

検定と信頼区間: 一般論

- 黒木玄
- 2022-05-31~2022-06-02

このノートでは[Julia言語 \(https://julialang.org/\)](https://julialang.org/)を使用している:

- [Julia言語のインストールの仕方の一例 \(https://nbviewer.org/github/genkuroki/msfd28/blob/master/install.ipynb\)](https://nbviewer.org/github/genkuroki/msfd28/blob/master/install.ipynb)

自明な誤りを見つけたら, 自分で訂正して読んで欲しい. 大文字と小文字の混同や書き直しが不完全な場合や符号のミスは非常によくある.

このノートに書いてある式を文字通りにそのまま読んで正しいと思ってしまうとひどい目に会う可能性が高い. しかし, 数が使われている文献には大抵の場合に文字通りに読むと間違っている式や主張が書いてあるので, 内容を理解した上で訂正しながら読んで利用しなければいけない. 実践的に数学を使う状況では他人が書いた式をそのまま信じていけない.

このノートの内容よりもさらに詳しいノートを自分で作ると勉強になるだろう. 膨大な時間を取られることになるが, このノートの内容に関係することで飯を食っていく可能性がある人にはそのためにかけた時間は無駄にならないと思われる.

目次

[1 お勧め解説動画](#)

▼ [2 まとめ](#)

[2.1 P値](#)

[2.2 検定](#)

[2.3 信頼区間](#)

[2.4 nuisanceパラメータがある場合](#)

▼ [3 P値の定義](#)

[3.1 統計モデルの設定](#)

[3.2 P値の定義](#)

▼ [3.3 データの数値以上に極端な値の定義の仕方](#)

[3.3.1 例\(二項分布モデル\)の場合](#)

[3.3.2 例\(正規分布の標本分布モデル\)の場合](#)

[4 P値を使った検定](#)

[5 P値関数を使った信頼区間](#)

[6 信頼区間と検定の表裏一体性](#)

[7 P値関数が「よい」かどうかの判断基準達](#)

[8 Neyman-Pearsonの補題](#)

[9 よくある誤解](#)

In [1]:

```
1 ENV["LINES"], ENV["COLUMNS"] = 100, 100
2 using BenchmarkTools
3 using Distributions
4 using LinearAlgebra
5 using Printf
6 using QuadGK
7 using Random
8 Random.seed!(4649373)
9 using Roots
10 using SpecialFunctions
11 using StaticArrays
12 using StatsBase
13 using StatsFuns
14 using StatsPlots
15 default(fmt = :png, titlefontsize = 10, size = (400, 250))
16 using SymPy
```

```
In [2]: 1 # Override the Base.show definition of SymPy.jl:
2 # https://github.com/JuliaPy/SymPy.jl/blob/29c5bfd1d10ac53014fa7fef468bc8deccadc2fc/src/types.
3
4 @eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::SymbolicObject)
5     print(io, as_markdown("\displaystyle " * sympy.latex(x, mode="plain", fold_short_frac=false)
6 end
7 @eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::AbstractArray{Sym})
8     function toeqnarray(x::Vector{Sym})
9         a = join(["\displaystyle " * sympy.latex(x[i]) for i in 1:length(x)], "\\")
10        "\"\\left[ \\begin{array}{r}$a\\end{array} \\right]\""
11    end
12    function toeqnarray(x::AbstractArray{Sym,2})
13        sz = size(x)
14        a = join([join("\displaystyle " .* map(sympy.latex, x[i,:]), "&") for i in 1:sz[1]],
15        "\"\\left[ \\begin{array}{c} * repeat(\"r\",sz[2]) * \"}\" * a * \"\\end{array}\\right]\""
16    end
17    print(io, as_markdown(toeqnarray(x)))
18 end
```

```
In [3]: 1 safemul(x, y) = x == 0 ? x : x*y
2 safediv(x, y) = x == 0 ? x : x/y
3
4 x ≲ y = x < y || x ≈ y
5
6 mypdf(dist, x) = pdf(dist, x)
7 mypdf(dist::DiscreteUnivariateDistribution, x) = pdf(dist, round(Int, x))
8
9 distname(dist::Distribution) = replace(string(dist), r"{.*}" => "")
10 myskewness(dist) = skewness(dist)
11 mykurtosis(dist) = kurtosis(dist)
12 function standardized_moment(dist::ContinuousUnivariateDistribution, m)
13     μ, σ = mean(dist), std(dist)
14     quadgk(x → (x - μ)^m * pdf(dist, x), extrema(dist)...)[1] / σ^m
15 end
16 myskewness(dist::MixtureModel{Univariate, Continuous}) = standardized_moment(dist, 3)
17 mykurtosis(dist::MixtureModel{Univariate, Continuous}) = standardized_moment(dist, 4) - 3
```

Out[3]: mykurtosis (generic function with 2 methods)

1 お勧め解説動画

P値と検定と信頼区間については次のリンク先の動画での解説が素晴らしいので、閲覧を推奨する:

- 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://youtu.be/vz9cZnB1d1c> (<https://youtu.be/vz9cZnB1d1c>)

信頼区間の解説は40分あたり以降にある。多くの入門的な解説が抱えているP値、検定、信頼区間の解説の難点は以下の2つに要約される:

- 複雑な現実と統計モデルを混同させるような解説が伝統的に普通になってしまっていること。
- 検定と信頼区間の表裏一体性(双対性)が解説されていないこと。

このことが原因がP値も 95% 信頼区間の 95% も数学的フィクションである統計モデル内での確率であることがクリアに説明されておらず、そのせいでP値と 95% 信頼区間の 95% についてまっとうな理解が得られ難くなっている。上で紹介した動画は教科書の説明がまずいことについて明瞭に言及しながら、伝統的な入門的解説が抱えている問題を解消しようとしている。

注意: 上の解説動画内で説明されている事柄を理解すれば、P値と検定と信頼区間について広まってしまった誤解を避けることができる。そして、その後は個別の場合について詳しく勉強するだけの問題になるだろう。

2 まとめ

以下のまとめは「理解が進むたびに何度もこのまとめに戻る」というような使い方をして欲しい。

2.1 P値

P値 (P-value)は以下を与えることによって定義される:

- 現実世界におけるデータの数値 x の生成のされ方に関するパラメータ θ を持つ統計モデル,
- 「データの数値以上に極端な」の意味の定義,
- さらに必要ならば近似計算法.

データの数値 x とパラメータの値 $\theta = \theta_0$ が与えられたとき, P値は

- データの数値 x 以上に極端な値が条件 $\theta = \theta_0$ の下での統計モデル内で生じる確率もしくはその近似値

と定義される. このとき, 条件 $\theta = \theta_0$ は **帰無仮説** (null hypothesis)と呼ばれることがある. (「データの数値以上に極端な」の定義は **対立仮説** (alternative hypothesis)の **集まり** を与えることによって与えらえると考えられる.)

このノートでは以上のように定義されたP値を

- データの数値 x に関する仮説 $\theta = \theta_0$ のP値

と呼ぶことにする.

P値は以下の2つの整合性の指標として使われる:

- 現実世界から得たデータの数値 x ,
- 統計モデル+パラメータの値 $\theta = \theta_0$.

P値が小さいほど, 統計モデル+パラメータの値 θ は現実世界から得たデータの数値 x と整合していないと考える.

2.2 検定

有意水準 と呼ばれる閾値(いきち, しきいち) $0 < \alpha < 1$ が与えられたとき, P値が α 未満ならば,

- 統計モデル+パラメータの値は現実世界から得たデータの数値と整合性がない

とみなす. このとき

- 統計モデル+パラメータの値は現実世界から得たデータによって **棄却** (reject)されたという.

この手続きを **検定** (test, testing, hypothesis testing)と呼ぶ. (P値が有意水準 α 以上になるとき, 「受容 (accept)された」ということがあるが, 混乱の原因になるので, このノートでは用いない. このノートでは「棄却された」「棄却されなかった」の組み合わせを一貫して使うことにする.)

棄却されなかった統計モデルとパラメータの値の組み合わせについては強い結論は何も出せない. 棄却されずにすんだ統計モデルとパラメータの値の組み合わせは単に閾値 α の設定で捨てられずにすんだだけなので, 「棄却されなかった統計モデルとパラメータの値の組み合わせは正しいと考えてよい」と考えることは典型的な誤解になる.

有意水準 α として 5% が非常によく使われているが, それは単に慣習的にそうなっているだけのことで, 5% の有意水準を使うことに科学的な合理性はない.

5% の有意水準の下での結果に一喜一憂することは非科学的な考え方である.

検定の手続きは「科学のお墨付きを得るための手段」ではない!

同じことは次の節で説明する信頼区間についても言える.

2.3 信頼区間

信頼区間 (confidence interval)の文脈で $1 - \alpha$ は **信頼度** (信頼係数)と呼ばれる.

統計モデルが実数パラメータ θ を持つとき, データの数値 x から決まる信頼度 $1 - \alpha$ の信頼区間は

- データの数値 x と有意水準 α で棄却されない統計モデルのパラメータ値 θ 全体の集合

として定義される. (この集合が区間にならない場合には信頼領域(confidence region)と呼んだりする. その集合を含む最小の区間を考える場合もある.)

信頼区間を使うことは, 検定の手続きを無数のパラメータ値 θ 達に対して適用することと同じである.

検定で棄却されなかった場合については強い結論は何も出せないので, 信頼区間は

- 正しさについて判断を保留すべきパラメータ値全体の集合

だとみなされる.

2.4 nuisanceパラメータがある場合

注意: この節の内容はこのノートの内容を超えて先走っている。

実際には統計モデルは興味があるパラメータ θ 以外にパラメータ η を含んでいることがある。(例えば平均パラメータ μ のみに興味があるときの正規分布 $\text{Normal}(\mu, \sigma)$ における標準偏差パラメータ σ .)

その場合にはパラメータ θ の値を決めても、統計モデルの確率分布は唯一に決まらず、パラメータ η の分だけ不定になる。

P値の定義は「データの数値 x 以上に極端な値が帰無仮説 $\theta = \theta_0$ の下での統計モデル内で生じる確率もしくはその近似値」であった。帰無仮説 $\theta = \theta_0$ で統計モデルの確率分布が唯一に決まらない場合にはこのP値の定義を単純に適用することが不可能になる。

このような状況のときに、余計なパラメータ η を **nuisanceパラメータ** (ニューサンスパラメータ, 局外パラメータ, 攪乱パラメータ, 迷惑パラメータ)と呼ぶ(nuisanceは迷惑や妨害を意味する名詞)。

大抵の場合にnuisanceパラメータが存在しているという問題は、P値を使う統計分析の基礎付けが複雑になる主な原因の1つになっている。

nuisanceパラメータへの対処法には例えば以下がある:

- nuisanceパラメータを動かして上限(supremum)を考える。
- 条件付き確率分布に移ってnuisanceパラメータの自由度を消す。
- 最尤法の漸近論を使ってnuisanceパラメータによらずに成立する近似を得る。
- Bayes法を使う。

3 P値の定義

3.1 統計モデルの設定

現実世界のデータの生じ方のモデル化として使用される統計モデルを考える。

例 (二項分布モデル):

- 現実世界のデータ: 当たりとはずれが出るルーレットを n 回まわしたときの当たりの回数 k
- 統計モデル: 二項分布モデル $\text{Binomial}(n, p)$

例 (正規分布の標本分布モデル):

- 現実世界のデータ: S市の中学3年生男子全体から n 人を無作為抽出して測った身長 x_1, \dots, x_n
- 統計モデル: 正規分布のサイズ n の標本分布モデル $\text{Normal}(\mu, \sigma)^n$

このように統計モデルはパラメータ付きの確率分布として与えられる。

3.2 P値の定義

データ x のモデル化になっているパラメータ θ を持つ統計モデルが与えられていると仮定する。

データの数値 x_0 とパラメータの数値 $\theta = \theta_0$ が与えられたとき、

- データの数値 x_0 から定まる仮説 $\theta = \theta_0$ のP値

を次によって定める:

- データの数値 x 以上に極端な値が仮説 $\theta = \theta_0$ 下の統計モデル内で生じる確率もしくはその近似値。

「仮説 $\theta = \theta_0$ 下の統計モデル内におけるデータの数値 x 以上に極端な値」の定義は目的ごとに別に与えられる。その概略については次の節を見よ。

3.3 データの数値以上に極端な値の定義の仕方

データの数値 x_0 とパラメータの数値 $\theta = \theta_0$ が与えられたとき、統計モデルとそのパラメータ値が与える確率分布において、 x がデータの数値 x_0 以上に極端な値であることに定義を目的に合わせて適切に定義する。

以下の条件が「仮説 $\theta = \theta_0$ の下で x がデータの数値 x_0 以上に極端な値であること」の定義としてよく使われる:

- x はデータの数値 x_0 以上(または以下)である。
- x が統計モデルの確率で生じる確率(もしくはその密度)がデータの数値 x_0 以下である。

x が実数値の変数ではなく、 \mathbb{R}^n に属する変数の場合には、 x の実数値関数 $S(x|\theta)$ を用意して、

- $S(x|\theta_0)$ はデータの数値での値 $S(x_0|\theta_0)$ 以上(もしくは以下)である.

という条件で「仮説 $\theta = \theta_0$ の下で x がデータの数値 x_0 以上に極端な値であること」を定義することが多い. 函数 $S(x|\theta_0)$ は **検定統計量** と呼ばれ, 目的ごとに適切に選択する必要がある.

3.3.1 例(二項分布モデル)の場合

データの数値「 n 回中 k_0 回成功」と成功確率パラメータの数値 $p = p_0$ が与えられているとする. このとき, 「仮説 $p = p_0$ の下での二項分布モデル内での成功回数 k がデータの数値 k_0 以上に極端であること」を以下のように様々な方法で定義できる:

(1) $k \geq k_0$ (または $k \leq k_0$). (この変種でClopper-Pearsonの信頼区間が得られる.)

(2) $P(k|n, p_0)$ を二項分布の確率質量函数とすると, $P(k|n, p_0) \leq P(k_0|n, p_0)$ (Sterneの信頼区間の場合).

この他にも正規分布近似(中心極限定理)を使って定義することもできる.

(3) 二項分布 $\text{Binomial}(m, p)$ に関する中心極限定理によれば, $(k - np)/\sqrt{np(1-p)}$ に対応する確率変数は近似的に標準正規分布に従う. 「標準正規分布 $\text{Normal}(0, 1)$ に従う確率変数 Z の値がデータの数値 k_0 以上に極端であること」を

$$|Z| \geq \frac{|k_0 - np_0|}{\sqrt{np_0(1-p_0)}}$$

という条件で定め, こうなる確率を標準正規分布を使って計算する(近似の一種, Wilsonの信頼区間の場合).

(3)' 上と同値な次の条件を使うこともある:

$$Z^2 \geq \frac{(k_0 - np_0)^2}{np_0(1-p_0)}.$$

$Z \sim \text{Normal}(0, 1)$ のとき, Z^2 は自由度 1 の χ^2 分布に従うので, こうなる確率を自由度 1 の χ^2 分布を用いて計算する(これも近似の一種). さらに上の条件は次とも同値である(Pearsonの χ^2 検定の場合):

$$Z^2 \geq \frac{(k_0 - np_0)^2}{np_0} + \frac{((n - k_0) - n(1 - p_0))^2}{n(1 - p_0)}.$$

この不等式の右辺を **Pearsonの χ^2 統計量** と呼ぶ. これはそのように呼ばれる検定統計量達

$$\chi^2 = \sum \frac{(\text{観測値} - \text{期待値})^2}{\text{期待値}}$$

の特別な場合に過ぎない.

(4) 上の(3)における右辺の分母の p_0 を $\hat{p} = k_0/n$ で置き換えて得られる次の条件で定義することもある(Waldの信頼区間の場合):

$$|Z| \geq \frac{|k_0 - np_0|}{\sqrt{n\hat{p}(1-\hat{p})}}.$$

以上のように, 同じ二項分布モデルのP値の定義の仕方も沢山ある. 基本的にどれを使ってもよい. ユーザー側は自分の目的に合わせて合理的だと考えられるものを自由に使えばよい.

3.3.2 例(正規分布の標本分布モデル)の場合

注意: この節の内容は別のノートで詳しく説明する. この段階では以下で説明する複雑な式を覚える必要はない.

データの数値 $x_0 = (x_1^{(0)}, \dots, x_n^{(0)}) \in \mathbb{R}^n$ と興味があるパラメータの数値 $\mu = \mu_0$ が与えられているとする. このとき, 「仮説 $\mu = \mu_0$ の下での正規分布の標本分布モデル内での $x \in \mathbb{R}^n$ の値がデータの数値 x_0 以上に極端であること」を以下の方法で定義すると便利であることが知られている.

まず, T 統計量 $T(x|\mu)$ を次のように定める:

$$T(x|\mu) = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

そして, 「仮説 $\mu = \mu_0$ の下で, x の値がデータの数値 x_0 以上に極端であること」を

$$|T(x|\mu_0)| \geq |T(x_0|\mu_0)|$$

という条件で定める. これは大雑把に言って,

- x が (μ_0, \dots, μ_0) からデータの数値 x_0 以上に離れている.

という意味の条件になっている.

これが便利なのは, 仮説 $\mu = \mu_0$ の下での正規分布の標本分布モデル内で $T(x|\mu_0)$ に対応する確率変数が自由度 $n - 1$ の t 分布に従うからである.

4 P値を使った検定

5 P値函数を使った信頼区間

6 信頼区間と検定の表裏一体性

7 P値函数が「よい」かどうかの判断基準達

8 Neyman-Pearsonの補題

9 よくある誤解

$$\sqrt{x}/\sqrt{2}$$

In []:

1	
---	--