古典的なノンパラメトリック検定はなにを仮定しているのか

2020 年 2 月 15 日

1　前置き

もともと工学部出身のぼくが, 医学系の研究室に行っておもしろかったことはいろいろある.

その中の一つが, 医学系の人のアーチファクト（artifact）という言葉の使い方だ.

アーチファクトというのは, 辞書的には人工物とか工芸品を指す.

ぼくからしたら「この結果はアーチファクトだよ」とか言われたら褒められているのかな？　みたいな感じがする.

でも違う. 医学系の言葉使いではアーチファクトというのは, 自然現象とか生命現象の本質じゃない人工的な混ざりもの, みたいなニュアンスだ.

ところで, ぼくは統計モデリングというのが好きだ. モデリングというのはデータを取った人からいろいろ話を聞いて, 関数とか確率分布とか微分方程式とか差分方程式とかを組み合わせて, データを取った人のイメージとか目的とかを統計の言葉に翻訳するような作業だ. モデルを作るというのは, ぼくの感覚では仮定を置くことに等しい.

それはもうアーチファクトのかたまりみたいな作業だ.

そのせいかどうかは知らないが, 医学系の人はモデリングよりも検定が好きな印象がある. 統計屋というと, いろんな検定を知っていて, 場合に応じて正しい検定を選べる人, みたいな認識をされることもある.

なかでもノンパラメトリック検定が好きなようだ.

ノンパラメトリック検定とは, 母集団に対して特定の確率分布を仮定しないで検定をする手法の総称とされる.

仮定が少ないほうが客観的な分析でえらいという感覚があるんだと思う.

そこがぼくの好みと相容れない.

でも好みじゃないのでノンパラメトリック検定に関する相談には乗りませんというのも大人げないので, これからちょっとノンパラメトリック検定について勉強していきたいと思っている.

この文書はそういったものだ.

この文書がこれからどういったものになるかは, ぼくもまだわかってない.

でも結びの言葉はなんとなくだけど決めてある.

「仮定を置かないんじゃなくて仮定を明確にして, アーチファクトをもってアーチファクトを制すのだ.」

2　ノンパラメトリック・モデル

ノンパラメトリックモデルという言葉とノンパラメトリック検定という言葉はまったく別物だと思ったほうがよさそうだ.

ノンパラメトリックモデルというのは, パラメータが多すぎるモデルのことを指す.

データのサイズにほぼ比例して, モデルのパラメータが増えるようなモデルはノンパラメトリックモデルと呼ばれる.

以下では, ノンパラメトリックモデルの話はしない.

ノンパラメトリック検定の話をする.

3　仮説検定一般に関する注意点

仮説検定は「確率版背理法」と例えられることもあるが, 現実のデータを扱う以上, 背理法のようにすっきりとはしていない. 仮説検定では, 「帰無仮説 $H_0 : \theta = 0$, 対立仮説 $H_1 : \theta \neq 0$」みたいな形で「仮説」を提示する. しかし, 帰無仮説 H_0 の確率を評価するのに, 帰無分布というのを作らなくちゃいけない.

仮説検定には, 「$H_0 : \theta = 0$」といった**意識されやすい仮定**と検定のやり方（何検定を使うか）を決めた時点で置かれている**意識されにくい仮定**が存在する. 例えば「○○検定を行った」と言った時点で, 「サンプルは正規分布に従う」とか「サンプルは独立同分布とする」とかの意識されにくい仮定はクリアされたことになっている.

仮説検定を行った結果, 帰無仮説が棄却されたというと意識されやすい仮定が棄却された感じがするが, 実は意識されにくい仮定のほうがもっとずれていたのかもしれない.

「H_0 でないから H_1 だ」という背理法ほど, 仮説検定はすっきりしていない.

4　符号検定（Sign Test）

検定というのは帰無分布を一つ決めて, そこからのずれを測っている.

帰無分布を決めなきゃいけないのに, 「特定の確率分布を仮定しない」なんてことができるのか.

どうやらできるようだ.

データ $X_1, X_2, \ldots, X_n$ を生成した分布が連続型の分布関数 F を持つということだけを仮定しよう. θ を F の中央値とする.

その上で母集団の中央値が 0 であるという帰無仮説を検定したい. まず, 帰無仮説 $H_0 : \theta = 0$, 対立仮説 $H_1 : \theta > 0$ の片側検定を考える.

そして,

$$S_n = \sum_{i=1}^{n} I_{\{X_i > 0\}} \tag{1}$$

という統計量を作ることにする. ここで I_A は指示関数（つまり $X_i > 0$ なら 1, さもなくば 0 を返す）である.

S_n は符号が $+$ になった回数を単に数えただけだ.

帰無仮説のもとで, S_n は二項分布 Binomial$(n, 1/2)$ に従う. 帰無仮説のもとで A という事象の起こる確率を $\Pr_{H_0}(A)$ と書くことにすると, この検定の有意水準 α での棄却域は, $\Pr_{H_0}(S_n > k) \leq \alpha$ を満たす k に対して $S_n > k$ となる範囲である.

4.1　シミュレーション

統計ソフト R で上記の符号検定の p 値を計算するには, 次のように書けばよい.

```
theta0 <- 0
n <- length(X)
Sn <- sum(X>theta0)
pbinom(Sn-1,n,0.5,lower.tail = FALSE)
```

4.2　t 検定に対する漸近相対効率

これについては書かないかもしれない.