

R.Q. (リサーチ・クエスチョン) 構築という視点から 伝統的検定手法とベイジアン推定を比較する

— rstan の生成量 (generated quantities) に注目しながら —

From the perspective of R.Q. (Research Question) construction
Comparing traditional test methods and Bayesian estimation:
With a focus on the amount of rstan generated

藤 本 一 男

Kazuo FUJIMOTO

This paper pointed out that there is an end-run around the statistical testing techniques that are available to researchers, which dictate the R.Q. (Research Question). The test theory of traditional statistics (NHST: null hypothesis significance test) is the cause of this problem, and that Bayesian estimation can be used as a clue to solve this problem, however, it is only a necessary condition, and when Bayesian estimation is performed using MCMC such as rstan, utilizing generated quantities It is possible to avoid the short-circuit of assuming that the research hypothesis has been proven by the adoption of the alternative hypothesis due to the rejection of the null hypothesis. This is illustrated by comparing Bayesian estimation with t-test and MCMC using historical data (Student's sleep data).

Based on this comparison, it became clear that the content of the R.Q. is related to the sophistication of the research hypothesis, which in turn supports the sophistication of the research hypothesis that is above the research hypothesis.

概要

まず、研究者が使用可能である統計的検定技法が、R.Q. (リサーチ・クエスション) を規定してしまう本末転倒の事態があることを指摘した。その原因に伝統的統計学の検定論 (NHST: 帰無仮説有意性検定) があるが、この問題を解決する手掛かりとして、ベイジアン推定を用いることができること、ただし、それは、`rstan` のような MCMC を用いてベイジアン推定を行う際に、生成量を活用することで、帰無仮説の棄却による対立仮説の採択をもって研究仮説が証明されたとする短絡を回避することが可能になることを述べた。こうしたことを、歴史的データ (Student の睡眠データ) をもちいて、 t -検定と MCMC をもちいたベイジアン推定を比較し例示した。

この比較を踏まえると、立てられる R.Q. の内容が調査仮説の検討の精緻さに関係していることが明らかとなり、ひいては、調査仮説の上位に位置する研究仮説の精緻さを支えるものであるということが明確になった。

キーワード: 調査仮説、R.Q.、リサーチクエスション、帰無仮説、有意性検定、 t -検定、ベイズ統計、MCMC、`rstan`

1.0 はじめに / 問題の所在

本稿は、R.Q. (リサーチ・クエスション) 構築という視点から、 t -検定とベイジアン推定を比較する。建前からすれば、本末転倒である、としても、R.Q. を構築する際に我々は採用する分析手法が有する「解決策」を前提としており、それが研究仮説の幅を制約してしまうことがある。その状況を具体的なデータを分析対象に用いて、 t -検定とベイジアン推定による result を比較することで考察していく。

ここで t -検定をとりあげるのは、それが伝統的推定論、Neyman=Pearson 体系での代表的検定手法であるからである。伝統的推定論 (頻度主義) とベイジアン推定の根本的な違いは、推定の構造である。伝統的推定論では、推定対象であるパラメータの「真の値」を定数として考え、手元にある標本データが確率的に変化すると考える。対して、ベイジアン推定では、それとは逆に、手元の標本データは定数で、推定対象であるパラメータが確率分布する、とする。

この違いは、信頼区間 (ベイジアンではこれと区別して信用区間という用

語をあてることもある)の理解の仕方の違いに厳然として現れる。

1.1 頻度主義による信頼区間解釈の不自然さ

伝統的統計学、頻度主義、Neyman=Pearson 体系での信頼区間は、検討しているパラメータが存在する確率を表現することはできない。(式に確率変数がない、ということは『統計学基礎 2015』p110-113 を参照)

この信頼区間を理解しようとするとき次ようになる。まず、母集団から標本をいくつも(沢山!)取得する。その標本ごとに「平均値」のような標本統計量を計算すると、その標本統計量は分布し、この標本の数が大きくなれば、(母集団の分布がなんであれ)その分布は、正規分布で近似できるようになるので(中心極限定理)、それを踏まえて母平均が存在する範囲=信頼区間を考えることができる(というよりも、このように理解しないといけない)。

実際には、(社会調査や実験では)標本は一回しか取得されていないので、取得された信頼区間は、標本が沢山取得されたとしたら、その標本セットのうちの95%は、「その範囲内」に真の母平均を含んでいる、というように理解する。頻度主義統計学で推定、検定を理解するためには、この標本分布という架空の分布を「リアル」に捉える能力が求められる¹⁾。

つまり、中心極限定理の仕組みを念頭に、沢山の標本を得られたとしたら、という実現しようもない仮定を思い描き、それを踏まえて、「95% 信頼区間」の「95%」の意味を考える、ということを強いられる。

頻度主義の前提を踏まえて「間違っていない」信頼区間の説明ができたとして、それからどのような発展が可能なのだろうか²⁾。

1.2 頻度主義の検定論

頻度主義を前提にした検定の理論もこの標本分布の理論を前提にしている。「5%水準」で、帰無仮説を棄却できるかどうかを判定し、帰無仮説が棄却されたら、対立仮説が採用される。この判断の定式化は確かに便利であるが、ASA 会長声明でも言明されているように、p 値によって帰無仮説が棄却され対立仮説が採択される、ということと、研究上の仮説が正しいということは別のことである(付録 A.1 APA2016 第3「原則」部分参照)。

ここで用いられるロジックは、証明したい仮説の否定形を帰無仮説として設定し、その帰無仮説が正しいとしたら、取得されたデータが実現する確率は非常に低い(5%以下!)である(かどうかを確認)。そうであれば、このことは、前提とした帰無仮説が「間違っている」ということを意味するので、

それを「棄却」rejectする。そこから、対立仮説を採択する、という「背理法」の手続きをへる。このように仮説を証明するために、設定されるハードルが、帰無仮説の棄却か否かという二分法になっているので、R.Q. もそれに引きずられるしかない。

ベジアン推定では、推定対象の母集団パラメータが確率分布をするので、得られた信用区間は、そのまま、パラメータが存在する確率を表している。信用区間下限 < パラメータ < 信用区間上限、はそのまま「95%の」確率でパラメータが存在する区間として理解すればいい。しかし、95%信用区間として自然な解釈を可能にすることと、R.Q.の制約問題の解決は、別のことである。

1.3 解決できないことは問としてたち現れない

この意味では頻度主義信頼区間/検定を使うのではなく、ベジアン推定を使えばいい、ということになるのだが、ここでは推定方法を選択するもう一つの側面に注目したい。つまり、ベジアン推定の採用は、問題解決の道ではあるが、必要条件でしかない。

ここで問題としたいのは、解決できる問題しかR.Q.として設定されない、という側面である³⁾。しかし、ここに、ASA2016が指摘する研究仮説の証明とp値による対立仮説の採択の取り違いや混乱の現実的根拠がある。

私たちは、統計的有意性検定の問題として解決できることを研究仮説にしないでだろうか。

2.0 伝統的検定手法としてのt-検定とベジアン推定を比べる

そこで、本稿では、データを用いて統計的に解決できること/できないことを具体的に比較してみる。

分析対象として、Stuendt⁴⁾のsleepデータを用いる。これは、Student 1908で事例を描くために使われているデータで、二つの薬剤（睡眠促進剤）⁵⁾は、Rで提供されているデータセットでは、(1)を1、(2)を2としているが、それを、A剤、B剤とした。以下に基本的な分布を図示する。

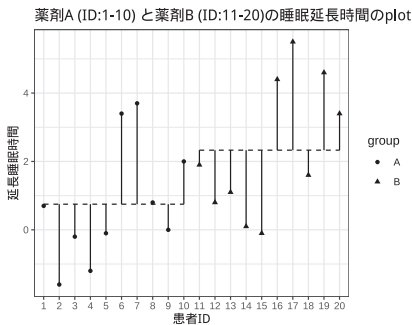


図 1 index plot

図 1 の index は、1-10、11-20 は対応している同一の被験者である。波線は、薬剤 A, B に対する平均値を plot している。

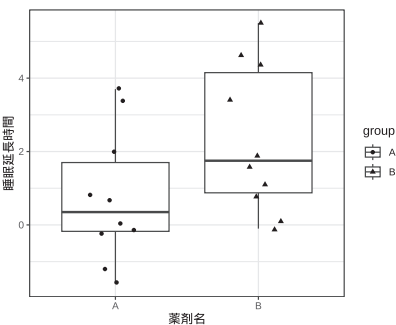


図 2 boxplot で A 剤、B 剤の比較

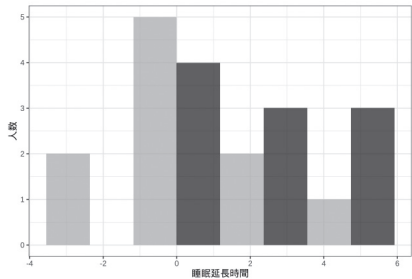


図 3 A 剤 B 剤のヒストグラム

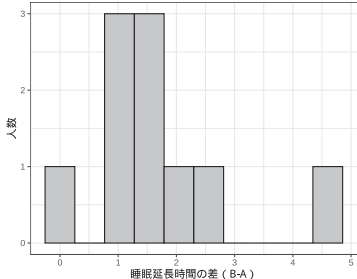


図 4 効果のヒストグラム

このサンプルデータから計算した基本統計量は以下の通りである。

表 1 sleep データの基本集計

変数	平均値	標本標準偏差
A	0.75	1.61
B	2.33	1.80
diff	1.58	1.11

2.1 t-検定による分析

このデータは、(A 剤、B 剤と略記)の睡眠延長効果を 10 人の被験者に対して、測定したデータである。A 剤、B 剤ともに、10 人に投与されているその効果が測定されているので、A、B の値は、それぞれ同一の被験者の A 剤、B 剤に対する反応を表している。

また、平均値は B の方が大きいので、睡眠延長効果は、 $A < B$ となっており、検定の役割としては、この差が統計的に有意であるかどうかを判定することになる。(これに対応する R.Q. は「A 剤より B 剤のほうが睡眠延長効果が長い」である。)

その際の対立仮説 H_1 は、「延長時間 $A < B$ 」つまり B 剤の方が A 剤より睡眠延長効果が大きい、であり、p 値で棄却するかどうかを判定する帰無仮説 H_0 は、「延長時間 $A \geq B$ 」となる。

図 1、2 をみれば、B 剤の方が睡眠延長効果は長い。その傾向が確認されたので、t-検定を用いる。検定は片側検定となる。

図 5 に、t-検定の function に、上述の option (片側検定など) を設置し、得られた result を示している⁶⁾。

```
```{r}
t.test(sleep.df$diff,alternative = "greater")
```
```

One Sample t-test

data: sleep.df\$diff
 $t = 4.0621$, $df = 9$, $p\text{-value} = 0.001416$
 alternative hypothesis: true mean is greater than 0
 95 percent confidence interval:
 0.8669947 Inf
 sample estimates:
 mean of x
 1.58

図 5 t-検定の result

分析対象は、 $\text{diff} = B - A$ である。検定は、片側検定 ($\text{alternative} = \text{"greater"}$)。Result をみると、p 値は、0.001416 で、 < 0.05 であるので、帰無仮説は棄却

され、対立仮説 (alternative hypothesis) が採択される⁷⁾。つまり、true mean is greater than 0 である。

そして、95% 信頼区間は、0.0867 ~ 無限大までとなり、
平均値は、1.58 と推定されている。

2.2 Sleep データのベジアン推定

つぎに、この Sleep データに対して、rstan を用いたベジアン推定を行ってみる。

Rstan のスクリプト (sleep.rstan) と Rmd (sleep.Rmd) は付録 A.2、A.3 を参照。

2.2.1 その 1 t-検定と同じことをする。つまり「片側検定で $B > A$ 」

MCMC での result である fit を表示すると以下ようになる。

```
```{r}
print(fit0, probs = c(0.05, 0.25, 0.5, 0.75, 0.95))
```
```

Inference for Stan model: anon_model.
3 chains, each with iter=11000; warmup=1000; thin=1;
post-warmup draws per chain=10000, total post-warmup draws=30000.

| | mean | se_mean | sd | 5% | 25% | 50% | 75% | 95% | n_eff | Rhat |
|------------|--------|---------|------|--------|--------|--------|--------|--------|-------|------|
| mu1 | 0.74 | 0.00 | 0.69 | -0.38 | 0.32 | 0.75 | 1.18 | 1.85 | 20671 | 1 |
| mu2 | 2.35 | 0.01 | 0.78 | 1.09 | 1.86 | 2.34 | 2.83 | 3.61 | 18896 | 1 |
| sigma1 | 2.11 | 0.00 | 0.62 | 1.37 | 1.68 | 1.99 | 2.39 | 3.25 | 18634 | 1 |
| sigma2 | 2.36 | 0.01 | 0.70 | 1.53 | 1.88 | 2.22 | 2.68 | 3.66 | 16978 | 1 |
| delta | 1.60 | 0.01 | 1.05 | -0.10 | 0.93 | 1.60 | 2.27 | 3.31 | 19284 | 1 |
| delta_over | 0.94 | 0.00 | 0.24 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 19018 | 1 |
| lp__ | -22.74 | 0.02 | 1.60 | -25.83 | -23.52 | -22.38 | -21.57 | -20.89 | 10568 | 1 |

Samples were drawn using NUTS(diag_e) at Wed Sep 25 21:07:16 2024.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).

図 6 MCMC の result (delta_over だけを計算)

delta_over は、 $B > A$ のサンプルの時は 1 が、そうではない ($A \geq B$) のときは、0 を生成量 (generated quantities) として、.rstan で定義している。これは、t-検定で確認した対立仮説がなりたつ確率を示している。平均値 0.94 は、94% を意味している⁸⁾。

では、ベイズの信用区間では解釈が自然というだけなのか。ここまでの比

較ではそういうことになる。

2.2.2 その2 生成量 (generated quantities) の設定による柔軟な判定

つぎに、生成量に、A 剤と B 剤の睡眠延長時間の差が1時間以上、1.5時間以上、2時間以上、3時間以上、という条件を計算してみる。ベイズ推定ならこういう計算が可能なのである。

.rstan ファイルの generated quantities のところが以下ようになる。(編みかけで表示)

```
generated quantities{
  real delta;          // 平均値B- 平均値A の変数
  real delta_over; // B-A が正である場合に、1、 else 0
  real delta_over1; // B-A が1 以上の場合に、1、 else 0
  real delta_over15;
  real delta_over2;
  real delta_over3;
  delta = mu2 - mu1;
  delta_over = step(delta);
  delta_over1 = delta > 1 ? 1 : 0;
  delta_over15 = delta > 1.5 ? 1 : 0;
  delta_over2 = delta > 2 ? 1 : 0;
  delta_over3 = delta > 3 ? 1 : 0;
}
```

.rstan をこのように修正して、再度スクリプトを run させてみる。Result は図7のようになる。delta_over1、delta_over15、delta_over2、delta_over3 の行が追加されている。

この result から明らかになるのは、B 剤の睡眠延長効果が1時間となる確率は、73%。さらに、1時間半 (delta_over15) は、54%、2時間差 (delta_over2) となるのは33%、そして3時間差 (delta_over3) は8%ということである。

このような計算は、伝統的検定では設定できない。P 値の扱いをめぐって、それを0.05 と比べて大小を比較するし帰無仮説の処遇を決めるのではなく、p 値自体の計算が簡単にできるのだから、その値や、検定統計量 (t 値、z 値、

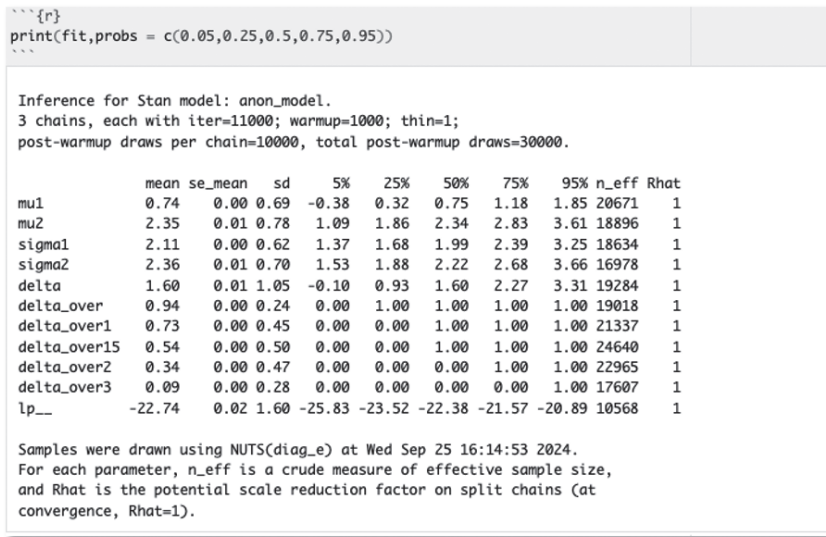


図 7 生成量 delta_over を 1, 1.5, 2, 3 と追加

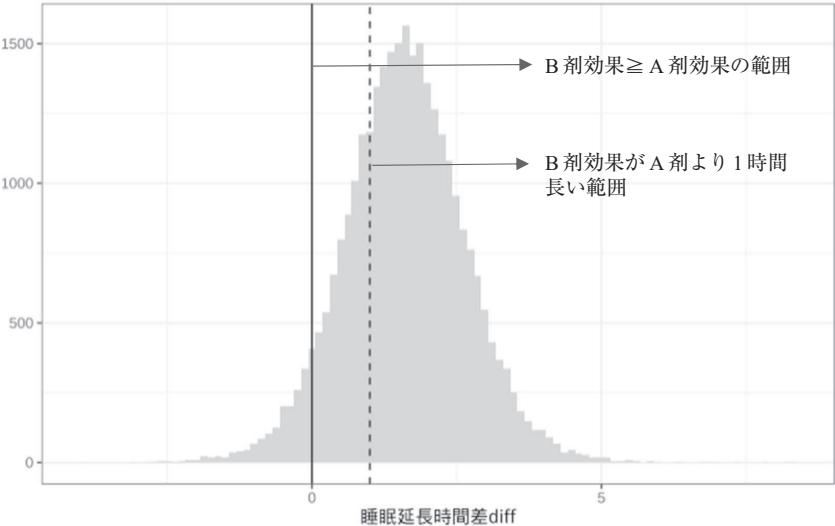


図 8 睡眠延長時間差の事後分布

カイ二乗値)も併記するように、という指針をみることもあるが、検定の原理からして、図7に示したような確率は原理的に論じられないので、これらを併記したところで算出することはできるものではない。

2.2.3 R.Q.の拡張

以上のようなベジアン推定がもつ計算可能性を踏まえると次のように柔軟に考えていける。

Sleep データで、B 剤の睡眠延長効果が大きい、ということを表現する R.Q. はどうなるだろうか。

- R.Q.1 睡眠延長時間が 95% の確率で $A < B$ である。
- R.Q.2 睡眠延長時間が、 $A < B$ で 1 時間以上の差がある確率はいくつか。
- R.Q.3 睡眠延長時間が、B は A の 1.5 倍となる確率はいくつか。
- などなど。

R.Q.1 は、 t -検定での 95% 信頼区間のベジアン信用区間版である。しかし、R.Q.2、3 は、 t -検定ではそもそも設定が不可能な問である。

3.0 t -検定とベジアン推定の比較で見えたもの

こうした比較を行うことで、適用可能な統計技法の持つ制約が R.Q. の立て方の制約になってしまう可能性を提示できたと思う。

ここでは、A 剤にくらべて B 剤での睡眠延長時間が少しでも大きい場合 (`delta_over`) と 1 時間多い場合 (`delta_over1`)などを生成量として generate している。

3.1 前提は、R.Q.の立てかたにある

研究仮説としては、そもそもどのような基準での比較を行なっているのかを厳密にたてておかなければならない。伝統的手法では「差があるか」という設問にすべてを投げ込んでいたわけだが、「どの程度の差」(たとえば 1 時間とか)があれば、B 剤の方が睡眠時間の延長にとって優位にある、と結論つけるのか、という基準が必要であったあることが、こうした処理が可能であり、また必要であることが浮かび上がってくる。

これは、実験のデザインの時点で明らかにされるべきことであるし、また、統計学の外側、研究している対象に属するものである。

3.2 ベイズ統計学利用の条件の成熟

確かに伝統的手法によるパターン化された検定手法は、貧弱な計算環境を前提に先人たちによって構築された体系である。それを、コンピューターが非常に高度に、そして普及した今日ベイズ統計学をはじめ伝統的手法とは異なったアプローチが可能になっている。ベイズ推定は、こうしたコンピューターパワーを活用して実用になった手法の一つである。

ただ、だからといってベイズ推定を使えば問題が解決するというわけではない。今回の睡眠促進剤としてのA剤とB剤の例で考えてみる。

たしかに、ベイズ推定による生成量を用いることで、t-検定では確認できなかった、1時間、1.5時間、2時間、3時間の睡眠時間延長の確率を得ることができた。ただし、これは、B剤が延長効果あり、という調査仮説を支えるピースの一つにすぎないはずである。

睡眠促進剤の薬学的な原理はまったくわからないので、内容的には想像の産物でしかないのであるが、B剤の優位性が、調査仮説としてたてられた時には、薬剤の内部構造に関する知見、また、マウスなどの動物実験での結果、などのB剤が優位を論証するいくつものエビデンスが用意されているはずで、それを構成する一つとして、10人の被験者を対象にした治験、つまり、A剤とB剤による睡眠延長時間の測定があったはずである。

そう整理すると、調査仮説は、統計学的にたてられるR.Q.とその分野（この例では薬学的治験）によるR.Q.の総合的なものとして用意されるものである。

「社会統計学」という分野があるが、そこでは、「統計学的手法」を社会学的知見からどう使うのか、という視点での講義構成が求められるべきである。つまり、R.Q.の立て方として講義が構成される必要がある。

3.3 統計的な推定を柔軟にできるようになることで見えてくるもの：研究仮説の精緻さ

伝統手法であれベイズ推定であれ、そのレイヤーで判定される仮説を調査仮説、リサーチクエスチョン(R.Q.)と呼ぶなら、その上位に、研究仮説と呼ぶレイヤーがある。そこには様々な統計的に解決されるピース以外にもその専門分野の問い、先行研究、仮説が位置している。調査仮説とそれを立証する統計的推論による判定はそこにハマる一つのピースであって、多くは、統計学の外側の「専門分野」の知見によって構成されている。先に扱った睡眠延長効果をめぐっては、比較している薬剤や動物実験での結果などである。

統計技法は、科学的論証の重要な役割を担っているだけに、それが主張する仮説の正当性の論理の緻密さは、全体の論証の質を規定している。睡眠促進剤の效能比較を「少しでも B 剤の方が長い」確率で展開するのか（荒い精度）、1 時間、1.5 時間、2 時間、3 時間の差をもたらす確率（精緻な精度）で展開するのは、論証の質の問題となる。

こう整理すると、5 % 水準で帰無仮説を棄却したからと、研究仮説が支持された、というのも乱暴な議論であるし、帰無仮説を棄却できなかったとして（先の例でいえば、1 時間の差は、73% なので、もしこの水準で帰無仮説を判断すると「棄却」されない）実験が無意味であった、ということにはならないだろう。この治験における B 剤の睡眠延長効果の程度は測定されているのだから、次の研究への足がかりが得られることになる。

以上の例は、睡眠促進剤についての専門家ではないために、統計的判断とは独立した薬学の専門領域を想定しているので現実的ではないかもしれないが、こうした統計学とは独立した領域での知見の重要性の指摘の例としてご理解いただきたい。

3.4 付記

なお、本稿では、ベイズ推定自体の正当性についての検討は省略している。伝統的な統計手法では、その定式化された簡便さによって、科学的な知見の確認というよりも、研究結果の「お作法」と化していることは APA2016 の警告に現れている。これと同じように、Stan など MCMC をもちいたベイズアンプローチもなにをみているのかの吟味を怠れば「お作法としてのベイズ」になる恐れがある。

こうした事態を予見してか、奥村は、伝統的技法かベイズ化の二分法ではなく、研究手法が一つ増えたという視点であつかうことを提案している（奥村 2018 あとがき）。

また、三中 2018 では、ベイズ手法を紹介しながら、計算機統計学の実展段階として歓迎しながらも、ベイズをめぐるなげかけられている理論的な 3 点は解決していないことに留意せよ、と警告している。

付録

A.1 ASA2016 会長声明から p 値の誤用に関する「3. 原則」の抜粋。

1. P 値はデータと特定の統計モデル（訳注：仮説も統計モデルの要素のひとつ）が矛盾する程度をしめす指標のひとつである。

P 値は、特定のデータとそのデータにあてはめたモデルとの矛盾する程度を要約するひとつのアプローチに過ぎない。最も一般的な内容は、一連の仮定のもとで構成され、いわゆる「帰無仮説」ともなうモデルである。多くの場合、帰無仮説では 2 グループ間に差がない、要因と結果の間に関係がない、というように効果がないことを仮定する。P 値が小さいほど、データと帰無仮説の統計的な矛盾の程度は大きくなる。ただし、P 値の計算の背後にある仮定がすべて正しければ、であるが。この矛盾の程度は帰無仮説を疑う、あるいは帰無仮説に反対する証拠としても解釈できるし、P 値の計算の背後にある仮定を疑う、あるいは反対する証拠としても解釈できる。

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えられた確率を測るものではない。

研究者は、しばしば P 値を帰無仮説が正しいという記述や、偶然の変動でデータが観察される確率に変えたがるが、P 値はそのどちらでもない。P 値は仮説やその計算の背後にある仮定に基づいたデータについての記述であり、仮説や背後にある仮定自身についての記述ではない。

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（訳注：有意水準）を超えたかどうかにかかわらずみにつくべきではない。

科学的な主張や結論を正当化するために、データ解析や科学的推論を機械的で明白なルール（「 $P \leq 0.05$ 」といった）に貶めるようなやり方は、誤った思いこみと貧弱な意思決定につながりかねない。二分割された一方の側で、結論が直ちに「真実」となったり、他方の側で「誤り」となったりすることはありえない。科学的推論を行う際、研究者はさまざまな背景情報を利用すべきであり、それには研究のデザイン、測定の実質、研究対象である事象のこれまでのエビデンス、データ解析の背後にある仮定の妥当性が含まれている。「可否」による二分類の決定は実用的ではあるが、P 値だけで決定が正しいかどうかを保証されるものではない。「統計的有意性」（通常「 $P \leq 0.05$ 」とされる）は、科学的結論（つまり真実であること）を主張するための保証として広く用いられているが、科学のプロセスを著しく損ねている。

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。

P値と関連した解析は選択して報告すべきではない。複数のデータ解析を実施して、そのうち特定のP値のみ（たいていは有意水準を下回った）を報告することは、報告されたP値を根本的に解釈不能としてしまう。見込みのありそうな結果をいいとこ取り——データのどぶさらい、有意症、有意クエスト、選択的推論、P値ハッキングとも呼ばれる——すると、出版された論文に統計的に有意な結果が誤って過剰に報告されるため、厳に避けなければならない。複数の統計的検定を行っていない場合でもこの問題は起こりうる。報告すべきことを研究者が統計的な結果に基づいて選択する場合、選択を行ったことと選択の根拠を読者がしらなければ、報告された結果の妥当な解釈は常に極めて難しくなる。研究の中で調べる仮説の数、データ収集の際に行ったすべての決定、実行したすべての統計解析、そして計算したすべてのP値を研究者は開示すべきである。少なくとも、どのような解析がいくつ行われたか、報告する際に解析とP値をどのように選んだのかをしらなければ、P値と関連した解析に基づいて妥当な科学的結論を導くことはできない。

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。

統計的有意性は科学や人間、経済にとって意味のあることとはことなる。P値が小さいからといって、必ずしも大きな、より重大な効果があることを意味しないし、P値が大きくても、重要ではないこと、あるいは効果がないことを意味しない。どんなに小さい効果でも、サンプルサイズが大きかったり測定精度が十分高ければ小さいP値となりうるし、大きな効果であっても、サンプルサイズが小さかったり測定精度が低ければ、大きなP値となることもある。同様に、効果の推定値がおなじ値であったとしても、推定値の精度がことなれば、ことなったP値となる。

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。

背景情報やほかのエビデンスがなければ、P値は限られた情報しか提供しないことを研究者は認識すべきである。たとえば、0.05に近いP値ひとつだけでは帰無仮説を否定する弱いエビデンスでしかない。同様に、比較的大きなP値であっても、帰無仮説を支持するエビデンスとはならない。ほかのたくさんの仮説が、帰無仮説と同等か、それ以上に観察されたデータと矛盾しない可能性がある。これらの理由から、P値以外のアプローチが適切かつ実施可能な場合は、P値を計算しただけでデータ解析を終えるべきではない。

A.2 rstan を実行する rmarkdown

```
scr<-"Rstan_sleep.stan" #
data <-list(N1 = 10,
           N2 = 10,
           x1 = sleep.df$A,
           x2 = sleep.df$B
          )
Par <-c("mul","mu2","sigma1","sigma2","delta",
        "delta_over","delta_over1")
war <- 1000
ite <- 11000
see <- 1234
dig <- 3
cha <- 3
fit<- stan(file = scr, data = data, iter=ite, seed=see,
           warmup=war,pars=par,chains=cha)
```

A.3 Rstan_sleep.rstan

```
// The input data is a vector 'y' of length 'N'.
data {
  int<lower=0> N1;
  int<lower=0> N2;
  real x1[N1];
  real x2[N2];
}

// The parameters accepted by the model. Our model
// accepts two parameters 'mu' and 'sigma'.

parameters {
  real mul;
  real mu2;
  real<lower=0> sigma1;
  real<lower=0> sigma2;
}
```

```

transformed parameters {
    real<lower=0> sigma1sq;
    real<lower=0> sigma2sq;

    sigma1sq = pow(sigma1,2);
    sigma2sq = pow(sigma2,2);
}

// The model to be estimated. We model the output
// 'y' to be normally distributed with mean 'mu'
// and standard deviation 'sigma'.

model {
    x1 ~ normal(mu1,sigma1);
    x2 ~ normal(mu2,sigma2);
}

generated quantities{
    real delta;
    real delta_over;
    real delta_over1;

    delta = mu2 - mu1;
    delta_over = step(delta);
    delta_over1 = delta > 1 ? 1 : 0;
}

```

謝辞

本稿の執筆に際しては、対応分析研究会（東京芸大 磯直樹先生主宰）での討議に助けられています。つたない発表に対して、さまざまな側面からアドバイスをくださった研究会参加のみなさまに感謝いたします。

また、ベイズ統計学に関しては、Tokyo.Rでもお世話になっているコグラフ株式会社の塩見登志和さんに、統計学全般については、東京女子大学情報処理センターの浅川伸一さんにアドバイスをいただきました。理解しきれていない部分も多々あると思います。文中の内容についての責任は、すべて私に

あることはいうまでもありません。

なお、本研究は、科研費基盤研究(C)「データの幾何学的配置に着目したカテゴリカルデータ分析手法の研究」(20K02162) および、基盤研究(B)「現代日本の文化と不平等に関する社会学的研究：社会調査を通じた理論構築」(22H00913)の助成を受けています。記して感謝いたします。

注

- ¹⁾ Schmuller 2017=2023:132 では「標本分布を理解しているかどうか統計を理解しているかどうかの重要な点である」としている。
- ²⁾ こう書くと、頻度主義アプローチは使い道がないという主張をしているように見えるが、この有意性検定の体系をつくった当のネイマンは、「いえることは、この方式で信頼区間を求めることを多数繰り返すときにそのうちの95%が真の μ を含むということであり、現実と与えられた観測値については、「なにもいえない」(実際、この手法の提案者であるJ. Neymanはそう答えている)」と述べているらしい(日本統計学会 2012, 2015:112)。つまりネイマン=ピアソン体系は、工場での生産ラインのように、「標本抽出」が繰り返し実施できる環境では有効だということである。では、一回しか抽出ができない条件ではどのように考えるのか。どうやら、ネイマンは「なにもいえない」と言っているらしい(p112)。出典が明記されていないし、出版社(統計協会)と監修の日本統計学会に出版を教えて欲しい旨のメールを書いたが、今にいたるまでご返信いただけてない。
- ³⁾ 「設定されないと断定して書いてしまうと、反論をいただくことは承知している。つまり、研究テーマは、統計手法とは独立のロジックを有していること、研究の前提であるからである。冒頭にも記したが、研究仮説と実験、検証方法の関係からいえば、本末転倒である。そのために、我々は解決不可能と思っている/感じていることはそもそも問題にしないのであるという傾向がある、という言い方にしてもいい。
- ⁴⁾ Student は、ギネスビールに勤務していた Gossett が使ったペンネームである。Gossett は、t-分布の発見者である。氏の伝記のエピソードは、SALSBURG, 2001=2006, 2010:54 などを参照。
- ⁵⁾ (1) D. hyoscyamine hydrobromide, と (2) L. hyoscyamine hydrobromide である。臭化水素酸ヒオスアミン L. と D。
- ⁶⁾ A、B を「対応のある2群」として分析してもよいし (paired=TRUE)、diff= B-A として、1 群 (diff) の平均値が0より大であることを検定してもよい。
- ⁷⁾ この「対立仮説」という用語にたいして、「研究仮説」を付与するアプローチもある。『数学嫌いのための社会統計学』のp165では、この Alternate hypothesis を「研究仮説と呼んでいる。『社会調査の応用』2012でも、p32の注で、通常は対立仮説と呼ばれることが多いが津島他(2010)にならって意味がわかりやすいこの(調査仮説)を使うことにする。」とある。
これでは、帰無仮説が棄却されれば、調査仮説が採択される、ということになってしまう。その意味で Alternate hypothesis を「研究仮説」と呼ぶことには賛成できない。
- ⁸⁾ この信頼区間とベイズ信用区間の比較については、日本統計学会 2015:112 や、奥村 2018:51 では、大きく違わない(信用区間の方が少し幅がある?)という説明がある。

参考文献

- アメリカ統計学会会長声明「統計の有意性と P 値に関する ASA 声明」(訳：日本計量生物学会)
<https://www.biometrics.gr.jp/news/all/ASA.pdf>
 原文は：Wasserstein RL, Lazar NA. Editorial: The ASA's statement on p-values: Context, process, and purpose. The American Statistician 2016; 70: 129-133.
- 金井・小林・渡邊, 2012,『社会調査の応用』弘文堂
- 日本統計学会編, 2015,『統計学基礎』改訂版、日本統計協会 (p112 - 3 コラム「信頼係数の解釈について」)
- 中澤港, 2003,『R による統計解析の基礎』ピアソン・エデュケーション
- 柳川暁, 2018,『p 値その正しい理解と適用』近代科学社
- 朝野熙彦, 2017,『ベイズ統計学 Excel から R へステップアップ』朝倉書店
- , 2018,『ベイズ統計学 Excel から RStan へステップアップ』朝倉書店
- 豊田秀樹, 2015,『基礎からのベイズ統計学 ハミルトニアンモンテカルロ法による実践的入門』朝倉書店
- , 2016,『はじめての統計データ分析 ベイズ的〈ポスト p 値時代〉の統計学』朝倉書店
- 松浦健太郎, 2016,『Stan と R でベイズ統計モデリング』, 共立出版
- 三中信宏, 2018,『統計思考の世界：曼荼羅で読み解くデータ解析の基礎』技術評論社
- 岡田謙介, 2017, ASA 声明とこれからの統計学の使われ方ー最近の心理統計分野の動向から,『社会と調査』No19:88-93, https://jasr.or.jp/wp/asr/asrpdf/asr19/asr19_060.pdf
- 奥村晴彦, 2018,『R で楽しむベイズ統計入門』技術評論社 (p51 3.6 信用区間とベイズ信用区間の比較)
- Rouanet 他, 2000, “New Ways in Statistical Methodology”, Peter Lang, Bern
- Schmuller, 2017=2023, “Statistical Analysis with R for Dummy” (訳：笠田実『R で基礎から学ぶ統計学』東京科学同人社)
- Student, 1908, “The Probable Error of a Mean”, Biometrika, Vol. 6, No. 1 (Mar., 1908), pp. 1-25, Oxford University Press on behalf of Biometrika URL: <https://www.jstor.org/stable/2331554>
 Accessed: 10-09-2024 13:09 UTC
- 津島・山口・田邊, 2023,『数学嫌いのための社会統計学 第3版』法律文化社