの条件の間で分散が異ることで、モデルの予測精度が低下しているからである。以上のことは、データとモデルが乖離しており、プールしたモデルを使わない方が良いという証拠の一つになる[*9]。

図 4.3.3 には、それぞれのモデルにおける確率密度関数を示した。現状の解析では、(c) の最尤モデルがもっともデータを説明できている。このモデルをなかったことにして、(d) を採用し、(d) 中の矢印間の距離を標準偏差で規格化した量を効果量とよぶことにする。効果量がデータを予測できないモデルの中心間距離を標準偏差で規格化した量になる。予測に適さないモデルに関する性質となっており、これでは効果量の意味を捉えにく

[*9] 実践 (論文) では、ここまで考察しているのかが府明瞭なことが多い。

表 4.2　モデルとデータの乖離に関する調査

条件 A	68%	95.4%
M_A^{ML}	694	950
M_A	604	902

条件 B	68%	95.4%
M_B^{ML}	671	958
M_B	892	999

い[*10][*11][*12][*13]。

正規モデル以外の場合

正規モデルではデータとの乖離が激しく、正規分布以外の分布形を仮定したモデルを使った場合を考える。例えば、指数分布を仮定した場合、分散をプールしてしまえば、平均も一致するので、効果量を定義通りに計算すると常に 0 である。正規モデルがデータと乖離していない場合には使える量である[*14]。

■ Normal(**正規**) **分布**

Normal(正規) 分布はもともとガウス分布という名前がついていた。同時期に複数人が Normal とよんだ。1873 年に哲学者のチャールズ · サンダー · パース、1877 年にドイツの統計学者ヴィルヘルム · レキシス、そして、先に紹介したゴールトン。

[*10] 多くの論文では、データが正規分布に近いのか、分散が等しいのか評価していないので、効果量が何を示しているのか判断つきにくい。ただし、コントロール郡と対照郡とで分散がほとんど同等であったという実験も多いため、分散の評価をしないのかもしれない。

[*11] さまざまなモデルで調べてみたという言い訳も考えられるが、明かにデータを説明できないモデルの性質を示すことに意味があるのだろうか

[*12] 効果量を大きくするにはどうしたらいいだろうか？観測データの中心から外れた大きめの値を抜けば、分散が小くなり、効果量は大きくなる。やりたいほうだいだ！

[*13] 問題:もっとデータに対して適切な分散は存在するだろうか？例えば、$\bar{z} = n\bar{x} + m\bar{y}$ として、分散を $\sigma^2 = \frac{\sum^n (x_i - \bar{z})^2 + \sum^m (y_i - \bar{z})^2}{n+m}$ とし、それぞれのモデルを構築してみてはどうだろう。データにフィットしているだろうか？不偏分散を採用して、モデルを構築したものだとどうだろう。データを説明できるだろうか？

[*14] ただの分散の等しい二つの統計モデルの中心間の距離が標準偏差何個分かを示しているにすぎない。

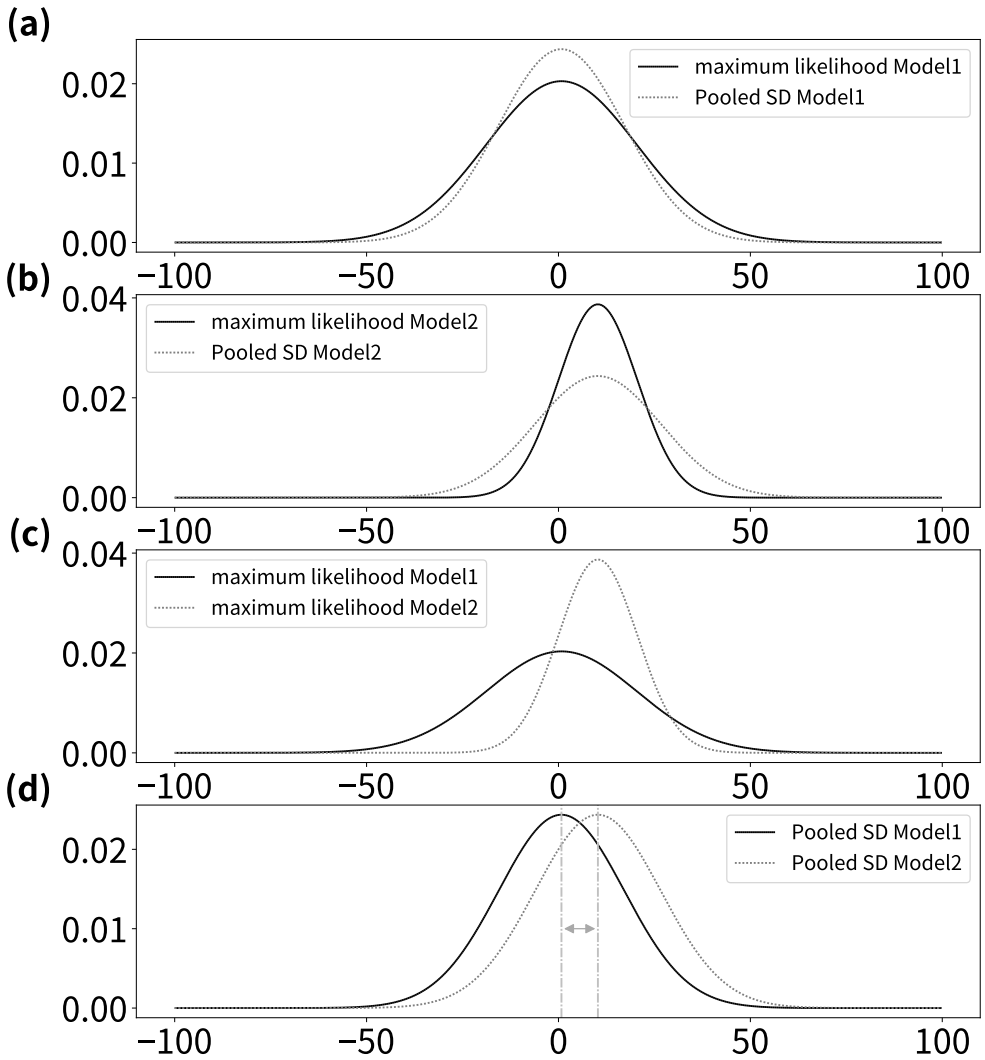


図 4.2　(a,b) 条件 A, B に対する最尤モデルとプールされた分散によるモデルの確率密度関数。(c) 条件 A, B の最尤モデルの確率密度関数。(d) プールされた分散のモデル。矢印間の距離を標準偏差で割った値が効果量。実際に知りたいのは、(c) における二つのモデルの性質の差異。

4.4. 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 D.3 である。

4.4.1. 信頼区間の近似

95% 信頼区間 (式 D.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 4.4.1)。$\lambda = 10$ 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が一致しやすくなる。

4.5. 対数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対数正規モデルと呼ぶ。

(1) $x_1, x_2, \cdots, x_n, i.i.d. \sim F$
(2) その分布 F は、対数正規分布
(3) 対数正規分布の母数 (平均と標準偏差) はそれぞれ μ, σ。

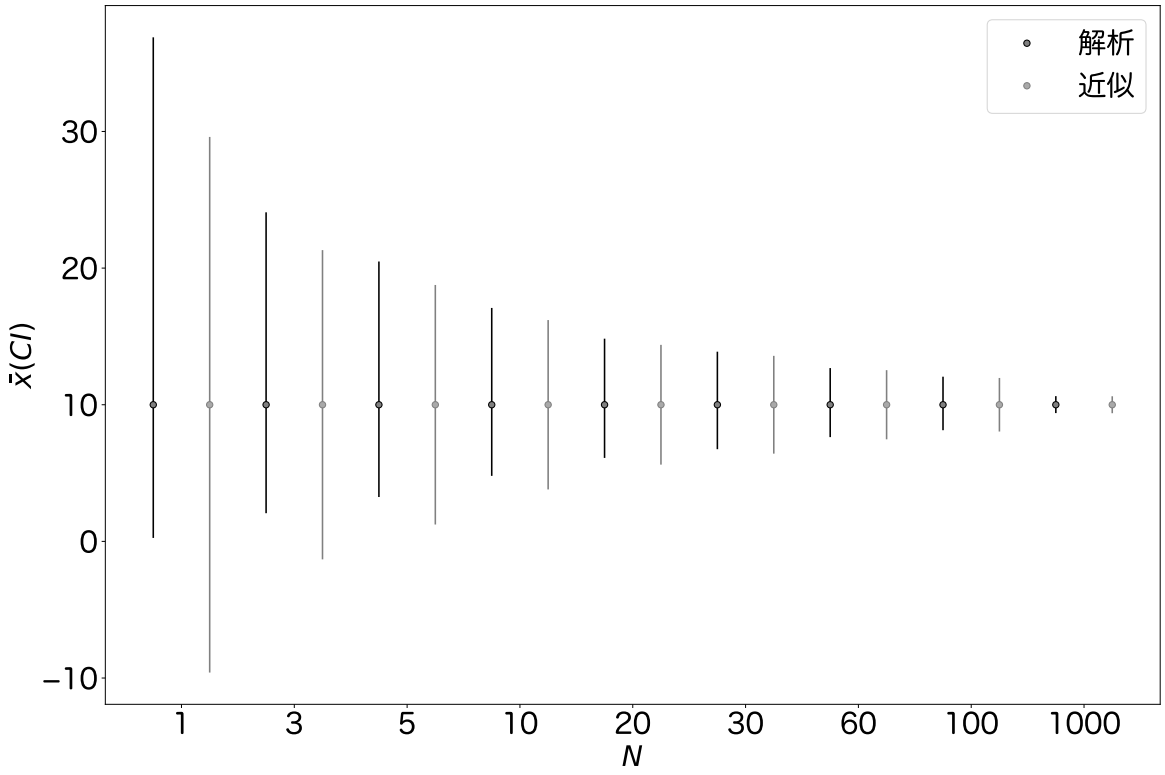


図 4.3　解析的な信頼区間と近似によって求められた信頼区間。$1/\lambda = 10$ とする。横軸にサンプルサイズ。縦軸に、平均値。エラーバーは信頼区間 (CI)。

この対数正規モデルを $M_{\log}(\mu, \sigma)$ と書く。最尤正規モデルの母数は、

$$\mu_{ML} = \frac{1}{n}\sum_{i=0}^{n} \log x_i$$

$$\sigma^2_{ML} = \frac{1}{n}\sum_{i=0}^{n} (\log x_i - \mu_{ML})^2$$

であり、そのモデルを $M^{ML}_{\log}(\mu_{ML}, \sigma_{ML})$ と書く。

4.6. モデルとデータの乖離を調べる

4.6.1. チェビシャフの不等式

補題 4.6.1. 確率変数 x が平均 μ, 分散 σ^2 であるとき、次の不等式が成り立つ。

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

ここで、k は任意の正の数。不等式の不等号を変えてやると、

$$P(|X-\mu| \leq k\sigma) \geq 1-\frac{1}{k^2}$$

この不等式を使えば、標準偏差をばらつきの基本単位として、その単位の中にどれくらいデータが入るのか推定できる。例えば、$k=2$ の場合を記述すると、

$$P(|X-\mu| \leq 2\sigma) \geq \frac{3}{4}$$

μ から 2σ の範囲の中にある確率変数は、75% 程度より多いことを示している。

```
1  x = np.arange(0,12,0.01)
2  y = np.sin(x/(2*np.pi)*20)
3  plt.plot(x,y)
4  plt.show()
5
6  np.random.shuffle(y) # やらなくてもよい
7
8  df = pd.DataFrame(y,columns=['x'])
9  sns.swarmplot(data=df, y="x"); plt.show()
10
11 mu,sigma = np.mean(df['x']),np.std(df['x'])
12 x = df['x']
13 print(mu,sigma)
14 print(np.sum( np.abs(x-mu)<=2*sigma )/len(x), np.sum( np.abs
       (x-mu)>=2*sigma )/len(x))
```

最後行の出力は、3/4 以上、1/4 以下であり、計算してみると、期待通り 1,0 が出力されている。

4.6.2. 正規モデルの場合

正規モデル $M(\mu)$ によりある現象を予測できるのかを考える。ここで、サンプルサイズ 1 の標本を使によって、モデルの良さを考察できるかを考える。$M(\mu)$ 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 (平均に対してデータが対称的に分布している)。
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。

4.6.3. 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モデルとしてみる。
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 4.6.3)。それぞれの分布の右側のエラーバーは、SD(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、正規モデルの予測と乖離する。このことは、図内の点が上下対称ではない点などで乖離していることを見わけることができる。

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 4.6.3)。図 4.6.3 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 1SD(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、データが予測区間に入っている割合が予測 (68%) と一致する。一方で、上でのべた様に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。このことが図 4.6.3 の Exponential から読み取ることは難しく、描画された範囲の中に 68% のデータが含まれていると考えてしまう。

エラーバーをみたとき、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示していると考えてしまう。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

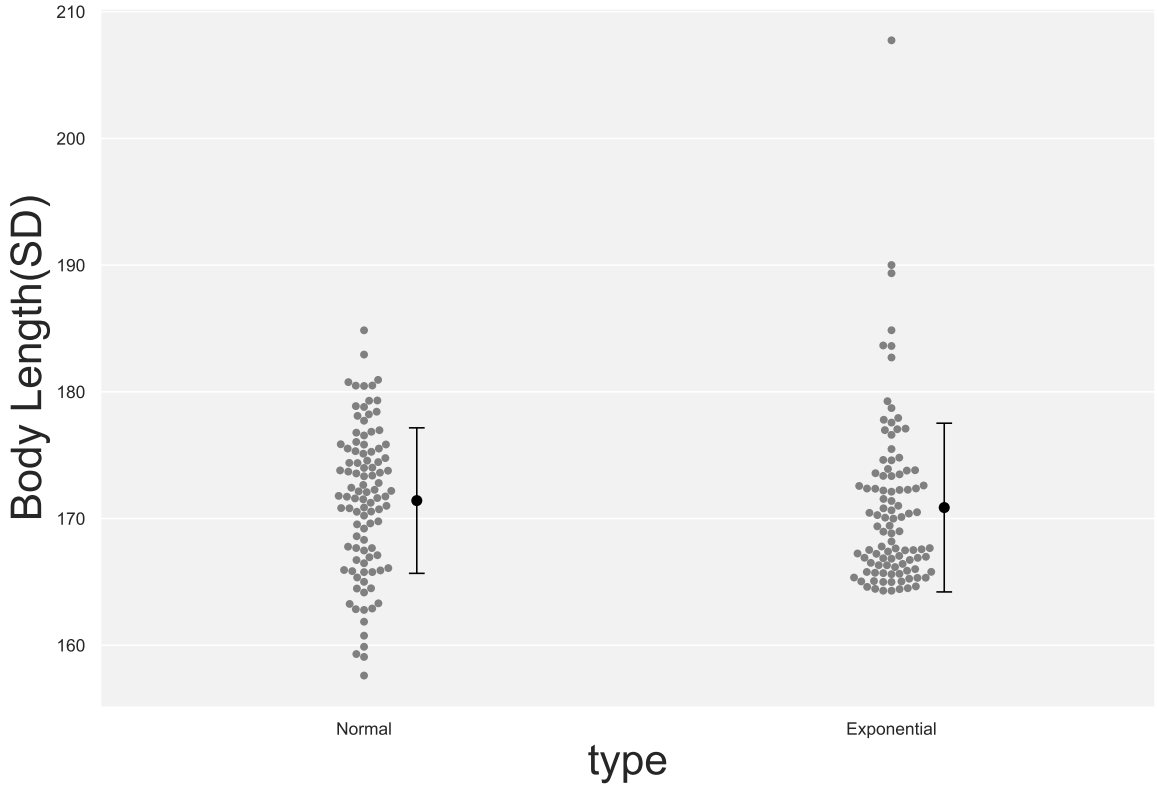


図 4.4　正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本をプロットした。それぞれの右側にあるエラーバーは、正規分布モデルが予測した 68% 予測区間。

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[*15]。
これでは、せっかく集めたデータを正しく伝えることができない。そこで、スワームプロットやボックスプロットなどの描画方法を使うことで、データのばらつきかたをより具体的に示すことができる。また、次の節で説明する方法により、より具体的な分布の形を特定することもある。

[*15] 読者にはデータが平均値を中心に対称に分布していると思わせることができるともいえる。

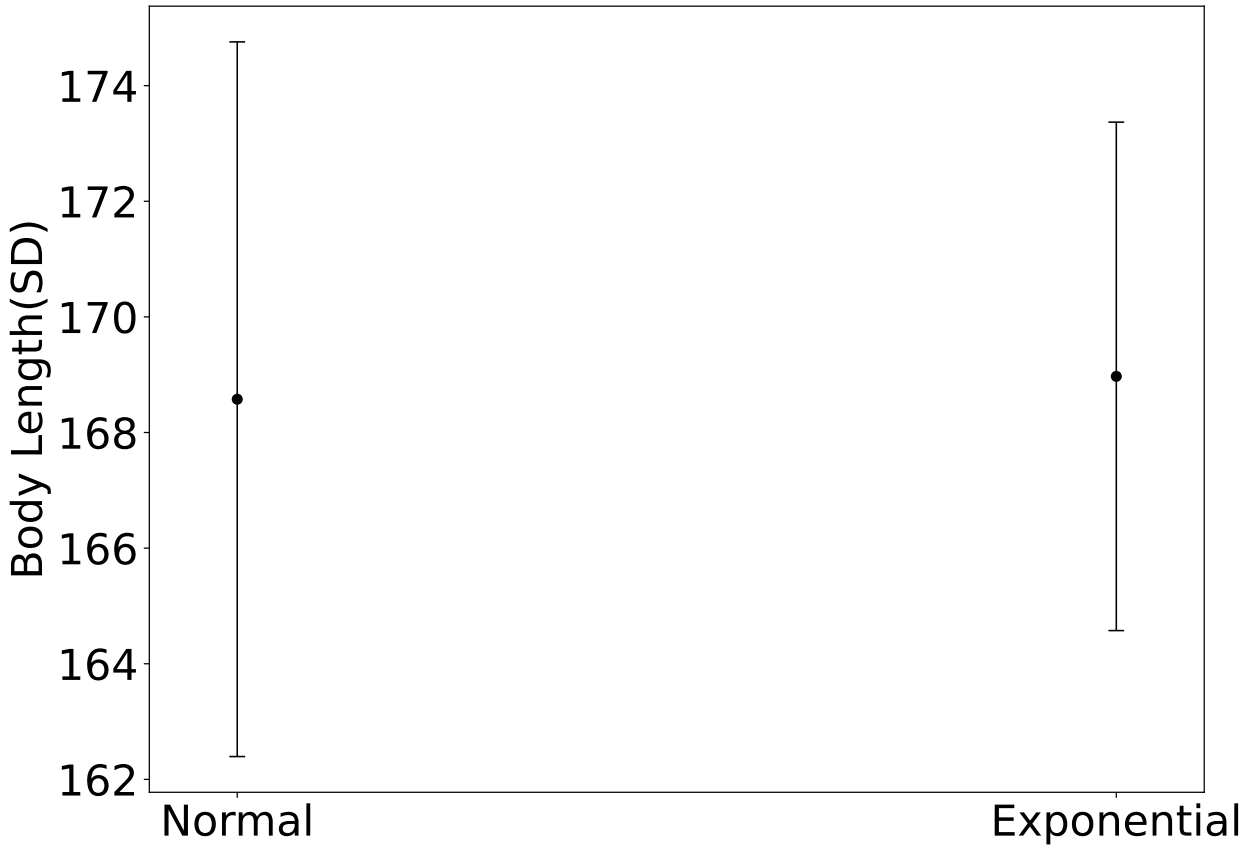


図 4.5　正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本を得た。その標本から推測される 68% 予測区間を描画した。

精度の良さを示すエラーバー

測定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、ばらつき方は大抵正規分布的である。また提案手法の SD がそれまでの計測方法よりも小さければ良い計測であるので、SD を表示している。

4.7. 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、

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

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

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

4.7.1. 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 4.7.1)。

4.7.2. データとモデルの比較

図 4.8 図 4.8 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 4.8 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

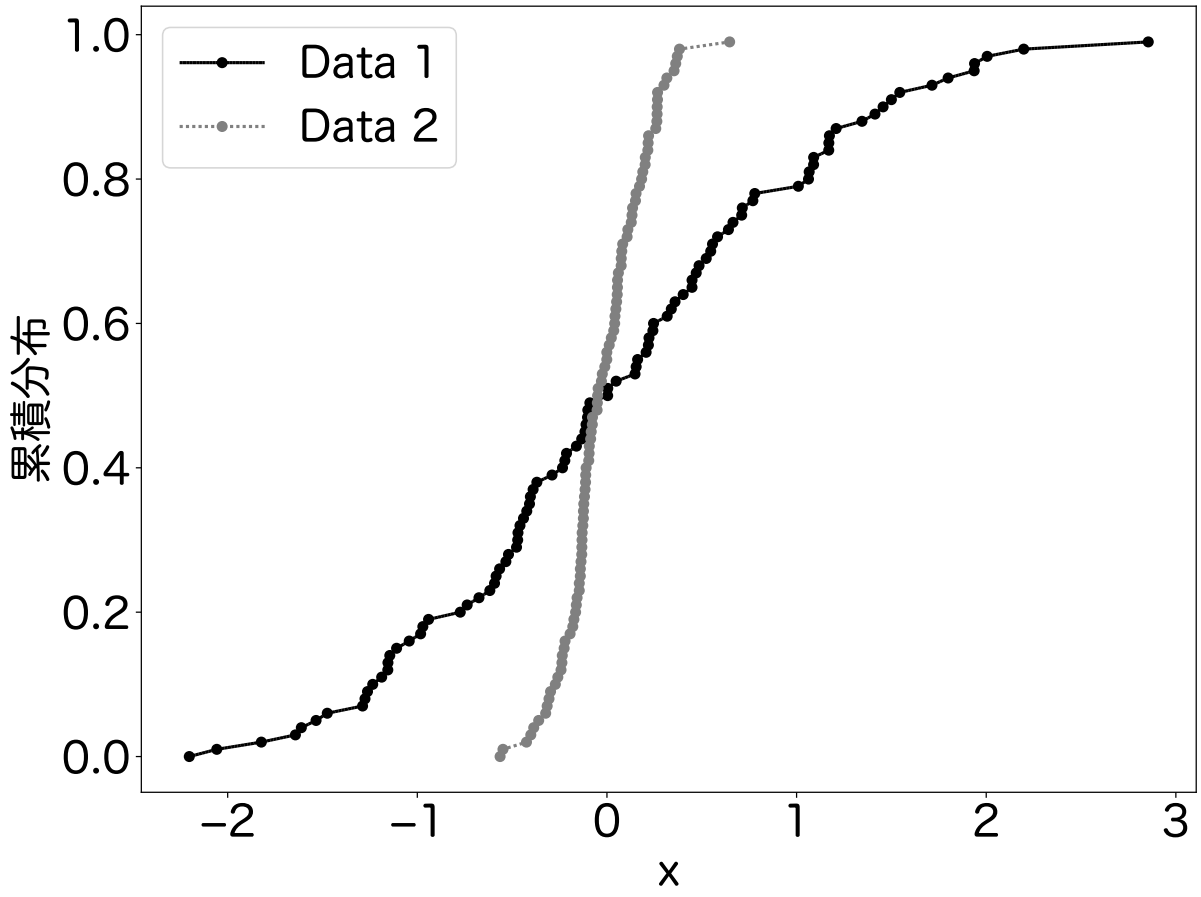


図 4.6　データの累積分布。Data 1 は、正規分布 $N(0,1)$、Data 2 は正規分布 $N(0,0.2)$ からサンプリングした。サンプルサイズは 100

て利用しにくくなる。

4.8.　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。まず上記式 4.7 について、j/n を、$F^{-1}(j/n)$ によって逆変換する。ここで、累積標準正規分布の逆関数を $F^{-1}(p)$ とする。そして、$X_{r(j)}$ と $F^{-1}(j/n)$ の組

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

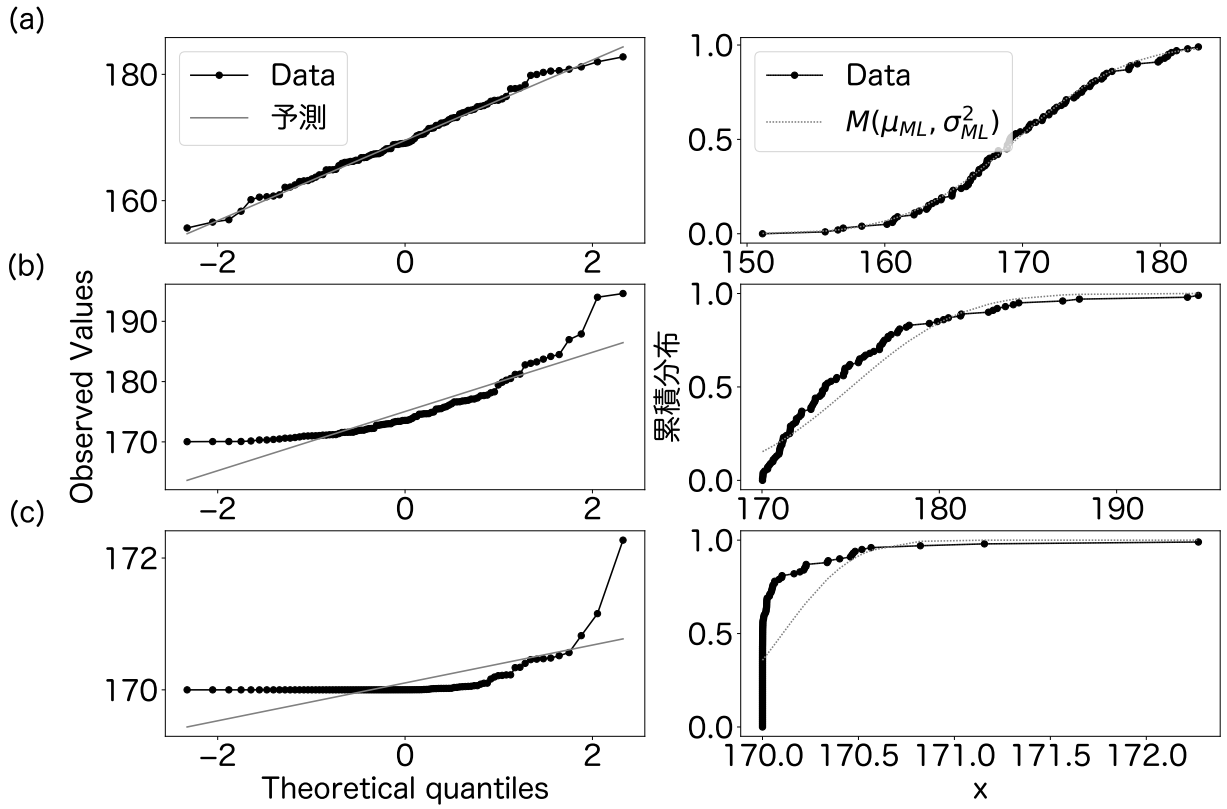


図 4.7　左には qq プロット、右は累積分布と最尤モデルの累積分布。サンプルサイズは 100(a) 正規分布 $N(170, 5.8)$(b) 指数分布 $\lambda = 5.8$(c) ガンマ分布 $s = 0.1$)

をプロットする。これが *qq* プロットである。

```
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 4.8 図 4.8 左側に描いた。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。

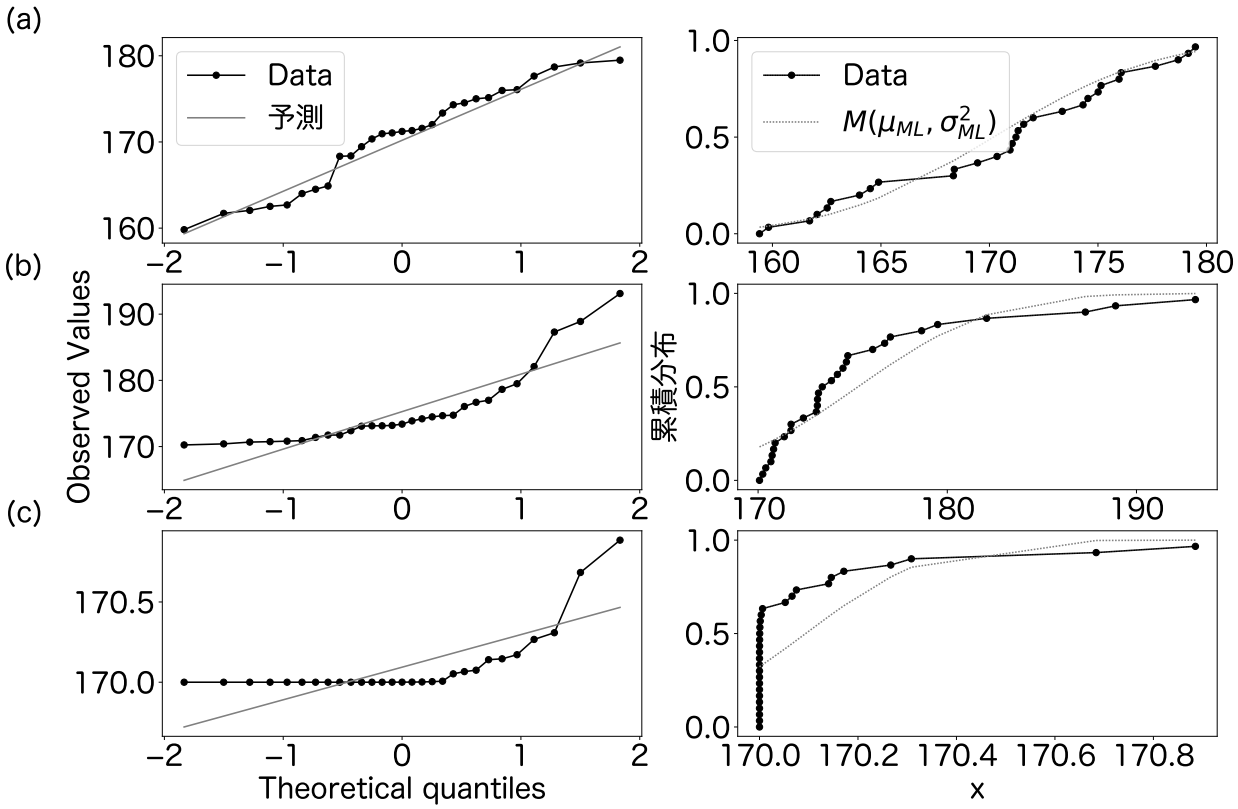


図 4.8　左には qq プロット、右は累積分布と最尤モデルの累積分布。サンプルサイズは 30(a) 正規分布 $N(170, 5.8)$(b) 指数分布 $\lambda = 5.8$(c) ガンマ分布 $s = 0.1$)

4.9.　AIC(相対的なモデルのデータへの適合具合)

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、良い予測をすることが常にそうなるともかぎらない。
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[*16]。例えば、過去の研究データから、平均 μ を決定し分散については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。

[*16] データ由来の母数 2 つあるので

第 5 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。この性質を利用して、モデルからサンプリングを行った標本について、統計量がより偏った値が一定値を下回るならば、このモデルからサンプリングされていないと判定を下し、ある特定の値より大きいならば、このモデルから得られた標本であると判断を下す。

5.1. 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。正規モデルの場合、以下の通りである。

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

この性質を利用することで、Z 値よりも偏った値がどの程度の確率で出現するかが計算できる。例えば、$Z = 0$ であれば、モデル内でこれ以上に偏った値が出現する確率は約 0.5 程度であり、よくある値であることがわかる。一方、$Z = 1.96$ であれば、モデル内でこれ以上に偏った値が得られる確率は約 0.025 程度となり、なかなかレアな値であると思うことができる。

定義 5.1.1. 統計モデルにおいて、ある統計量よりも偏った（大きいまたは小さな）統計量が得られる確率を p 値と呼ぶ。

あるモデルでの標本の得られにくさを、p 値に変換して検討することが可能である。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す。

5.1.1. 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の計算を練習してみよう。つまり、以下をコンピュータに計算させる。

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

正規モデル $M(\mu = 168; \sigma^2 = 6.8)$ から得た標本の統計量を、$\bar{x} = 172.4$、サンプルサイズを $n = 10$ とする。Z 値は $Z(\bar{x}, \mu) = 2.04$ となる。これを元に、以下のスクリプトを実行すると、p 値が約 0.04 程度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)
```

5.1.2. 自己標本の否定確率

正規モデル M において、M の標本を得てその検定統計量 Z が極端な値を取るとき、M の標本ではないと判定する。ここで、極端な値とは、Z が標準正規分布に従うことから、標準偏差の 2 よりも偏った値である場合に、M の標本ではないとする。このとき、標準正規分布表を参照すれば $P(|Z| > 2) = 0.02275 \times 2 = 0.046$ 程度となる。

標準偏差の代わりに確率を元に判定を行う。$P(|Z| > x) = 0.046$ はキリが悪いので代わりに、$P(|Z| > x) = 0.05$ を設定する。標準正規分布表を参照すれば $P(|Z| > x) = 0.05$ となるのは $x = 1.96$ 程度である。この値よりも大きな $|Z|$ となる標本は、M の標本ではないと判定する。

あるモデルの検定統計量がより偏った値を取る確率 0.05 を他の統計モデルでも適用すれば、さまざまなモデルで統一的に判定が行える。また、あるサンプルサイズの標本を 100 個モデルからサンプリングした場合、95 回はモデルから生成されたものと判断し、残りの

5 回についてはモデルから生成されていないと判断する。

定義 5.1.2. 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であることを判定する閾値[*1]である。

5.1.3. 母数平均の変化に応じた信頼区間

正規モデル M においてその検定統計量 Z について、式変形を行う。

$$\begin{aligned} & |Z| < z_{0.025}(= 1.96) \\ \rightarrow & \frac{\sqrt{n}|\bar{x} - \mu|}{\sigma} < z_{0.025} \\ \rightarrow & \mu - \frac{\sigma}{\sqrt{n}} z_{0.025} < \bar{x} < \mu + \frac{\sigma}{\sqrt{n}} z_{0.025} \end{aligned} \tag{5.1}$$

定義 5.1.3. 式 (5.1) の $\bar{x}$ の区間を信頼区間といい、次の式で定義される。

$$A = \{x; \mu - \frac{\sigma}{\sqrt{n}} z_{0.025} < x < \mu + \frac{\sigma}{\sqrt{n}} z_{0.025}\}$$

これ以外の区間を棄却域と言う。ここでは、$R = \mathbb{R} \backslash A$ が棄却域である。

信頼区間の範囲は、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。
まず、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 5.1 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の標本ではないと判定を下す。

5.2. 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。これは、M_a から M_b への統計量を元にしたモデル間類似度と言える。

[*1] 閾値（読み：いきち）＝限界値

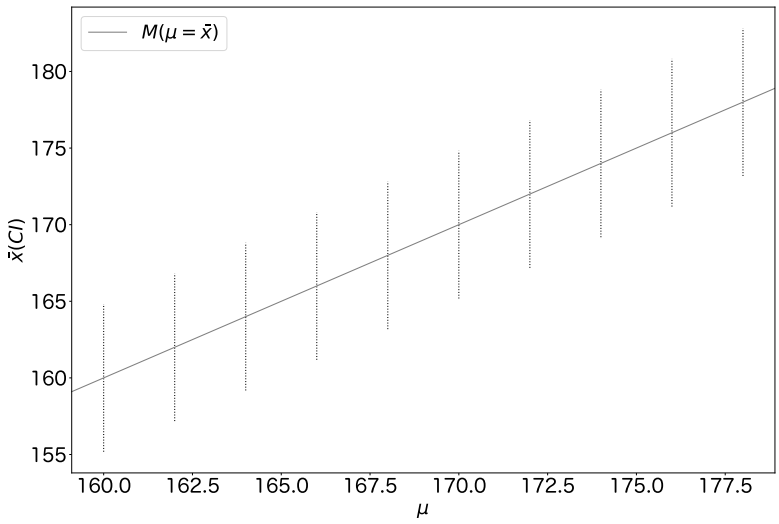


図 5.1　横軸にモデルの母数 μ、縦軸に、モデルが予測する平均値 $\bar{x}$、エラーバーに 95% 信頼区間を描いた。$N = 10, \sigma^2 = 6.8^2$

5.2.1. 検出力の定義

M_a におけるある統計量についてその信頼区間を A_a とするとき、M_b において A_a 内の統計量が出現する確率を β とする。具体的には以下の式で表される。

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

ここで、R_a, A_a はそれぞれ統計モデル M_a の棄却域、信頼区間を表し、P_a, P_b は、それぞれ統計モデル M_a, M_b におけるある統計量がしたがう分布の密度関数。$1 - \beta$ を検出力という[*2]。

検出力 $1 - \beta$ は、二つの異なるモデルを検定統計量を基準に比較するための指標である。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、また、M_a に対する M_a の検出力は、$1 - \alpha$ である。一方、モデルの母数が一致していないならば、β は 0 に近い値を取る。

[*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/

β は α を変数にする関数になっており、α を 0 に近づけていくと、信頼区間は徐々に大きくなり、$1-\beta$ は小さくなる。α を大きくすると、信頼区間は徐々に狭くなり、$1-\beta$ は大きくなる。

5.2.2. 正規分布モデルの検出力

正規モデルをつかって、$P_a(x \in R_a), P_b(x \in A_a)$ を計算する。σ^2 がすでに与えられた正規モデルを $M(\mu;\sigma^2)$ 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の平均値は、それぞれ $\bar{x}_a \sim N(\mu_a, \sigma^2/n)$、$\bar{x}_b \sim N(\mu_b, \sigma^2/n)$ である。また、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 $N(\mu_a, \sigma^2/n)$ の確率密度関数とすると、

$$P_a(x \in A_a) = 1 - \alpha$$

であるのは定義から明らか。また、P_b を $N(\mu_b, \sigma^2/n)$ の確率密度関数とすると、

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

である。
図 5.2 に検出力と α の領域を図示した。信頼区間は、図 5.2(a) において灰色で塗った x 軸の範囲である。α は図 5.2(d) 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 5.2(b) の濃い灰色の範囲である。図 (c) は、M_b の母数 μ_b を M_a の母数 μ_a に近付けたときの $\beta, 1-\beta$。

■サンプルサイズによる β の変化　α、M_a の母数平均 μ_a、M_b の母数平均 μ_b を固定したまま、サンプルサイズを変化させ, β の変化を図 5.3 に示す。$\bar{x}$ が従う分布 $(N(\mu, \sigma^2/n))$ 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をさくすると $1-\beta$ も小さくなる。

■モデルの母数による β の変化　μ_a を固定し、μ_b を変化させたときの検出力 $1-\beta$ を図 5.2.2 に示した。μ_a と、μ_b が一致していれば、$P_b(x \in A_a)$ は $1-\alpha$ になる。μ_b が μ_a から離れていくと、$P_b(x \in A_a) = 0$ に近づいていく。

■α による β の変化　α が小くなれば、信頼区間の幅が大きくなり、$1-\beta$ も小くなる。α が大きくなれば、信頼区間の幅が小くなり、$1-\beta$ も大きくなる。

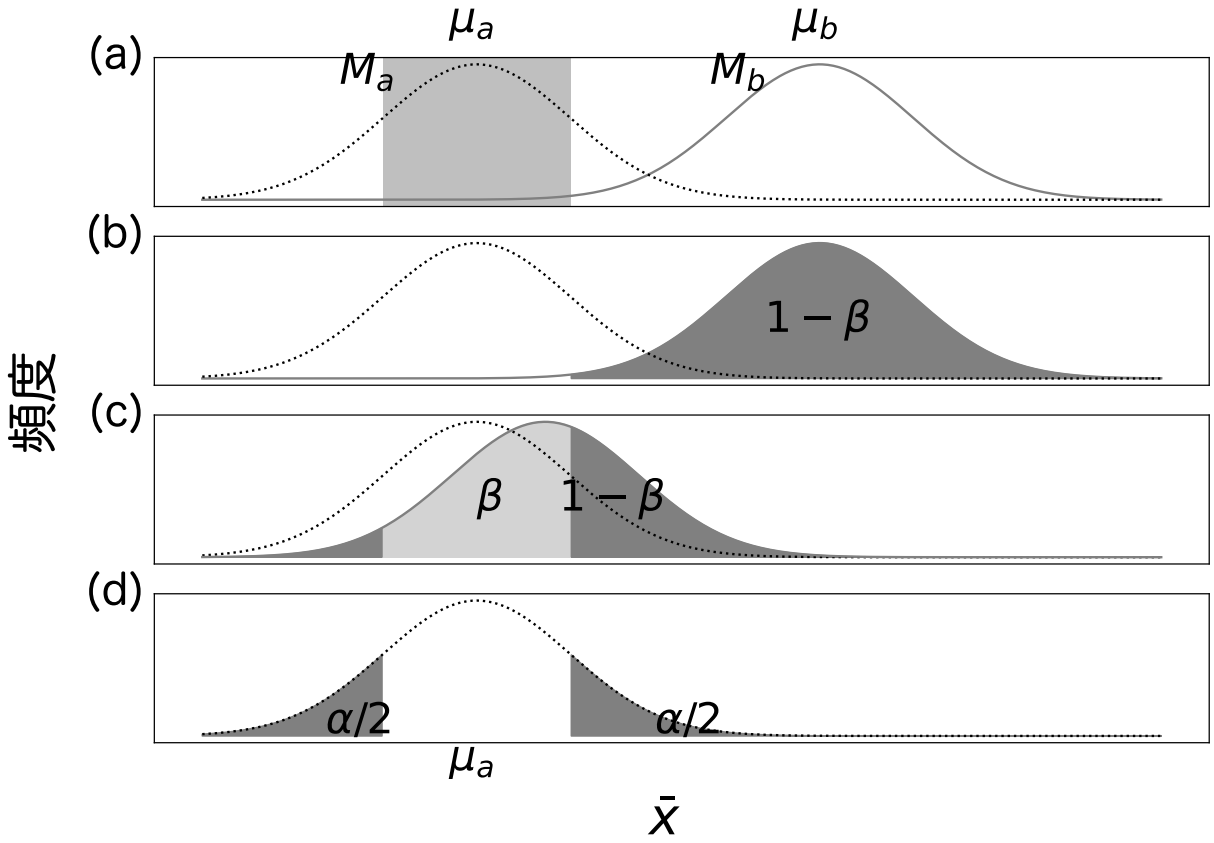


図 5.2　統計モデル M_a, M_b から計算された統計量 $\bar{x}$ の確率分布 P_a, P_b。(a) 灰色の範囲は M_a の信頼区間。(b) 灰色の領域は、$1-\beta$ の領域を示している。β の面積が非常に小さいので、グラフ上に描画できていない。(c)μ_b が μ_a に近いときの β と $1-\beta$ の領域。(d) 灰色の範囲の面積が α を示している。

5.2.3.　β の代数計算

正規モデル M_a, M_b を使って、β を計算する。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、

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

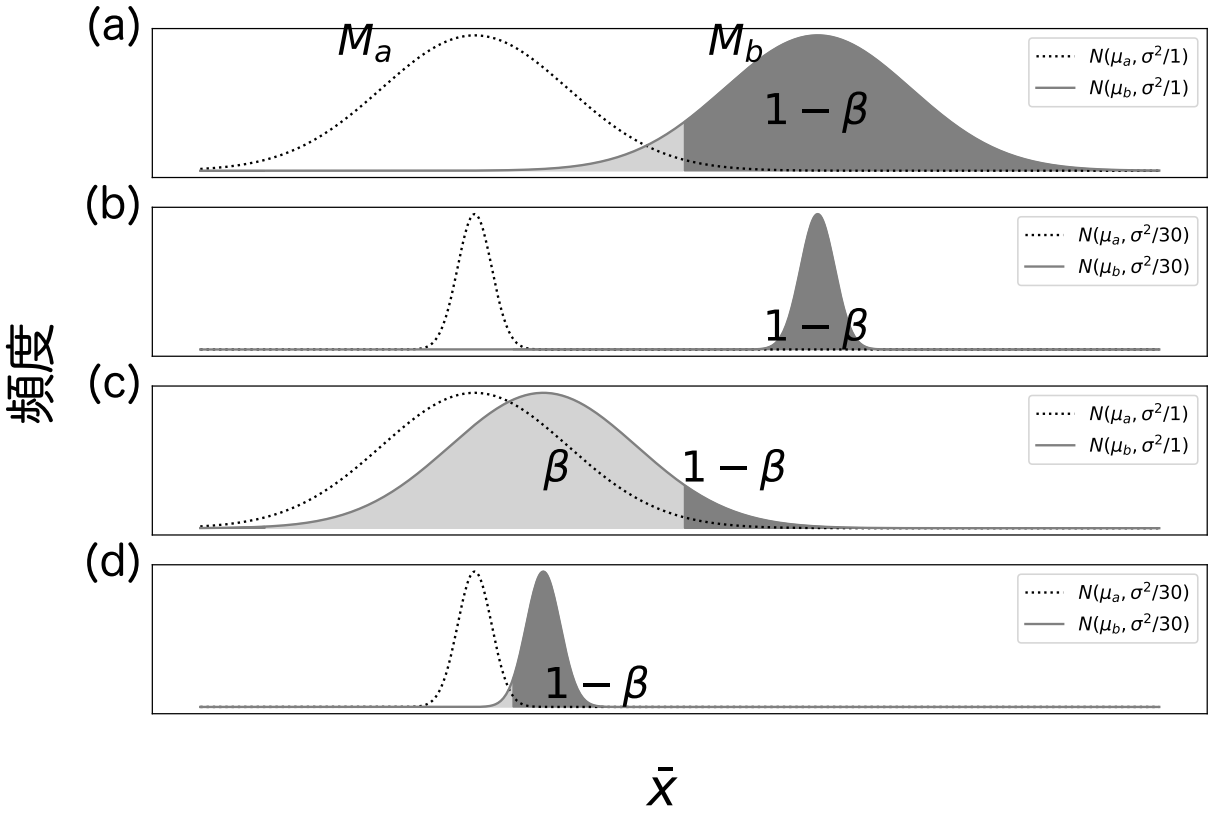


図 5.3　統計モデル M_a, M_b から計算された統計量 $\bar{x}$ の確率分布 P_a, P_b。(a)μ_a, μ_b のサンプルサイズ 1 の平均値がしたがう確率密度関数 $N(\mu_a, \sigma^2/1), N(\mu_a, \sigma^2/1)$。(b)(a) と同じ μ_a, μ_b に対して、サンプルサイズを 30 にした場合の確率密度関数。(c)μ_a, μ_b が (a) よりも近いときの $\bar{x}$ の確率密度関数。(d)(c) と同じ μ_a, μ_b に対してサンプルサイズを 30 にした場合の $\bar{x}$ の確率密度関数。

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 以外の確率変数である。M_b の標本平均 $\bar{x}_b$ は、$N(\mu_b, \frac{\sigma^2}{n})$ に従うので、その確率密度関数において、A_a が出現する確率が β 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、a, b を標準正規分布へ規格化したときの確率変数をそれぞれ A, B とする。すると、A は以下の計算式により求められる。

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

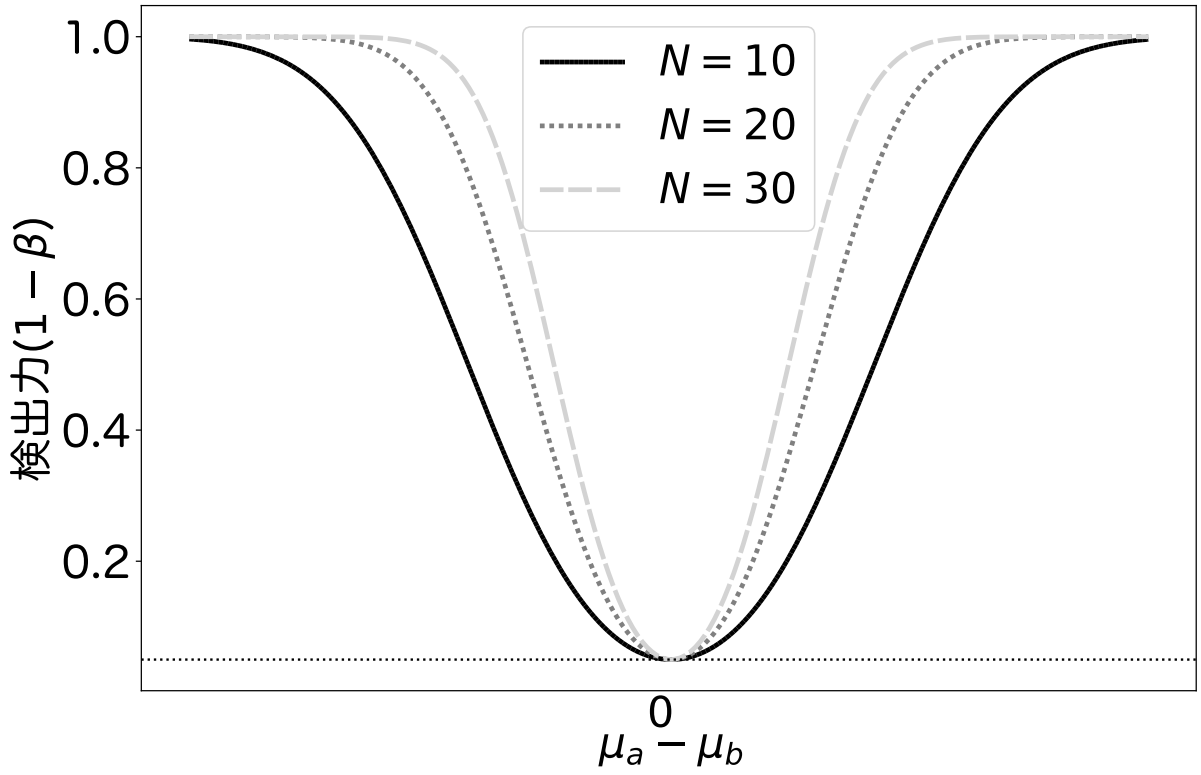


図 5.4　μ_a を変数にしたときの検出力 (検出力関数)。

同様に、B は以下の通り。

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

以上より、確率密度関数 $N(0,1)$ において、$-z_{\alpha/2}+\frac{\sqrt{n}}{\sigma}(\mu_a-\mu_b) \leq x \leq z_{\alpha/2}+\frac{\sqrt{n}}{\sigma}(\mu_a-\mu_b)$ の間での積分値が $1-\beta$ である。

■計算　$d = \frac{\mu_a-\mu_b}{\sigma}$ 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。

```
1 A,B = norm.interval(0.95,0.,1)
```

```
2 N = 9
3 d = 0.6
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)
```

答えは、0.564

5.2.4. 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、

```
1 A,B = norm.interval(0.95,0,1)
2 N = 100
3 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

5.3. 過誤のまとめ

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、偏った統計量ならば、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。この判断の間違いを第 1' の過誤と呼ぶ。

次に、モデル M_b からサンプリングによって得られた標本が、別のモデル M_a からサンプリングされものであるかどうかを判定することを考える。この場合、統計量が M_a の信頼区間含まれているかどうかを確認し、含まれていない場合には、モデル M_a からサンプ

リングされた標本ではないと判定する。問題は、統計量が信頼区間に含まれている場合である。この場合、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、誤って M_a からサンプリングされた標本であると判断することになる。この誤った判断を第 2' の過誤と呼ぶ。以上のことをまとめると、表 5.1 のようになる。

表 5.1　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)

■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 5.2)。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。
統計モデル M_a により、標本を M_a のものと M_b のものに分ける作業をおこなう。具体的には、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 において信頼区間に統計量が入っているかいなかをもとに判断する。この問題において回答が FP となったものが第 1' の過誤であり、FN となったものが第 2' の過誤である。

5.3.1. サンプルサイズの設定

統計モデル M_a により、標本を M_a のものと M_b のものに分ける作業をおこなう。このとき、なるべくなら第 2' の過誤を少くしたい。これは、$1-\beta$ をなるべく大きく設定ことで

表 5.2　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

解決できる。この作業では、すでにモデルが存在しているので、モデルの母数は固定である。また、有意水準についても α と決定しているものとする。$1-\beta$ を満せるように変更できる変数は、サンプルサイズのみである。ここでは、具体的な計算を行う。

サンプルサイズ

M_a と M_b の母数平均の差 d と検出力を指定したときに、M_a, M_b 間の検出力をある値以上にするための最小のサンプルサイズが計算できる。$\beta = 0.1, d = 0.8$ とし、この β を満たすように N を計算した。

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

二つのモデルの間の類似度を 0.9 以上にするために必要な最小のサンプルサイズは、18 であることがわかる。

5.4. 自己否定の誤推定

統計モデルからサンプリングされた標本を、統計量により元のモデルからサンプリングされたかどうかを判断するとき、そのモデルからサンプリングされた標本ではないと想定以

上に判断してしまうことがある。言い替えると、複数の標本のうち α をモデルの標本ではないと判断したいにもかかわらず、適切な処理や計算を行わないことで、α 以上におおくの標本にこのモデルの標本ではないと判断をくだすことになる。そのような処理や計算方法について説明をおこなう。

定義 5.4.1. p 値の分布が偏ることで、$p < \alpha$ となる確率が α ではないことを「有意水準 α で検定ができないと言う」。

5.4.1. 有意水準 α で検定が行えている

$x \sim N(\mu, \sigma^2)$ について、これを、累積分布関数により、x 以下の数値が出現する確率に変換する。F を確率密度関数とすると、$F(X < x)$ を計算する。これは、$0 \sim 1$ の間で一様に分布する。このことを確かめる。
正規分布から 10^5 回サンプリングを行い、その累積分布関数により、サンプリングした数値より小さな値が出現する確率を計算した。結果は、5.5 の通り $F(X < x)$ が、傾き 1 の直線上に乗る[*3]。
この性質を使うことで、有意水準 α で検定が行えていることを数値計算により確かめることができる。具体的には、検定統計量を計算し、これが従う分布形の累積分布関数により統計量より小さな値が出現する確率を計算する。これを複数の標本について行うと、計算された値が一様分布する。このことは、経験累積分布が傾き 1 の直線に乗ることを検証すればよい。一方で、有意水準 α で検定が行えないならば、経験累積分布が傾き 1 の直線から離れているはずである。

5.4.2. どんな統計モデルでも T 統計量で調べよう

正規モデルと統計量 T を使うと、T が信頼区間の中にある確率は α である。一方で、指数モデルを使い、統計量 T を使った場合、統計量 T が信頼区間の中にある確率は α よりも多くなる。ここでは、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のサンプルを一定の有意水準でモデルの物であるか検証可能かを調べる。
次の統計モデル $M_E(\lambda)$ を構築する。

[*3] 一様分布している

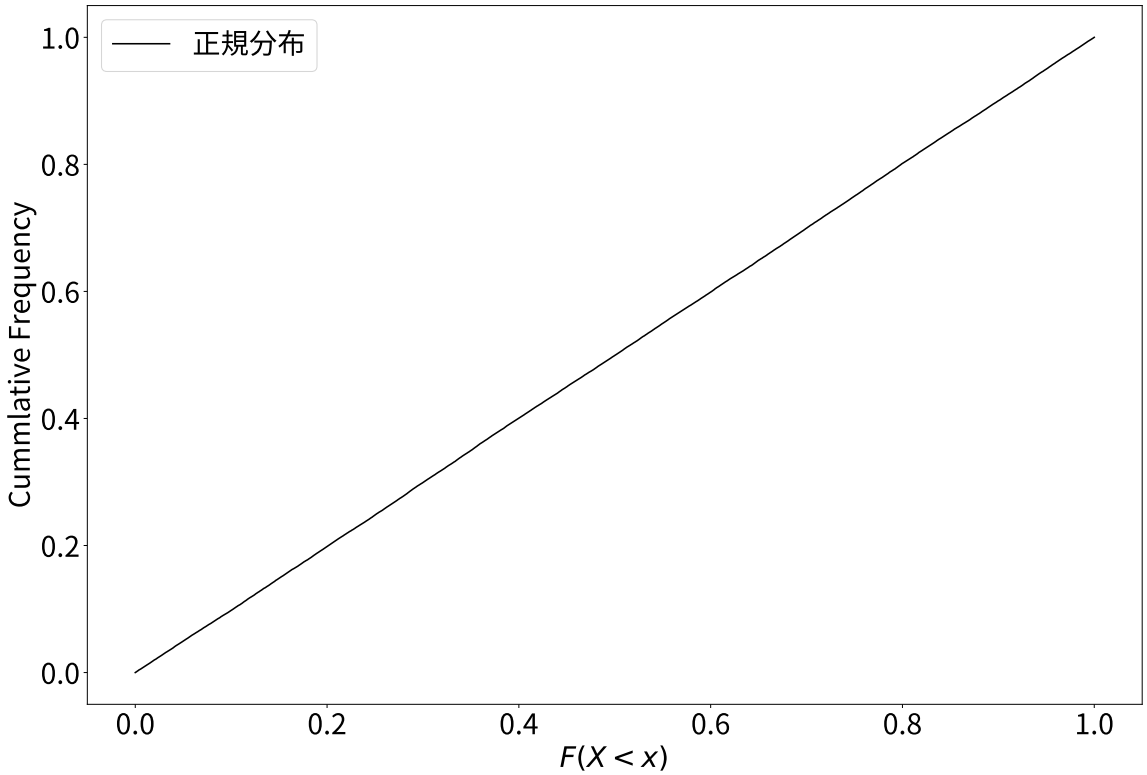


図 5.5　累積分布

1. $X_1, X_2, \cdots, X_n, i.i.d \sim F$
2. F は指数分布
3. 指数分布の母数は λ

母数 $\lambda = 1$ とした統計モデルを $M_E(1)$ とする。$M_E(1)$ から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。

$$
\begin{aligned}
T &= \frac{\bar{x} - 1}{\sqrt{\frac{\sigma^2}{n}}} \\
&= \sqrt{n}\frac{\bar{x} - 1}{\sigma}
\end{aligned}
$$

正規モデルから得た標本であれば、$T \sim t(n-1)$ である。今回は、指数モデルであるため、このようにはならないはずである。しかし、これが成り立つと考えて、計算をおこな

うとどのようになるだろうか。

数値計算

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。$M_E(1)$ または、正規モデル $M(\mu=0,\sigma^2=1)$ からある一定のサンプルサイズの標本を 10^5 回取得する。T 値を計算し、T 値より偏った値が得られる確率 p を計算する。以上の数値計算を、二つの条件において行う。

- (実験 1) サンプルサイズを $n=10$ とし、p 値を計算する。その p 値の分布の偏りを調べる。
- (実験 2) サンプルサイズ n を 4 つの条件 $n=(3,10,30,50,100)$ とし、それぞれのサンプルサイズ毎に、上で説明した数値計算を行い、p が想定している有意水準 $\alpha=0.05$ を越えない割合を計算する。

正規モデル $M(0,1)$ の場合、T 値は $t(n-1)$ 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、指数モデル $M_E(1)$ の場合、T は $t(n-1)$ 分布に従わない。このことから、実験 1 では、p 値の分布が偏り、一様分布からずれるそして、実験 2 では、$p<\alpha$ となる標本が、0.05 とは異ることが予想できる。

実験 1 の結果を図 5.6 に示した。正規モデルであれば、T 値は正規分布するので、p 値の分布は一様分布する。p 値の累積分布が傾き 1 の直線の上にあるので、このことが確かめられる。一方で、指数モデルにおける p 値の累積分布は、傾き 1 の直線の上にはないことがわかる。このことから、指数モデルでは期待した通りのことが起きないことが判った。

実験 2 の結果を図 5.7 に示した。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 (図 5.7 灰色の点)。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 (図 5.7 黒色の点)。

このことから、指数モデルのでは、設定した有意水準 0.05 よりも高い頻度で標本がモデル由来でないと判定される。

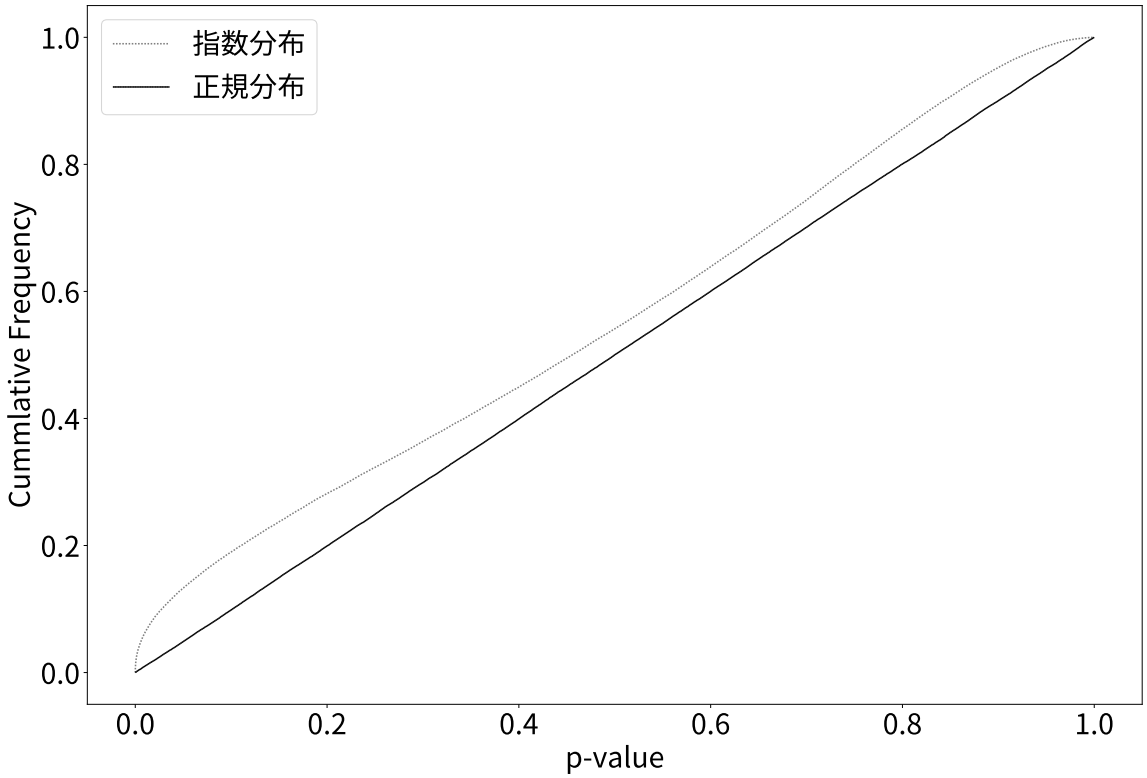


図 5.6　サンプルサイズ 10 での p 値の累積分布。実践は、正規モデル。破線は、指数モデル。

5.4.3. 検定を繰り返し使おう

正規モデル $M(\mu)$ において、100 個の標本を集め、それらの統計量が信頼区間に入っているかどうかを調べると、およそ 100α 個については信頼区間の外にある。ここで、標本を得るたびにその統計量が信頼区間にあるかどうかを調べてみると、やはり 100α 個が信頼区間のなかにある。

これを拡張し、二組の標本を 100 個得て、それら両方の統計量が信頼区間の中に入っているのは 100α 個だろうか。計算してみる。具体的には、正規モデル $M(\mu)$ において、複数の標本が信頼区間に入っているかをたしかめ、それらが正規モデルのサンプルであると考えることができるだろうか。言い替えると、複数の標本のうち少なくとも一つは信頼区間に入っていないなら、正規モデルの標本ではないと判断すると、それは想定通り複数の試

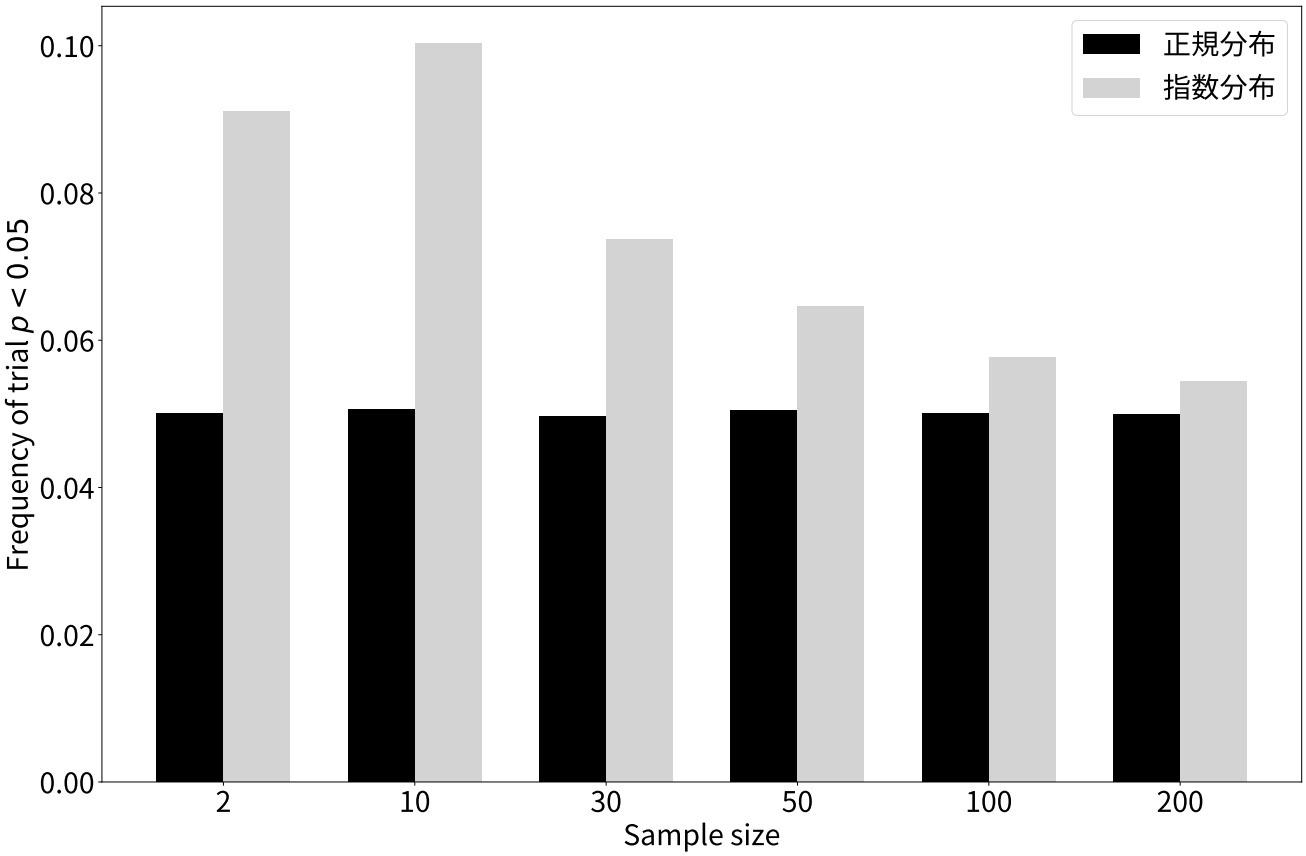


図 5.7　それぞれの分布から得た標本の T 値から計算した p 値で、$p < 0.05$ 以下になる割合。黒い破線が指数分布での数値計算の結果。灰色の破線が正規分布での数値計算の結果。

行の内 α 程度がその判断に引っ掛るだろうか。

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 $(1-\alpha) \times (1-\alpha) \times (1-\alpha) = (1-\alpha)^3$ である。一方で、棄却される確率 p' は、$1-(1-\alpha)^3$ である。

表 5.3 は、標本数に応じた p' である。標本数が大きくなるについれて、p' が大きくなることがわかる。これは、標本のペア数が多いと、標本がモデルのものではないと判定されやすくなることを示している。

表 5.3　標本数に応じた p'

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

数値計算

具体的に数値計算を行う。サンプルサイズ 10 の標本を 4 つ (標本数 4) 得る。この試行を 10^6 回繰り返す。各試行において、すくなくとも一つの標本の p 値が有意水準 α を超ていることを確かめる。

具体的に以下のコードを実行すればよい。最後の行で、1 から有意水準を超ていないサンプルの割合をひいて、少なくとも一つは有意水準を超ていた試行の割合を計算している。

```
1  repeats =10**6
2  sampleN = 4
3  N = 10
4  alpha = 0.05
5  mu=170
6  sigma = 5.8
7  sample = norm(mu,sigma).rvs(size=(sampleN,N,repeats)) #
       sample.shape = (sampleN, N , repeats)
8  normal_sample = np.sqrt(N)*(np.average(sample,axis=1).T-mu)/
       sigma # normal_sample.shape = (repeats, sampleN)
9  p_values = norm.cdf(normal_sample)
10 flag = (p_values < alpha*0.5)|(p_values > 1-alpha*0.5)
11 1-(np.sum(np.sum(flag,axis=1) == 0))/repeats
```

結果は、およそ 0.1849 程度になり、解析解と一致することがわかる。

5.4.4. 最小の p 値を採用しよう

固定のサンプルサイズの標本をいくつか得て、それぞれの標本において p 値を計算し、その中で最小の p となるものを採用するという操作を数値実験により行う。具体的には、平均分散をそれぞれ $\mu = 170, \sigma^2 = 5.8^2$ とし、正規分布からサンプルを生成する。またサンプルサイズを $N = 10$ とし、標本のペア数を 2 とする。それぞれのサンプルにおいて p 値を計算し、p 値の中で最小の p を採用する。これを 10^3 回繰り返す。

```
repeats =10**6
sampleN = 4
N = 10
alpha = 0.05
mu=170
sigma = 5.8
sample = norm(mu,sigma).rvs(size=(sampleN,N,repeats))
normal_sample = np.sqrt(N)*(np.average(sample,axis=1).T-mu)/
    sigma
p_values = norm.cdf(normal_sample)

min_p_values = np.min(p_values,axis=1)
first_p_values = p_values[:,0]
```

図 5.8 には、数値計算の結果を示した。通常 p 値は一様分布するので、累積分布は傾き 1 の直線の上にのる。実際に数値計算でも同様の結果が示されている。一方で、最小の p 値を選択すると、図 5.1 に示したように、傾き 1 の直線の上に乗らない。これは、p 値が一様ではなく、p 値が小さなものが通常よりも大きな頻度で生じていることを示唆している。このことから、優位水準 α を定めると、$p < \alpha$ となる p 値は α よりも大きな頻度で生じており、優位水準 α で検定ができないことが示される。

多重検定との関係

この処理は一つ前の節で説明した多重検定と一致する。具体的には、すくなくとも一つの標本において $p < \alpha$ となるは、最小の p 値を採用すると同じである。このことを数値計算により確かめておく。

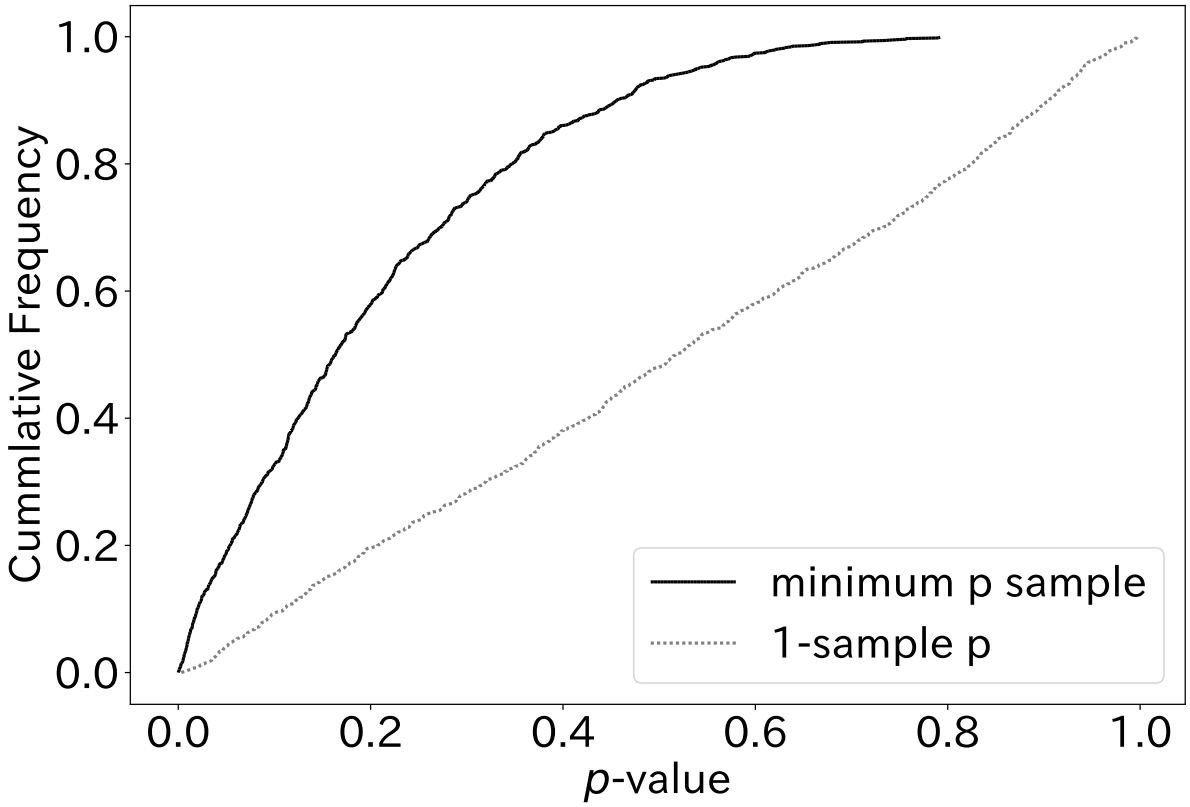


図 5.8　最小の p 値を採用したときの累積分布。実線は 2 つの標本の中から最小の p を選んだときの累積分布。破線は一つのサンプルから p を計算した場合の累積分布

```
1  repeats =10**6
2  sampleN = 4
3  N = 10
4  alpha = 0.05
5  mu=170
6  sigma = 5.8
7  sample = norm(mu,sigma).rvs(size=(sampleN,N,repeats))
8  normal_sample = np.sqrt(N)*(np.average(sample,axis=1).T-mu)/
       sigma
9  p_values = norm.cdf(normal_sample)
10
```

```
11 np.sum(np.min(p_values,axis=1)<alpha)/repeats
```

結果は、0.184 であり、解析解[*4]と一致する。

5.4.5. サンプル追加による p 値の変化

標本にサンプルを追加しながら検定をおこなうと、期待している結果が得られるだろうか (元ネタ [8])。ある標本において、サンプルを追加する毎に検定を実行する。この試行を複数回繰り返し、一度でも有意水準を下回った標本の個数を数える。ただし、$p < \alpha$ になるまで繰り返すと必ず有意となるので、最終的なサンプルのサイズはあらかじめ決定しておく。

数値実験を行う。初期サンプルサイズを $N = 10$ または $N = 20$、最終サンプルサイズ $N_{max} = 50$、標本数 10^6 とする。サンプルを Δs 個追加し、検定を実行する。Δs は、1,5,10 または 20 とする。それぞれの Δs に応じた検定回数は $41, 9, 5, 3$ 回である[*5]。

図 5.9 が数値実験の結果である。サンプルサイズを追加する毎に検定をすると、$p < \alpha$ となる頻度が α 以上になる。どの場合においても $p < \alpha$ となるのは α 程度であってほしいが、これは数値実験の結果と一致しない。

以下に数値実験用のコードを残しておく[*6]。

```
 1 sampleN = 10**6
 2 N = 10
 3 maxN = 50
 4 mu = 170
 5 sigma = 5.8
 6
 7 delta = 1
 8
 9 def nan_index(N,maxN,delta):
10     index = np.array([~(np.arange(maxN)>=i) for i in np.
           arange(N,maxN+delta,delta)] )
11     nan_index = np.ones(index.shape)*index
```

[*4] 解析解という用語は正しいのか不明

[*5] 多重検定とは条件が異なるので、$\Delta s = 41$ だとしても、$p < \alpha$ となる頻度が $1 - (1 - \alpha)^{41}$ とならない。

[*6] わかりにくいコードになってしまった。後で書き換えたい。

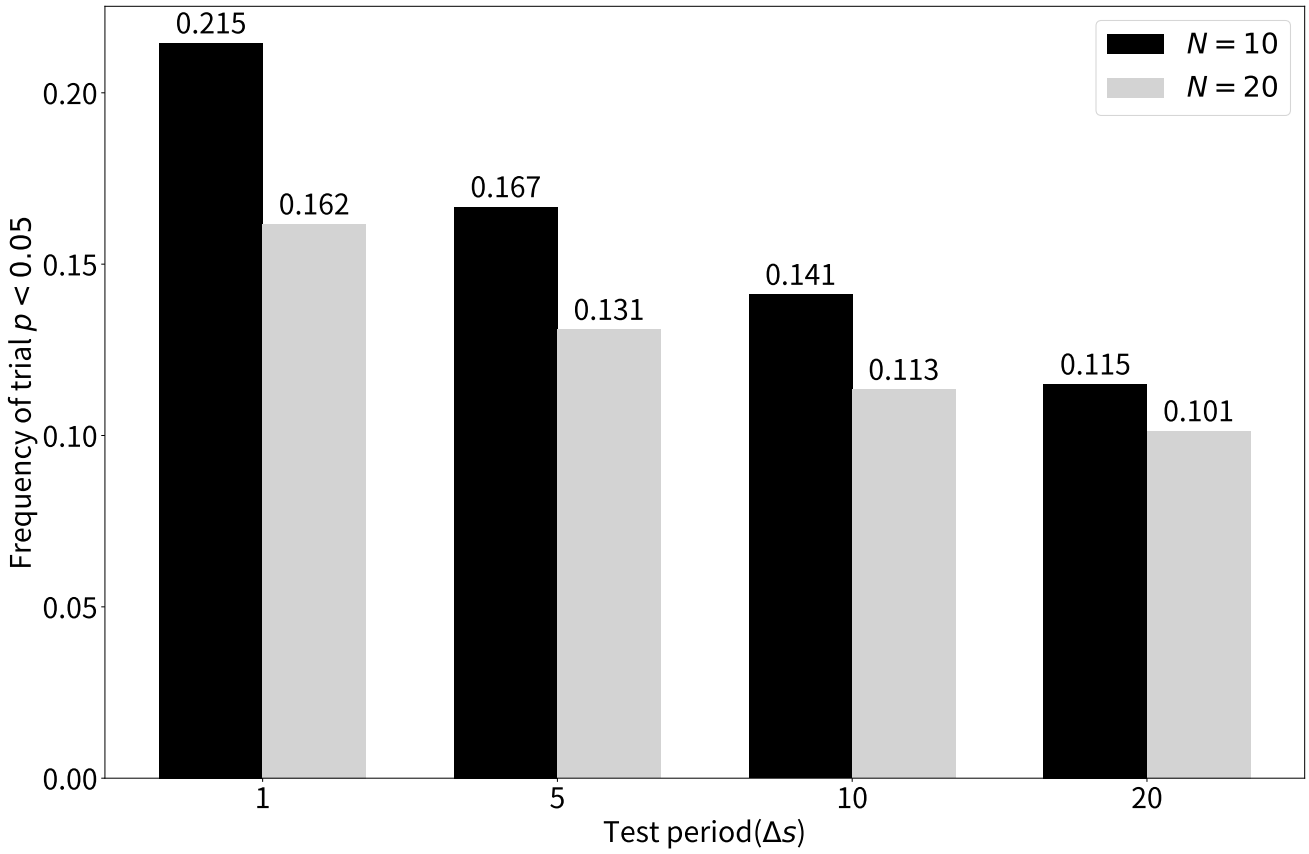


図 5.9　サンプルの追加個数と検定を実行。サンプルの追加個数を Δs としその違いによる $p < 0.05$ となる頻度。初期サンプルサイズが 10(黒色) と 20(灰色)。

```
12      binary_index = index.astype(np.float64)
13      binary_index[~index] = np.nan
14      return binary_index
15
16  def stopping_rule(delta):
17      nan_array = nan_index(N,maxN,delta)
18      sample_num = np.count_nonzero(~np.isnan(nan_array),axis
            =1)
19
20
21      # sample
```

```
22      norm_ = norm(mu, sigma)
23      sample = norm_.rvs(size=(sampleN,maxN))
24
25      rep_sample = np.tile( sample.reshape((-1,1,maxN)), reps=
            (len(nan_array),1))
26      restrict_sample = rep_sample*nan_array
27      sample_mean = np.nanmean(restrict_sample,axis=-1)
28
29      Z = np.sqrt(sample_num)*(sample_mean-mu)/sigma
30      p = norm.cdf(Z)
31      p_ = ((p < 0.025) | (p>0.975))
32      true_once = (np.cumsum(p_,axis=1)>=1)
33      return np.sum(true_once[:,-1])/sampleN
34  N = 10
35  result_10 = [stopping_rule(delta) for delta in [1,5,10,20]]
```

5.4.6. いつかは有意になる

非常にわずかなモデルのちがいでも検定を使うと、モデルの標本ではないと言えてしまうことを確認する。まず、二つの正規モデルを構築する $M_a = M(170, 5.8^2), M_b = M(171; 5.8^2)$ とする。母数平均が殆ど一致していることを、母数平均の規格化量 D を計算することで確認しておく。

$$D = \frac{|\mu_a - \mu_b|}{\sigma} = 1/5.8 = 0.172$$

これらのモデルはほぼ同じような予測を行うことは理解できる。
モデル M_a において生成したサンプルがサンプルサイズを大きくすることで、$p < \alpha$ になる様子を確認しておく。具体的には、M_a においてサンプルサイズ 300 の標本を 10 個生成する。それぞれの標本において、サンプルサイズを $1 \sim 300$ までにし、それぞれのサンプルサイズにおける p 値を計算する。

```
1   def nan_index(N,maxN,delta):
2       index = np.array([~(np.arange(maxN)>=i) for i in np.
            arange(N,maxN+delta,delta)] )
```

```
3       nan_index = np.ones(index.shape)*index
4       binary_index = index.astype(np.float64)
5       binary_index[~index] = np.nan
6       return binary_index
7
8   def calc_p():
9       maxN = sample_size
10      nan_array = nan_index(1,maxN,1)
11      sample_num = np.arange(1,maxN+1)
12      norm_ = norm(mu_a,sigma)
13      sample = norm_.rvs(size=(sample_size,num_of_sample))
14
15      rep_sample = np.tile( sample.reshape((-1,1,maxN)), reps=
            (len(nan_array),1))
16      restrict_sample = rep_sample*nan_array
17
18      sample_mean = np.nanmean(restrict_sample,axis=-1)
19      Z = np.sqrt(sample_size)*(sample_mean-mu_b)/sigma # モデル
            のサンプルか？M_b
20      return Z
21
22  nan_array = nan_index(1,sample_size,1)
23  mu_a=170
24  mu_b=171
25  sigma = 5.8
26  sample_size=300
27  num_of_sample = 10
28
29  Z = calc_p()
```

図 5.10 にはサンプルサイズに応じた p 値を表示した。サンプルを増やしていくことで、p 値が徐々に小くなることがわかる。殆ど同じようなモデルであってもサンプルサイズを増やしていけば、あるモデルのサンプルではないと主張できる。このことは、データをモデ

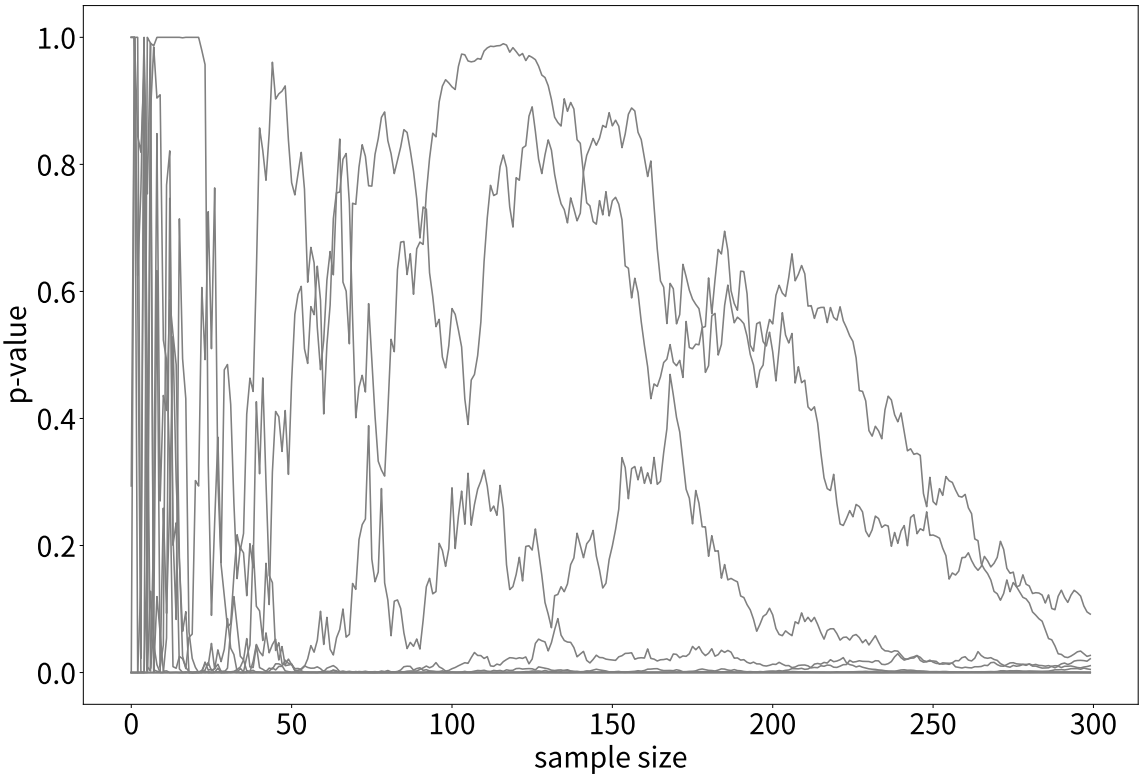


図 5.10　サンプル応じた p 値の変化

ルを比較するさいに非常に重要な事項である。

5.4.7. いつかは有意にならない

ある正規モデル M_a において、サンプルを生成し、一つづつサンプルを追加し、追加毎に p 値を計算する。結果を図 5.11 に示してある。この図の通り、サンプルサイズを追加すると、一時的には p 値があるしきい値 (図の点線 $\alpha = \frac{0.05}{2}$ または $1 - \frac{0.05}{2}$) を下回ることがあるが、その後サンプルを加えると、有意水準よりも大きな値となることがある。よく言われる「検定はサンプルサイズを大きくするといつかは有意水準を下回る」に反する[*7]。

[*7] こう言われているのは、理論と実験の話を混ぜているためである。正確には、モデルと実験を比較するさいに、実験でサンプルサイズを逐次追加すると、モデルと実際の乖離があきらかになりやすいため、有意になりがちである。既に前の小節で確認した。

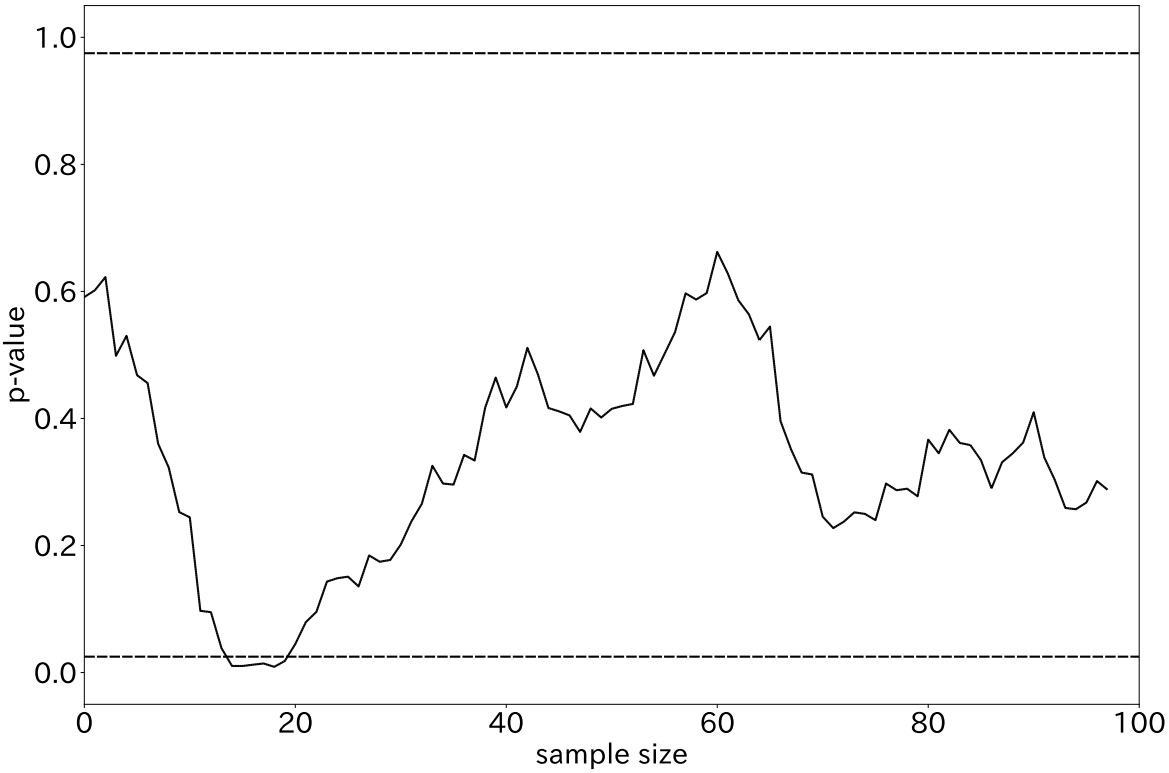


図 5.11　サンプルの追加に応じた p 値の変化

5.4.8. まとめ

ここまでで、検定統計量から複数の標本のうち $100\alpha\%$ 程度をはじきだす仕組みを学んだ。この仕組みでは、ある標本がモデル由来かどうかを調べることはできない。また、仕組みを適用するさいに、途中処理を加えてしまうと、複数の標本のうち $100\alpha\%$ 程度をはじきだすということができないことも明らかになった。モデルと我々のデータを比較するさいにもこの点に注意が必要になる[*8]。

母集団から得たデータはモデルから生成されていると考えることができないし、母集団から得た標本の $100\alpha\%$ を検出することが目的ではない。我々が考えている系は統計的仮説

[*8] 前述した通り、多重検定には注意をはらうが、他の留意点については特に気にとめない。特に、検定まえ検定が非難されるのは、まずは検定では何もわからないという点で、次に多重検定の問題がある。

検定の前提と異っており、得られるものと得たい物も完全に異っている。このことから、データとモデルを比較するための論理を作る必要がある。

5.5. 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 5.5.1. 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。

■検定統計量や p を計算するだけで解析完了

データとモデルの乖離具合を示す指標を計算するだけになってしまう。標本がモデルにより推測可能かを調べることで、より多くの予測を引き出すことができる。では、検出力 β やサンプルサイズ、中心間の距離の規格化量 D も記載すればいいということか？そうではない。

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 0 から遠いほど、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[*9]。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[*10]。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

[*9] p 値だけで判断してはいけない
[*10] ピーと鳴く鳥に p 値を例えたお話を作りだすことが目的で特に意味はない。将来的には以下の文章は消そう。

図 5.12　統計量を使ったモデルとデータの比較に関する概念図

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。

■統計的有意性

> 統計的有意性とは、ある影響が、偶然のみによって生ずるとは考えにくいことが統計的解析によって示されたことを意味します（生物学的有意性 を参照のこと）。有意性のレベルとは、その影響がどの程度偶然によって説明し得るかを示すものです。有意性が 0.05（5 %）のレベルとは、その影響が単なる偶然により生ずる可能性は 1/20 しかなく、0.01（1 %）のレベルとは、1/100 しかないことを意味します。ほとんどの生物学的現象は個々人すべてに必ず一様に起こるというわけではありませんから、一つの研究または実験で観察された影響は、ある程度の不確実性、あるいは不正確性を伴います。統計的解析においては、観察された影響をその確実性について評価し、それが偶然に生じ得る確率はどの程度か（有意性のレベル）を決定します。偶然によって生ずる確率が低い場合には「統計的に有意」と呼ばれ、真の影響を示すとみなされます。
> `https://www.rerf.or.jp/glossary/stats/`

本書では、モデル内での話として定義した。上記方針と本書は異る。

5.5.1. 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

p 値をもとに判断してはいけない。

■ 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[b]。
モデルの中で標本の統計量以上偏った値の出現確率を計算したものが p 値である。p 値が小さかったことは、モデル上でそのような統計量が出現しにくいということである。このことから、ある母数を持つモデルによりデータの平均値を予測しにくいことを示唆するのが p 値である。
正規分布や独立同分布ではないことを p 値は示唆しない。p 値によって、統計モデルの仮定の間違いを主張できるような値ではない。

---

[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
[b] 講義の録画のため先生の意見が正しく伝えきれてないというのもあるかもしれない

■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、その集団内でのばらつきを計測するためである。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。

[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

5.5.2. 有意水準 α で検定できない例

すでに説明したように、p 値使った判定には様々な制限がある。次のような限界がある。

- どんな母集団に対してでも特定の統計量を使う (母集団に関する知識の欠如)

- 複数の標本に対して検定を実行する
- 有意になるまでサンプルを取得する。
- 複数の標本のなかで最小の p になった標本を採用する

このような限界を無視した p 値の使用を p-Hacking と言う。70% 程度の研究者たちが 3 番目に挙げた方法でデータを取得したことを認めている [[9]]。数値実験により確かめたとおり、有意水準 α により検定ができない。ここに上げられていない方法でも p-Hacking は可能であり、文献 [10] にまとめられている。

5.5.3. p 値を使うことが常に最適な判断材料

p 値を使うことが常に最適な判断材料になることは非常に稀であり、p 値だけで結論が下せるようなことは生物学においては稀である。数理統計学で出ている結果は、全てモデルの中の話であり、現実がモデルと一致しているならば、モデルの予測通りの推論が行える。もちろんそんなことはない。また、モデルがデータの予測に利用できるということがわかっていれば、モデルの予測が現実の一部を捉えることができるという期待がもてる。このモデルとデータとの対比を全く行わずに検定は運用されている。p 値を使った、データとモデルの比較方法はすでに様々な論文において批判されている [11]。

5.5.4. いつかは有意になる

すでに小節 5.4.6 おいて説明したように、非常に僅かな違いでも、$p < \alpha$ となりやすいことを示しておく。TODO

■有意水準は 0.05 でよし

よくある受け答えを引用しておく [12]。

> Q: Why do so many colleges and grad schools teach p = 0.05?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use p = 0.05?
> A: Because that's what they were taught in college or grad school.

ここでの $p = 0.05$ は、$\alpha = 0.05$ のことで、有意水準を 0.05 で教える理由について聞いている。

本書では、モデルにおける計算を具体的に行うために $\alpha = 0.05$ を利用し、データとモデルを比較するさいにはこの基準を使わない。

■ある仮説を正しいと論証するより、正しくないと論証する方が簡単？

ある仮説を正しいと論証するよりも正しくないと論証する方が簡単という主張がある。
否定するのが簡単な命題は否定しやすいが、我々が考える命題は単純に否定することすら難しい。
次に、統計的仮説検定を使った枠組みにおいて、正しいか、正しくないかという二値的な判断を行えない。本書で扱う事象にたいしては、データとモデルを比較しているだけであり、モデルをデータの一部を説明するようにモデルを構築しているだけである。ここから直ちに仮説に対して真偽をあつかうことはできない。モデルの推定から、何のような傾向があるかなどを示すことはできる。

5.5.5. モデルの性能

ある基準 α を前もって設定できない。なぜなら、前情報がわからない状況で構築されたモデルによって標本を予測できないと判定を下すことができない。たとえ、設定したとして、$p \leq \alpha$ であったとしても、偶然棄却域にあったのかを区別できない。言い替えると、そのモデルで十分予測可能だったとしても、偶然できないと判定されることがある。

■ $p < 0.05$ なら差があるとする

すでに説明したとおり、$p < 0.05$ だとしても、それはなんらかに差があるということではないということもある。 ある特定の統計モデルにおいて、検定統計量が出現しにくいことから、そのモデルでは標本を予測するのは難しいのではないかと考える根拠の一つにすぎない。 生物統計学の教科書にでてくる「検定を使えば差があるかないかがわかる」という記述は、生物研究において p 値の取り扱いとしてこのコンセンサスがあるということを示しているだけである。(生物学上のコンセンサスとして) 差があるということにはできるが、それがなんらか意味があるのかについてはなんら調べがついていない。また、統計学的有意ということもあるが、前提が整っていない統計的仮説検定において統計的に有意と宣言するのはおかしなことである。

統計量として p 値のみが記述されているようなら、その論文に対し批判的に読むべきだろう。

■ 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 [13, 14]。
よい解釈として以下の 6 つの原則が示されている [13]

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 [15][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。
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 [16] 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

5.5.6. 何度も検定しよう

■なんども実験を繰り返す

> **Takami Sato:**
> 無知を晒すけど「p 値下げるために後からサンプルサイズ増やすな」って正直わかってない。後から論文を見て増やしたかどうか何てわからんし、そのデータに対する p 値としては計算上正しい訳だし。「小さい差に対しても有意になるから、p 値だけでなく値の差もしっかり議論しよう。」ならわかるんだが。
>
> https://twitter.com/tkm2261/status/1640467558813028352

> **Takami Sato:**
> 元からあるデータに足さずに改めて取り直したデータだけで計算すればなんも問題ないし積極的にやっていいそも、個人的には 5% のラインに意味はないんだから、後からサンプルサイズ足すくらいなら正直な p 値（6% でも 7% でも）で議論したらいいと思う
>
> https://twitter.com/Daphnia_t_ponyo/status/1640843352227848193

Ken McAlinn:

これは多重検定だからだめだよ

https://twitter.com/kenmcalinn/status/1640845051788967937

Ken McAlinn:

p 値の分布は一様だから 5% 棄却水準は (01 上の)0.05 の位置にあるんだけど、一回観測しちゃうと「そのまま」か「データを増やす」か条件がつくからどんどん分布が 0 のほうに歪んでく (有意じゃないとそのままを選ばないから)。そうすると 0.05 の位置までの確率が 5% 以上に増えるから過誤の確率が合わない。

https://twitter.com/kenmcalinn/status/1640486745505427465

Ken McAlinn:

だから計算的には追加でも最初から全部でも p 値は同じになるんだけど、追加の場合は有意水準を変えるか選択停止を考慮した p 値を計算しないと過誤の確率が合わない。p 値は 5% なのに過誤の確率が 80% になったりする。

https://twitter.com/kenmcalinn/status/1640488036164161540

Ken McAlinn:

根本的に言えば、尤度原理を満たさない p 値には問題があるから尤度原理を満たすベイズファクターとかを使いましょうっていうのは tm さんの言う通り。

https://twitter.com/kenmcalinn/status/1640491660919422976

tm:
データが最初から全部あるときの p 値と途中で増やしたときの p 値は定義に戻ると別の値になります。なので単純に「途中で追加したのにも関わらず最初から全部あったかのように計算するのは間違い」です。後から増やしてもわからないのは別の問題（隠蔽）で、これを気にするなら p 値は原理的にダメです。

https://twitter.com/tmaehara/status/1640474043038986240

tm(@tmaehara):
（どうデータをとったかによらず）どういうデータが得られたかだけで決まる手法は「尤度原理を満たす」といいます。尤度原理は成り立ってほしい性質ですが、帰無仮説検定はこれを満たしません。これは p 値に対する代表的な批判ポイントです。

https://twitter.com/tmaehara/status/1640480964882182145

5.6. サンプルサイズを決める

次の問題を考える。モデル $\mathrm{A}(M(\mu))$ により予測可能な集団に対して、ある薬品を与え、その後もう一度同じ量を計測する。その後の計測結果に対する想定されるモデル $\mathrm{B}(M(\mu+\Delta))$ とする。$M(\mu)$ と乖離していることを $\alpha = 0.05$ に設定し、検定統計量 T により検証する。また、二つのモデルの検出力は 0.8 であるようにする。すでにモデルについて調べたさいに明らかにしたように、これらのパラメータからサンプルサイズを決定することで、要求を満たすことができる。
しかし、この問題設定は、我々の研究では成立していない。なぜなら、以下が理由となる。

1. $\alpha = 0.05$ が良いという理由付ができていない
2. 検出力 $= 0.8$ が良いという理由付ができていない
3. Δ が決定できない。

これらの理由から我々の実験設定では、仮説検定の枠組みを用いてサンプルサイズを決定することができない。または決定したとしても、事前に決めた数値通りになることがまれ

であるだろう。
以上で議論した結果では、p 値によりなんらかを決定することができないことが明らかになっが、以下では、p 値によりなんらか意思決定が可能とし、統計的仮説検定の枠組みを用いてサンプルサイズを決定したときに生じる問題を議論する。

5.6.1. サンプルサイズを決めることができない

モデルからサンプリングした標本の検定統計量と比較して、実験により得られた標本の検定統計量はモデルの中で生じるのがまれであるから、そのモデルと標本が乖離しているという主張を行った。一方で、検出力とは、あるモデルの標本が別のモデル B の標本であることを見分けるために設定する値であった。これを実行するには、サンプルサイズを大きくすることで達成可能である。検出力を維持するために設定したサンプルサイズは比較的大きくなりがちである。ここでもう一度 p 値の話しに戻ると、ほんのわずかな違いであっても、サンプルサイズが大きくなれば、p は小さな値をとりやすくなる。設定した比較的大きなサンプルサイズの標本を用いれば、モデル A でも B でも p 値は小くなりやすい。言い替えれば、元のモデル A とも B とも乖離したと判断されやすくなる。
p あたいは、ある特定モデルとその標本に関する絶対評価を行う指標であり、検出力は、モデル間の距離をしめす指標である。検出力を高くするために、サンプルサイズを大きくすれば、想定しているモデルとの p 値は小くなりやすい。これは、モデル A とモデル B のどちらにたいしても p 値は小くなりやすい。

5.6.2. あたかも論理的にサンプルサイズを決める

サンプルサイズを $1-\beta = 0.8$ になるように決めたということが、論理的に妥当であると考えている場合、どのように実験を計画すれば、p 値が有意水準を下回りやすくなるだろうか。それは、モデル A と B の違いをなるべく小く見積ることである。例えば、正規モデルを仮定し、その母数平均を、μ と $\mu + \Delta$ のモデルを構築する。Δ が小ければ小さいほど、検出力を高くするためには、サンプルサイズを大きくする必要がある。研究者が Δ をあえて小くしてサンプルサイズを決定したとする。そのサンプルサイズで実験を行い、標本を得ると、非常に僅かな違いにより、モデル A でも B でも、標本の検定統計量は出現しにくくなる。つまり、$p < \alpha$ を得やすくなる。この方法で研究者の意図により、p 値が有意水準を下回ることで、仮説の棄却と採択を自由に操作できる。

■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 [17] がある。[18] 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。

5.6.3. 実験後の検出力

実験後の検出力の意味を考えてみよう。モデル M_A により予測が可能だった特徴が、M_B の方がより良さそうという結論が得られたとする。モデル M_B に対する M_A の検出力が計算する。α は、次の考えで決定できる。研究に利用した集団を予測できそうな M_B において、M_B の検定統計量のうちメジャーだと考えられる範囲または、その集団をなんどもサンプリングしたうち、標本をメジャーなものとして採用したい頻度を決定すれば良い。サンプルサイズについても、どのくらいのサンプリングを毎回おこなうのかを決定すればよい。このときの検出力は計算可能である。この検出力の意味は、サンプリングを繰り返したとき、母集団の標本であるのに、M_A の想定する母集団であると判断する頻度である。言い替えると、モデルの標本を元にした、モデル間の距離である。検出力が大きければ、検定統計量をもとにして、ふたつのモデルの違いがあるなぁと考えられる。
問題点は、以下の事であろう。

1. そのような操作を行いたい状況は想像できない。
2. このような実験を行っている研究がない

5.7. まとめ

p 値や信頼区間が論文に記述されていたとしても、

1. それだけでは、なにもわからない
2. 予測モデルとして、従来モデルがだめで、ほかにもっと良いモデルがあるかわからない
3. 分布形が正規分布だったのかわからない。常識的に考えて正規分布するはずだという研究者のバイアスに付け込んで誤読させる意図があるかもしれない

わかることは、その報告書の著者がどのような判断をすることにしたかということであろう。

第 6 章

モデルにおける尤度比の性質

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、モデルにおいてその尤度比以上の値の出現確立が計算可能である。さらに、特定のモデル M において、尤度比がありふれた値であるかどうかを考えることにより、モデルがデータを予測できるのかを考える。
前の章で統計的仮説検定をモデル外のデータに対して利用する方法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統計的性質なのかを説明する。

6.1. 概要

モデル $M^{'}$ により、実世界から無作意抽出して得た標本 $D^{'}$ を予測できるかを尤度比により検証する。パラメータの個数が k 個多いフルモデル M とし、モデル $M^{'}$ の尤度比を次で計算する。

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

$D \sim M^{'}$ ならば、母数の個が k 個多いモデル M との尤度比 $Dev(D, M', M)$ は、自由度 k の χ^2 分布に従う。Dev の定義は、式 A.9.1 を参照せよ。
ここで、モデルとは無関係なデータ D' について、D' を $M^{'}$ により説明しにくいなら、尤度比 $Dev(D', M, M')$ は比較的大きな値を取るはずである。言い替えれば、モデル $M^{'}$ の標本が取り得る尤度比と比較して、珍しくない尤度比であるならば、それは、$M^{'}$ において予測できると考えることにする。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 $M^{'}$ によって予測可能かなぁと考えられる。
もちろん、標本を要約した統計量による検証なので、標本とモデルの詳細な検討ができていないので、そのモデルを採用可能かはわからない。

6.1.1. 尤度比の従う分布の数値計算

例えば、正規モデル $M^{'} = M(170, 5.8^2)$ から標本 D を生成する。その μ に対する最尤モデル $M(\hat{\mu}\prime 5.8^2)$ とする。この正規モデル $M^{'}$ に対する尤度比は、次のようになる。

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

データ由来の母数は最尤推定を行った μ なので、その母数の個数は 1 であり、ゆえに尤度比は自由度 1 の χ^2 分布に従う。

実データへの適用

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_ = []
9  for item in data_:
10     #print(item.shape)
11     a = norm(np.average(item),5.8).pdf(item)
12     loglike2_.append(np.prod(a))
```

```
13
14 y = -2*np.log(loglike_/loglike2_)
15 x = sorted(y)
16 y_ = np.arange(len(y))/len(x)
17 plt.plot(x,y_)
18 plt.plot(x,chi2.cdf(x,df=1))
19 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 6.1(a) および (b) に示す。図 6.1(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 6.1(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。

6.1.2. 正規モデルにおける尤度比検定の計算

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。ここで、$\bar{x}$ は標本 X の標本平均。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、

$$L(\mu_0, X) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{\sum(x_i - \mu_0)^2}{2\sigma^2})$$
$$L(\bar{x}, X) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{\sum(x_i - \bar{x})^2}{2\sigma^2})$$

これらから $-2\log\lambda(X)$ を計算する。

$$\begin{aligned}
-2\log\lambda(X) &= -2\log\frac{L(\mu_0, X)}{L(\bar{x}, X)} \\
&= \frac{\sum(x_i - \mu_0)^2}{\sigma_0^2} - \frac{\sum(x_i - \bar{x})^2}{\sigma_0^2} \\
&= \frac{\sum(x_i^2 - x_i^2) - 2n\mu_0\bar{x} + n\mu_0^2 - n\bar{x}^2 + 2n\bar{x}^2}{\sigma_0^2} \\
&= \frac{n}{\sigma_0^2}(\mu_0^2 - 2\mu_0\bar{x} + \bar{x}^2) \\
&= \frac{n}{\sigma_0^2}(\bar{x} - \mu_0)^2
\end{aligned}$$

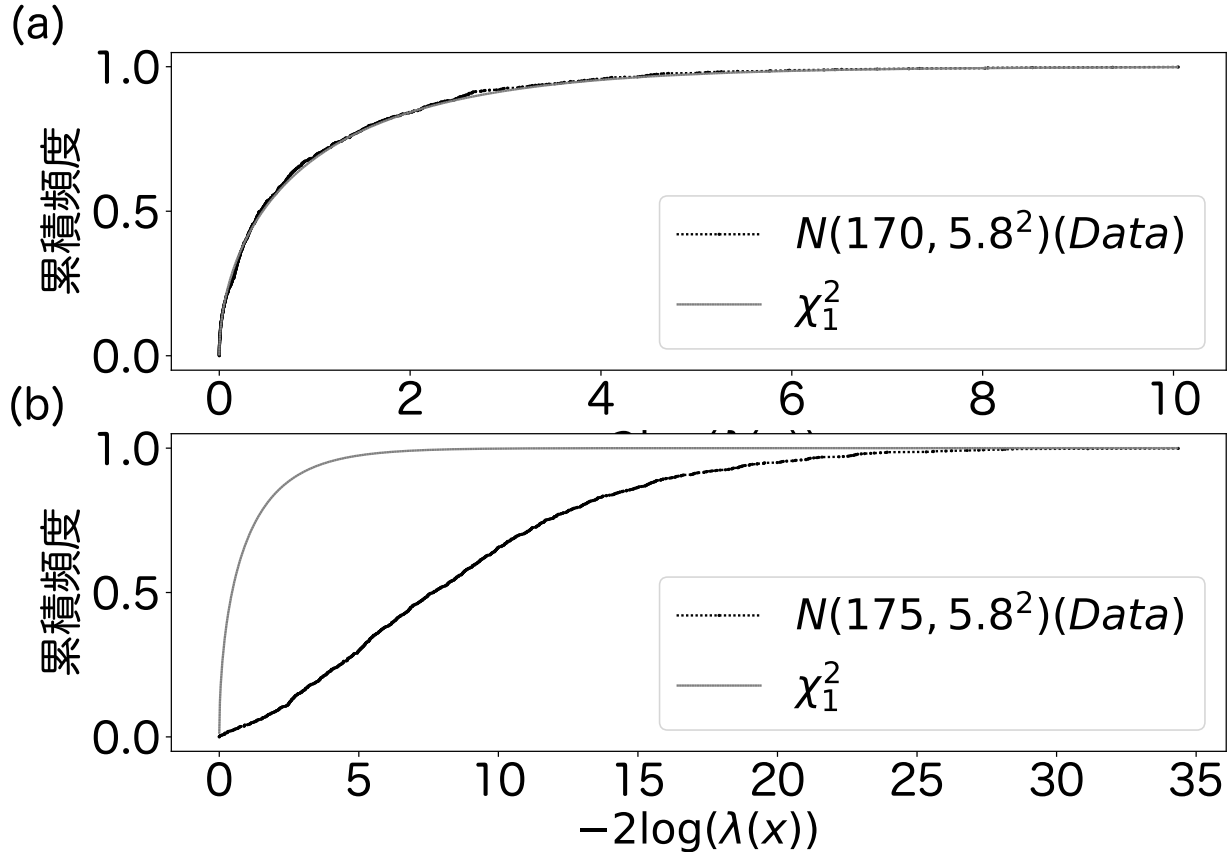


図 6.1　対数尤度比の累積頻度。モデルは正規モデル $M(170; \sigma^2 = 5.8^2)$。(a) 標本を $N(170, 5.8^2)$ からサンプリングした結果。(b) 標本を $N(175, 5.8^2)$ からサンプリングした標本。

これが、χ_1^2 に従う。

6.1.3. 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。

6.2. データと当てはめモデル $\hat{M}$ の比較

データを当てはめたモデル $\hat{M}$ とデータの比較。次を計算する。

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

ここで、標本 D' の最尤推定モデルを $\hat{M}$ とし、D を $\hat{M}$ の標本とし、$\hat{\hat{M}}$ は $\hat{M}$ の p 個の母数に対して D 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^2_p の中で、モデル由来ではない標本 D' に対する尤度比 $Dev(D',\hat{M},\hat{\hat{M}})$ 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。

6.2.1. 注意点

いくつか注意するべきことがある。

1. $\hat{M}$ が D' を予測しているかはこれだけでは不明 (小節 A.8.1)。
2. $\hat{\hat{M}}$ の方が良いとは言えてない。$\hat{M}$ に対する絶対評価[*1]。
3.「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。」などといわれることもあるが、本書とは異る方針の統計利用である。本書では、p が十分小ければ、$\hat{M}$ による予測はやめておいたほうがよいという証拠の 1 つとする。

■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り、相対的に M_2

[*1] 当てはまりの良さは、尤度の大小関係を見れば良い。フルモデルの方が大概の場合当てはまりが良くなる。

の方がより適合的であることを示唆していない。検定では、相対的なモデルのデータへの適合具合を示すことができない。
より適合的であることを示す量の一つとしては、尤度が挙げられる。複数のモデルにおいて、対数尤度が小さいモデルがデータに対して適合しているという判断ができる。

6.2.2. まぜると危険

尤度比検定を行えば、フルモデル[*2]が採用できるという説明がされることがある。これは以下の二つを混合している。

1. 特定モデルに対する標本の乖離具合い
2. モデル間のデータへの当てはまり具合に関する指標

まず、統計的仮説検定では、特定モデルの標本の統計要約量以上の値がそのモデルで出現する確率が計算できる。この道具を本書では、特定モデルとデータの乖離具合の指標の 1 つとして導入した。尤度比検定についても、特定モデルに関するデータの乖離具合の指標である。故に本書では、尤度比検定については、他のモデルの良さについては計れていないという立場をとる。
次に、モデル間のデータへの当てはまり具合について、式 A.9.1(Dev) を観察する。この式は、フルモデルの尤度と子モデルの差分を示している。Dev が 0 よりも小ければ、フルモデルの方が当てはまりが良く、0 より大きければ、子モデルの方が当てはまりがよいことを示す。基本的には、母数の個数が多いフルモデルの方が当てはまりが良いので、対数尤度が大きくなる。このことから、Dev を計算するまでもなく、フルモデルの方があてはまりとしては良い。
これら二つのことをまぜあわせ、結論として、フルモデルが良いという主張がなされていることがある。これら二つの評価とそして、得られた結果は別々に主張するべきことである[*3][*4]。

[*2] 尤度比の分母にくるモデル
[*3] フルモデルの方がいいことはわかるが、フルモデルが実際に使えるかという意味では検証されていないことは常に注意。
[*4] 生物統計学では、混ぜて議論を行っている。いまさら混ぜるなとしてもあまり意味がないだろう。

■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。 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[b]。本書が扱いたい科学において、 特殊なモデルで尤度比を使えば、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。

---

[a] 有意確率はおそらく p 値のこと

[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。

6.3. 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. $y_i \sim p(y_i; \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$

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。
2. M_1 における標本 X の尤度 L_1 を計算する。
3. M_2 において、標本 X を用いて、最尤モデルを構築し、そのモデルで尤度 L_2 を計算する。
4. $-2\log\frac{L_1}{L_2}$ を計算する。以上の手順を繰り返す。

以上を行うと、尤度比 $-2\log\frac{L_1}{L_2}$ が χ_1^2 に従うことがわかる。図 6.2a,b に結果を載せている。

数値計算

コードを書いておく。標本を 1000 個生成し、その尤度比を集め、chi^2 の累積分布関数と比較した。

```
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):
```

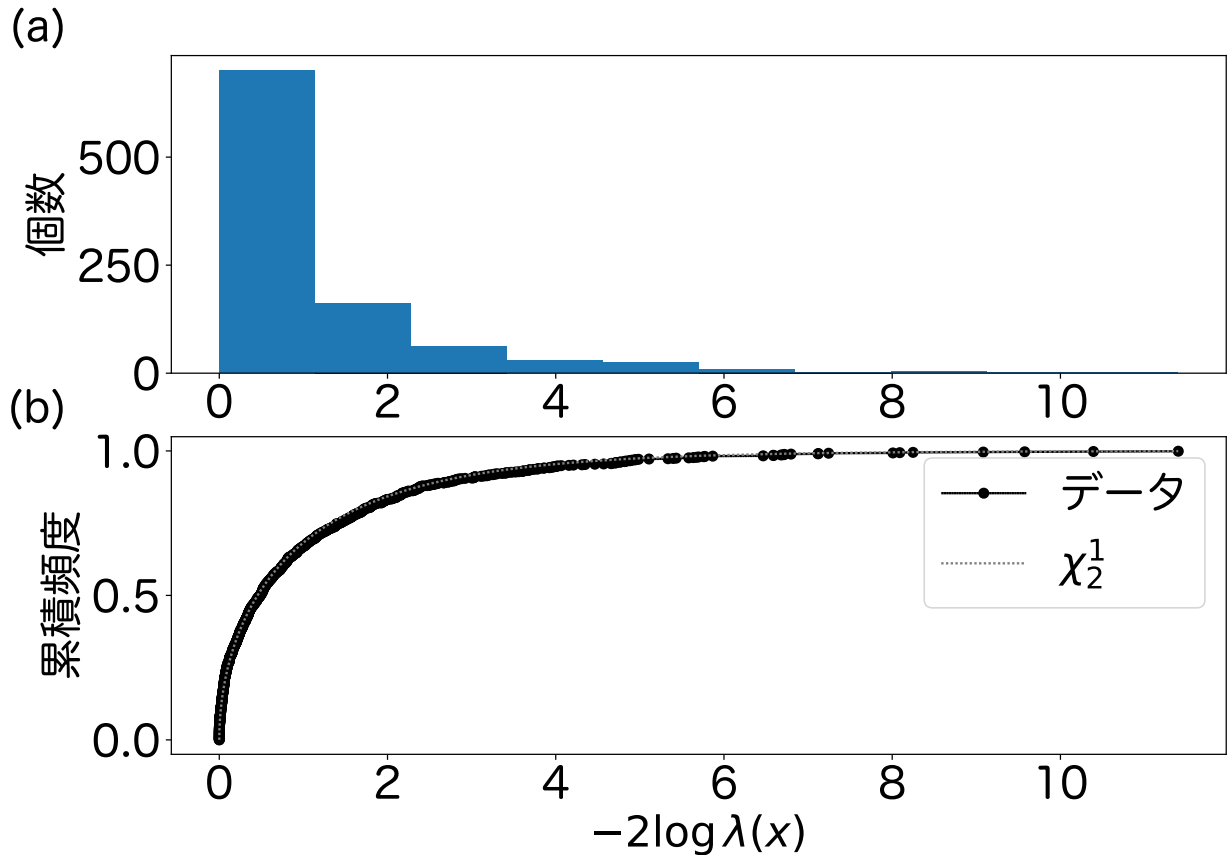


図 6.2　M_1 における対数尤度比の累積頻度。(a) ヒストグラム (b) 累積分布

```
15      l.append(get_dd(df))
16
17
18  x = sorted(l)
19  y = np.arange(len(l))/len(l)
20  plt.plot(x,y)
21  plt.plot(x, chi2.cdf(x,df = 1))
22  plt.show()
```

6.3.1. データとの比較

データと、最尤モデル M_1 との比較は同様に、

1. 無作意抽出した標本を x とする (モデルから得た標本ではない。)
2. 尤度 L_1 を M_1 上で計算する。
3. 標本 x の尤度 L_2 を M_2 上で計算する。
4. $-2\log\frac{L_2}{L_1}$ 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、$-2\log\frac{L_2}{L_1}$ が大きな値を取る。

■帰無仮説の取り方

通常の尤度比検定は、次の通りである。

定理 6.3.1. 帰無仮説:$M_a = M(\alpha)$
対立仮説:$M_b = M(\alpha, \beta \neq 0)$
とする。このとき次が自由度 *1* のカイ二乗分布に従う。

$$-2\log\frac{\text{帰無仮説における尤度}}{M_a\text{におけるサンプル } X \text{ を元にした最尤 } M(\alpha,\beta) \text{ での尤度}}$$

「データ解析のための統計モデリング入門」では定理として示されてはいないが、恣意的にまとめてみると次のことが成立していると考えられている。

定理 6.3.2. 帰無仮説:$M_a = M(\alpha)$
対立仮説:$M_b = M(\alpha, \beta \neq 0)$
とする。このとき、次が自由度 *1* のカイ二乗分布に従う。

$$-2\log\frac{\text{帰無仮説における尤度}}{M(\alpha,\beta\neq 0) \text{ の最尤モデルにおける尤度}}$$

尤度比が、数理統計学で紹介されるものとは異なっている。具体的には、分母における最尤推定される母数に関して制限がない物が通常の定理である。これが成立するのかは私にはわからないので、本書では採用していない。
また、本書と異なる点は、対立仮説を採択している点である。

> So we can state that 対立仮説 can be accepted.[a]

本書においては、尤度比検定ではモデルとの乖離を調べれるという方針なので、モデルとデータを比較していないのに、あるモデルを採択することはありえない[b]。生物統計では、モデルとデータの乖離を十分検証しなくても、そのモデルを採用するということがあるという点は留意するべきであろう[c]。

[a] https://kuboweb.github.io/-kubo/stat/2019/Ees/d/kubostat2019d.pdf
[b] 尤度比の性質は帰無仮説にがだめっぽいということを示している。どちらかのモデルの二者択一できるような物ではないとする。
[c] ネストされたモデルとデータの乖離具合を調査したのか？という疑問が残ることが多々ある。

6.4. One-way ANOVA

統計モデル $M_N = M(\mu, \sigma^2)$ に対し、これを拡張したモデルを考える。そのモデルは次の仮定をもつ。

1. $\mu_i (i = 1, 2, \cdots, n)$
2. $x_{ik} \sim N(\mu_i, \sigma^2)$

ここで、$m = \sum_{i=1}^{n} n_i$ と定義する。これを統計モデル $M_A = M(\mu_1, \mu_2, \cdots, \mu_n)$ とする。フルモデルと子モデルそれぞれについて、最尤モデルを検討する。

■子モデルの最尤推定　M_N から得られたサンプル x についてその最尤推定量は次のようになる。

$$\bar{\mu}_{\mathrm{ML}} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{k=0}^{n_i} x_{ik}, \bar{\sigma}^2_{\mathrm{ML}} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{k=1}^{n_i} (x_{ik} - \bar{\mu}_{ML})^2$$

こモデルの尤度は、式 6.1 を参考に考えると次のようになる。

$$\begin{aligned} L_N &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\bar{\sigma}^2_{\mathrm{ML}}}} \exp\left(-\frac{(x_{ij} - \bar{\mu}^2_{\mathrm{ML}})}{2\bar{\sigma}^2_{\mathrm{ML}}}\right) \\ &= (2\pi\bar{\sigma}^2_{\mathrm{ML}})^{-\frac{m}{2}} \exp\left(-m/2\right) \end{aligned}$$

■フルモデル　M_A からサンプリングされた標本 $x = (x_{ik})(k = 1, \cdots, n_i)$ とすると、このモデルの尤度は次のようになる。

$$L(\mu; x) = \prod_{\substack{i=1,2,\cdots,n \\ k=1,\cdots,n_i}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_{ik} - \mu_i)^2}{2\sigma^2})$$

最尤推定量は次のようになる。

$$\bar{\mu_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = \bar{x_i}, \bar{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2$$

尤度についても計算しておく。

$$\begin{aligned} L_A &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\bar{\sigma}^2}} \exp\left(-\frac{(x_{ij}-\bar{\mu_i})^2}{2\bar{\sigma}^2}\right) \\ &= (2\pi\bar{\sigma}^2)^{-\frac{m}{2}} \exp\left(-\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(x_{ij}-\bar{\mu_i})^2}{2\bar{\sigma}^2}\right) \\ &= (2\pi\bar{\sigma}^2)^{-\frac{m}{2}} \exp\left(-\frac{m\bar{\sigma}^2}{2\bar{\sigma}^2}\right) \\ &= (2\pi\bar{\sigma}^2)^{-\frac{m}{2}} \exp\left(-m/2\right) \end{aligned}$$

■尤度比　以上をもとに尤度比を計算する。

$$\begin{aligned} \lambda = \frac{L_N}{L_A} &= \frac{(2\pi\bar{\sigma}^2)^{-\frac{m}{2}} \exp\left(-m/2\right)}{(2\pi\bar{\sigma}^2_{\mathrm{ML}})^{-\frac{m}{2}} \exp\left(-m/2\right)} \\ &= \left(\frac{\bar{\sigma}^2}{\bar{\sigma}^2_{\mathrm{ML}}}\right)^{-\frac{m}{2}} \\ &= \left(\frac{\bar{\sigma}^2}{\bar{\sigma}^2_{\mathrm{ML}}}\right)^{-\frac{m}{2}} \\ &= \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_i})^2}{\sum_{i=1}^{n}\sum_{k=1}^{n_i}(x_{ik}-\bar{\mu}_{ML})^2}\right)^{-\frac{m}{2}} \end{aligned}$$

M_N の標本 x について、$-2\log\lambda(x)$ が χ^2 分布に従う。自由度は、M_N が 2 であり、M_A が、$n+1$ なので、χ^2 の自由度は $n+1-2=n-1$ である。
以上のことから、1 つの平均値を設定したモデルとデータが乖離していることを調べることができる。モデル M_N における統計的性質であることに注意。

■少なくとも一つは母数が違う

1-way ANOVA を使い、$p < \alpha$ を得たならば、少なくとも一つは母数が違う郡が入っている。言い換えれば、帰無仮説が棄却され、対立仮説を採択している。有意差検定では、対立仮説を採択することはないと説明している教科書でも、ANOVA では対立仮説を採択していることがある[a]。

本書では、M_N モデルとデータが乖離しているかもしれない程度しかわからないということにする。異なる母数で推測した方が良いという結論を得るには、検定以外の方法で、データを解析することが必要である。

[a] ハイブリッド検定が使われているので、その方針を採用しているだけである。

■数値計算　統計的性質が現れることを数値計算により確かめてみる。

```
mu = 170
sigma=5.8
n=10
sampleN = 200

norm_ = norm(mu,sigma)
l=[]
for i in range(10000):
    sample = norm_.rvs(size=(n,sampleN))

    ave_ = np.average(sample)
    sigma_ = np.std(sample)
    #sigma_, ave_
    ave_s = np.average(sample,axis=1)
    sigma_s = np.std(sample,axis=1)
    ave_s.shape,sigma_s.shape

    lam = np.sum( (sample - ave_)**2)/np.sum( (sample.T -
        ave_s)**2)
    l.append((sampleN*n)*np.log(lam))

x = np.sort(l)
y = np.arange(len(l))/len(l)
plt.plot(x,y)
plt.plot(x ,chi2.cdf(x,df = n-1))
```

```
25 plt.show()
```

第 7 章

身長を予測する統計モデル

7.1. 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、$\mu = 171.1\text{cm}$ 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。

[*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 ～ 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

本書では、恣意的に考えたモデルまたはこれまでの研究成果から建てられるモデルを使って推測をしてみるという考えに基づいて、統計モデルを構築する。これは実現可能性という考え方とは異なる。

7.2. 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。

7.2.1. ○○ 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、$P(x > 180) = 0.059$ ということが分かる。このことから、母集団から 100 人程度の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。

$160cm$ 以下の人割合も推測する。

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

結果、$P(x < 160) = 0.059$ と推測できる。

$P(x < 160)$ と $P(x > 180)$ が極めて近い値である理由は、この正規分布モデルが母平均 $\mu = 171.1$ を中心に対称に分布する関数であるためである。つまり、171.1 からおよそ

$10cm$ 離れた $160cm$ 以下の人と $180cm$ 以上の人は、両方とも平均から同じ距離だけ離れているため、正規モデルであれば同程度の確率で現れる。

7.2.2. 擬似的に無作為抽出を行う

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。

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

7.2.3. 母数によって変化する予測

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。$180cm$ 以上の人をモデル $M(168)$ により推測すると $P(x > 180) = 0.03$ 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 7.1 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。

■ $170cm$ 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 7.2 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 7.2 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 7.1)。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

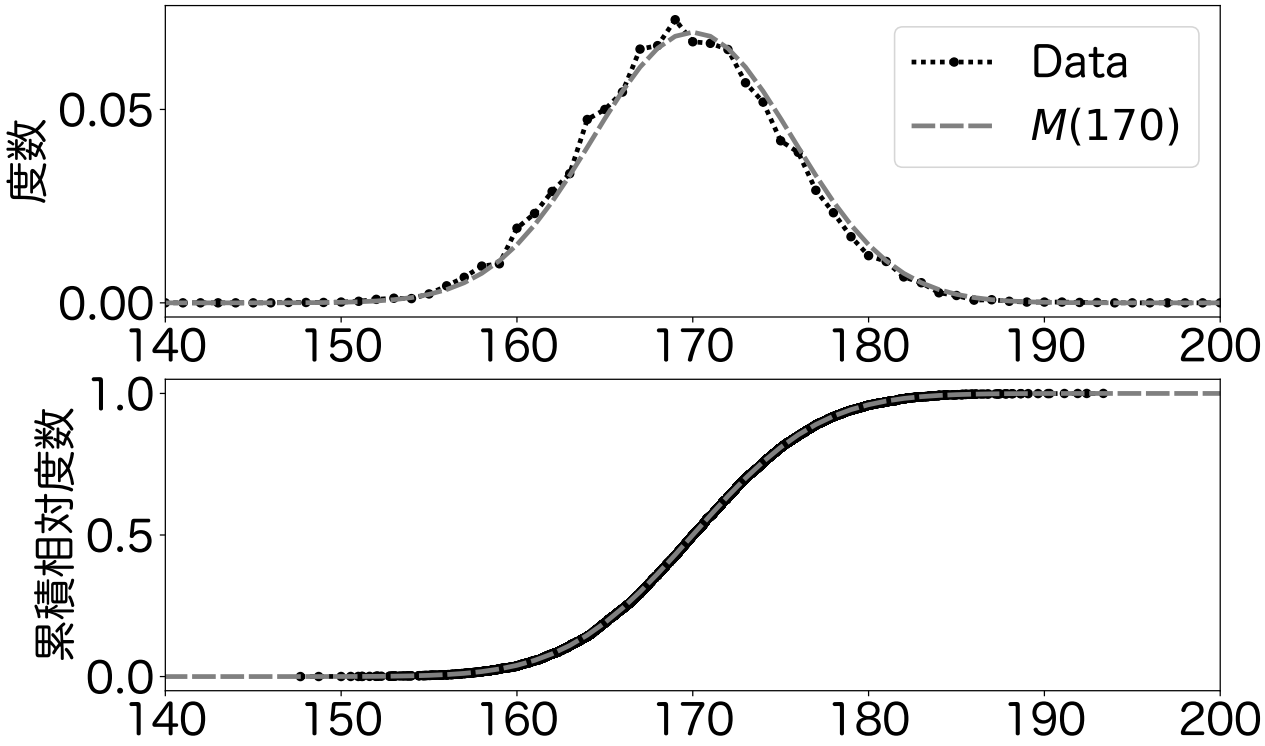


図 7.1　上:正規分布を含む統計モデル $M(170)$ によりサンプリングされた Data の頻度と、統計モデルの頻度。下:上と同じデータ・統計モデルの累積相対頻度

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。

7.3. 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。

7.3.1. 極端な値を使って調べる

統計モデルの極端な予想を使い、データと比較する。

サンプルサイズが大きい場合

データでは $180cm$ 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

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

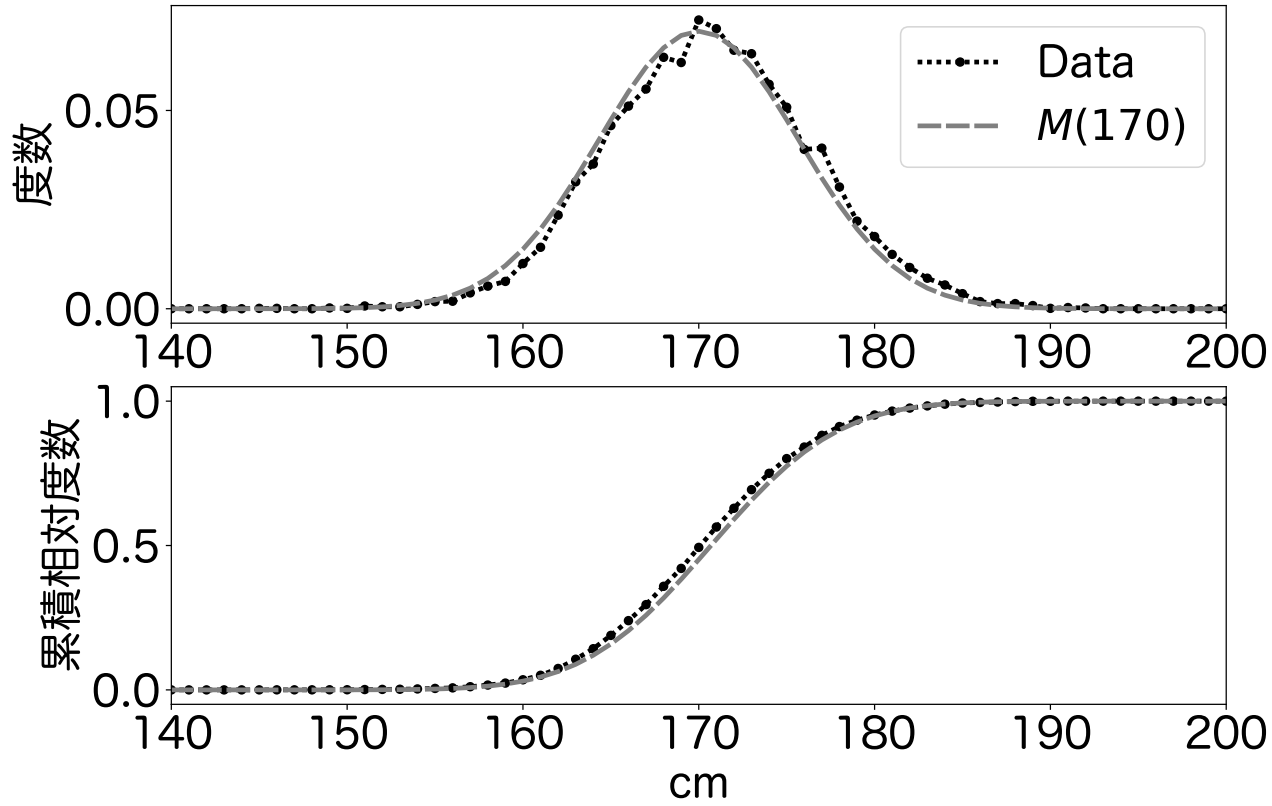


図 7.2　17 歳の男性から無作為抽出したデータ。上は、データと統計モデル $M(170)$ の度数。下は、データと統計モデル $M(170)$ の累積相対度数

の推測値 $P(x < 160) = 0.025$ と数値が近い。

表 7.1　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

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。

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や、二つのうち一方の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する（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。

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

表 7.2　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 7.2]。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。

7.4. 統計モデルとデータの比較 2

7.4.1. 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 7.3B 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 7.3A である。図 7.3A 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

7.5. 検定統計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標本の平均値 $\bar{x}$ 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。

7.5.1. データの検定統計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

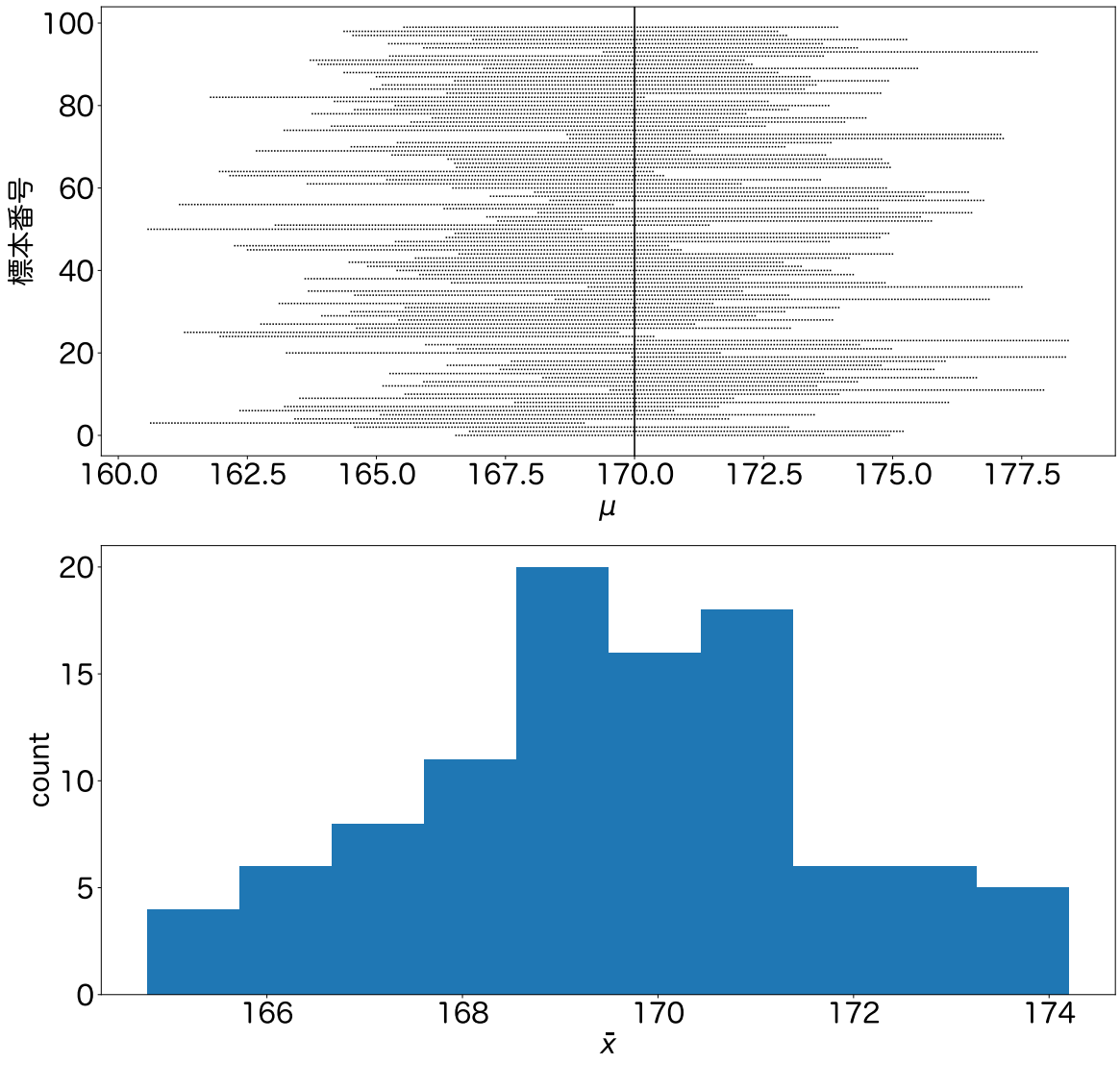


図 7.3　(A) モデルから標本を得て、その標本から信頼区間を計算し、表示したもの。
(B) 標本平均の分布

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

7.5.2. 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 7.4 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。

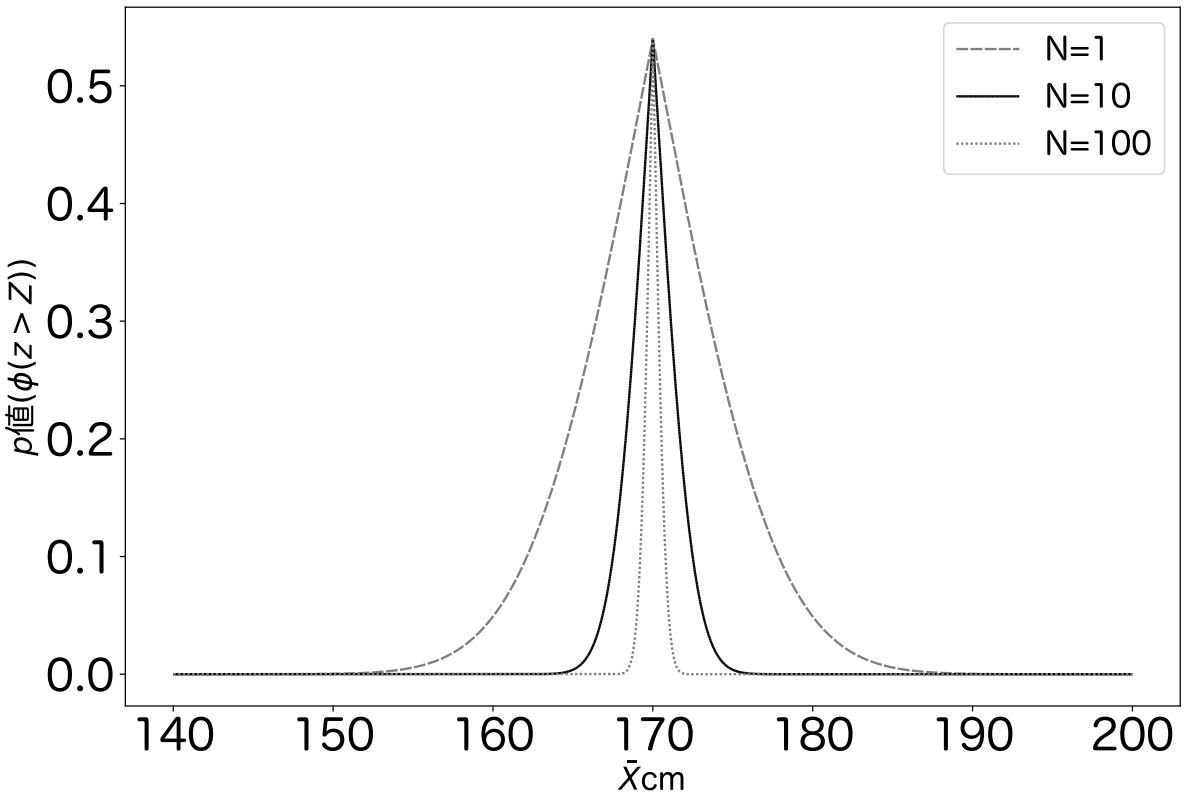


図 7.4　各標本平均 $\bar{X}$ が $M(\mu = 171)$ において得られる確率。$N = 1, 10, 100$ の場合。

第 8 章

モデルを使った研究の進め方

8.1. 指針

いくつか言葉を定義する。

定義 8.1.1. 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 8.1 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

[*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。

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

[*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

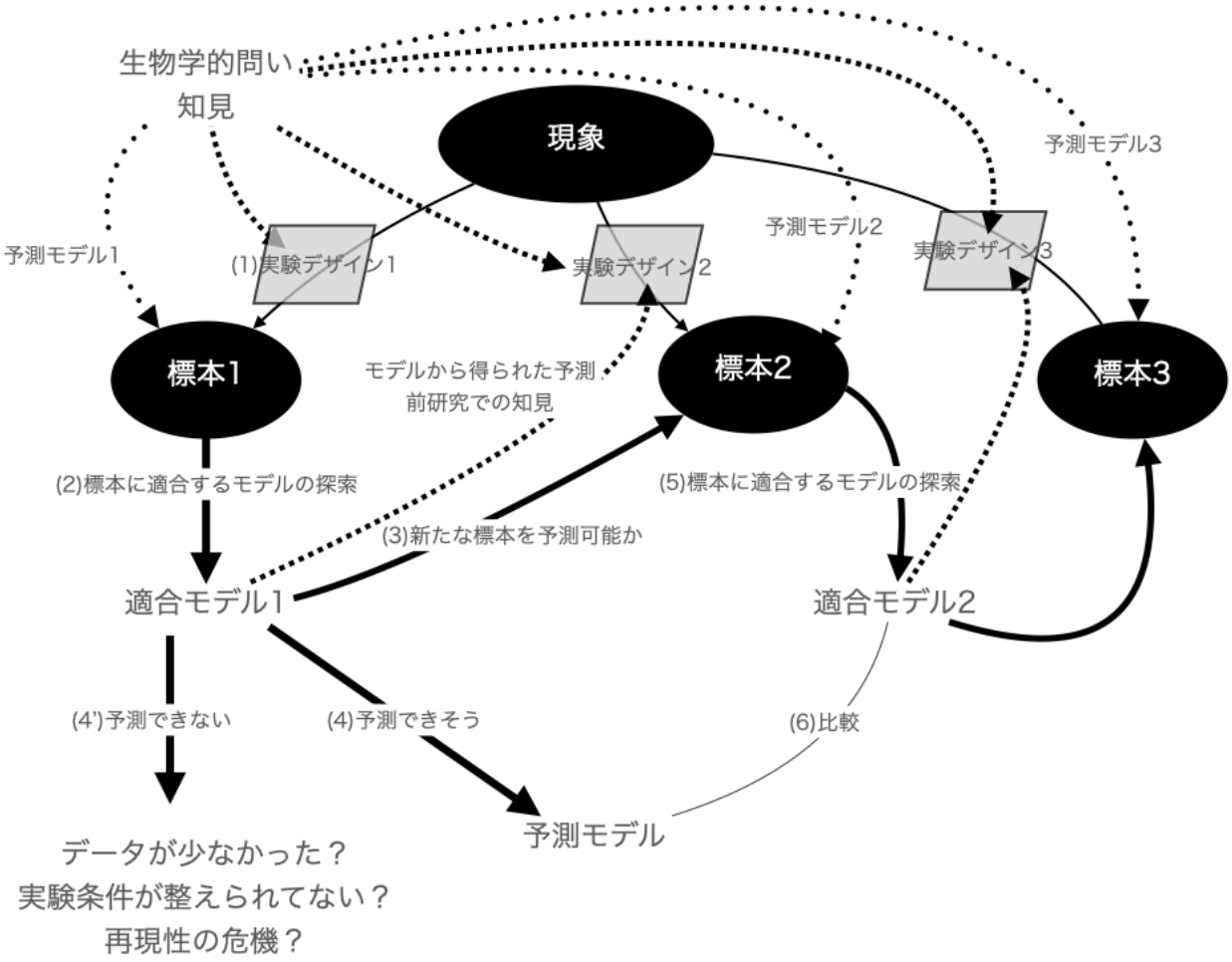


図 8.1　統計モデルを使った研究のフロー図。

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。

8.1.1. 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。

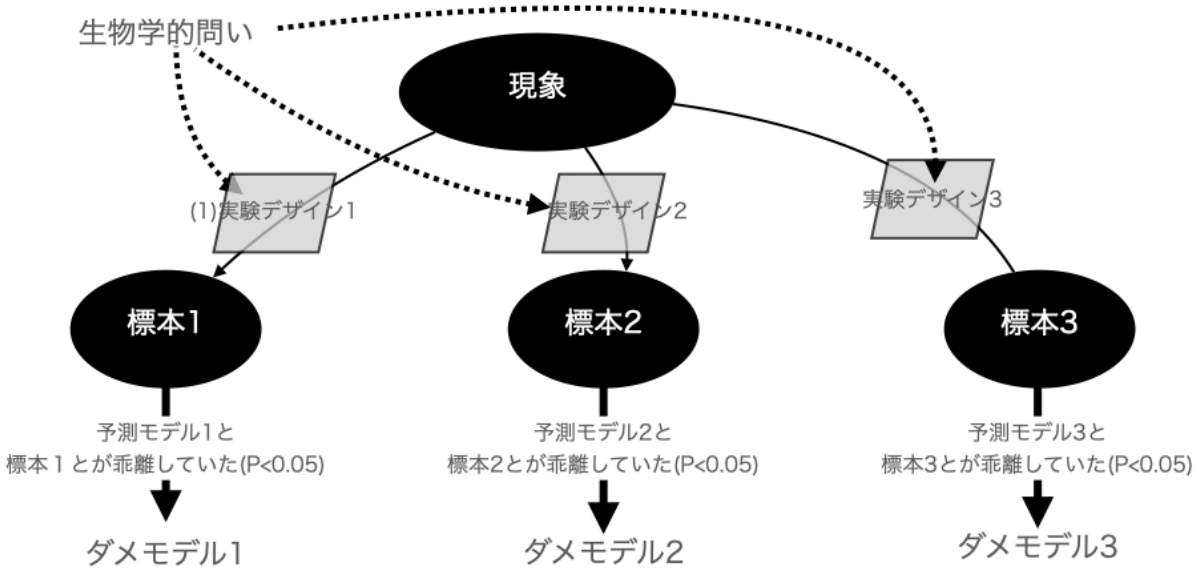


図 8.2　統計モデルを使った従来研究のフロー図。

8.2.　ダメモデルを羅列する研究例

図 8.2 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。

8.3. 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。

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。効果量 d が十分小さければ、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。

[*3] これだけだと何がどう違うのかは言及できてない

8.3.1. 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()
```

8.3.2. 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 8.1 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 8.1　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

図 8.3 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

[*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

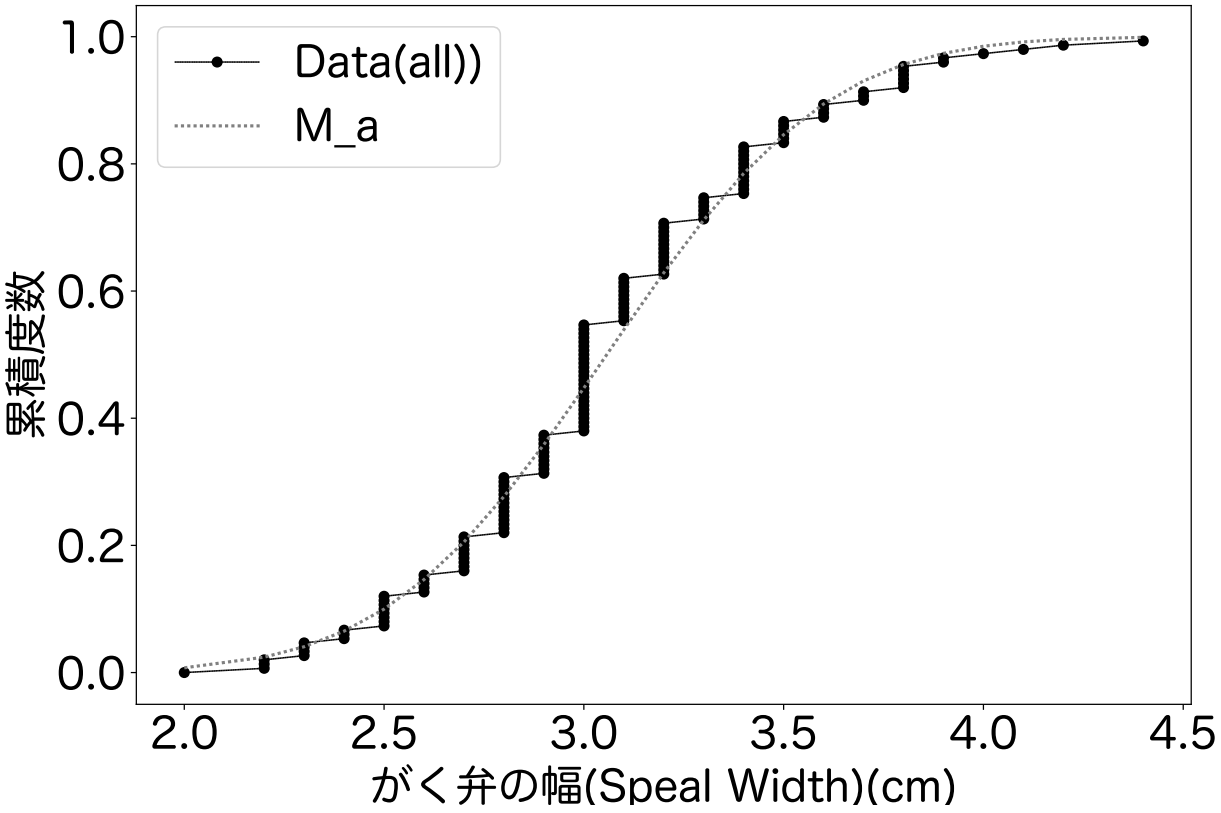


図 8.3　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M_a

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。

8.3.3. 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 8.2 がモデル M_a の予測

[*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

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 8.4 には、モデル M_a とデータの累積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 8.2　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 $M_a = M(\mu = 3.05, \sigma^2 = 0.434^2)$ における検定統計量も利用する。次のことがわかっている。

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

検定統計量 Z を計算した結果が表 8.3 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。

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。

表 8.3　検定統計量 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	-

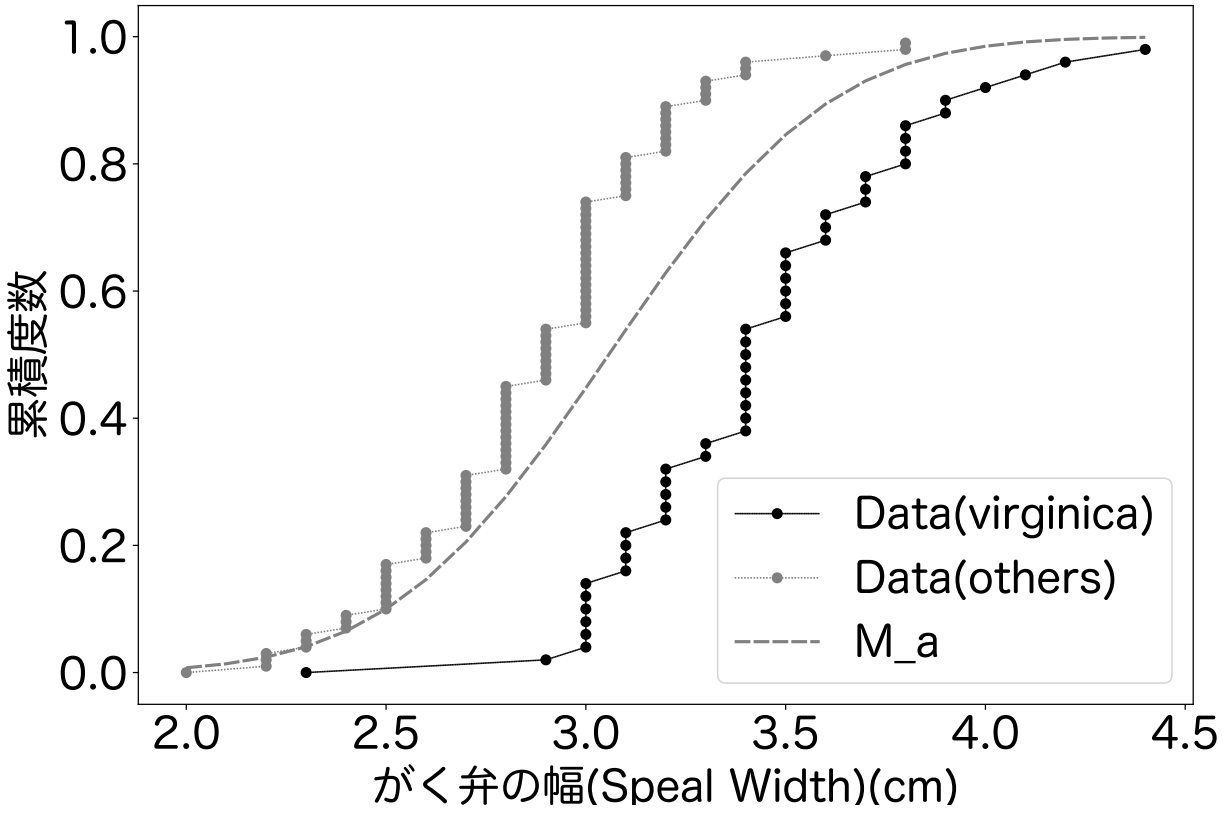


図 8.4　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M_a

8.3.4. 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 8.3 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 8.4 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 8.5 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

表 8.4　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

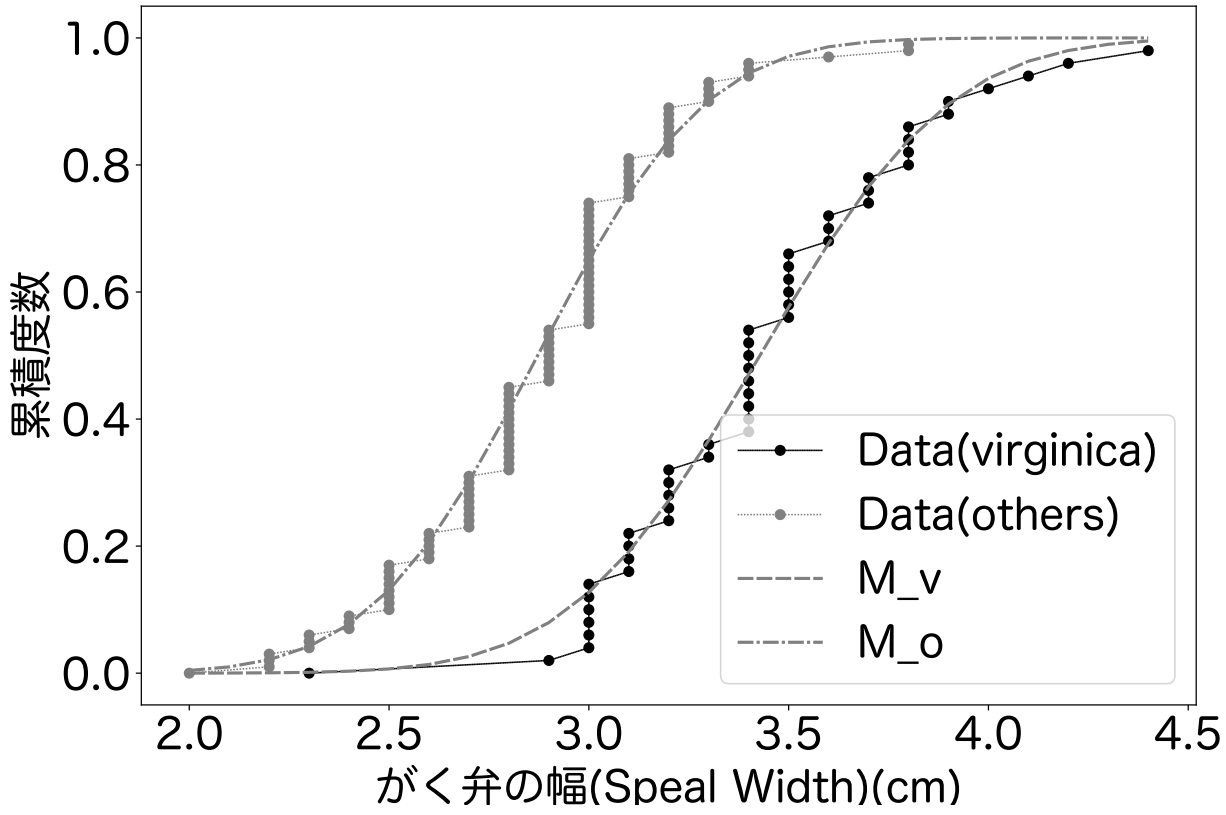


図 8.5　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M_v, M_o

8.3.5.　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

表 8.5　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

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 8.5 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。

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 8.6)。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。

[*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、分布形や検定統計量を変える。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。

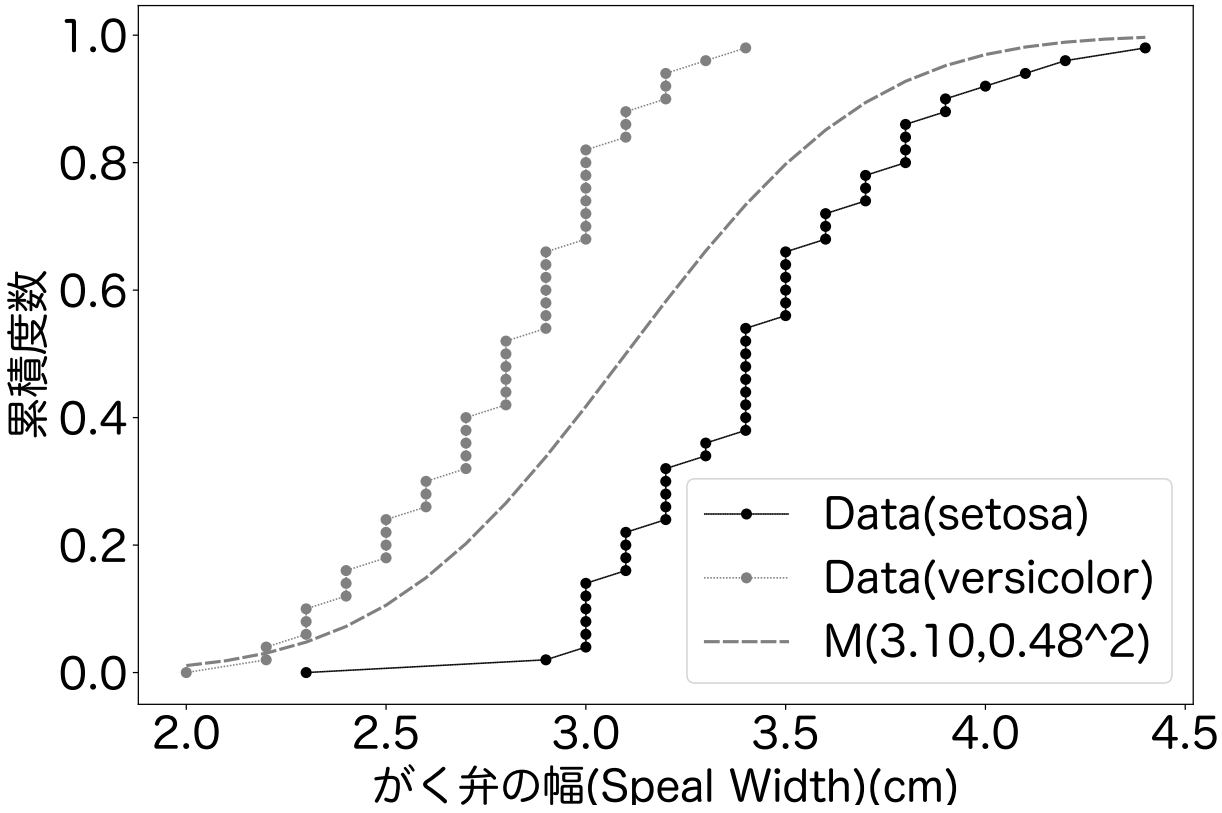


図 8.6　データのがく弁の幅の累積度数 (Data) と最尤モデルの累積度数 M

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。

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。

8.4. 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ないほうがよい。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。

第 9 章

親子の身長の研究

二つの要素に対する 3 つのモデルを取り上げる。まず、これらの性質について整理する。

9.1. 独立ではない変数を持つモデル

1. x, y を確率変数とする。
2. 平均を μ_x, μ_y 分散を σ_x^2, σ_y^2 とする。

このモデルでは、次の X と Y の共分散という量が定義できる。

$$\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$$

また、相関係数を次のように定義する。

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$(x - \mu_x)(y - \mu_y)$ という量を考えてみる。これは (μ_x, μ_y) を中心にし、

1. (x, y) が右上にあれば、$x - \mu_x$ と $y - \mu_y$ はともに正であり、その積も正
2. (x, y) が左下にあれば、$x - \mu_x$ と $y - \mu_y$ はともに負であり、その積は正
3. (x, y) が左上にあれば、$x - \mu_x$ は負そして、$y - \mu_y$ は正であり、その積は負
4. (x, y) が右下にあれば、$x - \mu_x$ は正そして、$y - \mu_y$ は負であり、その積は負

である。以上から、$(x - \mu_x)(y - \mu_y)$ は 2 次元平面上でのある中心からデータのばらつき方を示す量であること、そして、その平均は、平均的なデータのばらつきかたを示す。言い替えれば、平均的にデータが (μ_x, μ_y) を中心に右上から左下に多く存在すれば、共分

散は正の値をとることが期待される。共分散を σ_x, σ_y により規格化した量が相関係数である。
相関係数は、二つの変数が確率変数であることを仮定したモデルにより定義できる量である。

9.1.1. $r^2 \leq 1$ の証明

コーシーシュワルツの不等式 $a_1, a_2, \cdots, a_n, b_1, b2, \cdots, b_n$ を実数とする。

$$(\sum_{i=1}^{n} a_i b_i)^2 \leq (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$$

である。等号成立は、$a_i = 0$ または $b_i = 0$ または $b_1/a_1 = b_2/a_2 = \cdots = b_n/a_n$ が成り立つときである。
このことを利用する。$a_i = x_i - \bar{x}, b_i = y_i - \bar{y}$ とおく。

$$\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2 \leq \sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2$$

このことから、$r^2 \leq 1$ であり、$-1 \leq r \leq 1$ がわかる。

9.2. 2 変量正規分布

独立ではない二つの変数に対するモデルを構築する。

1. $(x_i, y_i) \sim F(i = 1, 2, \cdots, n)$
2. F は 2 変量正規分布
3. 母数は、平均値、分散および共分散

このモデルを 2 変量モデル $M_2(\mu, \sum)$ と呼ぶ。このモデルの最尤推定量は次の通り。

$$\mu_{\mathrm{ML}} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i\right)$$
$$\sum_{\mathrm{ML}} = \begin{pmatrix} \frac{1}{n}\sum(x_i - \bar{x})^2 & \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) & \frac{1}{n}\sum(y_i - \bar{y})^2 \end{pmatrix}$$

x_1 または x_2 を固定したときに得られる、x_2 または x_1 の期待値と分散は次の通り。

$$E[x_2|x_1] = \mu_2 + \frac{\sigma_{xy}}{\sigma_x^2}(x_1 - \mu_1)$$
$$Var[x_2|x_1] = \sigma_2^2(1 - \rho^2)$$
$$E[x_1|x_2] = \mu_1 + \frac{\sigma_{yx}}{\sigma_y^2}(x_2 - \mu_2)$$
$$Var[x_1|x_2] = \sigma_1^2(1 - \rho^2)$$

$E[x_2|x_1]$ を x_2 の x_1 に対する回帰という。この式は、平均二乗誤差を最小にするという良い性質をもっている。

2 変量正規モデルににおいて相関係数を Bravais-Pearson(ブラベー・ピアソン) の係数と呼ぶこともある。2 変量正規モデルその共分散が 0 のモデルにおいて、相関係数が次の分布に従う[*1]。

$$r \sim$$

このことを利用して、データと 2 変量正規モデル間の乖離具合を調べることができる。これは、相関係数に対する検定と呼ばれる作業に対応する。

9.2.1. 確率密度関数

このモデルの確率密度関数を確認する。

$$f(x,y) =) \frac{1}{\sqrt{(2\pi)^2|\sum|}} \exp\left(-\frac{1}{2}{}^t(\boldsymbol{x}-\boldsymbol{\mu})\sum{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$
$$= \frac{1}{\sqrt{(2\pi)^2|\sum|}} \exp\left(-\frac{1}{2}\left(\sigma_x^2(x-\mu_x)^2 - 2\sigma_{xy}(x-\mu_x)(y-\mu_y) + \sigma_y^2(y-\mu_y)^2\right)\right)$$

ここで、平均 $\mu = (\mu_x, \mu_y)$ 共分散行列 $\sum$ は

$$\sum = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

また、行列式は $|\sum| = |(\sigma_x^2\sigma_y^2 - \sigma_{xy}^2)|$ である。

確率密度関数の指数関数の内側の項について考察する。また、平均は $(0,0)$ の場合を考える。この式を一般性を失うことなく、式変形すると、

$${}^t\boldsymbol{x}A\boldsymbol{x} = {}^t\boldsymbol{x}\begin{pmatrix} a & b \\ b & c \end{pmatrix}\boldsymbol{x} = ax^2 + 2bxy + cy^2 = 1$$

[*1] 正規 2 モデル $M(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ において r がある分布に従うと言ってもいいが、相関係数を計算するとき、我々の頭のなかには 2 変量正規モデルでみてみようという考えがあり、正規モデルは頭にないはずである。

となる。この式は、ある条件を満すと楕円を表す式である。ここではある条件は満されているものとする。
ここで、行列 A の固有ベクトルを列ベクトルとする行列 P を定義する。

$$P = \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}$$

このとき、$A = {}^tPDP$ である。固有ベクトルの一次結合により $\boldsymbol{x}$ を表すと、$\boldsymbol{x} = P\boldsymbol{c}$ となる。これを ${}^t\boldsymbol{x}A\boldsymbol{x}$ に代入すると以下が得られる。

$$\lambda_1 c_1^2 + \lambda^2 c_2^2 = 1$$

このことから、楕円の長辺と短辺は固有値であり、その方向は、固有ベクトルであることがわかる。
行列 A の固有値と固有ベクトルを計算する。

■固有値　固有値は、行列式 $|A - \lambda E| = 0$ を満す λ のことである。これは、計算を行うと、

$$\begin{aligned} \left| \begin{pmatrix} a-\lambda & b \\ b & c-\lambda \end{pmatrix} \right| &= (\lambda - a)(\lambda - c) - b^2 \\ &= \lambda^2 - \lambda(a+c) + ac - b^2 = 0 \end{aligned}$$

より、この方程式の解は、

$$\lambda_1 = \frac{a + c + \sqrt{(a-c)^2 + 4b^2}}{2}$$
$$\lambda_2 = \frac{a + c - \sqrt{(a-c)^2 + 4b^2}}{2}$$

である。これが固有値である。

■固有ベクトル　固有ベクトルは、行列 $(\sum -\lambda E)$ をベクトル $(x\ y)$ に作用させたとき 0 になるようなベクトルのことである。これは、以下を解けばよい。

$$\begin{pmatrix} a-\lambda_1 & b \\ b & c-\lambda_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0$$

すると、

$$\boldsymbol{X_1} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} \frac{a-c+\sqrt{(a-c)^2+4b^2}}{2b} \\ 1 \end{pmatrix}$$

同様に、λ_2 についての固有ベクトルを求める。

$$\boldsymbol{X_2} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} \frac{a-c-\sqrt{(a-c)^2+4b^2}}{2b} \\ 1 \end{pmatrix}$$

ここで、固有ベクトルを列ベクトルとする行列 P を定義する。

$$P = \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}$$

この行列 P の転置行列は、P の逆行列と一致する。

$$^tP = P^{-1}$$

また、λ_1 と λ_2 が $\sum$ の固有値であるということから、次が成り立つ。

$$P^{-1} \sum P = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} (= D)$$

ここで、固有値を並べた行列を D とする。
以上から、次が成り立つ。

$$^t\boldsymbol{x} \sum \boldsymbol{x} = {}^t\boldsymbol{x} P D P^{-1} \boldsymbol{x}$$

■固有ベクトルの線形和による解　固有ベクトルの線形和は、$c_1X_1 + c_2X_2$ とかける。これを行列で表記する。

$$\boldsymbol{y} = \begin{pmatrix} c_1x_1 + c_2x_2 \\ c_1y_1 + c_2y_2 \end{pmatrix} = P \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = P\boldsymbol{c}$$

このベクトル y を楕円の方程式の x に代入する。

$$\begin{aligned} ^t\boldsymbol{y} \sum \boldsymbol{y} &= {}^t(P\boldsymbol{c}) \sum (P\boldsymbol{c}) \\ &= {}^t\boldsymbol{c} P^{-1} \sum P\boldsymbol{c} \\ &= {}^t\boldsymbol{c} D \boldsymbol{c} = \lambda_1 c_1^2 + \lambda_2 c_2^2 = 1 \end{aligned}$$

これを計算すると、$\lambda_1 c_1^2 + \lambda_2 c_2^2 = 1$ となり、これを満す (c_1, c_2) は円を x 方向に λ_1 倍、y 方向に λ_2 倍したものである。また、これを満す $(c_1, c_2) = (c_1\sqrt{\lambda_1}, c_2\sqrt{\lambda})$ は、円である。

■楕円の面積

$$\begin{aligned} \lambda_1 c_1^2 + \lambda_2 c_2^2 &= 1 \\ \left(\frac{1}{\sqrt{\lambda_1^{-1}}}\right)^2 c_1^2 + \left(\frac{1}{\sqrt{\lambda_2^{-1}}}\right)^2 c_2^2 &= 1 \end{aligned}$$

ここから、楕円の面積は、

$$S = \frac{\pi}{\sqrt{\lambda_1 \lambda_2}}$$

9.2.2. 決定係数 R^2 と相関係数 ρ^2 の関係

$E[x_2|x_2]$ による予測において、決定係数 R^2 と相関係数 ρ^2 の関係を調べる。これを用いて、残差平方和の計算をおこなう。

$$\begin{aligned} \sum_{i=1}^{n}(y_i - E[x_2|x_i])^2 &= \sum_{i=1}^{n}((y_i - \bar{y}) - \frac{\sigma_{xy}}{\sigma_x^2}(x_i - \bar{x})^2) \\ &= n\sigma_y^2 - n\frac{\sigma_{xy}^2}{\sigma_x^2} \\ &= n\sigma_y^2(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}) \\ &= n\sigma_y^2(1 - \rho^2) \end{aligned}$$

以上より、平均二乗誤差は、以下とも等しい。

$$RSS = (\sigma_y\sqrt{1-\rho^2})^2$$

このことを用いて、決定係数 R^2 について計算を行う。

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{n\sigma_y^2(1-\rho^2)}{n\sigma_y^2} \\ &= \left(\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right)^2 = \rho^2 \end{aligned}$$

このことから x_2 の x_1 に対する回帰においては、決定係数と相関係数の二乗が一致することがわかる。また、$0 \leq R^2 \leq 1$ であることもわかる[*2]。

[*2] 直感的に、平均よりも悪くなることはないので、$0 \leq R^2$ も明らか

■決定係数の意味　決定係数は、平均値による予測と比べて、提案した誤差モデルでの予測がどれくらい良くなったのかを示す指標である。第二項は、分母に、平均と観測値の差の二乗、分子に予測値と観測値の差の二乗をしたものである。予測性能が良ければ、分子が小くなり、第二項が 0 に近くなり、決定係数は 1 に近付く。予測性能がわるければ、分子が大きくなり、平均よりも悪い予測をおこなうならば、決定係数は負になる。

9.2.3. 独立な２変数モデルの回帰平均二乗誤差

ここで、正規モデル $M(\mu,\sigma)$ を正規モデル $M'(\mu,\sigma)$ により予測することを考える。モデル M, M' での確率変数をそれぞれ x, y とする。x を y により予測したとき、その誤差を計算する。

$$RMSE^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

$$\begin{aligned} RMSE^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left((x_i-\mu)-(y_i-\mu)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{2}{n}\sum_{i=1}^{n}(x_i-\mu)(y_i-\mu) \\ &= Var[x] + Var[y] - 2cov[x,y] \\ &= 2Var[x] \end{aligned}$$

以上より、

$$RMSE_{random} = \sqrt{2}\sigma$$

がわかる。これは、確率変数を別の確率変数により予測しようとすると、二乗誤差は分散の $\sqrt{2}$ 倍である[*3]。

決定係数が負になる

以上の議論を踏まえ、決定係数が負になることをみておこう。

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{n}(x_i-y_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ &= 1 - \frac{2n\sigma}{n\sigma} \\ &= -1 \end{aligned}$$

[*3] 平均により予測すると、二乗誤差は σ なので、この予測モデルはそれよりも悪い。

平均値より予測精度が悪いことを決定係数が示している。

9.2.4. 相関係数が 0.8 のとき

相関係数が 0.8 のとき、$RMSE$ は、$RMSE^2 = \sigma_y\sqrt{1-\rho^2}$ より、

$$RMSE = 0.6\sigma$$

である。これは、回帰により二乗誤差が σ からその 0.6 倍改善されたことを示している。また、確率変数による予測は $\sqrt{2}\sigma$ なので、このモデルと比較すると、

$$0.6\sigma = 0.428(\sqrt{2})\sigma$$

より、0.42 倍 改善されている[*4] ランダムモデルの $RMSE$ の 58% 割引ともいえる。

9.3. 誤差モデル I

ここで、誤差項が確率変数であることを仮定してモデル Model I を構築する。

1. x_i は、与えられた定数
2. a, b を実数の定数
3. $u_i = y_i - ax_i - b$
4. $u_i \sim N(0, \sigma^2)$

この仮定により構築されるモデルを $M_I(a, b; x)$ または $M_I(a, b)$ と表記する[*5]。また、最尤推定量を挿入した最尤モデルを $M_I(\hat{a}, \hat{b})$ と表記する。この最尤推定量は次の通りである。

$$\hat{a} = \frac{Q_{xy}}{Q_{xx}}$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

[*4] https://mathlog.info/articles/2936

[*5] 正規性の仮定の代わりに、無相関、等しい分散、平均が 0 の仮定を与えるのが簡単な定義である。ここでは、正規性を仮定しておいた

予測値と観測値の差分を残差 e_i、またその二乗和を残差平方和とよぶ。それぞれ、以下の通り。

$$e_i = y_i - (ax_i + b) \quad (i = 1, 2, \cdots, n)$$
$$RSS = \sum_{i=1}^{n} e_i^2$$

そして分散の推定値を以下のように定義する。

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

モデル $M_I(a, b)$ において、次のことも解っている。

$$\frac{(\hat{a} - a)}{\left(\frac{\hat{\sigma}^2}{Q_{xx}}\right)^{\frac{1}{2}}} \sim t_{n-2}$$

また、このことから、a の信頼区間は、次の通り。

$$|a| \leq \hat{a} \pm (\hat{\sigma}^2 / Q_{xx})^{1/2} t_{n-2, \frac{\alpha}{2}}$$

y の予測区間は、以下の通りである。

$$y_0 \pm \left[\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Q_{xx}}\right)\right]^{1/2} t_{n-2,\alpha/2}$$

9.4. 誤差モデル II

x に対する予測値との差が正規分布にしたがうことを仮定したモデルも構築する。

1. y_i は、与えられた定数
2. a, b を実数の定数
3. $v_i = \frac{1}{a}(y_i - ax_i - b)$
4. $v_i \sim N(0, \sigma^2)$

9.5. 観測点を直線により予測する

実数のペア $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ が次の線形な関係を持つとする。

$$y_i = ax_i + b + u_i, \quad (i = 1, 2, \cdots, n)$$

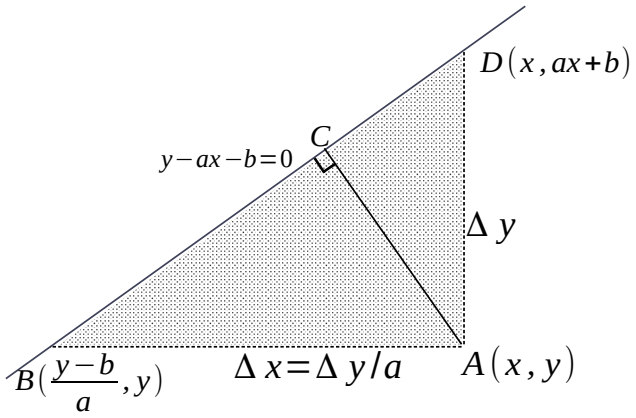


図 9.1　点 A と直線 $Y = aX + b$ の関係

点 A(x, y) と直線の関係を図 9.1 に示す。点 A と同じ x 座標の直線上の点 D は、$(x, ax + b)$ である。点 A と同じ y 座標の直線上の点 B は、$(\frac{y-b}{a}, y)$ である。DA 間の距離を $|\Delta y$、BA 間の距離を $|\Delta x|$ で表す。BA 間の距離は、$|\Delta x| = \frac{|\Delta y|}{a}$ である。これは、

$$\begin{aligned}\Delta x &= x - \frac{y-b}{a} \\ &= \frac{ax + b - y}{a} \\ &= \frac{\Delta y}{a}\end{aligned}$$

より明らかである。また、点 A から直線に向けて垂直に下した点を C とする。AC 間の距離は、$\frac{|y-ax-b|}{\sqrt{a^2+1}}$ である。また、三角形 ABD の面積の倍は $\Delta x \Delta y = \frac{\Delta x^2}{a}$ である。
以上から点と直線の距離について 4 つの方法で定義ができることがわかる。また、各点についてそれぞれの式を和にすると、以下の通りである。

1. 点と直線の最短距離 AC:$E_3 = \sum(\frac{\Delta y_i}{\sqrt{1+a^2}})^2$
2. y 軸に関する距離 AD:$E_1 = \sum \Delta y_i^2$
3. x 軸に関する距離 AB:$E_2 = \sum(\frac{\Delta y_i}{a})^2 = \sum(\Delta x)^2$
4. 面積を元にした距離 $AB \times AD$:$2 \times E_4 = \sum(\frac{\Delta y_i}{\sqrt{a}})^2$

ある点と直線への距離が離れていれば、その点への予測ができていないことを示し、近ければ、それなりによい予測をしているだろう。このことから、E_j が小ければ、それぞれ

の距離の意味で、各々の点と直線が近いはずである。そこで、まず E_j が最も小さくなるように直線のパラメータ $(a_j, b_j)(j = 1, 2, 3, 4)$ を定める。さらに、それぞれの直線の性質についてしらべる。

■点と直線の最短距離　点と直線の距離について証明を行う。大抵の高校数学の教科書には記述されているはずである。点 $A(x, y)$ から直線 $Y - aX - b = 0$ への直線距離 d の関係を求める。点 B は、点 A を x 方向に移動させたとき、直線と交わる点である。つまり点 B は、$(\frac{y-b}{a}, y)$ である。また点 D は、点 A を y 方向に移動させたとき、直線と交わる点である。つまり点 D は、$(x, ax + b)$ である。点 C は、点 A を直線 $Y - aX - b = 0$ へ垂直に下ろした点である。この AC 間の距離を d とする。直線 DA と直線 AC のなす角度を θ とする。このとき、次の関係が求められる。

$$\begin{aligned}\sin\theta &= \frac{AC}{BA} \\ &= \frac{d}{x - \frac{y-b}{a}} \\ \cos\theta &= \frac{AC}{DA} \\ &= \frac{d}{ax + b - y}\end{aligned}$$

また、$\cos^2\theta + \sin^2\theta$ を計算する。

$$\begin{aligned}\cos^2\theta + \sin^2\theta &= \frac{d^2}{(y - \frac{y-b}{a})^2} + \frac{d^2}{(ax + b - y)^2} \\ &= \frac{d^2(a^2 + 1)}{(ax + b - y)^2} \\ &= 1\end{aligned}$$

この式を d について解く。

$$d^2 = \frac{(y - ax - b)^2}{a^2 + 1}$$

■$\sum \Delta y_i^2$ に関する計算

$$E_1 = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \tag{9.1}$$

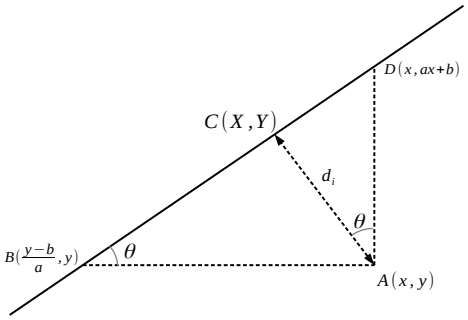


図 9.2　点 A から直線 $Y = aX + b$ への直線距離 d の関係

式 9.1 にいくらか式変形を行う。

$$\begin{aligned}\sum_{i=1}^{n}(y_i - ax_i - b)^2 &= \sum_{i=1}^{n}\{(y_i - \bar{y}) - a(x_i - \bar{x}) + (\bar{y} - b - a\bar{x})\}^2 \\ &= \sum_{i=1}^{n}\{(y_i - \bar{y}) - a(x_i - \bar{x})\}^2 + n(\bar{y} - b - a\bar{x})^2 \\ &= Q_{xx}a^2 - 2Q_{xy}a + Q_{yy} + n(\bar{y} - b - a\bar{x})^2\end{aligned}$$

ここで、以下の式を定義しておく。

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ Q_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ Q_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ Q_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\end{aligned}$$

■E_1　まず、式 9.1 を、b について偏微分を行う。

$$\frac{\partial E_1}{\partial b} = -2(\bar{y} - b - a\bar{x})$$

この式が 0 となる b について解くと、次が求まる。

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

また、E_1 を a について偏微分を行う。

$$\frac{\partial E_1}{\partial a} = 2aQ_{xx} - 2Q_{xy}$$

以上から、a が求められる。

$$\hat{a} = \frac{Q_{xy}}{Q_{xx}}$$

■E_2　a, b に関係する部分は、以下の式である。

$$E_2 = \frac{1}{a^2}(Q_{xx}a^2 - 2Q_{xy}a + Q_{yy} + n(\bar{y} - b - a\bar{x})^2)$$

E_2 の b に関する偏微分が 0 になる点は、以下の式である。

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

また、E_2 の a に関する偏微分を計算する。

$$\begin{aligned}\frac{\partial E_2}{\partial a} = &-\frac{2}{a^3}(a^2Q_{xx} - 2Q_{xy}a + Q_{yy}) \\ &+\frac{1}{a^2}(2aQ_{xx} - 2Q_{xy})\end{aligned}$$

この式を a についてとくと、最尤推定量がもとめられる。

$$\hat{a} = \frac{Q_{yy}}{Q_{xy}}$$

■E_3　変数 a, b に関連のある項を計算する。

$$h(a, b) = \sum_{i=1}^{n}(x_i - u_i)^2 + (y_i - v_i)^2$$

対数尤度を最大化するかつ $v_i - au_i - b = 0$ を満すものを求める。

これは難しいので、幾何学的な考察を行う。$h(a,b)$ は、(x_i, y_i) から (u_i, v_i) 上への距離の和を示している。これを最小化するのは、(x_i, y_i) から直線 $v_i - au_i - b = 0$ への直線距離を最小化しているのに等しい。このことから、直線から点への距離の公式から、その和は、次の式で表すことができる。

$$E_3 = \sum_{i=1}^{n} \frac{(y_i - b - ax_i)^2}{1 + a^2}$$

E_1 との違いは、分母に $(1 + a^2)$ の項が加わったことである。これが、推定量に違いを生じさせる。式 E_3 を展開していく。

$$(1 + a^2)E_3 = Q_{yy} + a^2 Q_{xx} - 2aQ_{xy} + n(\bar{y} - a\bar{x} - b)^2$$

この式を最小化する。まず b により偏微分を行う。

$$\frac{\partial E_3}{\partial b} = -2n(\bar{y} - a\bar{x} - b)$$

これが 0 になるので、最尤推定した $\hat{b}$ は次の式となる。

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

次に、a について偏微分をおこなう[*6]。

$$\frac{\partial E_3}{\partial a} = \frac{(2aQ_{xx} - 2Q_{xy})(1 + a^2) - 2a(Q_{yy} + a^2 Q_{xx} - 2aQ_{xy})}{(1 + a^2)^2}$$

分子を整理すると、次の式となる。

$$Q_{xy}a^2 - a(Q_{yy} - Q_{xx}) - Q_{xy}$$

$\frac{\partial E_3}{\partial a} = 0$ より、a について解く。上式は、a に関する二次方程式なので、a を解く。

$$\hat{a} = \frac{Q_{yy} - Q_{xx} + \sqrt{(Q_{yy} - Q_{xx})^2 + 4Q_{xy}}}{2Q_{xy}}$$

■E_4

$$\begin{aligned} E_4 = \sum_{i=1}^{n} u_i^2 &= \sum_{i=1}^{n} \frac{1}{a}(y_i - ax_i - b)^2 \\ &= \frac{1}{a}(Q_{xx}a^2 - 2Q_{xy}a + Q_{yy} + n(\bar{y} - b - a\bar{x})^2) \end{aligned}$$

[*6] $(f/g)' = \frac{f'g - fg'}{g^2}$

ここで、E_4 の b に関する偏微分が 0 となる b を求める。

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

同様に、E_4 の a に関する偏微分が 0 となる a を求める。

$$\begin{aligned}\frac{\partial E_4}{\partial b} &= \frac{1}{a}(2aQ_{xx} - 2Q_{xy}) \\ &\quad -\frac{1}{a^2}(Q_{xx}a^2 - 2Q_{xy}a + Q_{yy}) = 0 \\ \rightarrow &\quad a^2Q_{xx} - Q_{yy} = 0\end{aligned}$$

以上から、最尤推定量が求められる。

$$\hat{a} = \sqrt{\frac{Q_{yy}}{Q_{xx}}}$$

この式は、$\frac{Q_{xy}}{Q_{xx}}$ と $\frac{Q_{yy}}{Q_{xy}}$ の幾何平均と一致する。ここで、幾何平均は、0 より大きな数 $a_1, a_2, \cdots, a_n$ について、次の量のことである。

$$(a_1 a_2 \cdots a_n)^{\frac{1}{n}}$$

9.5.1. まとめ

実際のところ、E_1 の中にも a に係る項が入っているので、a がどのような値になるかは推測しにくい。気持ちとして以下のようになることが考えられる。E_2, E_3 は、分母に傾き a が入っている。この項が 1 より大きければ、E_2, E_3 を小くし、1 より小ければ、E_2, E_3 は大きくなる。このことから、E_2, E_4 において a は 1 よりも大きくなりがちであることが予想される。E_3 については、分母に $1 + a^2$ の項があるため、任意の a において、E_3 を小くしてくれそうである。

■a_1 と a_2 の大小関係

$$\begin{aligned}\frac{1}{a_2}a_1 &= \frac{Q_{xy}}{Q_{yy}}\frac{Q_{xy}}{Q_{xx}} \\ &= \frac{Q_{xy}^2}{Q_{xx}Q_{yy}}\end{aligned}$$

ここで、r^2 を以下のように定める。

$$r^2 = \frac{Q_{xy}^2}{Q_{xx}Q_{yy}}$$

この r^2 は 1 以下であることから、

$$a_1 \leq a_2$$

であることがわかる。これは、E_2 の直線の傾きは E_1 の直線の傾きよりも急であることを示唆している。

■a_1, a_2, a_4 **の関係**　相加相乗平均とは次のことである。実数 $a, b > 0$ について、次が成り立つ。

$$\frac{a+b}{2} \geq \sqrt{ab}$$

ここで、$a = a_1, b = a_2$ とおくと次がわかる。

$$\frac{a_1 + a_2}{2} \geq \sqrt{a_1 a_2} = a_4$$

このことから、a_4 は a_1, a_2 の平均値よりも小さい。

	a	b
E_1	$\frac{Q_{xy}}{Q_{xx}}$	$\bar{y} - \hat{a}\bar{x}$
$E_2 = \frac{1}{a^2} E_1$	$\frac{Q_{yy}}{Q_{xy}}$	$\bar{y} - \hat{a}\bar{x}$
$E_3 = \frac{1}{1+a^2} E_1$	$\frac{Q_{yy} - Q_{xx} + \sqrt{(Q_{yy} - Q_{xx})^2 + 4Q_{xy}^2}}{2Q_{xy}}$	$\bar{y} - \hat{a}\bar{x}$
$E_4 = \frac{1}{a} E_1$	$\sqrt{\frac{Q_{yy}}{Q_{xx}}}$	$\bar{y} - \hat{a}\bar{x}$

第 10 章

親子の身長の関係

Pearson と Lee らによる親と子供の身長に関するデータを利用した[*1]。

	子供	親
個数	179	179
平均	68.33	67.08
標準偏差	4.53	4.03
最小値	59.50	58.50
25%	64.50	64.00
50%	68.50	67.50
75%	71.50	70.50
最大値	78.50	74.50

	$\hat{a}$	$\hat{b}$
E_1	0.59	29.02
E_2	2.16	-76.69
E_3	1.25	-15.76
E_4	1.13	-7.18

[*1] https://vincentarelbundock.github.io/Rdatasets/datasets.html

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.1. 基本的な統計量

なんらかの標本のサンプルを実数の数列とする。つまり、そのサンプルを $x_1, x_2, \cdots, x_n$ とすると、それぞれの i について $x_i \in R$ である。

A.1.1. 平均分散

平均と分散を次の式で定義する。

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

A.1.2. 逐次更新

新たなサンプル x_{n+1} を得たとき、元の実数列 $x_1, x_2, \cdots, x_n$ に関する平均と $\bar{x}, \sigma_x^2$ により、平均と分散を計算できる。平均値は、

$$\bar{x}_{n+1} = \frac{1}{n+1}(n\bar{x} + x_{n+1})$$

であることがわかる。また、分散は、n までの分散を、σ_n^2 と書くことにすると、次のように式を変形できる。

$$\begin{aligned}
\sigma_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \frac{1}{n}(\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2) \\
&\rightarrow n\sigma_n^2 + n\bar{x}_n^2 = \sum_{i=1}^{n} x_i^2
\end{aligned}$$

これらを使い、x_{n+1} を得たときの σ_{n+1}^2 を計算する。

$$\begin{aligned}
\sigma_{n+1}^2 &= \frac{1}{n+1}(\sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^2 + (x_{n+1} - \bar{x}_{n+1})^2) \\
&= \frac{1}{n+1}(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}_n\bar{x}_{n+1} + n\bar{x}_{n+1}^2 + \\
&\qquad x_{n+1}^2 - 2x_{n+1}\bar{x}_{n+1} + \bar{x}_{n+1}^2) \\
&= \frac{1}{n+1}(x_{n+1}^2 + n\sigma_n^2 + n\bar{x}_n^2 - (n+1)x_{n+1}^2) \\
&= \frac{1}{n+1}(n(\sigma_n^2 + \bar{x}_n^2) + x_{n+1}^2) - x_{n+1}^2
\end{aligned}$$

A.1.3. 標本の追加による平均値分散の更新

新たな標本 $y_1, y_2, \cdots, y_m$ を得たとする。このとき、二つの標本の平均 $\bar{z}$ と分散 σ_z^2 を計算する。平均は、次の通り。

$$\bar{z} = \frac{1}{n+m}(n\bar{x} + m\bar{y})$$

分散は次の通り。

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

この式は、$x_1, x_2, \cdots, x_n$ や $y_1, y_2, \cdots, y_m$ を知らなくても、それぞれの平均と分散とサンプルサイズから、これらサンプル全体の平均、分散が計算できることを示している。

A.2. 確率変数

A.2.1. 確率変数がある分布関数に従う

確率変数は、変数だけでなくその出現頻度をあらわす変数のことである。
確率変数 x が、ある分布関数に従うとは、

定義 A.2.1. X を連続型の確率変数とする。このとき、R 上に定義された次の関数 $f_X(x)$ を X の確率密度関数または単に確率密度という。

1. $f_X(x) \leq 0$, $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
2. $a < X \leq b$ となる事象の確率 $P(a < X \leq b)$ は、

$$P(a < X \leq b) = \int_a^b f_X(x)dx. \tag{A.1}$$

$f_X(x)$ を単に $f(x)$ ともかく。
X の累積分布関数 *(*または分布関数*)* を、

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

と表す。

定義 A.2.2. $X_1, X_2, X_3, \cdots, X_n$ を確率変数とする。R の任意のボレル集合に対して、

$$P(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2)\cdots P(X_n \in A_n)$$

が成り立つならば、$X_1, X_2, \cdots, X_n$ は独立であるという。

定理 A.2.1. $X_1, X_2, \cdots, X_n$ が連続型確率変数のとき、結合密度 $f(x_1, x_2, \cdots, x_n)$ を持つとして、それぞれの周辺密度を $f_{X_i}(x_i), i = 1, 2, \cdots n$ とする。このとき、$X_1, X_2, \cdots, X_n$ が独立であることは次が成り立つことと同値である。すべての $x_1, x_2, \cdots, x_n \in R$ にたいして

$$f(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n).$$

定義 A.2.3. $X_1, X_2, \cdots$ が独立同分布 *(i.i.d)* な確率変数であるとは、次の性質を満たすときをいう。

- 任意の整数 $n > 0$ に対して、$X_1, X_2, \cdots, X_n$ は独立な確率変数である。
- 任意の整数 $i, j > 0$ に対して、X_i と X_j は同じ分布をもつ。

定義 A.2.4. 1 から累積分布関数を引いたものを、相補累積分布関数といい、

$$G(x) = 1 - F(x) = f(X > x) \tag{A.2}$$

$$= \int_x^\infty f(z)dz. \tag{A.3}$$

である。

累積分布関数は、ある値よりも小さな値を得る確率を示す。相補累積分布関数は、ある値よりも大きな値を得る確率を示す。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。

A.2.2. 平均・分散

定義 A.2.5. 連続型確率変数 X の平均値を、次で定義する。

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx$$

$E[X]$ の値を μ_X などと表す。

定義 A.2.6. 連続型確率変数 X の分散 $V[X]$ を、次で定義する。

$$V[X] = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)dx$$

A.3. 正規分布

正規分布の確率密度関数は、

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{A.4}$$

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

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) \tag{A.5}$$

であり、図 A.1(a) である。標準正規分布の累積分布関数は、

$$\Phi(x) = p(X < x; 0, 1) \tag{A.6}$$

$$= \int_{-\infty}^{x} \phi(z)dz \tag{A.7}$$

$$= \frac{1}{2}\left(1 + \mathrm{erf}\frac{\mathrm{x}-\mu}{\sqrt{2\sigma^2}}\right) \tag{A.8}$$

であり、図 A.1(b) である。

相補累積分布関数は、

$$1 - \Phi(x) = p(X > x; 0, 1) \tag{A.9}$$

$$= \int_{x}^{\infty} \phi(z)dz. \tag{A.10}$$

A.3.1. 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。

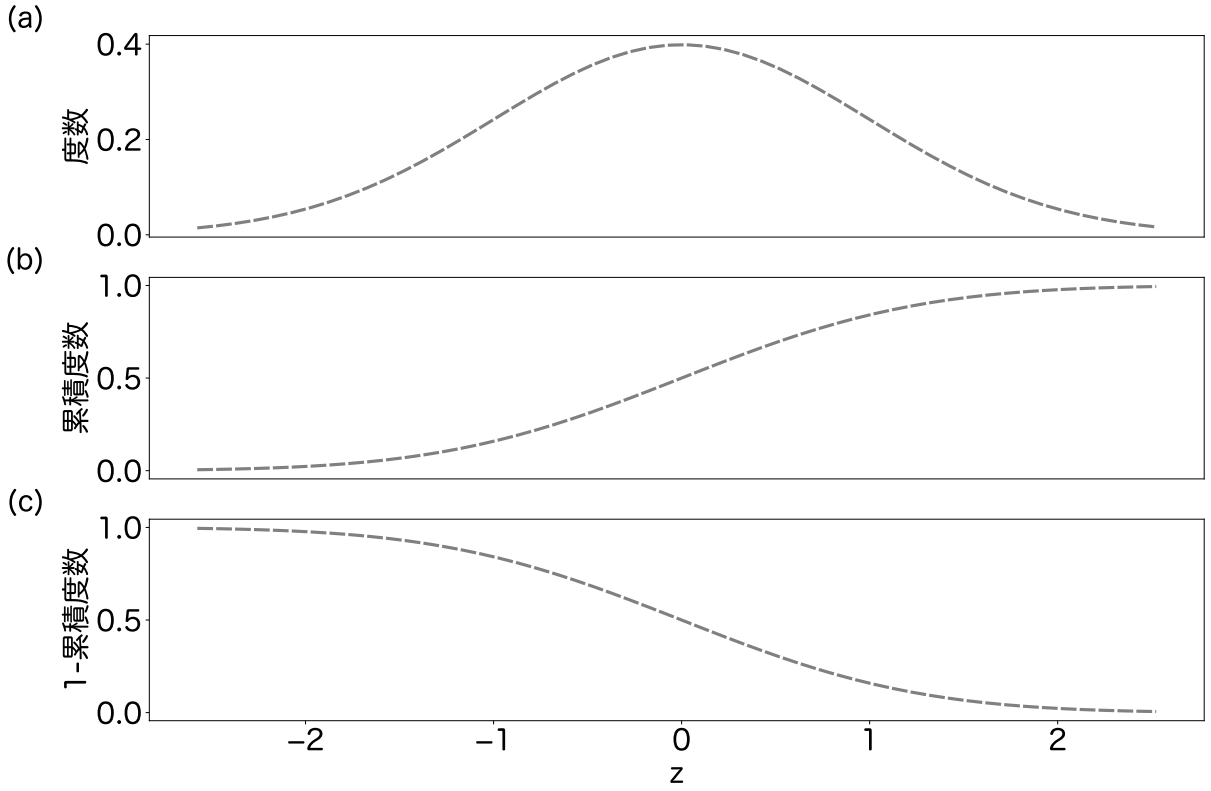


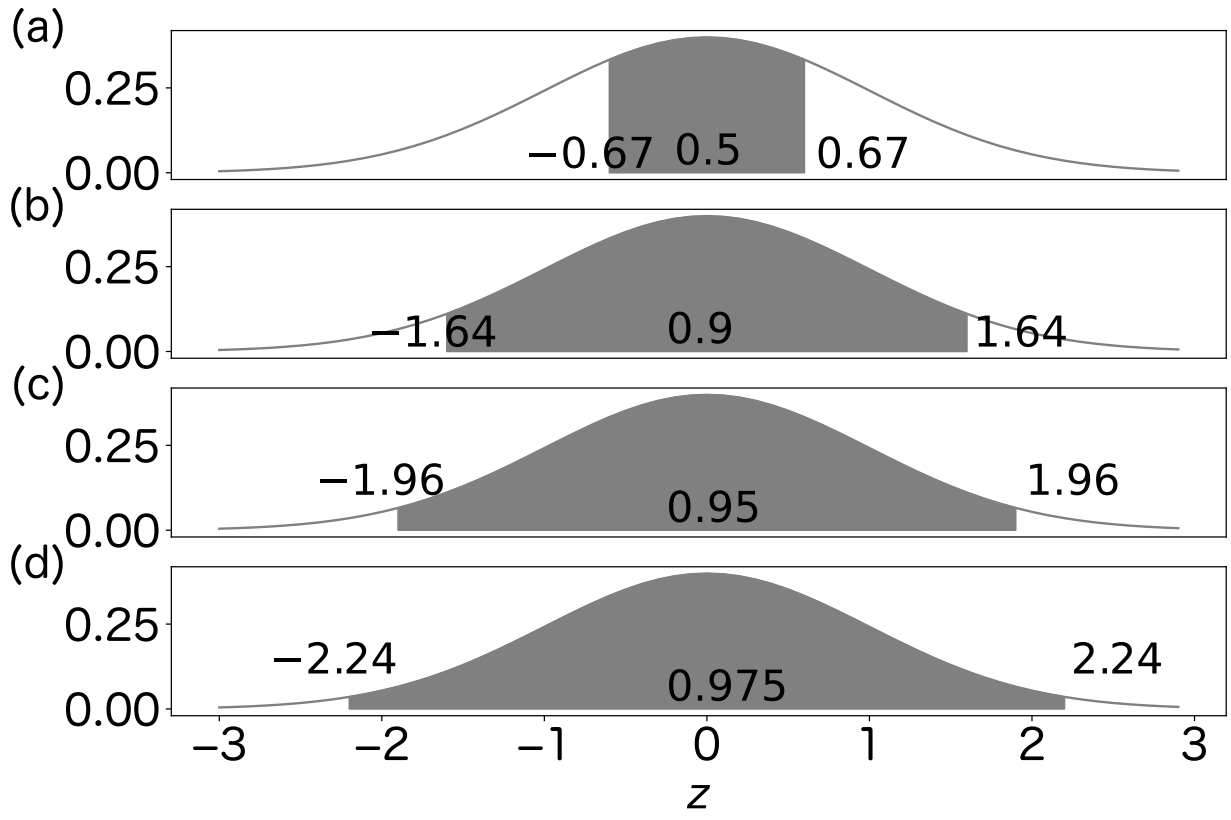
図 A.1　標準正規分布 (a) 確率密度関数 (b) 累積度数分布 (c)1-累積度数分布

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。

$$\int_{-a}^{a} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) dz = 0.95 \tag{A.11}$$

```
1 b,a = norm.interval(0.95,0,1) # 積分値がになる範囲を計算0.95
2 print(norm.cdf(b, loc=0, scale=1)-norm.cdf(a, loc=0, scale=1)
      ) # になるかを確認
      0.95
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、

$$\int_{-\infty}^{a_1} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})dx = \frac{\alpha}{2}. \tag{A.12}$$

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.3.1. 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.3.2. $\alpha(0 \leq \alpha \leq 1)$ 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.3.2. 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。

$$P(|x| > z) = 1 - \Phi(|x|) + \Phi(-|x|) \tag{A.13}$$

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。

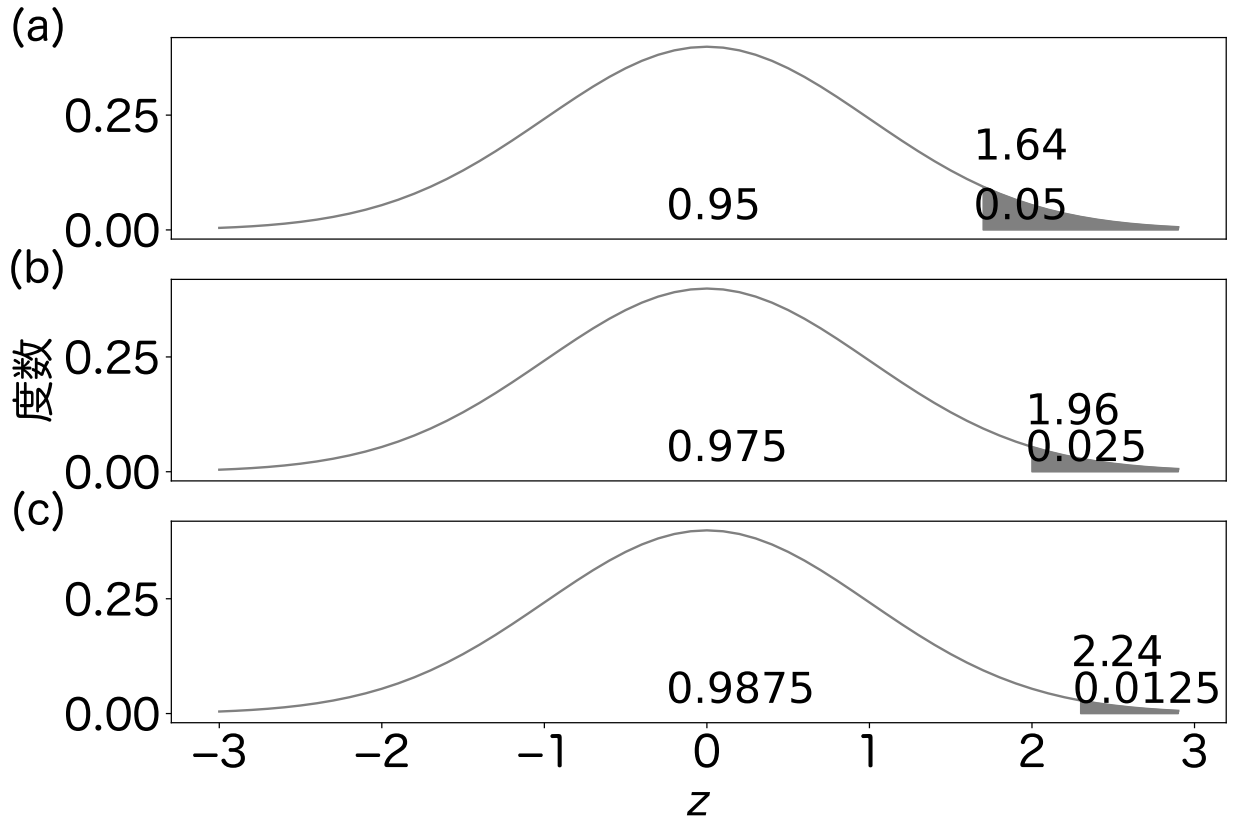


図 A.2　標準正規分布におけるより大きな値 (より偏った値) を取る確率。(a)$z = 1.64$ より大きな値を取る確率は 0.05。(b)$z = 1.96$ より大きな値を取る確率は 0.025。(c)$z = 2.24$ よりも大きな値を取る確率は 0.0125

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.3.3.　$N(0,1)$ 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 $z = 1.64$ 以上の値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)
```

A.3.4. $N(1.96,1)$ 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.3.5. 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.3.5 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。

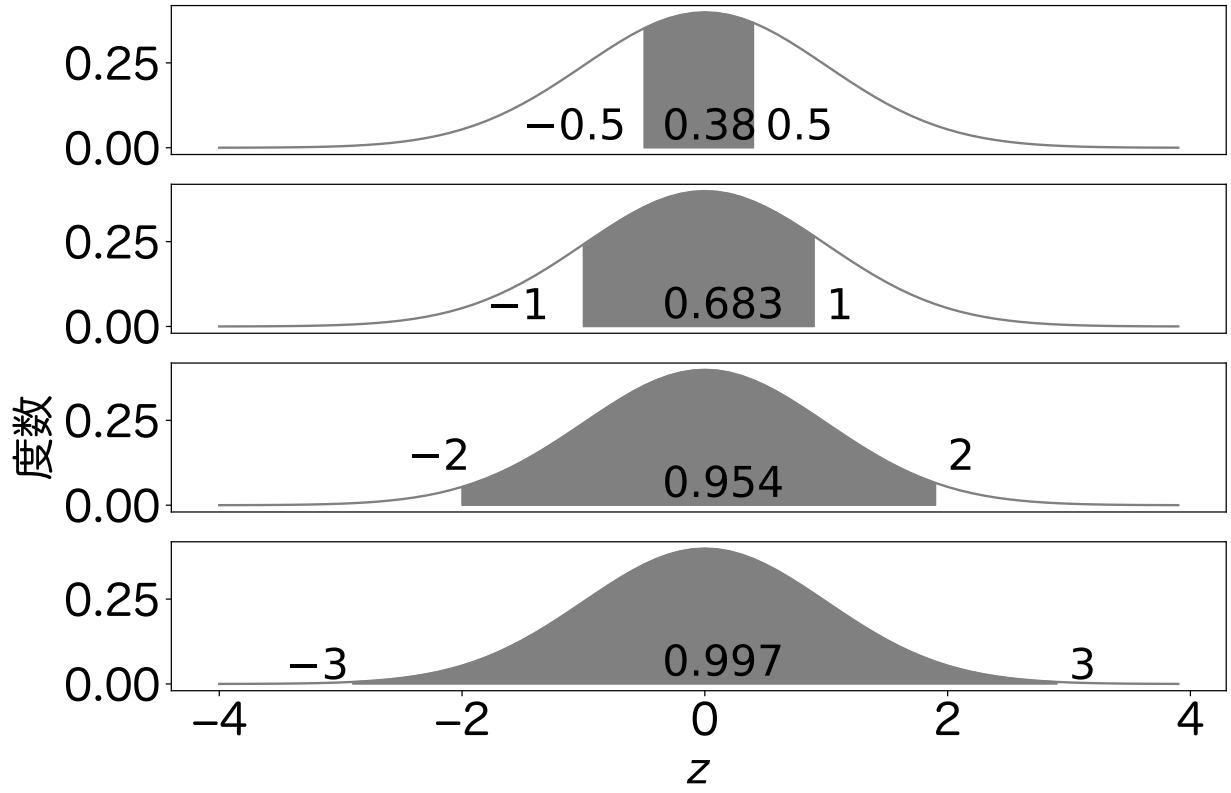
表 A.1　σ を基準にした値の出やすさ

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.4. 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、

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



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。

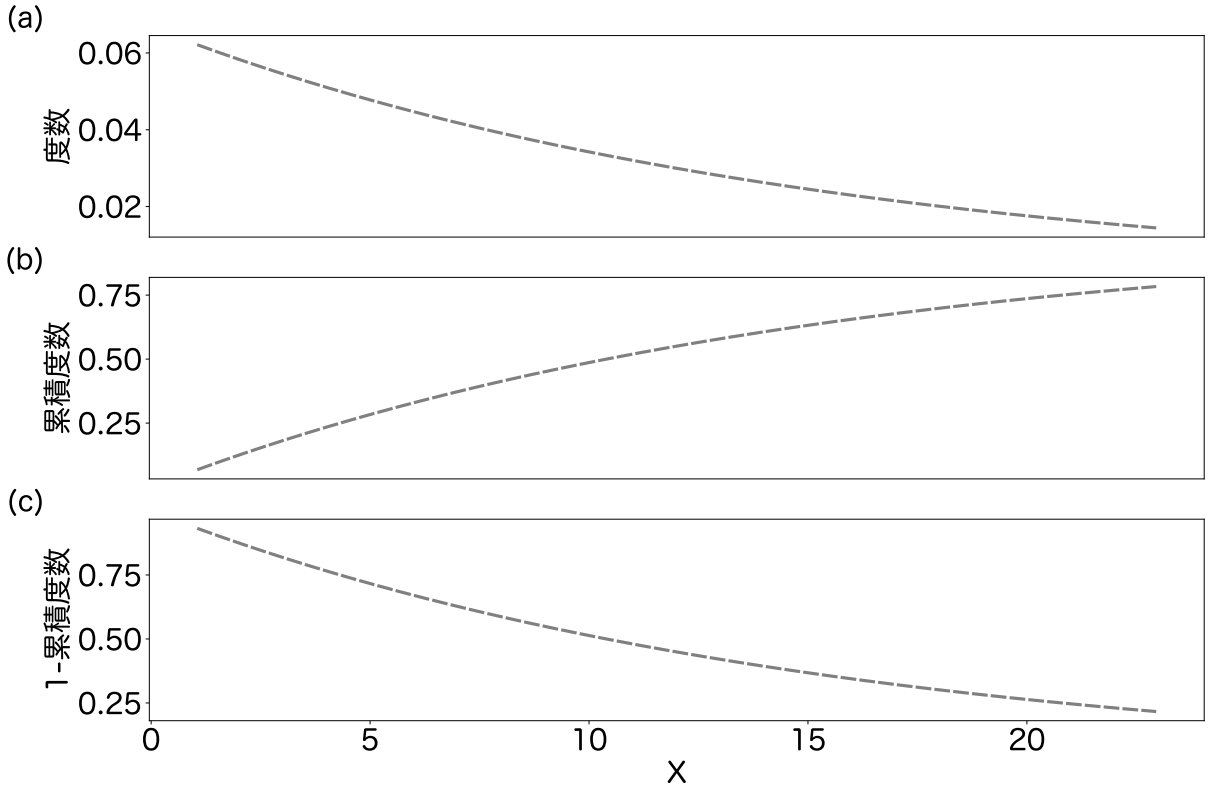


図 A.3　指数分布 $\lambda = 1/15$(a) 確率密度関数 (b) 累積度数分布 (c) 相補累積度数分布

A.4.1.　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、

$$\left[\frac{1}{\lambda} \log \frac{1}{1 - \alpha/2}, -\frac{1}{\lambda} \log \frac{\alpha}{2} \right] \tag{A.14}$$

である。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この区間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図 A.4 に示した。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.4.2. 指数分布に従う確率変数の予測区間

確率変数の出現しやすい区間は、一般には予測区間として定義される。そして、予測区間は、ある割合 $1-\alpha$ で確率変数が出現する最小の区間であるとする。ここでは、A.14 が最小ではないことを示す。具体的には、$1-\alpha$ の割合で確率変数が出現する区間を導出し、この区間が式 A.14 より小さいことを示す。$[0, a]$ をその区間とし、

$$1 - \exp(-\lambda a) = 1 - \alpha$$

を解く。するとこの区間は、

$$\left[0, \frac{1}{\lambda} \log \frac{1}{\alpha} \right]$$

である。区間 A.14 と、区間 A.4.2 の大小関係を比較する。区間の大きさの差を計算する[*1]。

$$\lambda \left((\frac{1}{\lambda} \log \frac{2}{\lambda} - \frac{1}{\lambda} \log \frac{1}{1 - \frac{\alpha}{2}}) - \frac{1}{\lambda} \log \frac{1}{\alpha} \right)$$
$$= -\log \frac{\alpha}{2} + \log \left(1 - \frac{\alpha}{2} \right) + \log \alpha$$
$$= \log 2 \left(1 - \frac{\alpha}{2} \right) = \log(2 - \alpha)$$

[*1] 対数の計算 $\log A + \log B = \log AB, \log \frac{A}{B} = \log A - \log B$,$\log \frac{1}{A} = -\log A$

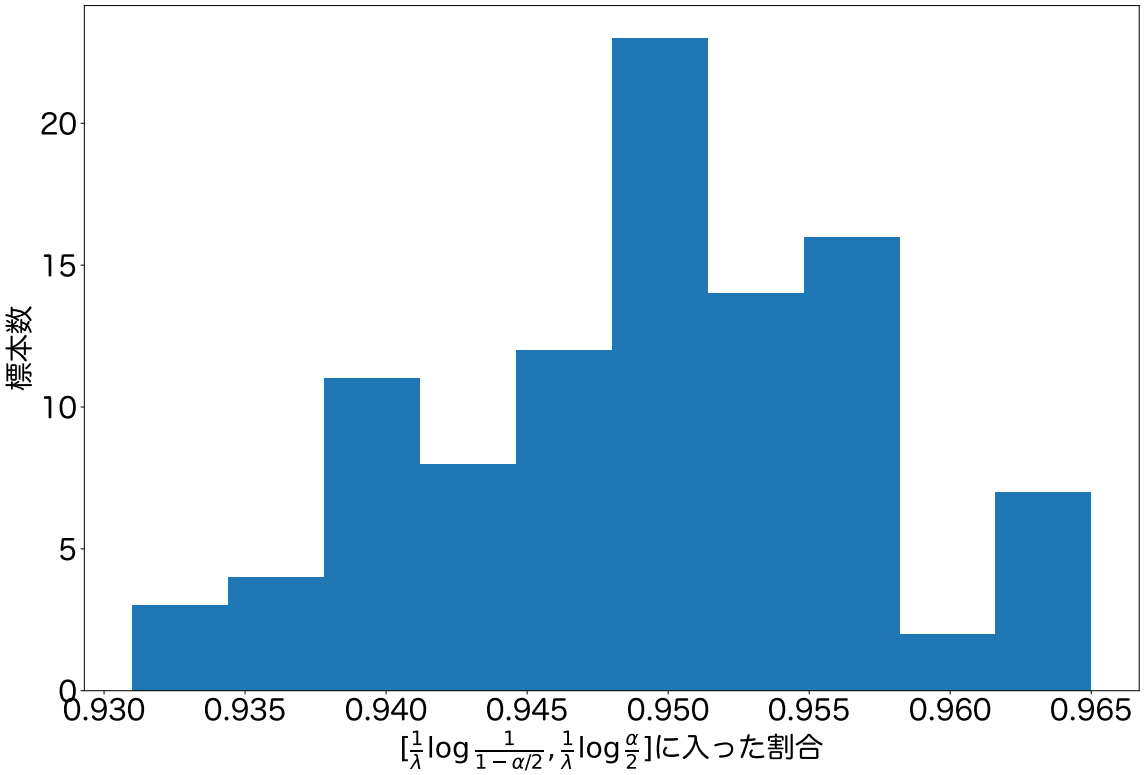


図 A.4　指数分布 $\lambda = 1/10$ からサンプルサイズ 1000 の標本を 100 回シミュレーションし、各標本においてデータが区間 $[\frac{1}{\lambda}\log\frac{1}{1-\alpha/2}, -\frac{1}{\lambda}\log\frac{\alpha}{2}]$ に入った割合を計算した。そのヒストグラム。

また、最後の式は、常に正なので、区間 A.14 の方が大きいことがわかる。よって、区間 A.4.2 のほうが区間 A.14 より小さい区間である[*2][*3]。

[*2] 問題:最小の区間であろうか？

[*3] どっちをつかってもそんなに差がでないので、どっちでを使ってもいい。おおまかに入っていることを確認するためだけにつかう。数学的なことにはこれ以上考えない。

A.5. 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[*4]。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[*5]。この関数も左右非対称である。

A.5.1. 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。

表 A.2　$\alpha = 0.05$

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

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

[*5] $\gamma(a, x) = \int_0^x t^{a-1} \exp^{-t} dt$ 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。

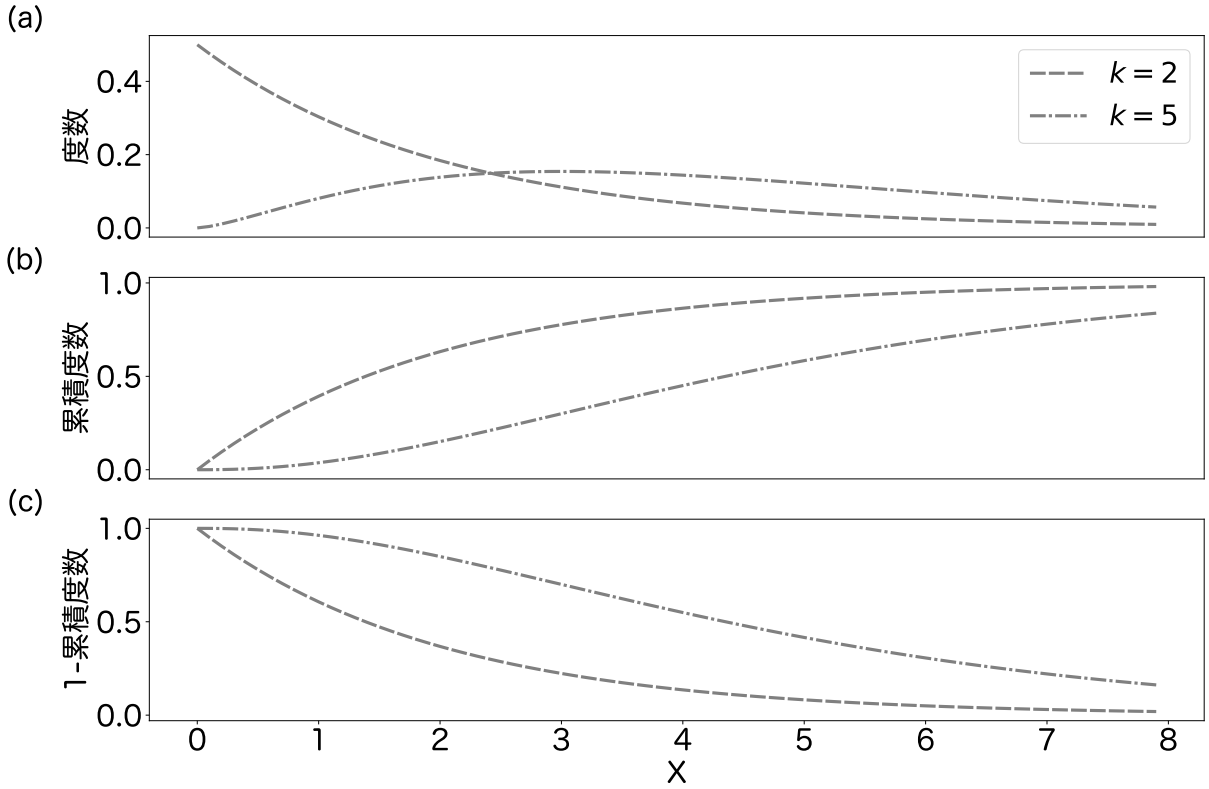


図 A.5　カイ二乗分布

A.6.　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。

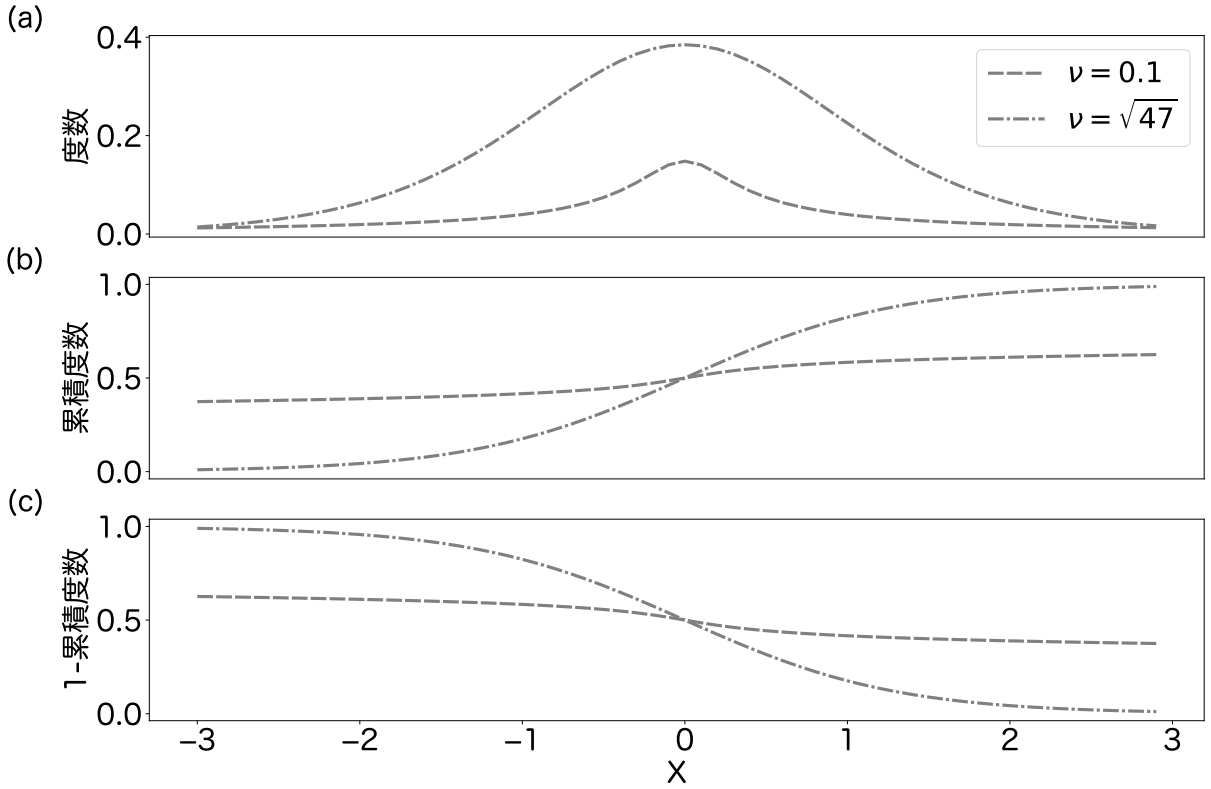


図 A.6　t 分布

A.6.1.　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.7. 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.7.1. 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、

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

がわかる。以上をまとめておく。

定理 A.7.1. $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ 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、$\mu - z_{0.025}\sqrt{\frac{\sigma^2}{n}} < \bar{x} < \mu + z_{0.025}\sqrt{\frac{\sigma^2}{n}}$ である。

定理 A.7.2. $X_1, X_2, \cdots, X_{n_1} \sim N(\mu_1, \sigma_1^2), Y_1, Y_2, \cdots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ 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.7.2. 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.7.3. $X_1, X_2, \cdots, X_n \sim Ga(1, \frac{1}{\lambda})$ 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

$$M_{n\bar{X}} = M_{X_1+X_2+\cdots+X_n} = M_{X_1} M_{X_2} \cdots M_{X_n} \tag{A.15}$$

$$= (1 - \frac{1}{\lambda}t)^{-1}(1 - \frac{1}{\lambda}t)^{-1} \cdots (1 - \frac{1}{\lambda}t)^{-1} \tag{A.16}$$

$$= (1 - \frac{1}{\lambda}t)^{-n} \tag{A.17}$$

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

A.8. 尤度・対数尤度・AIC

定義 A.8.1. 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[*6] [*7]。

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

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。

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

[*6] 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.

[*7] 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。

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。

$$l(\theta) = \sum_{i=0}^{100} \log f(x_i|\theta) \tag{A.18}$$

$$= \sum_{i=0}^{100} \log \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i-\theta)^2}{2}\right) \tag{A.19}$$

$$= -\frac{100}{2}\log(2\pi) + \sum_{i=0}^{100} \frac{(x_i-\theta)^2}{2} \tag{A.20}$$

この式より、2 次関数であることは明らかである。

A.8.1. 最尤推定

定義 A.8.2. 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する。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。

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

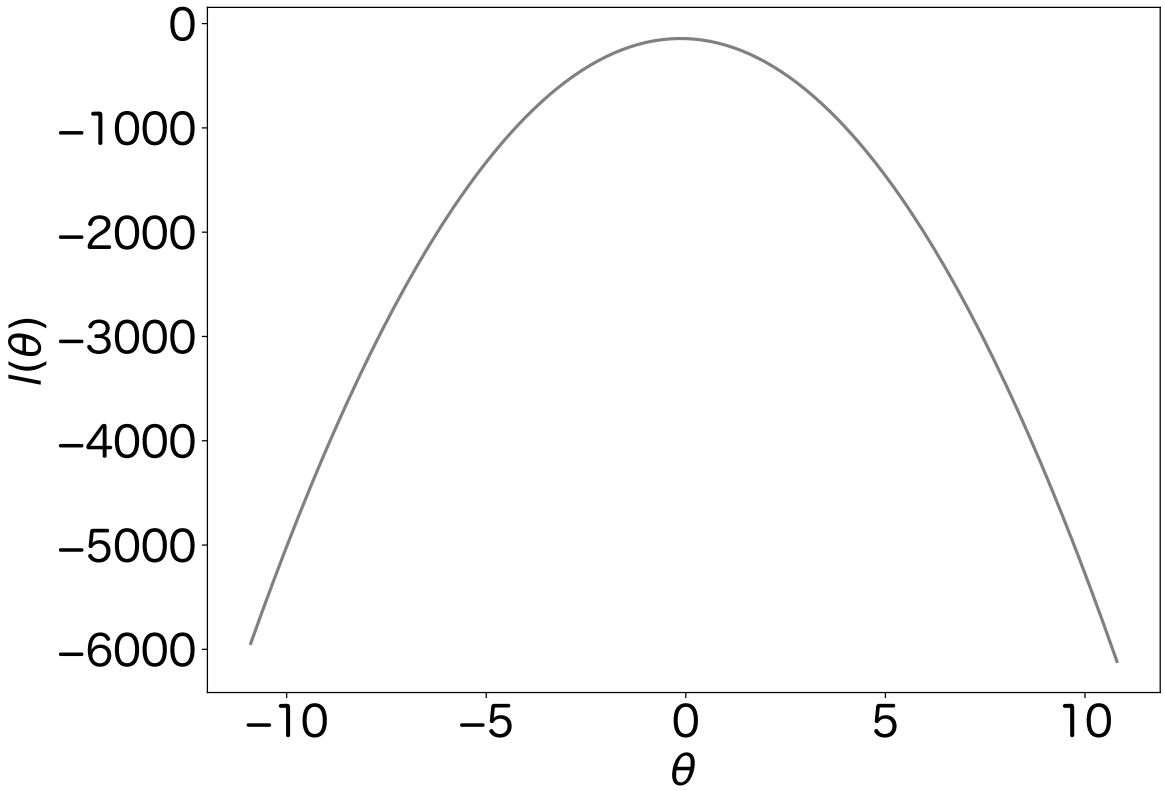


図 A.7　$N(\theta, 1)$ における対数尤度関数。確率変数は、$N(0, 1)$ からサンプリングした。

$\frac{\partial l(\mu,\sigma^2)}{\partial \mu} = 0$ 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[*8]。

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。

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

[*8] 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

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。

最尤推定したモデルだけど、予測には使えないモデル

最尤推定した母数を使った最尤モデルだけど、予測には使えないモデルを紹介する。データがあるその平均値の周りに非対称に分布している場合を考える。このデータに対して、正規モデルを採用し、その母数を最尤推定したモデルを構築する。このモデルは、明らかに予測に利用しにくい。

最尤モデルとして、その母数などが論文に記述されていたとする。この場合、著者らがデータの性質がモデルにより説明できることを示したと思わせられる。実際のところ検証したのかは、読者にはわからないことがある。単に最尤モデルが記述されていたとしても、そのモデルが良かったとは考えないほうが良さそうである。

例えば、平均と分散などの統計的な要約量がかかれていて、その後、正規モデルを利用してその性質を調べているということがある。実際のデータを調べると、正規モデルの特徴をデータがもっていないこともありえる。

A.8.2. 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.9. 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。

A.9.1. 尤度比

フルモデルモデル M_2 に対しその子モデルを M_1 とする。M_2 の母数を (θ_1,θ_2) ただし、$\theta_1 \in R^n, \theta_2 \in R^k$ とし、M_1 の母数を θ_1 とする。また、D を M_1 のサンプルサイズ d の標本とする。

定義 A.9.1. 母数の個数が異なる統計モデル間の最尤推定する前の尤度比を次のように定義する[*9]。

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

式変形を行えば[*10]、

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

となる。自由度の高いモデル M_2 とそれよりも自由度の低いモデル M_1 の対数尤度の差の 2 倍である。
また、M_2 における母数 θ_2 を最尤推定したときのモデルを $M_2^{ML}(D)$ とする。このとき、次を尤度比と呼ぶ。

$$Dev(D,M_1,M_2^{ML}(D))$$

このとき、次のことが解っている[*11]。

定理 A.9.1.

$$Dev(D,M_1,M_2^{ML}(D)) \sim \chi_n^2$$

分母のモデルにおけるデータ由来の母数の個数から $(n+k)$、分子のモデルにおけるデータ由来の母数の個数 (k) を引いたものが自由度 (n)。

式が複雑であるが、4 つのステップで説明できる。

1. 子モデル M_1 の標本を生成。

[*9] 厳密な尤度比の定義は数理統計学の教科書を参照せよ。
[*10] $\log\frac{a}{b} = \log a - \log b$
[*11] ただしいくらかの条件がある。詳しくは数理統計学のテキストを参照するべき

2. それ以上に高い自由度を持つ対するフルモデル M_2 で、標本を予測しようと最尤 M_2 モデルを構築する。
3. M_1 における最尤モデルも構築する
4. 最尤モデルにおける対数尤度の差が、χ^2_n に従う (なんどもこの操作をすると)。

A.9.2. データから推定したモデルの尤度比

定理 A.9.1 より直ちに次がわかる。あるデータ D' に対して、なんらかの推定方法で母数を推定したモデル $\hat{M}_1$ について、

補題 A.9.1.

$$Dev(D, \hat{M}_1, \hat{\hat{M}}_2) = -2\log \frac{\hat{M}_1 \text{における } D \text{ の尤度}}{\hat{\hat{M}}_2 \text{における } D \text{ の尤度}} \sim \chi^2_k$$

が成り立つ。ここで D は、$\hat{M}_1$ の標本であり、$\hat{\hat{M}}_2$ は、標本 D を使って最尤推定を行なったモデル、$\hat{M}_1$ における最尤推定した母数の個数を $n(=0)$ とすると、$\hat{\hat{M}}_2$ における最尤推定した母数の個数を $n+k(=k)$ とする。

A.9.3. 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。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。

A.9.4. 中心極限定理

定理 A.9.2 (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。

付録 B

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。

B.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。ただし、$\mu_2 \geq \mu_1$

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 とする。

B.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、$\frac{\sigma_1}{\sigma_2} \sim 1$ の場合、次に $\frac{\sigma_1}{\sigma_2} \gg 1$ それぞれにおける信頼区間および検出力を考える。

B.2.1. 信頼区間

統計モデル $M(\mu_1, \mu_2; \sigma^2, \sigma^2)$ により、$x_1, x_2, \cdots, x_n, i.i.d. \sim N(\mu, \sigma^2), y_1, y_2, \cdots, y_m, i.i.d. \sim N(\mu_2, \sigma^2)$ 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。有意水準 α を $\alpha = 0.05$ 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、

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

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 = 30
3 sigma = 5
4 m = 20
5 n=10
6
7 norm1 = norm(mu1, sigma)
8 norm2 = norm(mu2, sigma)
9 N=10**5
```

```
sample1 = norm1.rvs(size=(m,N))
sample2 = norm2.rvs(size=(n,N))

xbar1 = np.mean(sample1, axis=0)
xbar2 = np.mean(sample2, axis=0)

U = np.sqrt(m*n)/(sigma*np.sqrt(m+n))
Z = ((xbar1-mu1)-(xbar2-mu2))*U

print(1-np.sum(np.abs(Z)<1.96)/N)
```

およそ 95% 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。

B.2.2. 検出力

$M(\mu_1, \mu_2)$ における統計量 Z が、もう一つの統計モデル $M(\mu, \mu; \sigma^2, \sigma^2)$ 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、次が成り立つ。

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

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、

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

である。$N(\frac{\mu_2-\mu_1}{U}, 1)$ の 95% 予測区間の端をそれぞれ A, B とする。A, B は次の式で求められる。

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

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。

数値計算 1

正規 2 モデル $M(20,10;5)$ とモデル $M(10,10;5)$ に統計量 Z を基にしてそれらそれらモデル間の距離は (検出力) は、次のように計算できる。

```
1  mu1 = 20
2  mu2 = 10
3  sigma = 5
4  m = 20
5  n=10
6
7  U = (sigma*np.sqrt(m+n))/np.sqrt(m*n)
8
9  A=-1.96+(mu1-mu2)/U
10 B=1.96+(mu1-mu2)/U
11 1-(norm.cdf(B)-norm.cdf(A))
```

プログラムの出力は $1-\beta = 0.9993$ であり、これらモデルは Z 基準でかなり異るように考えられる。

数値計算 2

平均母数が同一のモデル $M(\mu,\mu;5)$ から次の条件で標本を生成する。サンプルサイズ 20 と 10 の標本を 10^5 個生成する。この標本から統計量 Z を計算する。次に、平均母数が異る正規 2 モデル $M(\mu+10,\mu;5)$ から同一条件で標本を生成し、統計量 Z を計算する。

```
1   mu1 = 10
2  mu2 = 10
3  sigma = 5
4  m = 20
5  n=10
6  N=10**5
7
8  norm1 = norm(mu1,sigma)
9  norm2 = norm(mu2,sigma)
10 sample1 = norm1.rvs(size=(m,N))
```

```
sample2 = norm2.rvs(size=(n,N))

xbar1 = np.mean(sample1,axis=0)
xbar2 = np.mean(sample2,axis=0)

U = (sigma*np.sqrt(m+n))/np.sqrt(m*n)
Z0 = ((xbar1-xbar2))/U
```

```
mu1 = 20
mu2 = 10
sigma = 5
m = 20
n=10

norm1 = norm(mu1,sigma)
norm2 = norm(mu2,sigma)
N=10**5
sample1 = norm1.rvs(size=(m,N))
sample2 = norm2.rvs(size=(n,N))

xbar1 = np.mean(sample1,axis=0)
xbar2 = np.mean(sample2,axis=0)

U = (sigma*np.sqrt(m+n))/np.sqrt(m*n)

Z1 = ((xbar1-xbar2))/U
```

$Z1$ が同一母数平均モデルでの信頼区間において出現する確率を計算する。

```
A,B=np.quantile(a=Z0, q=[0.025,0.975])
A1,B1=np.quantile(a=Z1, q=[0.025,0.975])
print(1-np.sum((Z1<A) | (Z1>B))/N)
```

最後の出力は、0.99929 であり、前節での計算結果とおおよそ一致する。

B.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 $\frac{(\bar{x}-\mu_1)-(\bar{y}-\mu_2)}{U} \sim N(\mu_2 - \mu_1, 1)$ であることから、

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

同様に、

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

よって、

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

である。

B.3. 母分散の事前知識がないときの統計モデル

σ について具体的な知識がない状況を想定し、正規モデル $M(\mu_1, \mu_2, \sigma^2, \sigma^2)$ について考える。

B.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、

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

である。以上より、信頼区間は、

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

である。

B.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。

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

$\lambda = \sqrt{\frac{nm}{n+m}}\Delta$、$\Delta = \frac{\mu_2 - \mu_1}{\sigma}$ 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 [19][*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。

[*1] 私は証明を読んでいない。いつか読む。

[19] より、例題を解いてみる。$\alpha = 0.05, n = 10, m = 8, \mu_1 = 5.6, \mu_2 = 5.0, \sigma = 1.0, n+m-2 = 16, \lambda = \sqrt{n \times m/(n+m)} \times \frac{\mu_2 - \mu_1}{\sigma} = 1.265$ 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。モデル A($M(5.6, 5.6; \sigma^2 = 1.0^2)$) からサンプルサイズ $N = 10^4$ の標本を生成し、統計量 $t0$ を計算する。その 95% 信頼区間 $[A, B]$ を求める。モデル B($M(5.6, 5.0; \sigma^2 = 1.0^2)$) からサンプルサイズ N の標本を生成し、統計量 $t0$ を計算する。$t0$ 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。

```
n=10
m=8
mu1=5.6
mu2=5.0
sigma = 1.0
phi = n+m-2
N = 10000

sample1 = norm(mu1,sigma).rvs(size = (n,N))
sample2 = norm(mu1,sigma).rvs(size = (m,N))
xbar = np.average(sample1,axis=0)
ybar = np.average(sample2,axis=0)
S2 = (np.sum((sample1-xbar)**2,axis=0)+np.sum((sample2-ybar)
    **2,axis=0))/float(n+m-2)
S = np.sqrt(S2)
t0 = (xbar-ybar)/(S*np.sqrt(1/n+1/m))
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。

B.4. 自己否定の誤推定

設定した有意水準 α で検定が出来無い例をいくつか挙げる。但し、議論は正規モデルの場合と同じであるので、多重検定についてのみ説明を行う。

B.4.1. 検定を繰返し使おう

正規 2 モデルの標本 $x = (x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_m), z = (z_1, z_2, \cdots, z_n, w_1, w_2, \cdots, w_m)$ についてこれらをペアにする。これらを検定統計量を用いて、それが信頼区間に含まれるかを判別する。信頼区間に含まれる頻度は $1 - \alpha$ 程度であるだろうか。
このように、想定した有意水準 α での検定ができていないことがわかる。

付録 C

仮説検定の実践

実際に利用されている仮説検定について説明する。すでに p 値の利用方法について否定的に批判している論文が多数でているので、何か付け加えることはなにもない。

C.1. 仮説検定における手順

仮説検定とは、仮説を採用するかを決定する方法である[*1]。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]。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[*4]。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[*5]。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

[*1] すでに示したように、p 値がある値を超たかどうかのみによって科学的な結論や政策の決定をおこなうべきではない
[*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とされることがある
[*4] 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
[*5] 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

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、$p > \alpha$ であれば、帰無仮説を採用する[*6][*7][*8]。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[*9][*10][*11]。

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。
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)。

[*6] 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。

[*7] 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。

[*8] 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。

[*9] 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。

[*10] 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。

[*11] 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

付録 D

検定とモデル

様々な検定が利用されており、それらにおいては何んらかのモデルから導出される統計量に関する性質が利用されている。ここでは、検定と元々のモデルとの対応関係についてまとめておく。検定とモデルとで同じ統計的性質を使っているので、まったく同じことを記述することになる。検定では元のモデルが指定されていないので、何を検証したのかがわからない。元のモデルが何かを意識できるようにした。

D.1. 正規モデル

正規モデル $M_N(\mu, \sigma^2)$ について、次が成り立つ。

■母分散 σ が不明な場合　$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ とする。統計量 T を、

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

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。

■母平均 μ が既知の場合

定理 D.1.1. $x_1, x_2, \cdots, x_n, i.i.d. \sim N(\mu, \sigma^2)$ について、次が成り立つ。

$$(n-1)\left(\frac{S_x}{\sigma}\right)^2 \sim \chi_n^2$$

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

■母平均 μ が不明な場合　分散の変化によって、標本がはずれていることを示す[*1]。

定理 D.1.2. $x_1, x_2, \cdots, x_n, i.i.d. \sim N(\mu, \sigma^2)$ について、次が成り立つ。

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

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。

D.2. 正規 2 モデル

■検定統計量 T について　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。

D.3. 独立性の検定

D.3.1. 検定

k 種類の事象 $A_1, A_2, \cdots, A_k$ はそれぞれおよそ $p_1, p_2, \cdots, p_k$ の割合で出現するとする。このとき、各事象を観測した回数は $n_1, n_2, \cdots, n_k$ であった。観測結果が理論から得られ

[*1] 分散の検定

る期待回数と適合することを検討したい。そこで、帰無仮説 H_0「期待回数と観測回数は等しい」を検定したい。

表 D.1　2 項検定

事象	A_1	A_1	$\cdots$	A_k	計
観測回数	n_1/m	n_2/m	$\cdots$	n_k/m	$m = \sum_{i=0}^{k} n_i$
期待回数	Np_1	Np_2	$\cdots$	Np_k	N

H_0 のもと、次の統計量を考える。

$$\sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

ここで、事象 A_i が表れる回数を確率変数 X_i である[*2]。この統計量は、$Np_i \geq 5(i = 1, 2, \cdots, k)$ であるとき自由度 $k-1$ の χ^2 分布にしたがう。
このことから、データとの乖離をしらべるには、まず次の統計量を計算する。

$$F = \sum_{i=1}^{k} \frac{(n_1/m - Np_i)^2}{Np_i}$$

そして、$F \geq \chi^2_{k-1}(\alpha)$ ならば H_0 を棄却する。
さて、これはどのようなモデルから導出された統計量とデータを比較しているのだろうか。確率変数 X_i はどのような分布に従っているのかわかるだろうか[*3]。

D.3.2. モデル

モデルを構築する。

1. $Z_i \sim B(n, p_i)(i = 0, 1, \cdots, k)$ ここで、$B(n, p_i)$ は二項分布。

このモデルを M_B とする。このとき、次の統計量 X_2 を定義する。

$$X_2 = \sum_{i=0}^{k} \frac{Z_i - np_i}{np_i}$$

[*2] この確率変数がなにに従うのかの記述がないことが多いが、なくてもいいものなのだろうか
[*3] 以上のような記述では、分布系がわからない。

X_2 は、$n \to \infty$ のとき、自由度 $(k-1)$ の χ^2 分布に従う。ただし、$n \to \infty$ ではなく $np_igeq5(i=0,1,\cdots,k))$ で十分である [20]。適合度検定をおこなうということは、モデル M_B から演繹的に導出される X_2 の性質とデータとを比較するということである。p 値が小ければ M_B で推測することは止めておいたほうがよさそうかもしれないと考え、このモデルとデータをさらに詳しく調べる必要がある。
このモデル M_B において、尤度比に関する統計量をつかうこともできる。

D.4. 独立性の検定

ある事象 A, B はそれぞれ k, c 個に分類され、その事象を $A_1, A_2, \cdots, A_k$ および $B_1, B_2, \cdots, B_c$ とする。このとき、事象の組 (A_i, B_j) が、得られた回数を $n_{i,j}$ とする。例えば、n 人の成人を無作意に選び、その人のパートナーの有無を A とし、パートナーがいるを A_1 いないを A_2 とする。また、対象者の年収を B とし、100 万から 200 万を B_1、200 万から 300 万を B_2 などとする。パートナーがいて、年収が 100 万から 200 万の人が 100 人いれば、$n_{1,1} = 100$ である。また、パートナーがいて、年収が 200 万から 300 万の人が 1000 人いれば、$n_{1,1} = 1000$ である。
これを表にまとめると、次の様になる。ここで、$n_j^* = \sum_{i=1}^{k} n_{j,i} (j = 1, 2, \cdots, k), n_*^j =$

表 D.2　2 項検定

事象	A_1	A_2	$\cdots$	A_k	計
B_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,k}$	n_1^*
B_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,k}$	n_2^*
B_c	$n_{c,1}$	$n_{c,2}$	$\cdots$	$n_{c,k}$	n_c^*
	n_*^1	n_*^2	$\cdots$	n_*^k	n

$\sum_{i=1}^{c} n_{i,j} (j = 1, 2, \cdots, c), n = \sum_{j=1}^{k} n_j^* + \sum_{j=1}^{c} n_*^j$ である。

D.4.1. 検定

D.4.2. モデル

モデル M_B を構築する。

1. $X_i \sim B(n, p_i)$

2. $Y_i \sim B(n, q_i)$
3. ($X_{i,j} \sim B(n, p_i q_j)(i = 1, 2, \cdots, k, j = 1, 2, \cdots, c)$))

以上の仮定は確率変数はそれぞれ独立した 2 項分布に従うことを意味する。
このモデルにおいて、次の検定統計量 $Q(X)$ に関する性質を得る。

$$Q(X) = \sum^{k}\sum^{c} \frac{X_{i,j} - \frac{p_i q_j}{n}}{\frac{p_i q_j}{n}}$$

これは、漸近的に自由度 $(k-1)(c-1)$ の χ^2 分布に従うことが知られる。
上記のデータであれば、次の Q を計算する。

$$Q = \sum_{i}^{k}\sum_{j}^{c} \frac{n_{i,j} - n\frac{n_i^* n_*^j}{n^2}}{n\frac{n_i^* n_*^j}{n^2}}$$

D.4.3. 計算例

実際の例をみる。

表 D.3　2 項検定

	治った患者	治らなかった患者	合計
新薬を投与された患者	45	15	60
偽薬を投与された患者	20	20	40
合計	65	35	100

次のモデル M_B を構築する。

1. $X_{i,j} \sim B(n, p_i q_j)(i = 1, 2, j = 1, 2)$
2. 母数 p_i, q_j は不明

このデータとの乖離を検定統計量 Q を計算することで確かめてみる。

$$\begin{aligned} Q = &\frac{(45 - 60 * 20/100)^2}{60 * 20/100} + \frac{(15 - 60 * 35/100)^2}{60 * 35/100} + \\ &\frac{(20 - 40 * 65/100)^2}{40 * 65/100} + \frac{(20 - 40 * 35/100)^2}{40 * 35/100} \\ = &6.5934 \end{aligned}$$

これは、自由度 1 の $\chi^2(0.05)$ は、3.84 である。$\chi^2(0.01) = 6.64$ 程度なので、モデル M_B では珍しい値である気分になってくる。それぞれが独立にきまるモデル、モデル M_B による推測はやめておいたほうがよさそうである。この解析だけでは他の良いモデルがあることは示されていない[*4]。

D.5. 指数分布を含むモデル

定理 D.5.1. $x_1, x_2, \cdots, x_n, i.i.d. \sim \mathrm{Exp}(\lambda)$ とする。このとき $x_1 + x_2 + \cdots + x_n \sim Ga(n, \lambda)$ である。

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

定理 D.5.2. n を自然数とする。$G \sim Ga(\frac{n}{2}, \beta), Y_n \sim \chi^2_n$ 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、

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

を満たし、λ の上限は、

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

[*4] 別のモデルが良いとは言い切れない

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、

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

D.6. 指数 2 モデル

指数分布を含んだモデルを構築する。

1. $x_1, x_2, \cdots, x_n, i.i.d. \sim \mathrm{Exp}(\theta_1)$
2. $y_1, y_2, \cdots, y_n, i.i.d. \sim \mathrm{Exp}(\theta_2)$

これを、$M_E(\theta_1, \theta_2)$ とする。$\theta_1 = \theta_2$ のモデルを $M_N(\theta)$ とする。

■$M_N(\theta)$ における尤度関数　$M_N(\theta)$ において、この尤度関数を計算する。

$$L_0 = \theta^{-n_1-n_2} \exp\{-\theta^{-1}T\}$$

ただし、$T = \sum_{i=0}^{n} x_i + \sum_{i=0}^{n} y_i$ である。尤度が最大 $\frac{\partial L_1}{\partial\theta} = 0$ となる θ を計算する。

$$\frac{\partial L_0}{\partial\theta} = \{-(n_1+n_2) + \theta^{-1}T\}\theta^{-n_1-n_2-1}\exp(-\theta^{-1}T). \tag{D.4}$$

より、$\theta_0 = \frac{T}{n_1+n_2}$ である。θ_0 を L_0 に代入すると、

$$L_0 = \theta_0^{-n_1-n_2}\exp(-n_1-n_2). \tag{D.5}$$

である。

■$M_N(\theta_1, \theta_2)$ における尤度関数　同様に、対立仮説のもとで、尤度関数 L_1 は、

$$L_1 = \theta_1^{-n_1} \exp\left(-\frac{n_1}{\theta_1}\bar{x}\right) \theta_2^{-n_2} \exp\left(-\frac{n_2}{\theta_2}\bar{y}\right) \tag{D.6}$$

$\frac{\partial L_1}{\partial \theta} = 0$ となる θ_1 を計算する。

$$\frac{\partial L_1}{\partial \theta_1} = \left(-n_1\theta_1^{-n_1-1} \exp\left(-\frac{n_1}{\theta_1}\bar{x}\right) + n_1\bar{x}\theta_1^{-n_1-2} \exp\left(-\frac{n_1}{\theta_1}\bar{x}\right)\right) \theta_2^{-n_2} \exp\left(-\frac{n_2}{\theta_2}\bar{y}\right). \tag{D.7}$$

$\frac{\partial L_1}{\partial \theta_1} = 0$ 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_1 は、

$$L_1(\hat{\theta}_1, \hat{\theta}_2) = (\hat{\theta}_1)^{-n_1} \exp(-n_1)(\hat{\theta_2})^{-n_2} \exp(-n_2) \tag{D.8}$$

である。

■尤度比

$$\Lambda = \frac{L_0}{L_1} = \frac{\theta_0^{-n_1-n_2} \exp(-n_2 - n_2)}{(\hat{\theta}_1)^{-n_1}(\hat{\theta}_2)^{-n_2} \exp(-n_1 - n_2)} \tag{D.9}$$

$$= \left(\frac{\hat{\theta_0}}{\theta_0}\right)^{n_1} \left(\frac{\hat{\theta_1}}{\theta_0}\right)^{n_2} \tag{D.10}$$

尤度比検定より、$-2\log\Lambda \sim \chi_1^2$ である[*5]。

[*5] いくらか条件がある

参考文献

[1] Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, Vol. 70, No. 2, pp. 129–133, 2016.

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 赤池弘次. Aic と mdl と bic. オペレーションズ リサーチ, Vol. 1996, pp. 375–378, 1996.

[8] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. 2016.

[9] Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, Vol. 23, No. 5, pp. 524–532, 2012.

[10] Angelika M Stefan and Felix D Schönbrodt. Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, Vol. 10, No. 2, p. 220346, 2023.

[11] Points of significance. *Nature Human Behaviour*, Vol. 7, No. 3, pp. 293–294, Mar 2023.

[12] Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, Vol. 31, pp. 337–350, 2016.

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

[14] 医療統計解析使いこなし実践ガイド:. 羊土社, 2020.

[15] 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.

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

[17] 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.

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

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

[20] マスミナカジマ, 眞澄中嶋. χ^2 適合度検定の初等的証明. 鹿児島経済論集, Vol. 61, No. 3, pp. 123–131, 12 2020.