

シミュレーションプログラミングを用いた統計教育の試み

順天堂大学 数理・データ科学教育研究センター 大橋 真也
Shinya OHASHI, Juntendo University

1 はじめに

本年度より、本学に新設された健康データサイエンス学部の1年生向けに「確率と統計I/II」を前・後期を通じて担当することになった。この授業では、伝統的な統計学とデータサイエンスとの違いを教えるとともに、統計処理ソフトウェアであるRを活用し、プログラミングを通して統計学や機械学習の基本について学ぶ試みを行っている。実際に、統計的検定、回帰、分類、機械学習などの理論に関して、数学的な理論だけではなく、シミュレーションによるプログラミングにより、その理論を理解するだけでなく、自己の経験として学ぶことにより、その理解の助けとなると考えた。

この授業を実施しようと考えたのには、本学初のデータサイエンス学部であり、統計学の中でデータサイエンスとして必要である項目や理論を、効果的に学ぶことができる方法はないかを模索した結果である。実際、学生は高等学校の授業などとは異なるプログラミング等を含めた理論や実習であるため、戸惑う学生も当初は多くいた。高等学校数学科の「数学B」等で学ばれる初歩的な統計学を未履修である学生が多いだけではなく、高等学校情報科等においてもプログラミングを学習する経験がない学生が多かったためであった。

2 統計教育にプログラミングは必要か

統計学とプログラミングは、大学の一般教育の中でも別々な授業で実施されていることもあり、統計学の授業で、プログラミングを教えることに対する批判も一部あるときいている。統計学やデータサイエンスを教える上で、プログラミング教育は有効であるのか、どのような繋がりがあるのかを、過去の講演や論文等から検討を行った。

2.1 David Donoho の講演

David Donoho は、2015年にプリンストンで開催されたTukey100周年ワークショップの中で、50年以上も前にJohn Tukeyが統計学から「データ分析」という新たな科学を生まれ、その後さまざまな統計学者により、MITやミシガン大学などで新しい科学分野としての「データサイエンスイニシアティブ」として、カリキュラムが編成され、今日の「データサイエンス」が生まれた、という講演を行った[1]。「データサイエンス」は古典的な統計学の領域を超え、統計的モデリングよりもデータの準備や提示、予測に重点を置くように考えられた。また、データサイエンスは、統計学と機械学習のスーパー

セットであるともこの講演の中で言っている。すなわち、統計学は、データサイエンスと密接に接続されているということである。さらにデータサイエンスとは、「ビッグデータ」への「スケールアップ」のための技術を追加したものであり、現在の科学そのもののすべてが、まもなくマイニング可能なデータになるからである、とも言っている。

2.1.1 データサイエンスとは何か

先に挙げたミシガン大学で策定された Michigan Institute for Data Science (MIDAS) では、データサイエンスは、次のように定義されている。

Data Science is “This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and inter-disciplinary applications.”

それ以外にもデータサイエンスは統計学と等価であるとする考え方や、統計学をリブランディングした学問であるとする主張など、さまざまな考え方が巷にあふれ、IBMやAmazonなどの企業でもデータサイエンスを科学の一分野としての位置づける考え方で定義している状況が見られる [2][3]。

2.1.2 データサイエンスにプログラミングは必要か

ミシガン大学の MIDAS をもとにした GDS(Greater Data Science) と呼ばれるデータサイエンスの拡大された定義の中では、データサイエンスを次の6つの区分で分類している。

1. Data Gathering, Preparation, and Exploration
2. Data Representation and Transformation
3. Computing with Data
4. Data Modeling
5. Data Visualization and Presentation
6. Science about Data Science

この中の3つ目の”Computing with Data”は、まさにデータサイエンスの中でのプログラミングの必要性を次のように説明している。

Every data scientist should know and use several languages for data analysis and data processing. These can include popular languages like R and Python, but also specific languages for transforming and manipulating text,

and for managing complex computational pipelines. It is not surprising to be involved in ambitious projects using a half dozen languages in concert.

Beyond basic knowledge of languages, data scientists need to keep current on new idioms for efficiently using those languages and need to understand the deeper issues associated with computational efficiency.

また、4つ目の”Data Visualization and Presentation”でもEDA(探索的データ解析)の視点から、データを可視化する技術は必要であると説明している。それ以外でもデータサイエンスにおいては、直接はプログラミングではないが、コンピュータを活用したモデリング等の技術が必須であるとしている。

すなわち、データサイエンスを学習するには、コンピュータの利用も必要であり、内容によってはプログラミングを学習することで有効に学ぶことができると考えることができる。

2.1.3 統計学とデータサイエンスの違い?

統計学を新しい科学として、発展した分野がデータサイエンスという考え方が多くあり、この考え方は伝統的な統計学の手法を否定するものではない。しかしながら、統計学とデータサイエンスでは、それぞれの分野の研究が進み、専門用語などに関しても使い分けが必要となっている。今回この授業で教科書として活用している「Practical Statistics for Data Scientists」(邦訳「データサイエンスのための統計学入門」[7])では、この用語の混乱に関して、次のような「サンプル」という用語の例を挙げている。

Terminology for rectangular data can be confusing. Statisticians and data scientists use different terms for the same thing. For a statistician, predictor variables are used in a model to predict a response or dependent variable. For a data scientist, features are used to predict a target. One synonym is particularly confusing: computer scientists will use the term sample for a single row; a sample to a statistician means a collection of rows.

2.2 ASAの声明から変化したこと

統計学とデータサイエンスは、共通部分が多いが、異なる学問であるとされる主張も多くあり、データサイエンスを教えることは、統計学を教えることと同義であるとは言えない。しかしながら昨今、統計学にも大きな動きがある。ASA(米国統計協会)の2016年の声明[4]である。

The ASA's Statement on p -Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's 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.

Cobb's concern was a long-worn circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do, we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "The scientific distillate: The scientific method of testing hypotheses by statistical analysis stands on a shaky foundation." A November 2013 article in Physics Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) on February 7, 2014, said: "Statistical techniques for testing hypotheses... have more flaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jeff Leek responded: "The problem is not that people use P -values poorly. Leek wrote: "It is that the vast majority of data analysis is used

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The ASA statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of P -values as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on p -values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provide the community a new view. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wassenaar with assembling a group of experts representing a wide variety of points of view. The bulk of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in effect, that if there was going to be a dis-

図 1: ASA's Statement

ASA は、この声明で伝統的な p 値の正しい使い方を説明し、その誤用や誤解を解消し、 p 値至上主義を脱しようとしている。また、従来の仮説検定をそのまま実施することには疑問を投げ掛けている。ASA の声明にある p 値に関する原則はの 6 つである。

1. 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。
2. p 値は、調べている仮説が正しい確率やデータが偶然のみで得られた確率を測る。
3. 科学的な結論やビジネス、政策における決定は、 p 値がある値を超えたかどうかにかのみ基づくべきではない。
4. 適正な推測のためには、すべてを報告する透明性が必要である。
5. p 値や統計的有意性は、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。

また、ASA は、"Moving to a World Beyond $p < 0.05$ " [5] として、統計的有意性に関する正しい考え方を主張している。

この声明により、統計学を用いるさまざまな分野で影響があり、統計的有意性に関する議論が高まってきている。ただし残念ながら、日本においては、一部の学会のみがこれに着目しており、この声明に関して知らない人たちも多くいる。

ASA の声明の一部に関して、次のようなことが言える。

- チャンピオンデータや p 値ハッキングのようなことをするのはやめる。
- 仮説検定において、適正な統計的モデリングを行い、それに関して仮説検定を行う必要がある。
- p 値は、正しい確率を表してはおらず、確実性の優劣を論じることはできない。

この1つ目に関しては、まさに p 値に関する誤用であり、現在も学術論文等に数多く見受けられる。膨大探索効果 (vast search effect) と呼ばれるものもその一つであり、多重比較やデータモデルを繰り返し使うことにより、バイアスや再現不能な結果を生じさせることができるということである。野球のスタジアム等で、2万人の観客にコイントスを10回行わせれば、10回連続で表が出るデータを多く集められるということである。

3つ目に関しては、仮説検定においては現在は Neyman-Pearson 流の有意水準 α 値を設定し、その値を超えるかどうかで有意性を判断するという考え方である。ただし、適正な統計的モデリングを行うことにより、Fisher 流の p 値を指標とするデータ分析も考えることができる。コンピュータがない時代では、正確な p 値を計算することができずに、それに合った適当な分布の有意水準をもとにすることしかなかったからである。ただしコンピュータが進んだ時代に、正確な p 値が算出できたとしても、そのデータが従うとされる分布は、理論的な分布であり、その考え方は近似でもある。そのため、 p 値を指標にすることは単純に行うことはできない。すなわちコンピュータ時代に適正な仮説検定を教えることも必要なのである。

2.2.1 仮説検定の考え方

2022年度より、高等学校では新しい学習指導要領により教育課程が始まり、「数学I」では、「仮説検定の考え方」が導入された。単に事象を適当な分布に当てはめるのではなく、二項分布などの実際の事象の確率を統計的確率として求め、その事象を当てはめる妥当性を検討し、 p 値を算出するという考え方である。

この考え方は、ASA 声明などの適正な p 値を考えさせるための学習指導となると考えられる。また「数学B」では、この考えを受けて信頼区間や仮説検定を導入している。

ただし、仮説検定や p 値についての誤用や誤解は、いまでも後を絶たない。学術論文だけではなく、NHKの高校講座 数学Iや医師国家試験 [6] などにおいても誤った出題がなされている。

有意性検定についての正しい議論を、日本の統計教育でもさらに進めていきたいものである。

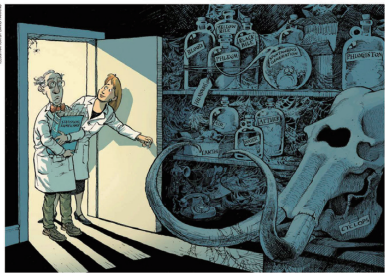
付 録
A. 医師国家試験 105 回 (平成 23 年) 問題 43

43. 新しく発売された抗菌薬 A の肺炎に対する治療効果を調べるために、新たに入院する肺炎患者を対象として、抗菌薬 A を投与した群 (A 群) と既存の抗菌薬 B を投与した群 (B 群) とに割り付けて、治療効果を入院期間で比較検討した。得られた結果を表に示す。

	A 群	B 群	P 値
対象者数	198 人	101 人	
入院期間 (平均)	8.1 日	3.6 日	0.036

この結果の解釈について正しいのはどれか。

- a A 群は B 群に比べて入院期間が平均で 3.6 日短い。
- b A 群の入院期間の平均値の誤差は 3.6 % 以内である。
- c A 群の方が B 群よりも入院期間が短くなる確率は 3.6 % である。
- d A 群の 96.4 % の患者は入院期間が B 群の平均入院期間よりも短い。
- e A 群と B 群とで入院期間に差がないのに、誤って差があるとする確率は 3.6 % である。



Retire statistical significance
Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

図 2: 医師国家試験での出題と、有意性検定に関する記事の例

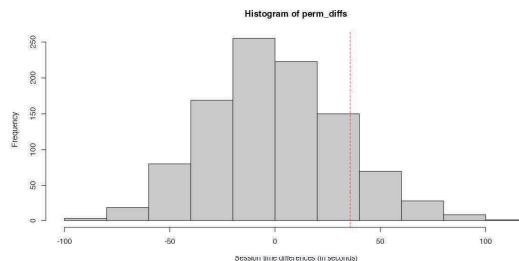
2.2.2 有意性検定はプログラミングで教える？

大学の統計学においても有意性検定に関しては、同様のアプローチが必要ではないだろうか。 p 値が何を意味するのかの理解を助けるために、リサンプリング検定を用いて、実施した例が以下の例である。Web 粘性性 (A のサイトと B のサイトではどちらのサイトに長くいてくれるかを A/B テストで行っている。) について、元々あるデータからリサンプリング 1000 回を行い、その状況をヒストグラムにして、実際に起きた事象がどの程度の確率で起こりうるのかを考えさせようという実習である。

```

1 perm_fun<-function(x,nA,nB){
2   n<-nA+nB
3   idx_b<-sample(1:n,nB)
4   idx_a<-setdiff(1:n,idx_b)
5   mean_diff<-mean(x[idx_b])-mean(x[idx_a])
6   return(mean_diff)
7 }
8 perm_diffs<-rep(0,1000)
9 for(i in 1:1000){
10  perm_diffs[i]=perm_fun(session_times$Time,21,15)
11 }
12 hist(perm_diffs,xlab='Session time differences (in seconds)')
13 abline(v=mean_b-mean_a,col='red')
14 mean(perm_diffs>(mean_b-mean_a))

```



等確率の事象ではなく、元々ある分布からリサンプリングを行うことにも意味がある。単純にどのような二項分布に従っているのかを考えさせるのではなく、元々の分布を使用することから、その意味が理解しやすいだろう。プログラミングに関してはまだ慣れていない時期に行った実習であることから学生はかなり苦戦はしていたが、人によって微妙に異なる p 値にも関心を持ったようである。

2.3 統計学にもプログラミングは必要か

前述した有意性検定だけではなく、さまざまな統計学の手法をコンピュータの力なしで行うことが現在は難しい。さらに複雑なデータ分析や機械学習などの手法を行うのに、高等学校数学科の教科書で学ぶような手計算で行うのは、実態に即していないと考える。そのため、必要な統計的処理に関しては、コンピュータを用いたプログラミングや可視化を活用して、学生の直感的な理解につなげることが必要であると考えます。

3 統計学や機械学習をプログラミングで教えること

統計学や機械学習について学習する際に、実際のデータも必要であるし、その結果は処理することで初めてわかることもある。また手計算や理論的な理解だけでは難しい内容も多く出てくる。この授業では、Peter Bruce らの書いた「Practical Statistics for Data Scientists」[7] の日本語訳を教科書として用いている。

3.1 数式で統計学を教えること

統計学のさまざまな理論を用いて教えるために、大学の初年時に微分積分や線形代数を学び、それらを応用することで理解を深めることができる。しかしながら、特定の理工系分野の学生ならば、これらの数学を履修しているが、それらを履修していない学生も多くいる。また、高等学校数学科において「数学 III」を学んでいない学生については、理工系の学生でも微分積分や線形代数の理解が精一杯であり、それらを応用できる学生も限られてくる。

本来の数学的な理論を後回しにしても、実際の統計学や機械学習の手法を理解させることは、プログラミングをすることによって理解させることができる。確かに一部はブラックボックスになってしまうが、その手法に関してマクロ的な理解は可能である。学年が上がり、専門的な科目を学ぶ時点で、ミクロ的な理解を深めてもよいのではないだろうか。

3.2 分析の手順を教える

データの分析においては、PPDAC サイクルや CRISP-DM などのさまざまな手順を示す標準的なサイクルがある。基本として、それらを学ばせるとともに、実際のデータに対して、その手順を実施することができるように、プログラミングを活用している。

授業の中で説明した理論や手法とともに、実際のデータを用いた実習を行い、それを応用して別なデータで処理を行う体験を行うために、プログラミングを実施している。以下は、ある回(ランダムフォレスト)のレポート課題として出題している問題例である。

penguins データについて、次の処理を行いなさい。

- (1) penguins データの 1,3,4,5,6 列目だけを取り出し、その欠損値を除去したデータを pen という名前にしなさい。
- (2) pen で目的変数を species として、ランダムフォレストを行いなさい。
(OOB をもとにした混同行列が表示されているか、確認する)
- (3) OOB 誤差のグラフを作成しなさい。
- (4) type1,2 の 2 つの変数重要度のグラフを作成しなさい。

この回では、OOB(Out-Of-Bag) データを活用させるために、トレーニングデータやテストデータへのデータの分割を行っていないが、多くの統計的な分析や機械学習などで、検証のためにテストデータの作成なども手順の一つとして行わせている。

毎回、似たような手順を実施することが理解され、標準的な統計分析の手法を理解させることにプログラミングが役に立っている。

4 シミュレーションをプログラミングする

プログラミングさせる素材は、先に挙げたりサンプリング検定などに代表されるシミュレーションや回帰、分類、教師なし学習などの統計学的機械学習の分野である。前期の授業ではこれらの基本についての理論を学ばせ、可視化等の手法から R の基本的な構文になれさせることにより、後期はシミュレーションを基本とした統計処理を手順(ストーリー)のある内容として進めることができている。

4.1 実際の例

実際の例を2つほど挙げる。1つ目の例は ISLR というパッケージにある、ローン返済に関するデータである。ローンを借りた人の属性により、ローンを返済できたかどうかに関しての分類をロジスティック回帰を用いて行い。仮定の例示のデータでそれの評価するというものである。この例を実施している段階では、トレーニングデータとテストデータの分割に関して、専用の関数 (tidyverse の `sample_frac` など) を用いずに分割を行っている。また混同行列に関してもその理解が深まるように、ある程度手順を追わせるように作成させている。

```

1 library(ISLR)
2 head(Default)
3 nrow(Default)
4 train_n<-sample(1:10000,8000)
5 test_n<-setdiff(1:10000,train_n)
6 train_df<-Default[train_n,]
7 test_df<-Default[test_n,]
8 df_glm<-glm(default~student+balance+income,
9             family="binomial",data=train_df)
10 summary(df_glm)
11 predict(df_glm,
12        data.frame(student="Yes",balance=1000,income=30000))
13 df_pred<-predict(df_glm,test_df)
14 df_pred_yn<-ifelse(df_pred>0,"Yes","No")
15 cm<-table(test_df$default,df_pred_yn)
16 cm

```

2つ目の例は、前のセクションで問題を例示したランダムフォレストの例である。OOB データの活用なども含め、学生にはこの授業の中でなじみのある penguins データ (重回帰などでかなり活用した) を用いて実施している。ここでは OOB による混同行列を出力するだけでなく、その誤認識率のグラフの作成や変数重要度のグラフの作成なども行わせている。

```

1 library(palmerpenguins)
2 library(tidyverse)
3 library(randomForest)
4 str(penguins)
5 pen<-penguins[,c(1,3,4,5,6)]
6 pen<-na.omit(pen)
7 str(pen)
8 pen_rf<-randomForest(
9   species~.,data=pen,importance=T)
10 pen_rf
11 err_df<-data.frame(err_rate=pen_rf$err.rate[, 'OOB'],
12                   num_trees=1:pen_rf$ntree)
13 graph<-ggplot(err_df,aes(x=num_trees,y=err_rate))+
14   geom_line()+
15   theme_bw()
16 graph
17 varImpPlot(pen_rf,type=1)
18 varImpPlot(pen_rf,type=2)

```

4.2 シミュレーションプログラミングのメリット

シミュレーションプログラミングを統計教育として取り入れたことによるメリットは多い。本授業実践でも以下のようなメリットが挙げられる。

- 乱数シミュレーションであるため、学生ごとに結果が微妙に異なり、他学生の課題からの複製不正などを行うことが少なくなる。
- 結果が異なることから、その結果に関して学生間でのコミュニケーションがはかることができる。また、結果が上手く出るかどうかに関してのワクワク感が生まれる。
- 先にも挙げたように、統計手法の手順をステップごとに組み立てていくことを理解し、実際の統計学的な理論の理解の助けになっている。
- 目的が明確であるので、プログラムのエラーや修正などに関しても学生が積極的に取り組んでいる。また目的のあるプログラミングであることから、プログラミングスキルの向上にも結びついている。
- 表計算ソフトウェアや専用ソフトウェアのようにメニュー構成などを覚えることに注視することもなく、自由な記述により実習が可能である。また細かな修正などに関しても自由度は高い。

4.3 シミュレーションプログラミングのデメリット

ただし、メリットばかりではない。このシミュレーションプログラミングを統計教育に取り入れるに当たっては、次のような問題点も考えられる。

- 結果は異なるが、他人のコードを複製して、課題の提出を行う学生もいる。

- 統計手法の手順がある程度身につくと、その違いのみをきちんと着目することができるようになる一方で、課題が簡単に処理できてしまう。
- 統計手法に関して、実践的な理解はできているが、数学的な理解を育成することはできていない。
- 多くの学生のシミュレーションによる、膨大探索効果や多重比較になっている点の問題点の解消はできていない。すなわち、特異な結果が出てもそれを信じてしまう。
- プログラミングがハードルになり、本来の内容を理解することができていない学生もいる。

これらの問題点以外にもいくつかの問題が出てきているが、これらの問題に関しては、徐々に解消していく予定である。

4.4 中等教育での活用

高等学校においても、プログラミング教育が進められ、数学科や情報科においてデータサイエンス分野が重視されている。理論の理解もちろん必要ではあるが、折角プログラミングを学ぶ機会を得ているのであることから、カリキュラム・マネジメントを活用することで、シミュレーション、プログラミング、統計教育を有機的に結びつけていくことができると考える。教科書に掲載されていないからと言って、実施しないのでは、統計教育の学習にとって機会を逃しているのではないだろうか。

大学初年度の学生にとってもワクワク感を育てることができている内容でもあるので、高等学校でもより教科書の題材に沿った学習内容としてのシミュレーションプログラミングの活用を考えていきたい。今後そのような素材や例を開発していく予定である。

5 おわりに

統計教育というと、理論中心の授業になりがちであったが、可視化やシミュレーションにおいて、プログラミングを活用することによって学生のモチベーションを上げることができる。一般には統計学習における ICT 活用と言われれば、プログラミングよりハードルは低いと、表計算ソフトウェアや BI ツールなどの方がはるかに活用されてきているだろう。しかし、今後ともその自由度の大きさからプログラミングによる統計学習を進めていき、さまざまなコンテンツを作成していきたいと考えている。また、高等学校新学習指導要領で学んできた学生が入学してくる 2025 年度からは、プログラミングの扱いや統計学の扱いに関して、一部見直しを行い、高等学校との学習内容の接続に着目して、内容の精選を行っていききたいと考えている。

参考文献

- [1] 50 Years of Data Science, David Donoho, 2017,
<https://www.tandfonline.com/doi/epdf/10.1080/10618600.2017.1384734?needAccess=true&role=button>
- [2] Data science without statistics is possible, even desirable, Vincent Granville,
<https://vincentgranvillesposts.quora.com/Data-science-without-statistics-is-possible-even-desirable>
- [3] Statistics is the least important part of data science, Andrew, 2013,
<https://statmodeling.stat.columbia.edu/2013/11/14/statistics-least-important-part-data-science/>
- [4] The ASA's Statement on p-Values: Context, Process, and Purpose, ASA, 2016,
<https://www.tandfonline.com/doi/epdf/10.1080/00031305.2016.1154108?needAccess=true&role=button>
- [5] Moving to a World Beyond “ $p < 0.05$ ”, ASA, 2019,
<https://www.tandfonline.com/doi/epdf/10.1080/00031305.2019.1583913?needAccess=true&role=button>
- [6] p-value is a Useful and Excellent Measure for Reporting Results of Statistical Analysis on Clinical Data, Takashi Yanagawa, 2017
- [7] Practical Statistics for Data Scientists(邦訳「データサイエンスのための統計学入門」第2版(オライリージャパン)), Peter Bruce, Andrew Bruce, Peter Gedeck, 2020