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/)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)

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

In [1]:

```
ENV["LINES"], ENV["COLUMNS"] = 100, 100
using BenchmarkTools
using Distributions
using LinearAlgebra
using Printf
using QuadGK
using Random
Random.seed!(4649373)
using Roots
using SpecialFunctions
using StaticArrays
using StatsBase
using StatsFuns
using StatsPlots
default(fmt = :png, titlefontsize = 10, size = (400, 250))
using SymPy
```

In [2]:
```
# Override the Base.show definition of SymPy.jl:
# https://github.com/JuliaPy/SymPy.jl/blob/29c5bfd1d10ac53014fa7fef468bc8deccadc2fc/src/types.

@eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::SymbolicObject)
    print(io, as_markdown("\\displaystyle " * sympy.latex(x, mode="plain", fold_short_frac=fal
end
@eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::AbstractArray{Sym})
    function toeqnarray(x::Vector{Sym})
        a = join(["\\displaystyle " * sympy.latex(x[i]) for i in 1:length(x)], "\\\\")
        """\\left[ \\begin{array}{r}$a\\end{array} \\right]"""
    end
    function toeqnarray(x::AbstractArray{Sym,2})
        sz = size(x)
        a = join([join("\\displaystyle " .* map(sympy.latex, x[i,:]), "&") for i in 1:sz[1]],
        "\\left[ \\begin{array}{" * repeat("r",sz[2]) * "}" * a * "\\end{array}\\right]"
    end
    print(io, as_markdown(toeqnarray(x)))
end
```

In [3]:
```
safemul(x, y) = x == 0 ? x : x*y
safediv(x, y) = x == 0 ? x : x/y

x ≶ y = x < y || x ≈ y

mypdf(dist, x) = pdf(dist, x)
mypdf(dist::DiscreteUnivariateDistribution, x) = pdf(dist, round(Int, x))

distname(dist::Distribution) = replace(string(dist), r"{.*}" => "")
myskewness(dist) = skewness(dist)
mykurtosis(dist) = kurtosis(dist)
function standardized_moment(dist::ContinuousUnivariateDistribution, m)
    μ, σ = mean(dist), std(dist)
    quadgk(x -> (x - μ)^m * pdf(dist, x), extrema(dist)...)[1] / σ^m
end
myskewness(dist::MixtureModel{Univariate, Continuous}) = standardized_moment(dist, 3)
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 $\theta = \theta_0$ が与えらえたとき, 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「科学的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る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 $\mathrm{Normal}(\mu, \sigma)$ 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つに決まらず, 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つ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,　撹乱パラメータ, 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 $\mathrm{Binomial}(n, p)$

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 $x_1, \ldots, x_n$
- 統計モデル: 正規分布のサイズ n の標本分布モデル $\mathrm{Normal}(\mu, \sigma)^n$

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 の生成のされ方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 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 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

P値の定義を確定させるためには, データの数値 x 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従う確率変数 X の値がデータの数値 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設定する必要がある.

以下の条件が「仮説 $\theta = \theta_0$ の下での統計モデルに従う確率変数 X がデータの数値 x 以上に極端な値であること」の定義としてよく使われる:

(1) $X \geq x$ (または $X \leq x$).

(1)' $X \geq x$ と $X \leq x$ の確率が小さい方の条件.

(2) モデル内での X の値が生じる確率(もしくはその密度)がデータの数値 x 以下である.

X が実数値の確率変数ではなく, $\mathbb{R}^n$ 値の確率変数の場合には, X の実数値函数 $S(X|\theta_0)$ を用意して,

(3) $S(X|\theta_0) \geq S(x|\theta_0)$ (もしくは $S(X|\theta_0) \leq S(x|\theta_0)$)

という条件で「仮説 $\theta = \theta_0$ の下での統計モデルに従う確率変数 X がデータの数値 x 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.

他にも正規分布近似を使う方法も多用される.

3.3.1 例(二項分布モデル)の場合

データの数値「n 回中 k 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. さらに, 仮説 $p = p_0$ 下の二項分布 $\mathrm{Binomial}(n, p_0)$ に従う確率変数 K を用意する:

$$K \sim \mathrm{Binomial}(n, p_0).$$

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 の値がデータの数値 k 以上に極端であること」を以下のように, 互いに同値でない様々な方法で定義できる:

(1) $K \geq k$ (もしくは $K \leq k$)という条件で「k 以上に極端」の意味を定義する. この定義は仮説 $p \leq p_0$ (もしくは $p \geq p_0$)の **片側検定** (one-tailed test, one-sided test) で使われる.

(1)' 仮説 $p = p_0$ の下での二項分布モデル内での $K \geq k$ と $K \leq k$ の確率の小さい方の2倍(と 1 の小さい方)を **両側検定** のP値として使う(Clopper-Pearsonの信頼区間の場合). 以下の定義はどれも両側検定の場合になっている. **検定は通常両側検定を使用する.**

(2) K の値がモデル内で生じる確率がデータの数値 k がモデル内で生じる確率以下になる(二項分布の確率質量函数を $P(k|n, p_0)$ と書くときの $P(K|n, p_0) \leq P(k|n, p_0)$)という条件で「$k$ 以上に極端」の意味を定義する(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 $\mathrm{Binomial}(m, p_0)$ に関する中心極限定理によれば, $(K - np_0)/\sqrt{np_0(1-p_0)}$ は np と $n(1-p)$ が十分に大きければ近似的に標準正規分布に従う. 「標準正規分布 $\mathrm{Normal}(0, 1)$ に従う確率変数 Z の値がデータの数値以上に極端であること」を

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

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してP値とする(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:

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

$Z \sim \mathrm{Normal}(0, 1)$ のとき, Z^2 は自由度 1 のχ²分布に従うので, こうなる確率を自由度 1 のχ²分布を用いて計算してP値とする(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χ²検定の場合):

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

これの不等式の右辺を **Pearsonのχ²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 をパラメータの推定量 $\hat{p} = k/n$ で置き換えて得られる次の条件で「データの数値以上に極端であること」を定義することもある(Waldの信頼区間の場合):

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

以上のように, 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 $x_0 = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right) \in \mathbb{R}^n$ 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 $(\mu_0, \ldots, \mu_0)$ 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.

3.4 P値は帰無仮説下の統計モデルのデータの数値との整合性の指標

データの数値 x から計算される帰無仮説 $\theta = \theta_0$ のP値は, 仮説 $\theta = \theta_0$ 下の統計モデルとデータの数値 x の整合性の指標である.

P値が小さいことは, 仮説 $\theta = \theta_0$ 下の統計モデルとデータの数値 x があまり整合していないと考える.

4 P値を使った検定

データ x の生成のされ方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.

さらに有意水準と呼ばれる **閾値**(いきち, しきいち) $0 < \alpha < 1$ が与えられていると仮定する. α は目的に合わせて適当に小さな値としておく. (有意水準として 5% がよく用いられているが, そのことに科学的な合理性はない.)

データの数値 x とパラメータの数値 $\theta = \theta_0$ が与えられているとき, 仮説 $\theta = \theta_0$ のP値を求め, P値が α 未満になるとき, 仮説 $\theta = \theta_0$ 下の統計モデルは **棄却** (reject)されたという.

この手続きを **仮説検定** (Hypothesis tesitng)もしくは単に **検定** と呼ぶ.

すなわち, 検定とは, ある閾値を設けて, その閾値以上に整合性がないモデルのパラメータ値を捨て去る手続きのことである.

ただし, 閾値を設けて捨て去る行為なので間違う危険性がある. 有意水準が小さなほどそのリスクは小さくなるが, その分だけ, 科学的に興味深い結果を見逃してしまうリスクが増える.

5 P値函数を使った信頼区間

有意水準 α の検定の手続きをパラメータ θ のすべての値に適用したとき, 棄却されなかったパラメータ値全体の集合をパラメータ θ に関する **信頼度** (信頼係数) $1-\alpha$ の **信頼区間** (confidence interval)と呼ぶ. ($\alpha = 5\%$ のとき, 信頼度 $1-\alpha$ の信頼区間を 95% 信頼区間と呼ぶことが多い.)

有意水準 α の検定の手続きでパラメータの値 $\theta = \theta_0$ が棄却されることは, 仮説 $\theta = \theta_0$ 下の統計モデルのデータの値 x との整合性(P値)が閾値 α 未満になることであった. そのような状況を

- 仮説 $\theta = \theta_0$ 下の統計モデルとデータの値 x との整合性が無さすぎる

と言うことにしよう. このスタイルの下では, 信頼区間は

- 統計モデルの下で, データの数値 x との整合性が無さすぎないパラメータ値全体の集合

であると言える.

「整合性が無さすぎること」は「正しいこと」を意味しないし, 「正しい可能性が高いこと」も意味しない. 信頼区間に含まれるパラメータ値の下での統計モデルの妥当性については判断を保留しなければいけない.

例: 例えば, パラメータ θ がある治療法の治療効果を意味するパラメータであったとしよう. そのとき, 現実世界における調査で得たデータの数値 x に関する θ の信頼区間は

- その区間に含まれる治療効果の数値の各々については現実の正しさについては判断を保留する.
- その区間のどれかの値が真の治療効果であっても大丈夫なようにしておく.
- 得られたデータの数値へのその統計モデルの使用が妥当でない可能性についても常に注意を払う.

のような使い方をすることが妥当だと思われる.

6 信頼区間と検定の表裏一体性

6.1 検定における棄却領域の合併と信頼区間全体の合併は互いに相手の補集合

データ x の生成のされ方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設定されていると仮定し, データの数値 x に関する仮説 $\theta = \theta_0$ のP値 $\mathrm{pvalue}(x|\theta_0)$ が定義されていると仮定し, 有意水準 α が与えられているとする.

データの数値 x が与えらえたとき, パラメータの数値 $\theta = \theta_0$ をP値 $\mathrm{pvalue}(x|\theta_0)$ に対応させる函数

$$\theta_0 \mapsto \mathrm{pvalue}(x|theta_0)$$

を **P値函数** (P-value function)と呼ぶ. P値函数は無数の仮説 $\theta = \theta_0$ 達の検定結果の情報をすべて持っているので, 信頼区間の情報もそこに含まれる.

このとき, 仮説 $\theta = \theta_0$ を有意水準 α で棄却するようなデータの数値全体の集合を **棄却領域** (rejection region)と呼び, 次のように書くことにする:

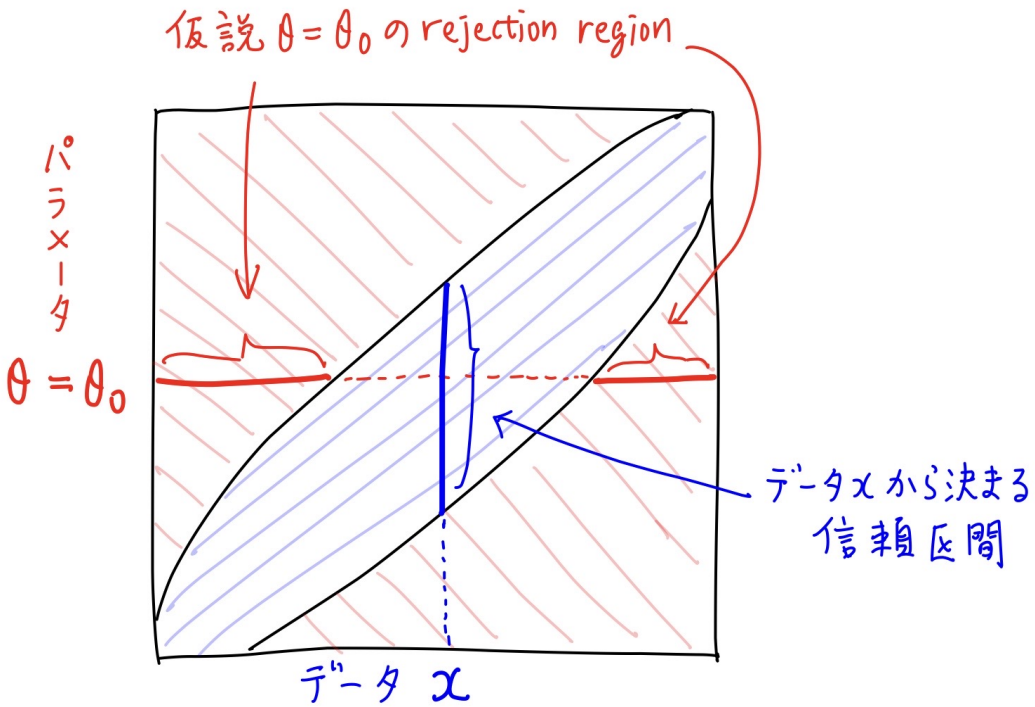
$$R_\alpha(\theta_0) = \{\, x \mid \mathrm{pvalue}(x|\theta_0) < \alpha \,\}.$$

信頼度 $1-\alpha$ の信頼区間 $C_\alpha(x)$ は

$$C_\alpha(x) = \{\, \theta_0 \mid \mathrm{pvalue}(x|\theta_0) \geq \alpha \,\}$$

と書ける. これらの関係は以下の図のようになっている.

以下の図を見れば, すべてのデータの数値 x について信頼区間 $C_\alpha(x)$ を与えることと, すべてのパラメータの数値 $\theta = \theta_0$ についての棄却領域 $R_\alpha(\theta_0)$ を与えることが同じであることがわかる. そして, 図中の赤の領域と青の領域の境界上でのP値 $\mathrm{pvalue}(x|\theta_0)$ の値が有意水準 α になる.



6.2 仮説 $\theta = \theta_0$ 下の統計モデル内でパラメータ値 $\theta = \theta_0$ が信頼区間に含まれる確率

(x, θ_0) がこの図中の(信頼区間を含む)青の領域の点であることは, 仮説 $\theta = \theta_0$ がデータの数値 x によって棄却されないことと同値であり, さらに, データの数値 x から決まる信頼区間にパラメータ値 $\theta = \theta_0$ が含まれることとも同値である.

X がパラメータ値 $\theta = \theta_0$ の統計モデルに従う確率変数であると仮定する. X は仮説 $\theta = \theta_0$ 下の統計モデル内で生成された仮想的なデータの数値だと考えられる.

P値の定義より $\mathrm{pvalue}(X|\theta_0) < \alpha$ となる確率は α もしくはその近似になるので, $\mathrm{pvalue}(X|\theta_0) \geq \alpha$ となる確率は $1 - \alpha$ もしくはその近似値になる.

そして, $\mathrm{pvalue}(X|\theta_0) \geq \alpha$ という条件は仮説 $\theta = \theta_0$ 下の統計モデル内で生成されたデータの値 X から決まる信頼度 $1 - \alpha$ の信頼区間に θ_0 が含まれることと同値である.

ゆえに, 仮説 $\theta = \theta_0$ 下の統計モデル内で生成されたデータの値 X から決まる信頼度 $1 - \alpha$ の信頼区間にパラメータ値 $\theta = \theta_0$ が含まれる確率は $1 - \alpha$ またはその近似値になる.

例えば, 大雑把に言うと, 95% 信頼区間に統計モデル内でデータを生成したパラメータ値が含まれる確率は 95% またはその近似値になる.

注意: この事実はP値の定義より, 95% 信頼区間の 95% が確率とみなされることを意味している. ただし, その確率は数学的フィクションである **統計モデル内で測った確率** に過ぎない. 何らかの特別な理由があって, 現実のデータの数値 x の生成のされ方を統計モデルが忠実に再現していると考えられるならば, 95% という数値は現実においても意味を持ち得る. しかし, そうでない場合いは, 単なる **モデル内確率** であることに十分な注意を払う必要がある.

注意: 「95% 信頼区間の 95% は確率ではなく, 割合である」というようなことが多くの教科書に書かれているようだが, 以上の説明を読めば分かるように自明に誤りである.

注意: 「95% 信頼区間の 95% は確率ではなく, 割合である」と言いたい人達は, データの数値 x はすでに確定した数値になっており, 確率的に揺らがないので, データの数値 x から計算される信頼区間も確率的に揺らがないので, パラメータ値 $\theta = \theta_0$ が信頼区間に含まれる確率を考えることはできない, というようなことを言いたいのかもしれないが, 上の説明を読めば分かるように的を外している. 95% 信頼区間の 95% は数学的フィクションであるモデル内確率であり, モデル内で生成されたデータの数値は確率変数になっている. 現実とモデルを混同するという典型的に非科学的な考え方に陥っている疑いもある.

文献: P値函数の使い方については次の文献が詳しい:

- Timothy L. Lash, Tyler J. VanderWeele, Sebastien Haneuse, and Kenneth J. Rothman.
  Modern Epidemiology, 4th edition, 2020. Google (https://www.google.com/search?q=Modern+Epidemiology+4th)

第4版よりも古い版でもよい.

6.3 二項分布モデルでの視覚化

In [4]:

```
# Clopper-Pearson

function pvalue_clopper_pearson(n, k, p)
    bin = Binomial(n, p)
    min(1, 2cdf(bin, k), 2ccdf(bin, k-1))
end

# Sterne

_pdf_le(x, (dist, y)) =  pdf(dist, x) ≲ y

function _search_boundary(f, x0, Δx, param)
    x = x0
    if f(x, param)
        while f(x - Δx, param) x -= Δx end
    else
        x += Δx
        while !f(x, param) x += Δx end
    end
    x
end

function pvalue_sterne(dist::DiscreteUnivariateDistribution, x)
    Px = pdf(dist, x)
    Px == 0 && return Px
    Px == 1 && return Px
    m = mode(dist)
    Px ≈ pdf(dist, m) && return one(Px)
    if x < m
        y = _search_boundary(_pdf_le, 2m - x, 1, (dist, Px))
        cdf(dist, x) + ccdf(dist, y-1)
    else # x > m
        y = _search_boundary(_pdf_le, 2m - x, -1, (dist, Px))
        cdf(dist, y) + ccdf(dist, x-1)
    end
end
pvalue_sterne(n, k, p) = pvalue_sterne(Binomial(n, p), k)

# Wilson

function pvalue_wilson(n, k, p)
    z = safediv(k - n*p, √(n*p*(1-p)))
    2ccdf(Normal(), abs(z))
end

# Wald

function pvalue_wald(n, k, p)
    z = safediv(k - n*p, √(k*(n-k)/n))
    2ccdf(Normal(), abs(z))
end
```

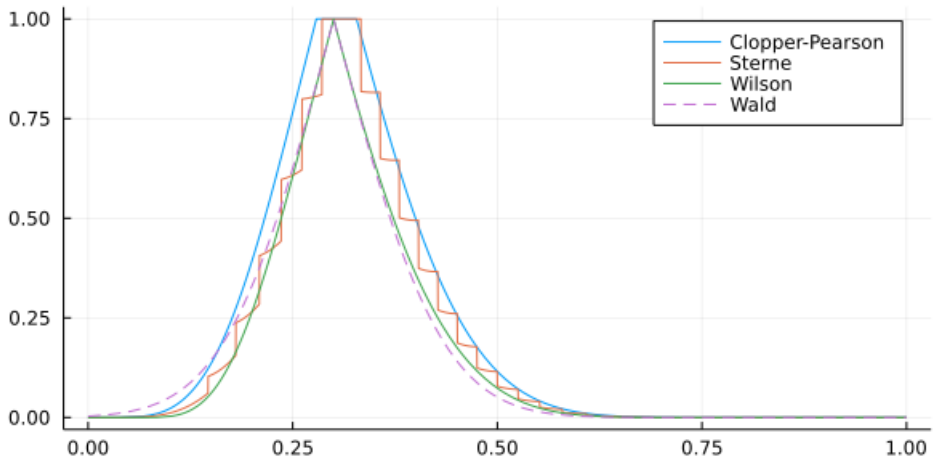
Out[4]: pvalue_wald (generic function with 1 method)

In [5]:

```
# P値函数（データ k を固定した場合）

n, k = 20, 6
plot(p -> pvalue_clopper_pearson(n, k, p), 0, 1; label="Clopper-Pearson")
plot!(p -> pvalue_sterne(n, k, p), 0, 1; label="Sterne")
plot!(p -> pvalue_wilson(n, k, p), 0, 1; label="Wilson")
plot!(p -> pvalue_wald(n, k, p), 0, 1; label="Wald", ls=:dash)
plot!(; size=(600, 300))
```

Out[5]:



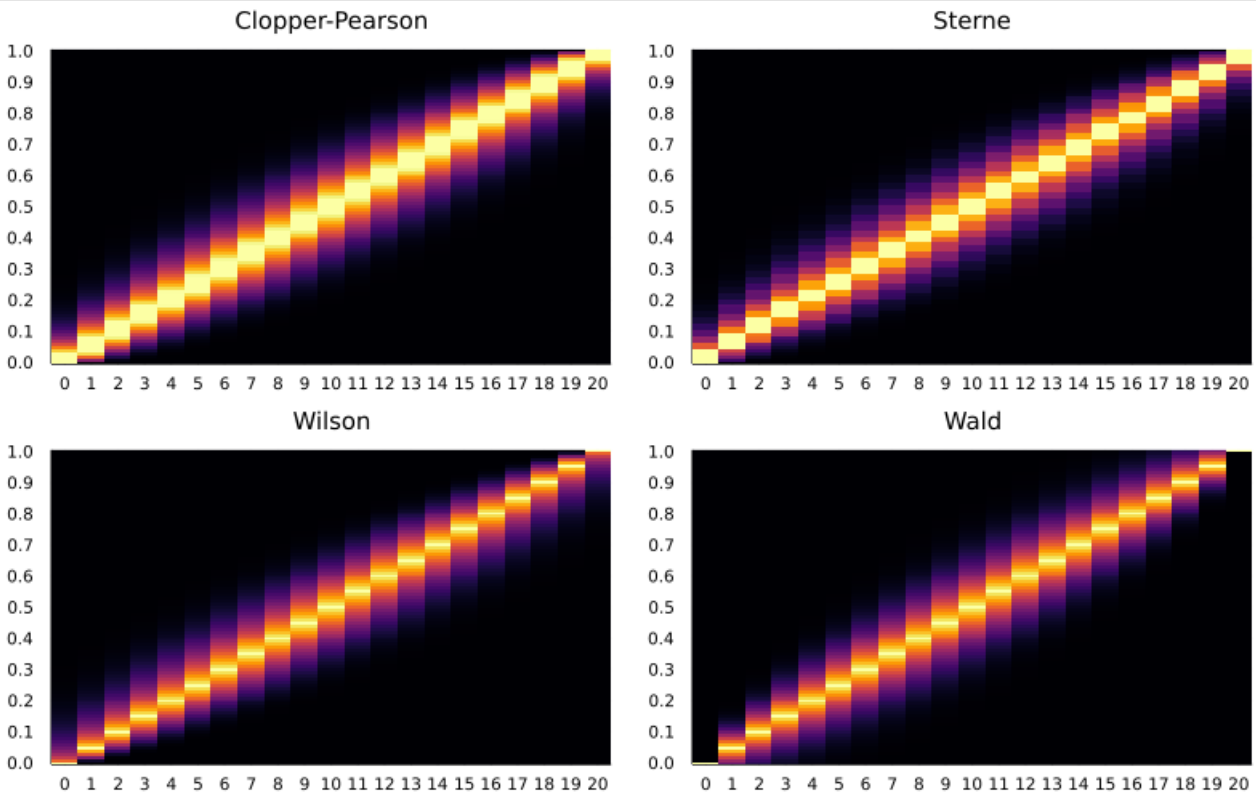
In [6]:

```
# pvalue(x|p) 達のヒートマップ

n = 20
k = 0:n
p = 0:0.01:1

P1 = heatmap(k, p, (k, p)->pvalue_clopper_pearson(n, k, p);
    colorbar=false, title="Clopper-Pearson")
P2 = heatmap(k, p, (k, p)->pvalue_sterne(n, k, p);
    colorbar=false, title="Sterne")
P3 = heatmap(k, p, (k, p)->pvalue_wilson(n, k, p);
    colorbar=false, title="Wilson")
P4 = heatmap(k, p, (k, p)->pvalue_wald(n, k, p);
    colorbar=false, title="Wald")
plot(P1, P2, P3, P4; size=(800, 500), layout=(2, 2),
    xtick=0:20, ytick=0:0.1:1, tickfontsize=7)
```

Out[6]:

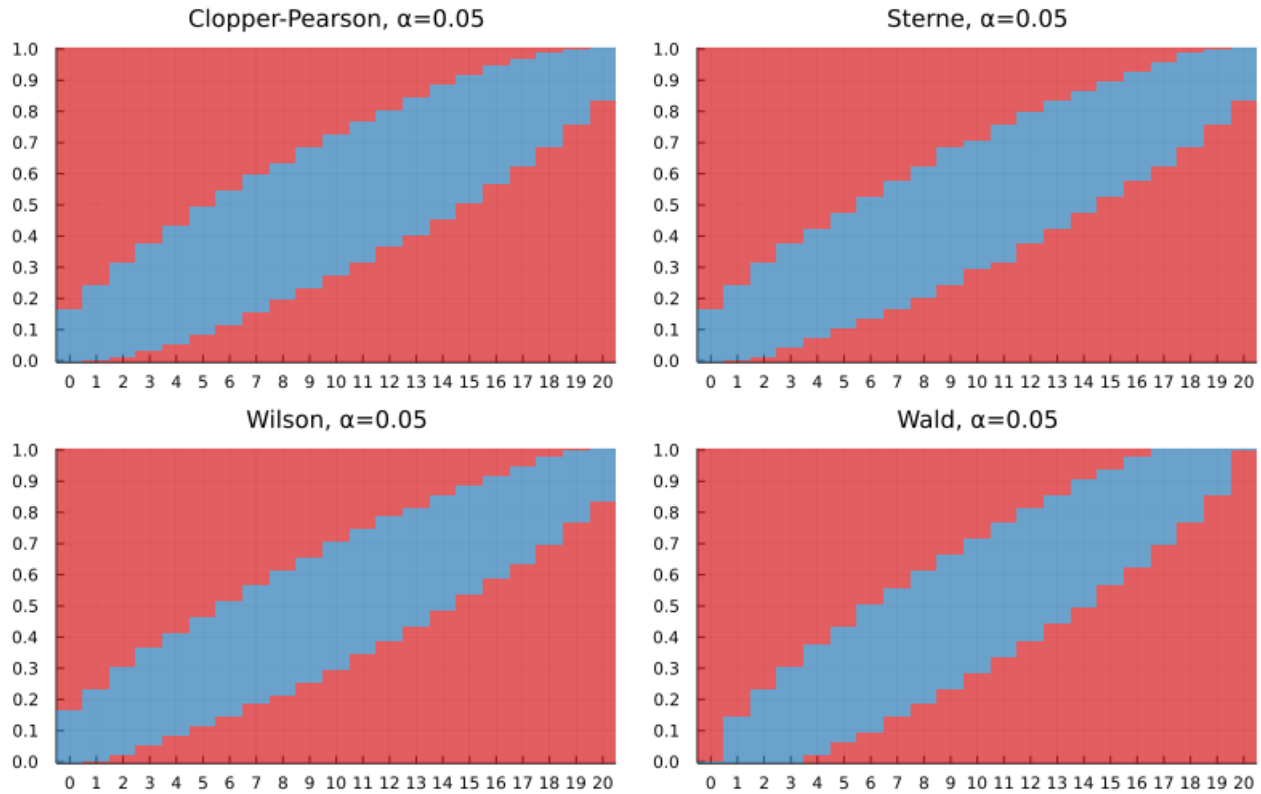


In [7]:
```
# pvalue(x|p) ≥ α のヒートマップ

α = 0.05
n = 20
k = 0:n
p = 0:0.01:1

P1 = heatmap(k, p, (k, p)->pvalue_clopper_pearson(n, k, p) ≥ α;
    colorbar=false, title="Clopper-Pearson, α=$α", c=:RdYlBu_4, alpha=0.7)
P2 = heatmap(k, p, (k, p)->pvalue_sterne(n, k, p) ≥ α;
    colorbar=false, title="Sterne, α=$α", c=:RdYlBu_4, alpha=0.7)
P3 = heatmap(k, p, (k, p)->pvalue_wilson(n, k, p) ≥ α;
    colorbar=false, title="Wilson, α=$α", c=:RdYlBu_4, alpha=0.7)
P4 = heatmap(k, p, (k, p)->pvalue_wald(n, k, p) ≥ α;
    colorbar=false, title="Wald, α=$α", c=:RdYlBu_4, alpha=0.7)
plot(P1, P2, P3, P4; size=(800, 500), layout=(2, 2),
    xtick=0:20, ytick=0:0.1:1, tickfontsize=7)
```

Out[7]:



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

In [ ]:
```
1
```