

P値と仮説検定と信頼区間: 一般論

- 黒木玄
- 2022-05-31～2022-06-05, 2024-01-06, 2024-05-28, 2025-05-16

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 \(https://nbviewer.org/github/genkuroki/msfd28/blob/master/install.ipynb\)](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、数学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。

このノートブックはGoogle Colabで実行できる

(<https://colab.research.google.com/github/genkuroki/Statistics/blob/master/2022/08%20Hypothesis%20testing%20and%20confidence%20General%20theory.ipynb>).

目次

- ▼ 1 お勧め解説動画とお勧め文献
 - 1.1 お勧め解説動画
 - 1.2 お勧め文献: P値に関するASA声明
- ▼ 2 まとめ
 - 2.1 P値
 - 2.2 S値
 - 2.3 検定
 - 2.4 信頼区間
 - 2.5 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」の意味の定義の仕方
 - 3.4 例(二項分布モデルの場合)
 - 3.5 例(正規分布の標本分布モデルの場合)
 - 3.6 P値は帰無仮説下の統計モデルのデータの数値との相性の良さ(compatibility)の指標
- 4 検定
- ▼ 5 第一種の過誤(αエラー)
 - 5.1 第一種の過誤(αエラー)の定義
 - 5.2 第一種の過誤(αエラー)の確率
 - 5.3 二項分布モデルの4種のP値に関する第一種の過誤の確率のグラフ
- ▼ 6 信頼区間
 - 6.1 信頼区間のP値もしくは検定を用いた定義
 - 6.2 信頼区間の使い方
- ▼ 7 信頼区間と検定の表裏一体性
 - 7.1 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
 - 7.2 信頼区間と検定の表裏一体性について書かれた教科書の例
 - 7.3 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
 - 7.4 二項分布モデルでのP値函数の視覚化
- ▼ 8 統計モデルやP値函数が「よい」かどうかの判断基準達
 - 8.1 計算方法はシンプルな方がよい
 - 8.2 頑健な方がよい
 - 8.3 第一種の過誤の確率は有意水準に近い方がよい
 - 8.4 第一種の過誤の確率は有意水準以下である方がよい
 - 8.5 検出力は高い方がよい
- ▼ 9 Neyman-Pearsonの仮説検定
 - 9.1 パラメータを持たない2つの統計モデルに関する仮説検定
 - 9.2 Neyman-Pearsonの補題
 - 9.3 Neyman-Pearsonの補題の証明
 - 9.4 尤度比検定の例: 正規分布モデルの場合
 - 9.5 Neyman-Pearsonの仮説検定に関する極端な解釈の普及の問題

9.6 このノートの立場

▼ 10 よくある誤解

- [10.1 P値の定義と使い方の復習](#)
- [10.2 問題: P値に関するよくある誤解](#)
- [10.3 信頼区間の定義の復習と使い方の復習](#)
- [10.4 問題: 信頼区間にに関するよく見る誤解](#)
- [▼ 10.5 平均の信頼区間達の視覚化](#)
 - [10.5.1 正規分布の標本達から得られる平均の信頼区間達](#)
 - [10.5.2 ガンマ分布の標本から得られる平均の信頼区間達](#)
 - [10.5.3 対数正規分布の標本達から得られる平均の信頼区間達](#)
- [10.6 問題解答例: P値に関するよくある誤解](#)
- [10.7 問題解答例: 信頼区間にに関するよく見る誤解](#)

In [1]:  # Google Colabと自分のパソコンの両方で使えるようにするための工夫

```
1 # Google Colabと自分のパソコンの両方で使えるようにするための工夫
2
3 import Pkg
4
5 """すでにPkg.add済みのパッケージのリスト"""
6 const packages_added = [info.name for (uuid, info) in Pkg.dependencies() if info.is_direct]
7
8 """必要ならPkg.assした後にusingしてくれる関数"""
9 function _using(pkg::AbstractString)
10    if pkg in packages_added
11        println("# $(pkg).jl is already added.")
12    else
13        println("# $(pkg).jl is not added yet, so let's add it.")
14        Pkg.add(pkg)
15    end
16    println("> using $(pkg)")
17    @eval using $(Symbol(pkg))
18 end
19
20 """必要ならPkg.addした後にusingしてくれるマクロ"""
21 macro _using(pkg) :(_using($(string(pkg)))) end
22
23 isdir("images") || mkdir("images")
24 ENV["LINES"], ENV["COLUMNS"] = 100, 100
25 using Base.Threads
26 ## @_using BenchmarkTools
27 @_using Distributions
28 using LinearAlgebra
29 using Printf
30 @_using QuadGK
31 ## @_using RCall # requires the R language.
32 ## @_rlibrary exactci # requires the package exactci of R.
33 using Random
34 Random.seed!(4649373)
35 ## @_using Roots
36 @_using SpecialFunctions
37 @_using StaticArrays
38 @_using StatsBase
39 @_using StatsFuns
40 @_using StatsPlots
41 default(fmt = :png, size = (400, 250),
42         titlefontsize = 10, plot_titlefontsize = 10)
43 ## @_using SymPy
```

```
# Distributions.jl is already added.
> using Distributions
# QuadGK.jl is already added.
> using QuadGK
# SpecialFunctions.jl is already added.
> using SpecialFunctions
# StaticArrays.jl is already added.
> using StaticArrays
# StatsBase.jl is already added.
> using StatsBase
# StatsFuns.jl is already added.
> using StatsFuns
# StatsPlots.jl is already added.
> using StatsPlots
```

```
In [2]: # Override https://github.com/jverzani/SymPyCore.jl/blob/main/src/SymPy/show_sympy.jl#L31-
#=#
3 @eval SymPy begin
4     function Base.show(io::IO, ::MIME"text/latex", x::SymbolicObject)
5         out = _sympy_.latex(↓(x), mode="inline", fold_short_frac=false)
6         out = replace(out, r"\frac{⇒}{\dfrac{}}")
7         print(io, string(out))
8     end
9 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)

In [4]:

```

1 # 二項分布モデルの4種のP値函数
2
3 # Clopper-Pearson
4
5 function pvalue_clopper_pearson(n, k, p)
6     bin = Binomial(n, p)
7     min(1, 2cdf(bin, k), 2ccdf(bin, k-1))
8 end
9
10 # Sterne
11
12 _pdf_le(x, (dist, y)) = pdf(dist, x) ≈ y
13
14 function _search_boundary(f, x0, Δx, param; maxiters=10^7)
15     x = x0
16     if f(x, param)
17         for _ in 1:maxiters
18             !f(x - Δx, param) && return x
19             x -= Δx
20         end
21     else # if !f(x, param)
22         for _ in 1:maxiters
23             x += Δx
24             f(x, param) && return x
25         end
26     end
27     error("""
28         _search_boundary($f, $x0, $Δx, param = $param; maxiters = $maxiters) \
29         has exceeded the maximum number $maxiters of iterations."")
30 end
31
32 function pvalue_sterne(dist::DiscreteUnivariateDistribution, x)
33     Px = pdf(dist, x)
34     0 < Px < 1 || return Px
35     m = mode(dist)
36     Px ≈ pdf(dist, m) && return one(Px)
37     if x < m
38         y = _search_boundary(_pdf_le, 2m - x, 1, (dist, Px))
39         cdf(dist, x) + ccdf(dist, y-1)
40     else # x > m
41         y = _search_boundary(_pdf_le, 2m - x, -1, (dist, Px))
42         cdf(dist, y) + ccdf(dist, x-1)
43     end
44 end
45
46 pvalue_sterne(n, k, p) = pvalue_sterne(Binomial(n, p), k)
47
48 # Wilson
49
50 function pvalue_wilson(n, k, p)
51     z = safediv(k - n*p, √(n*p*(1-p)))
52     2ccdf(Normal(), abs(z))
53 end
54
55 # Wald
56
57 function pvalue_wald(n, k, p)
58     z = safediv(k - n*p, √(k*(n-k)/n))
59     2ccdf(Normal(), abs(z))
60 end

```

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

```

In [5]: ┌─ 1 ── function illustrate_binomial_pvalues()
  2   n = 20
  3   p = 0.25
  4   k = 9
  5
  6   p̂ = k/n
  7   σ² = n*p̂*(1-p̂)
  8
  9   null = Binomial(n, p)
 10  μ, σ² = mean(null), var(null)
 11  nullname = distname(null)
 12
 13  f(x) = mypdf(null, x)
 14  xlim = (-2, n+1)
 15  xlimk = (k-0.5, n+0.5)
 16
 17  pval = @sprintf "%.03f" pvalue_clopper_pearson(n, k, p)
 18  cval = @sprintf "%.03f" ccdf(null, k-1)
 19  P1 = plot(title=(1) Clopper-Pearson")
 20  plot!(f, xlim...; label="", c=1)
 21  plot!(f, xlimk...; label="", fillrange=0, c=1, fc=:red, fa=0.5)
 22  vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
 23  annotate!(k+5, 4f(k), text("P-value = $pval", 10, :red))
 24  plot!([k+7, k+7], [2.4f(k), 3.6f(k)]; arrow=true, c=:red, label="")
 25  annotate!(k+6, 3f(k), text("2x", 10, :red, :right))
 26  annotate!(k+5, 2f(k), text("one-tailed = $cval", 10, :red))
 27  plot!([k+7, k+1], [1.5f(k), 1.5f(k+1)]; arrow=true, c=:red, label="")
 28  plot!(; xtick=0:n, tickfontsize=7)
 29
 30  pval = @sprintf "%.03f" pvalue_sterne(n, k, p)
 31  j = _search_boundary(_pdf_le, 2mode(null)-k, -1, (null, pdf(null, k)))
 32  xlimj = (-0.5, j+0.5)
 33  P2 = plot(title=(2) Sterne")
 34  plot!(f, xlim...; label="", c=1)
 35  plot!(f, xlimk...; label="", fillrange=0, c=1, fc=:red, fa=0.5)
 36  plot!(f, xlimj...; label="", fillrange=0, c=1, fc=:red, fa=0.5)
 37  vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
 38  plot!([k-0.5, j+0.5], [f(k), f(k)]; label="", c=2, ls=:dash)
 39  annotate!(k+5, 3f(k), text("P-value = $pval", 10, :red))
 40  plot!([k+5, k+1], [2.5f(k), 0.8f(k)]; arrow=true, c=:red, label="")
 41  plot!([k+5, j+1], [2.5f(k), 0.8f(j)]; arrow=true, c=:red, label="")
 42  plot!(; xtick=0:n, tickfontsize=7)
 43
 44  pval = @sprintf "%.03f" pvalue_wilson(n, k, p)
 45  normal = Normal(μ, √σ²)
 46  g(x) = pdf(normal, x)
 47  P3 = plot(title=(3) Wilson")
 48  plot!(f, xlim...; label="", c=1)
 49  vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
 50  plot!(g, xlim...; ls=:dash, label="Normal(μ=$μ, σ=√$σ²)", c=2)
 51  plot!(g, k, n; label="", fillrange=0, c=2, fc=:red, fa=0.5)
 52  plot!(g, -2, 2μ-k; label="", fillrange=0, c=2, fc=:red, fa=0.5)
 53  annotate!(k+5, 3f(k), text("P-value = $pval", 10, :red))
 54  plot!([k+5, k+0.5], [2.5f(k), 0.8g(k)]; arrow=true, c=:red, label="")
 55  plot!([k+5, j+0.2], [2.5f(k), 0.7g(j)]; arrow=true, c=:red, label="")
 56  plot!(; xtick=0:n, tickfontsize=7)
 57
 58  pval = @sprintf "%.03f" pvalue_wald(n, k, p)
 59  normal = Normal(μ, √σ²)
 60  h(x) = pdf(normal, x)
 61  P4 = plot(title=(4) Wald")
 62  plot!(f, xlim...; label="", c=1)
 63  vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
 64  plot!(h, xlim...; ls=:dash, label="Normal(μ=$μ, σ=√$σ²)", c=2)
 65  plot!(h, k, n; label="", fillrange=0, c=2, fc=:red, fa=0.5)
 66  plot!(h, -2, 2μ-k; label="", fillrange=0, c=2, fc=:red, fa=0.5)
 67  annotate!(k+5, 3f(k), text("P-value = $pval", 10, :red))
 68  plot!([k+5, k+0.5], [2.5f(k), 0.8g(k)]; arrow=true, c=:red, label="")
 69  plot!([k+5, j+0.2], [2.5f(k), 0.7g(j)]; arrow=true, c=:red, label="")
 70  plot!(; xtick=0:n, tickfontsize=7)
 71
 72  plot(P1, P2, P3, P4; size=(800, 520), layout=(2, 2))
 73  plot!(; plot_title="model: $nullname, data: n=$n, k=$k")
 74 end
 75
 76 #illustrate_binomial_pvalues()

```

Out[5]: illustrate_binomial_pvalues (generic function with 1 method)

In [6]:

```
1 # 二項分布モデルの4種のP値函数のプロット
2
3 function plot_binom_pvaluefunctions(; n = 20, k = 6)
4     P1 = plot(p → pvalue_clopper_pearson(n, k, p), 0, 1;
5             label="", title="Clopper-Pearson", c=1)
6     P2 = plot(p → pvalue_sterne(n, k, p), 0, 1;
7             label="", title="Sterne", c=2)
8     P3 = plot(p → pvalue_wilson(n, k, p), 0, 1;
9             label="", title="Wilson", c=3)
10    P4 = plot(p → pvalue_wald(n, k, p), 0, 1;
11        label="", title="Wald", ls=:dash, c=4)
12    plot(P1, P2, P3, P4; size=(800, 520), layout=(2, 2))
13    plot!(; xtick=0:0.1:1, ytick=0:0.1:1)
14    plot!(; xguide="success rate parameter p",
15          yguide="P-value", guidefontsize=10)
16    plot!(; plot_title="data: n = $n, k = $k",
17          plot_titlefontsize=12)
18    plot!(; titlefontsize=12)
19 end
20
21 #plot_binomial_pvaluefunctions(; n = 20, k = 6)
```

Out[6]: plot_binom_pvaluefunctions (generic function with 1 method)

In [7]:

```
1 # 二項分布モデルのP値函数の動画 (データ k を動かしてアニメ化)
2
3 function gif_binomtest(; n = 20, fn = "images/pvaluefunction20.gif", fps = 5, step=1)
4     anim = @animate for k in [0:step:n; n-1:-step:1]
5         plot(p → pvalue_clopper_pearson(n, k, p), 0, 1; label="Clopper-Pearson")
6         plot!(p → pvalue_sterne(n, k, p), 0, 1; label="Sterne")
7         plot!(p → pvalue_wilson(n, k, p), 0, 1; label="Wilson")
8         plot!(p → pvalue_wald(n, k, p), 0, 1; label="Wald", ls=:dash)
9         plot!(; xtick=0:0.1:1, ytick=0:0.1:1)
10        plot!(; xguide="success rate parameter p", yguide="P-value")
11        title!("data: n = $n, k = $k")
12        plot!(; size=(600, 300))
13        2k > n && plot!(; legend=:topleft)
14    end
15    gif(anim, fn; fps)
16 end
```

Out[7]: gif_binomtest (generic function with 1 method)

In [8]:

```
1 # pvalue(x|p) 達のヒートマップ (明るい所ほど値が大きい)
2
3 function heatmap_binom_pvaluefunctions(; n = 20, xtick = 0:n)
4     k = 0:n
5     p = 0:0.01:1
6
7     P1 = heatmap(k, p, (k, p)→pvalue_clopper_pearson(n, k, p);
8                 colorbar=false, title="(1) Clopper-Pearson")
9     P2 = heatmap(k, p, (k, p)→pvalue_sterne(n, k, p);
10                colorbar=false, title="(2) Sterne")
11     P3 = heatmap(k, p, (k, p)→pvalue_wilson(n, k, p);
12                colorbar=false, title="(3) Wilson")
13     P4 = heatmap(k, p, (k, p)→pvalue_wald(n, k, p);
14                colorbar=false, title="(4) Wald")
15     plot(P1, P2, P3, P4; size=(800, 810), layout=(2, 2),
16          xtick, ytick=0:0.1:1, tickfontsize=7,
17          xguide="k", yguide="p")
18     plot!(; plot_title="data size: n = $n", plot_titlefontsize=12)
19 end
20
21 #heatmap_binom_pvaluefunctions(; n = 100, xtick=0:10:100)
```

Out[8]: heatmap_binom_pvaluefunctions (generic function with 1 method)

```
In [9]: # pvalue(x|p) ≥ α のヒートマップ
1
2
3 function heatmap_binom_rejectionregions(; α = 0.05, n = 20, xtick = 0:n)
4     k = 0:n
5     p = 0:0.01:1
6
7     c = cgrad([colorant"red", colorant"blue"])
8     alpha = 0.5
9     P1 = heatmap(k, p, (k, p)→pvalue_clopper_pearson(n, k, p) ≥ α;
10        colorbar=false, title=(1) Clopper-Pearson, c, alpha)
11     P2 = heatmap(k, p, (k, p)→pvalue_sterne(n, k, p) ≥ α;
12        colorbar=false, title=(2) Sterne, c, alpha)
13     P3 = heatmap(k, p, (k, p)→pvalue_wilson(n, k, p) ≥ α;
14        colorbar=false, title=(3) Wilson, c, alpha)
15     P4 = heatmap(k, p, (k, p)→pvalue_wald(n, k, p) ≥ α;
16        colorbar=false, title=(4) Wald, c, alpha)
17     plot(P1, P2, P3, P4; size=(800, 800), layout=(2, 2),
18         xtick, ytick=0:0.1:1, tickfontsize=7,
19         xguide="k", yguide="p")
20     plot!(; plot_title="significance level: α = $α, \
21           data size: n = $n", plot_titlefontsize=12)
22 end
23
24 #heatmap_binom_rejectionregions(; α = 0.05, n = 100, xtick=0:10:100)
```

Out[9]: heatmap_binom_rejectionregions (generic function with 1 method)

```
In [10]: # 二項分布モデルの4種のP値函数に関するシミュレーション
1
2
3 function sim_binomtest(n, p; L=10^5)
4     bin = Binomial(n, p)
5     p_clopper_pearson = Vector{Float64}(undef, L)
6     p_sterne = Vector{Float64}(undef, L)
7     p_wilson = Vector{Float64}(undef, L)
8     p_wald = Vector{Float64}(undef, L)
9     @threads for i in 1:L
10        k = rand(bin)
11        p_clopper_pearson[i] = pvalue_clopper_pearson(n, k, p)
12        p_sterne[i] = pvalue_sterne(n, k, p)
13        p_wilson[i] = pvalue_wilson(n, k, p)
14        p_wald[i] = pvalue_wald(n, k, p)
15    end
16    F_clopper_pearson = ecdf(p_clopper_pearson)
17    F_sterne = ecdf(p_sterne)
18    F_wilson = ecdf(p_wilson)
19    F_wald = ecdf(p_wald)
20    (; F_clopper_pearson, F_sterne, F_wilson, F_wald)
21 end
22
23 function plot_sim_binomtest(n, p; L=10^5, xmax=1)
24     (; F_clopper_pearson, F_sterne, F_wilson, F_wald) = sim_binomtest(n, p; L)
25     x = range(0, xmax, 1001)
26     P1 = plot(x, x→F_clopper_pearson(x);
27        label="", title=(1) Clopper-Pearson, c=1)
28     plot!(x, x; label="", ls=:dot, c=:black)
29     P2 = plot(x, x→F_sterne(x);
30        label="", title=(2) Sterne, c=2)
31     plot!(x, x; label="", ls=:dot, c=:black)
32     P3 = plot(x, x→F_wilson(x);
33        label="", title=(3) Wilson, c=3)
34     plot!(x, x; label="", ls=:dot, c=:black)
35     P4 = plot(x, x→F_wald(x);
36        label="", title=(4) Wald, c=4)
37     plot!(x, x; label="", ls=:dot, c=:black)
38
39     plot(P1, P2, P3, P4; size=(700, 720), layout=(2,2))
40     plot!(; xtick=0:0.1xmax:1, ytick=0:0.1xmax:1, tickfontsize=7)
41     plot!(; xlim=(-0.02xmax, 1.02xmax), ylim=(-0.02xmax, 1.02xmax))
42     plot!(; xguide="significance level α", yguide="probability of type I error")
43     plot!(; plot_title="data: n = $n, p = p₀ = $p", plot_titlefontsize=12)
44 end
```

Out[10]: plot_sim_binomtest (generic function with 1 method)

```
In [11]: 1 function pvalue_t(X, μ)
2     n = length(X)
3     X̄ = mean(X)
4     S² = var(X)
5     T = (X̄ - μ)/√(S²/n)
6     2ccdf(TDist(n-1), abs(T))
7 end
8
9 function confint_t(X; α = 0.05)
10    n = length(X)
11    X̄ = mean(X)
12    S² = var(X)
13    SE = √(S²/n)
14    t = quantile(TDist(n-1), 1 - α/2)
15    (X̄ - t*SE, X̄ + t*SE)
16 end
```

Out[11]: confint_t (generic function with 1 method)

```
In [12]: 1 function illustrate_ttest();
2     x = [10, 4, 20, 10, 18, 7, 5, 8, 4, 16, 8])
3     n = length(x)
4     X̄ = mean(x)
5     S² = var(x)
6     μ = 7
7     t = (X̄ - μ)/√(S²/n)
8     pval = 2ccdf(TDist(n-1), abs(t))
9     pval_str = @sprintf "%.03f" pval
10
11    println("null hypothesis: μ = ", μ)
12    println("data: x = ", x)
13    println("data size: n = ", n)
14    println("summary statistics: X̄ = ", X̄, ", S² = ", S²,
15            ", √S² = ", round(√S²; digits=1))
16    println("t-statistics: t = (X̄ - μ)/√(S²/n) = ", t)
17    println("degree of freedom: n - 1 = ", n - 1)
18    println("P-value: 2(1 - cdf(TDist(n-1), abs(t))) = ", pval)
19
20    P1 = plot()
21    plot!(TDist(n-1), -5, 5; label="TDist($(n-1))", c=1)
22    plot!(TDist(n-1), abs(t), 5; label="", c=1, fillrange=0, fc=:blue, fa=0.3)
23    plot!(TDist(n-1), -5, -abs(t); label="", c=1, fillrange=0, fc=:blue, fa=0.3)
24    plot!([t, t], [0, 3pdf(TDist(n-1), t)]; label="t = $(round(t; digits=2))", c=:blue)
25    plot!(Normal(), -5, 5; label="Normal(0,1)", ls=:dash, c=2)
26    plot!(; xtick=-10:10)
27    annotate!(-5, 0.25, text("P-value = $pval_str", 10, :blue, :left))
28    plot!([-2, -2], [0.23, 0.08]; arrow=true, label="", c=:blue)
29    plot!([-2, 1.7], [0.23, 0.04]; arrow=true, label="", c=:blue)
30
31    ccdf_t = @sprintf "%.03f" ccdf(TDist(n-1), t)
32    ccdf_n = @sprintf "%.03f" ccdf(Normal(), t)
33    P2 = plot()
34    plot!(TDist(n-1), 1, 5; label="TDist($(n-1))", c=1)
35    plot!(TDist(n-1), abs(t), 5; label="", c=1, fillrange=0, fc=:blue, fa=0.3)
36    plot!([t, t], [0, 3pdf(TDist(n-1), t)]; label="t = $(round(t; digits=2))", c=:blue)
37    plot!(Normal(), 1, 5; label="Normal(0,1)", ls=:dash, c=2)
38    plot!(; xtick=-10:10)
39    annotate!(5, 0.15, text("1 - cdf(TDist($(n-1)), t) = $ccdf_t", 10, :blue, :right))
40    annotate!(5, 0.12, text("1 - cdf(Normal(0,1), t) = $ccdf_n", 10, :red, :right))
41
42    plot_title = """
43        null hypothesis: μ = $μ, \
44        data: x = [10, 4, 20, 10, 18, 7, 5, 8, 4, 16, 8], \
45        size: n = $n
46        """
47    plot(P1, P2; size=(800, 270))
48    plot!(; plot_title)
49 end
50
51 #illustrate_ttest()
```

Out[12]: illustrate_ttest (generic function with 1 method)

```
In [13]: 1 function illustrate_confintmean(; dist = Normal(), n = 30, α = 0.05)
2     μ_true = mean(dist)
3
4     plot()
5     for i in 1:200
6         X = rand(dist, n)
7         ci = collect(confint_t(X))
8         c = pvalue_t(X, mean(dist)) < α ? 2 : 1
9         plot!([i, i], ci; label="", lw=2, c)
10    end
11    plot!(size=(800, 220))
12    hline!([mean(dist)]; label="", c=:black, lw=0.3)
13    title!("
14        $(100(1-α))% confidence intervals for mean of \
15        size-$n samples of $(distname(dist))
16        ")
17 end
18
19 #illustrate_confintmean(dist = LogNormal())
```

Out[13]: illustrate_confintmean (generic function with 1 method)

```
In [14]: 1 # 正規分布モデルで定めた平均のt検定のシミュレーション
2
3 function sim_ttest(; dist=Normal(), n=20, μ=mean(dist), L=10^5)
4     pvals = Vector{Float64} (undef, L)
5     tmp = [Vector{eltype(dist)} (undef, n) for _ in 1:nthreads()]
6     @threads for i in 1:L
7         X = rand!(dist, tmp[threadid()])
8         pvals[i] = pvalue_t(X, μ)
9     end
10    ecdf_pvals = ecdf(pvals)
11 end
12
13 function plot_sim_ttest(; dist=Normal(), n=30, μ=mean(dist),
14     dist_str=distname(dist), L=10^5)
15     ecdf_pvals = sim_ttest(; dist, n, μ, L)
16     f(x) = ecdf_pvals(x)
17     plot(; legend=:topleft)
18     plot!(f, 0, 1; label="")
19     plot!(identity, 0, 1; label="", c=:black, lw=0.3, ls=:dash)
20     plot!(; xlim=(-0.005, 0.155), ylim=(-0.005, 0.155))
21     plot!(; xtick=0:0.01:1, ytick=0:0.01:1, xrotation=45,
22             tickfontsize=7)
23     plot!(; xguide="α", yguide="probability of P-value < α")
24     title!("
25         ecdf of t-test P-values for mean
26         sample: $(dist_str), n=$n\
27         ")
28     plot!(; size=(400, 415))
29 end
30
31 #plot_sim_ttest(; dist=LogNormal())
```

Out[14]: plot_sim_ttest (generic function with 1 method)

```
In [15]: 1 Random.seed!(4649373)
```

Out[15]: TaskLocalRNG()

1 お勧め解説動画とお勧め文献

1.1 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主に以下の2つである：

- 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.

そのせいで、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の確率であることがクリアに説明されておらず、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。

1.2 お勧め文献: P値に関するASA声明

- 統計的有意性と P 値に関する ASA 声明 (日本語訳) [PDF (<https://www.biometrics.gr.jp/news/all/ASA.pdf>)]

P値に関する解説ではこれが非常によい。しかし、その第2節の

2. P値とは?

おおざっぱにいうと、P 値とは特定の統計モデルのもとで、データの統計的要約（たとえば、2 グループ比較での標本平均の差）が観察された値と等しいか、それよりも極端な値をとる確率である。

という説明における「データ」という用語の解釈には注意を要する。この意味での「データ」は現実世界での観察で得たデータのことではなく、統計モデル内で生成された仮想的なデータのことである。

現実世界で得たデータは決まった数値になるが、数学的フィクションである統計モデル内で生成されたデータは確率変数とみなせ、それが現実世界での観察で得たデータの数値以上の極端な値を取る確率を考えることができる。

統計学は、現実世界でデータを得る活動と数学的世界で統計モデルを考えることを行ったり来たりするので、現実とモデルを混同しないように注意しなければいけない。

P値は現実において常に意味を持つ確率またはその近似値ではなく、数学的フィクションである統計モデル内部で測った確率に過ぎない。

- 佐藤俊哉著、『宇宙怪人しまりす 統計よりも重要なことを学ぶ』、朝倉書店、2024 [Web Site (https://www.asakura.co.jp/detail.php?book_code=12297)]

この本も非常にお勧めである。統計学の解説で「統計よりも重要なことを学ぶ」というタイトルの本を勧めるのはちょっとおかしいように感じるかもしれないが、私の個人的な意見では、広い意味での統計学は「統計よりも重要なこと」も含んでいる。この本を読めば、統計学には確率論の応用を超えた部分が明瞭にあることを理解できるだろう。

- Sander Greenland, Advancing statistics reform: Ways to improve thinking and practice in the face of resistance, 2022 [PDF (https://biostat.ucdavis.edu/sites/g/files/dgynsk4966/files/inline-files/Greenland_Advancing%20statistics%20reform%2C%20part%204_Slides%201-110%2C%2001%20June%202022.pdf)]

このノートの解説ではGreenlandさんの考え方を取り入れている。

2 まとめ

以下のまとめは「理解が進むたびに繰り返しこのまとめに戻る」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) と呼ばれることが多い。(「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れると考えられる。)

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

と呼ぶこととする。

P値は使用している統計モデルの下で以下の2つの **相性の良さ** (compatibility, **compatibility**)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+パラメータの値の設定 $\theta = \theta_0$ は現実世界で得られたデータの値 x と相性が悪いと考える。

2.2 S値

P値 P の底が2の対数の-1倍 $S = -\log_2 P$ を **S値** (S-value, **surprisal value**, **意外さの値**)と呼ぶ。単位は **ビット** (bit)だとみなされる。

S値は統計モデル+パラメータの値の設定の側から見たデータの値の **意外さ** 指標である。

S値はP値をコイン投げで何回続けて表が出た場合の確率に相当するかになっている。

例: $P = 1$ に対応するS値 $S = 0$ は意外さが 0 であることを意味する。

例: $P = 1/2^4 = 6.25\%$ はコイン投げで4回続けて表が出る確率に一致し、対応するS値は $S = 4$ になる。すなわち、P値 $P = 6.25\%$ が表す意外さはコイン投げで4回続けて表が出た場合の意外さに等しい。

例: $P = 1/2^5 = 3.125\%$ はコイン投げで5回続けて表が出る確率に一致し、対応するS値は $S = 5$ になる。すなわち、P値 $P = 3.125\%$ が表す意外さはコイン投げで5回続けて表が出た程度の意外さに等しい。 $P = 6.25\%$ とその半分の $P = 3.125\%$ の意外さの違いはちょうど 1 ビットである。

例: $P = 5\%$ のとき $S = -\log_2 P \approx 4.3$ なので、P値 $P = 5\%$ が表す意外さはコイン投げで約4.3回続けて表が出た場合の意外さに等しい。

文献: S値については次の文献を参照せよ:

- Z. Rafi, S. Greenland, Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise, BMC medical research methodology, 2020. <https://scholar.google.co.jp/scholar?cluster=591718439647960929> (<https://scholar.google.co.jp/scholar?cluster=591718439647960929>)

2.3 検定

有意水準 (significance level)と呼ばれる閾値(いきち、しきいち) $0 < \alpha \leq 1$ 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相性が悪い

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)されたと言ったりする。

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妥当である」のように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、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。

2.4 信頼区間

信頼区間 (confidence interval)の文脈で $1 - \alpha$ は **信頼水準** (confidence level)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つ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

- 妥当性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。

閾値を設けることへの警告: 検定や信頼区間には、人為的な閾値を必要によって、2值的な判断を強制するという欠点がある。閾値を設けずにP値を直接使えば閾値の設定は必要でなくなる。

信頼区間の別名: P値は、統計モデル+パラメータの値の設定 $\theta = \theta_0$ と現実世界で得られたデータの値 x のコンパチビリティ (相性の良さ, compatibility)の指標であった。その解釈を信頼区間の場合に適用して、「信頼」という誤解を招きがちな言い方を排して、**コンパチビリティ区間 (compatibility interval, 相性が良さの様子を表している区間)**と呼ぶべきだという有力な意見がある。

100(1 - α)% コンパチビリティ区間は「閾値 α によってデータの値と相性が良いとみなされた統計モデルのパラメータ θ の値の範囲」だと解釈される。

2.5 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 $Normal(\mu, \sigma)$ 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、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、搅乱パラメータ、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れば以下がある:

- 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

まず、現実世界におけるデータの生成のされ方のモデル化として使わ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 $Binomial(n, p)$

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の数値達 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 $Normal(\mu, \sigma)^n$

このように統計モデルは多くの場合に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後の方の説明を参照せよ。

仮説 $\theta = \theta_0$ は 帰無仮説 (null hypothesis)と呼ばれることが多いが、このノートでは単に「仮説」と呼ぶことが多い。

まとめ: P値=帰無仮説下のモデル内でデータの数値以上に極端な値が生じる確率(の近似値)

補足: 「 x が x' 以上に極端な値である」という条件は **反射性**(reflexivity)と **推移性** (transitivity)と **比較可能性**(comparability)の3つの条件を満たしていることを要請する。すなわち「 x が x' 以上に極端な値である」ことを $x \geq x'$ と書くと、以下が成立していると仮定する:

- 反射性(reflexivity): $x \geq x$,
- 推移性(transitivity): $x \geq x'$ and $x' \geq x'' \implies x \geq x''$,
- 比較可能性(comparability): $x \geq x'$ or $x' \geq x$.

注意: 実際にはこの基本に忠実にP値が定義されるとは限らない。

3.3 「データの数値以上に極端な値」の意味の定義の仕方

注意: 以下の説明は抽象的過ぎて分かりにくいので、後の節の具体例の方を先に参照した方がよいと思われる。

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の定義を、統計分析の目的に合わせて適切に設定する必要がある。

「仮説 $\theta = \theta_0$ 下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、以下のような条件がよく使われる:

(0) $X \geq x$ (または $X \leq x$).

この(0)の条件は例えば χ^2 検定で使われる。

さらにこの(0)の条件は帰無仮説 $\theta \leq \theta_0$ を(θ の値を θ_0 以上のどれかとする)を対立仮説 $\theta > \theta_0$ (θ の値を θ_0 未満のどれかとする)と比較する **片側検定** で使用される。

しかし、主に使われるのは(一部に見かけ上の例外もある(例: χ^2 検定))、帰無仮説 $\theta = \theta_0$ を対立仮説 $\theta \neq \theta_0$ (θ の値を θ_0 以外のどれかの値とする)と比較する **両側検定** の場合である。以下はすべて両側検定の場合である。(用語「対立仮説」については Neyman-Pearsonの仮説検定の節を参照せよ。)

(1) $X \geq x$ と $X \leq x$ の確率が小さい方の条件。(この場合にP値はその確率の2倍にする。)

(2) モデル内の 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。

(4) 他にも正規分布近似を使う方法も多用される。

3.4 例(二項分布モデルの場合)

注意: 以下の文章による説明はわかりにくいので、下の方に用意してある図を先に見た方が良いかもしない。

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し、仮説 $p = p_0$ 下の二項分布 $\text{Binomial}(n, p_0)$ に従う確率変数 K を用意する:

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

このとき、「仮説 $p = p_0$ 下の二項分布モデル内の成功回数 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:

(0) $K \geq k$ (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。この定義はそれぞれ仮説 $p \leq p_0$ (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) **Clopper-Pearsonの信頼区間の場合:** 仮説 $p = p_0$ の下での二項分布モデル内の

$K \geq k$ 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.

以下の定義達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) **Sterneの信頼区間の場合:** 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という条件で, すなわち,

二項分布の確率質量函数を $P(k|n, p_0)$ と書くとき, $P(K|n, p_0) \leq P(k|n, p_0)$

という条件で「 k 以上に極端」の意味を定義する. すなわち,

仮説 $p = p_0$ 下の統計モデル内でデータの数値 k 以上に確率的に珍しいこと

を「 k 以上に極端」の定義とする.

以下のように正規分布近似(中心極限定理)を使って定義することもできる.

(3) **Wilsonの信頼区間の場合:** 二項分布 $\text{Binomial}(n, p_0)$ 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 $\text{Normal}(0, 1)$ 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).

(3)' **Pearsonの χ^2 検定の場合:** 上の(3)と同値な次の条件を使うこともある:

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

$Z \sim \text{Normal}(0, 1)$ 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:

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

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

$$\chi^2 = \sum \frac{(\text{現実における観測値} - \text{帰無仮説下の統計モデル内の期待値})^2}{\text{帰無仮説下の統計モデル内の期待値}}$$

の特別な場合に過ぎない. 上の場合に「現実における観測地」は n 回中の成功回数 k と失敗回数 $n - k$ であり, 「帰無仮説下の統計モデル内の期待値」は成功回数の期待値 np_0 と失敗回数の期待値 $n(1 - p_0)$ である. (より進んだ注意: Pearsonの χ^2 統計量はスコア検定における検定統計量になっている.)

(4) **Waldの信頼区間の場合:** 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:

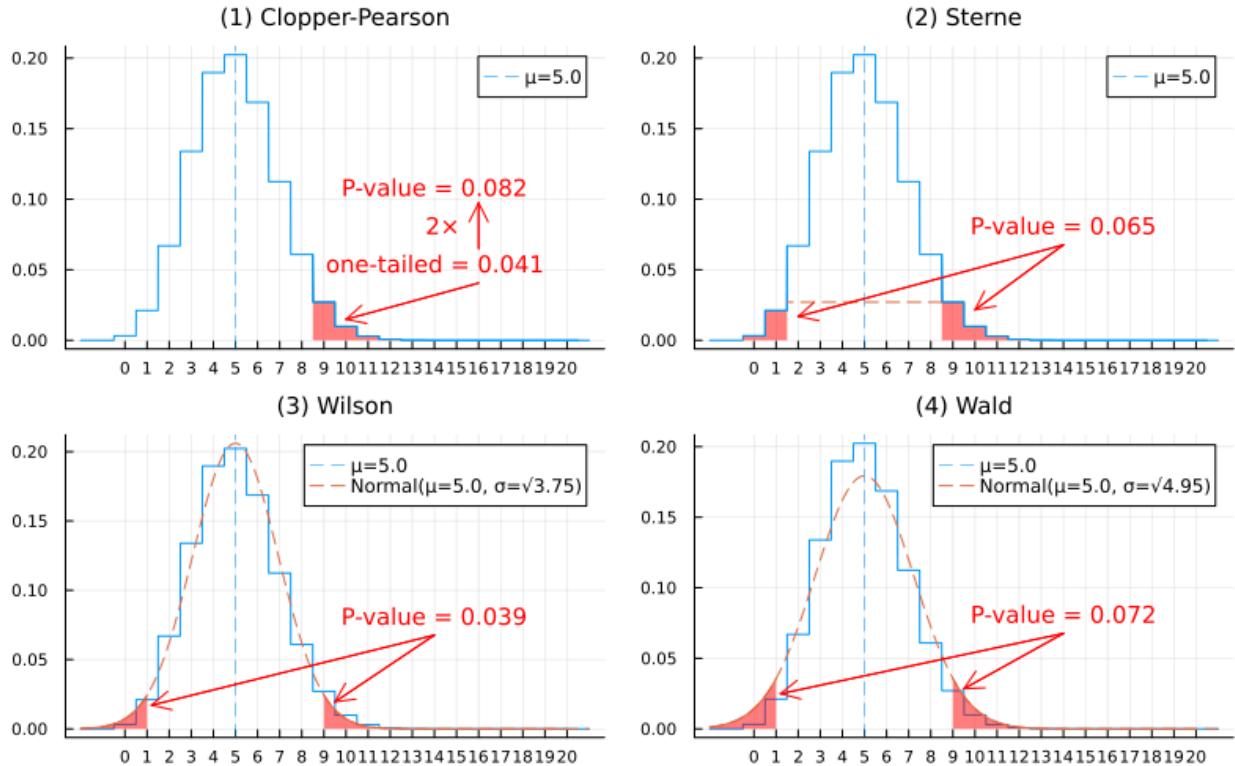
$$|Z| \geq \frac{|k - np_0|}{\sqrt{n\hat{p}(1 - \hat{p})}}.$$

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.

In [16]: 1 illustrate_binomial_pvalues()

Out[16]:

model: Binomial(n=20, p=0.25), data: n=20, k=9



注意: Waldの信頼区間を与えるP値の計算で使っている正規分布近似では、正規分布の分散としてモデルの分散 $np(1 - p) = 3.75$ ではなく、データ「 $n = 20$ 回中 $k = 9$ 回成功」から推定した $k(n - k)/n = 4.95$ を使っているせいで、正規分布近似の精度が低くなっている。精度が高いのはWilsonの信頼区間を与えるP値の方である。しかし、信頼区間の計算はWaldの側がずっと簡単になる。

注意: ここでは詳しく説明できないが、Sterneの信頼区間の方がClopper-Pearsonの信頼区間よりも被覆確率という基準で精度が高いことが知られている。原論文でもそのことが指摘されている:

- Theodore E. Sterne, Some Remarks on Confidence or Fiducial Limits, Biometrika, Vol. 41, No. 1/2 (Jun., 1954), pp. 275-278
[link (<https://www.jstor.org/stable/2333026>)]

しかし、計算法はClopper-Pearsonの信頼区間の方がずっとシンプルである。

注意: P値の定義の仕方の各々には利点と欠点がある。

3.5 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 $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ 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に従う確率変数 $X = (X_1, \dots, X_n)$ の値がデータの数値 $x = (x_1, \dots, x_n)$ 以上に極端であること」を以下の方法で定義するとよい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 $\mu = \mu_0$ の下で確率変数 X の値がデータの数値 x 以上に極端であること」を

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

という条件で定める。これは大雑把に言って、

- モデル内で生成された仮想的なデータ X の標本平均 \bar{X} が $\mu = \mu_0$ からデータの標本平均 \bar{x} 以上に離れている。

という意味の条件になっている。ただし、分散を推定して適当にスケールして比較している。

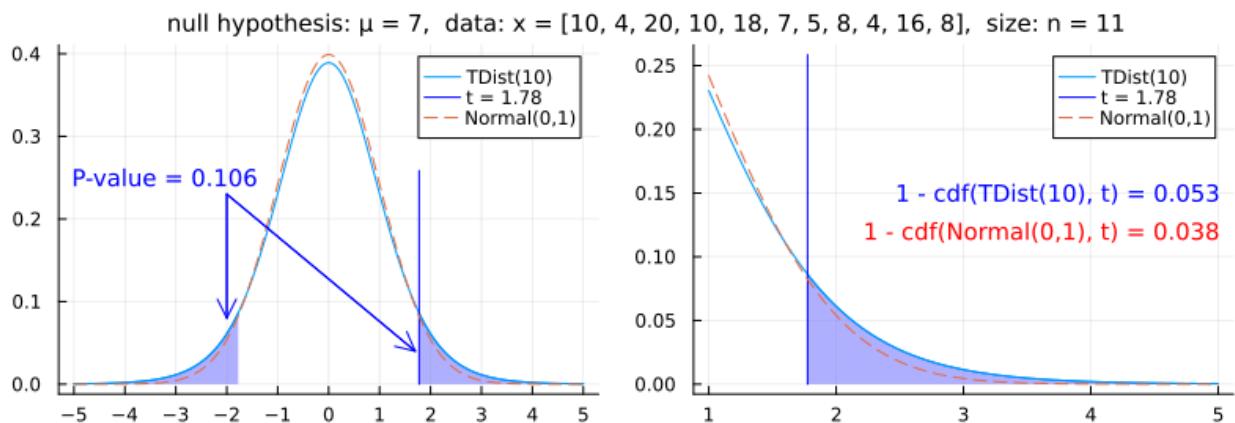
これが便利なのは、仮説 $\mu = \mu_0$ の下での正規分布の標本分布モデル内で、分散パラメータ σ^2 がどんな値であっても、確率変数 $T(X|\mu_0)$ が自由度 $n - 1$ の t 分布に従うからである。

注意: 統計モデルが θ 以外にパラメータ η を持っていて、帰無仮説 $\theta = \theta_0$ を課しても、パラメータ η の分だけ統計モデルの確率分布が唯一つに決まらないとき、 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、搅乱パラメータ、迷惑パラメータ) と呼ぶ。

In [17]: 1 illustrate_ttest()

```
null hypothesis: mu = 7
data: x = [10, 4, 20, 10, 18, 7, 5, 8, 4, 16, 8]
data size: n = 11
summary statistics; x̄ = 10.0, s² = 31.4, √s² = 5.6
t-statistics: t = (x̄ - μ)/√(s²/n) = 1.7756312235479013
degree of freedom: n - 1 = 10
P-value: 2(1 - cdf(TDist(n-1), abs(t))) = 0.106177674255521
```

Out[17]:



注意: 自由度 $n - 1 = 10$ の t 分布 $TDist(10)$ と標準正規分布 $Normal(0, 1)$ は非常に近いように見えるが、その裾野の部分の確率の大きさで違いが生じていることに注意せよ。

問題: サイズ $n = 11$ のデータ

```
x = [10, 4, 20, 10, 18, 7, 5, 8, 4, 16, 8]
```

と仮説 $\mu = 7$ について、標本平均 \bar{x} と不偏分散 s^2 と $t = (\bar{x} - \mu)/\sqrt{s^2/2}$ を自分で計算してみて、上の計算例と一致することを確認せよ。さらに、インターネットで検索して、 t 検定のP値を求める方法を調べ、実際に計算してみて、上の計算例と一致することを確認せよ。

3.6 P値は帰無仮説下の統計モデルのデータの数値との相性の良さ(compatibility)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、データの数値 x と仮説 $\theta = \theta_0$ の下での統計モデルの相性の良さ(compatibility)の指標である。

P値が小さいとき、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。

これがP値の基本的な使い方である。

まとめ: P値はデータの数値と帰無仮説下のモデルの相性の良さ(compatibility)の指標。

4 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) $0 \leq \alpha \leq 1$ 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、(帰無)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。(実際には「帰無仮説は棄却された」と略した言い方をすることが多い。)

この手続きを **仮説検定** (Hypothesis testing) もしくは単に **検定** (test) と呼ぶ。

すなわち、検定とは、ある閾値を設けて、データの数値との相性の良さ(compatibility)の指標が閾値未満のモデルのパラメータの値を捨て去る手続きのことである。

ただし、検定は閾値を設けて捨てる行為なので、捨て去り過ぎてしまう誤りを犯すリスクがある。

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。

まとめ: 検定 = 閾値を設けてデータの数値との相性の良さ(compatibility)の指標が低過ぎるパラメータ値を捨て去る手続き.

5 第一種の過誤(αエラー)

5.1 第一種の過誤(αエラー)の定義

X はパラメータの数値を $\theta = \theta_0$ に設定した統計モデルに従う確率変数であると仮定する.

データの数値がその確率変数 X の値であるときの仮説 $\theta = \theta_0$ のP値が有意水準 α 未満になる確率を仮説 $\theta = \theta_0$ に関する **第一種の過誤の確率** または **αエラーの確率** と呼ぶ.

すなわち, 仮説 $\theta = \theta_0$ 下の統計モデル内部において, その仮説 $\theta = \theta_0$ が検定の手続きで棄却されてしまうことを **第一種の過誤** (Type I error, αエラー, α-error) と呼び, その確率を **第一種の過誤の確率** と呼ぶ.

第一種の過誤は検定する仮説 $\theta = \theta_0$ が成立しているモデル内部においてその仮説が棄却されてしまうことを意味しており, その確率は小さい方がよい. しかし, 第一種の過誤の確率を小さくすることが原因で生じる害(後で述べる検出力の低下)もあるので, 目的に合わせてバランスを取ることが必要である.

注意: 第一種の過誤は「無実のものを有罪にしてしまう誤り」に例えられることがある.

注意: 第一種の過誤は統計モデルが自分自身を確率的に否定してしまうことであると言ふこともできる.

注意: 現実世界から得たデータの数値 x は値が確定した定数だが, 数学的フィクションである統計モデル内ではデータの数値が確率変数 X としてランダムに生成されていると考えることができる. 多くの教科書でこの辺について誤解に誘導するような説明があるので注意して欲しい.

まとめ: 第一種の過誤 = 帰無仮説下の統計モデル内で帰無仮説が棄却されること.

5.2 第一種の過誤(αエラー)の確率

定理: 有意水準 α の下での第一種の過誤の確率の値は α で近似される.

証明: x の値が x' 以上に極端であるという条件を $x \geq x'$ と書き, X はパラメータ値 $\theta = \theta_0$ の統計モデルに従う確率変数であると仮定する.

このとき, P値の定義(X の値が x 以上に極端な値になる確率もしくはその近似値)より,

$$(データの数値 x に関する仮説 \theta = \theta_0 のP値) \approx P(X \geq x).$$

$P(X \geq x_\alpha) \approx \alpha$ となる x_α を取る. このとき,

$$\begin{aligned} & (\text{データの数値 } x \text{ について仮説 } \theta = \theta_0 \text{ のP値が } \alpha \text{ 未満になるという条件}) \\ & \approx (P(X \geq x) < \alpha \text{ という条件}) \\ & \approx (x \geq x_\alpha \text{ という条件}). \end{aligned}$$

したがって,

$$\begin{aligned} & (\text{仮説 } \theta = \theta_0 \text{ に関する第一種の過誤の確率}) \\ & = (\text{データの数値が確率変数 } X \text{ の値であるときに 仮説 } \theta = \theta_0 \text{ のP値が } \alpha \text{ 未満になる確率}) \\ & \approx (X \geq x_\alpha \text{ となる確率}) = P(X \geq x_\alpha) \approx \alpha. \end{aligned}$$

証明終

注意: 上の問題の結果は, 仮説 $\theta = \theta_0$ 下の統計モデル内部において, 同仮説 $\theta = \theta_0$ のP値の分布が一様分布で近似されることを意味している. 次の節の計算例を参照せよ.

注意: 実際にはP値を基本に忠実でない方法で定義することもあり, 上の定理の証明が適用できない場合もあるのだが, 検定の理論では第一種の過誤の確率が有意水準に近くなることを要請することが普通である.

まとめ: 第一種の過誤の確率は有意水準に等しいかそれに近い値になる.

これは実際には「要請」であり, 証明されるべき結果ではないとみなされる.

5.3 二項分布モデルの4種のP値に関する第一種の過誤の確率のグラフ

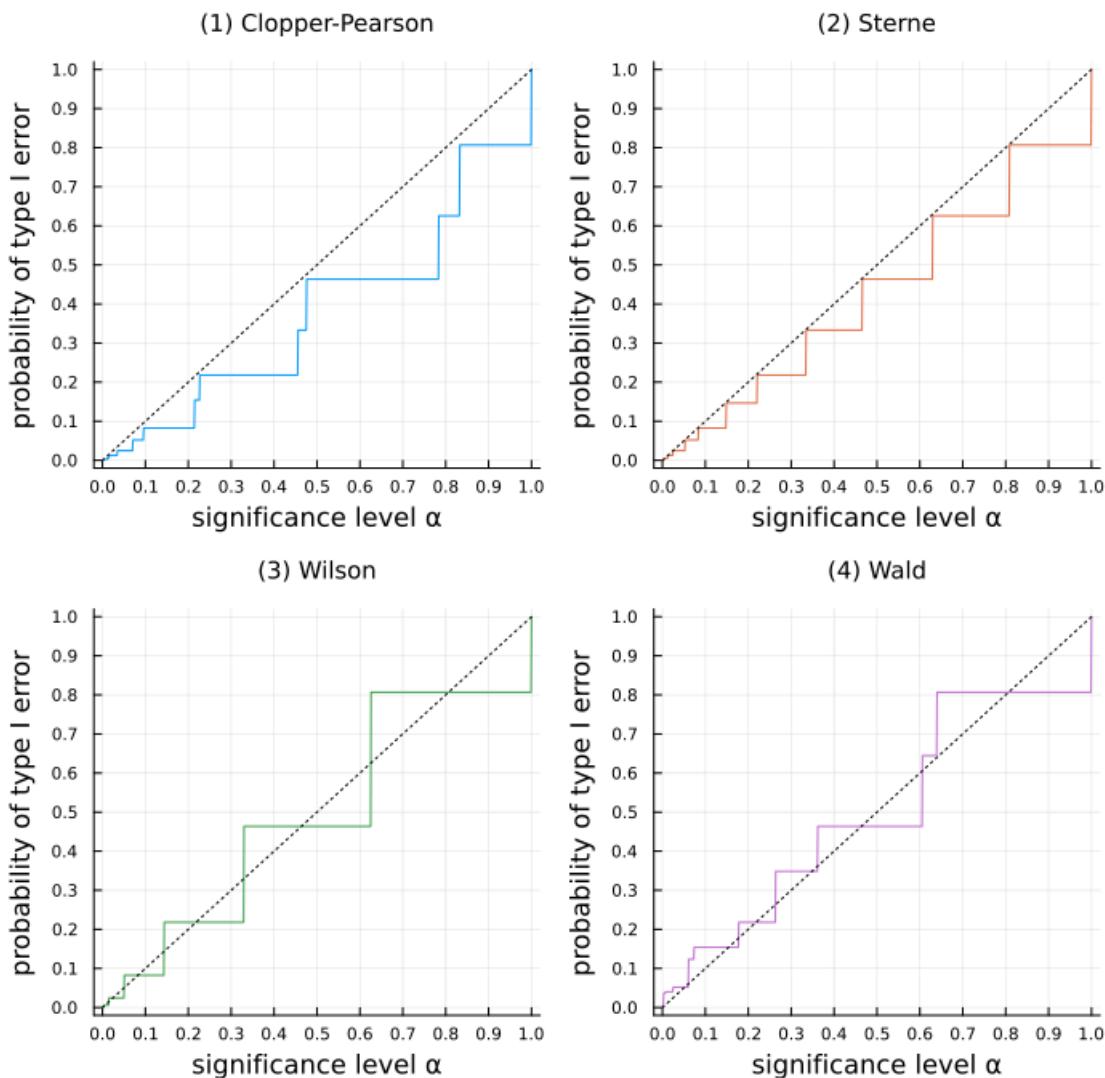
二項分布モデルの4種のP値について, 0 から 1 のあいだの有意水準 α 達について第一種の過誤の確率を計算してグラフを描いてみよう.

そのグラフは、仮説 $p = p_0$ 下の二項分布モデル内部における同仮説 $p = p_0$ のP値の分布の累積分布函数のグラフだと言ってよいので、そのグラフが「45度線」に近ければ近いほど、その分布は一様分布に近いということになる。

In [18]: 1 plot_sim_binomtest(20, 0.3)

Out[18]:

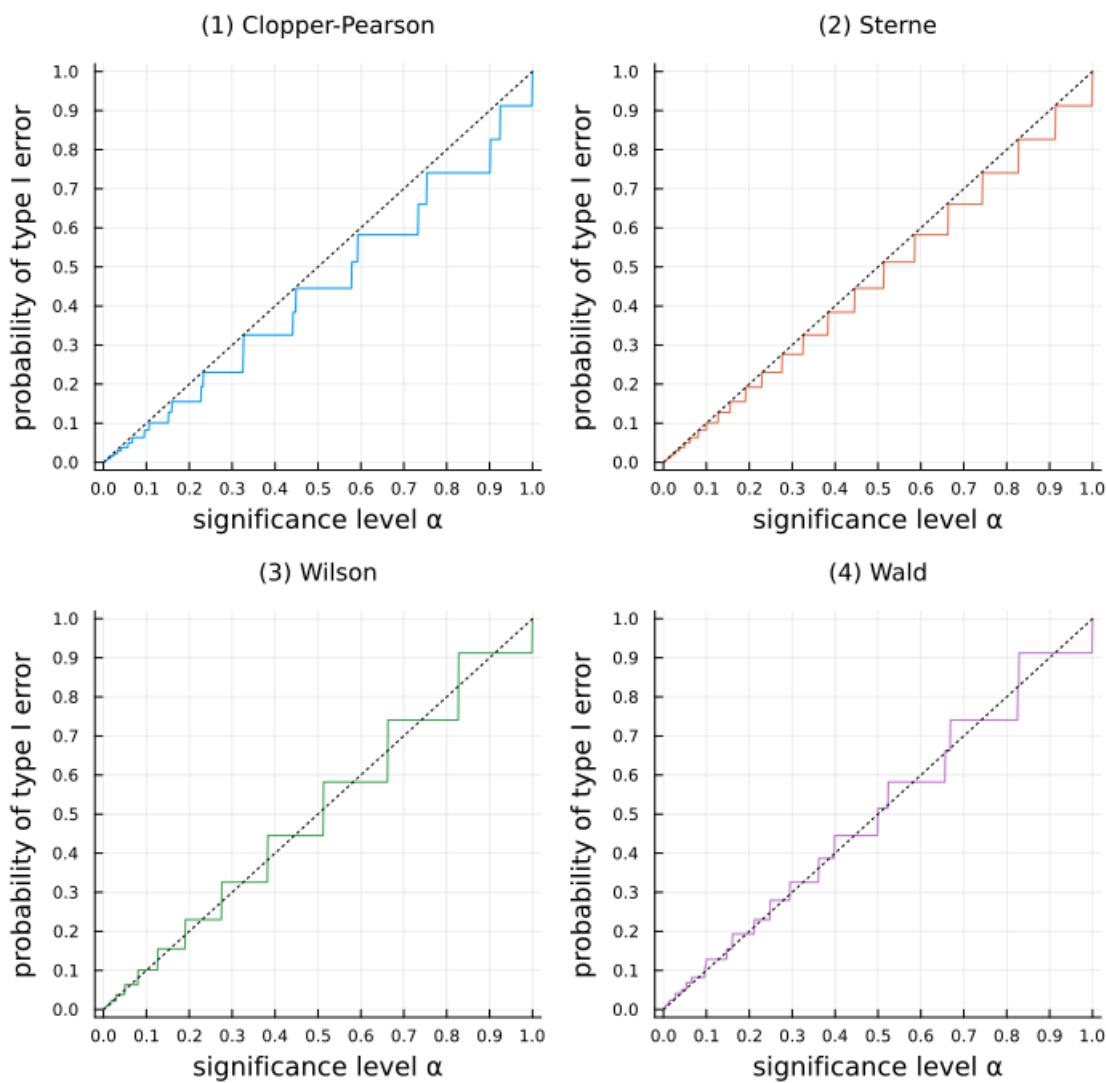
data: n = 20, p = p₀ = 0.3



In [19]: 1 plot_sim_binomtest(100, 0.3)

Out[19]:

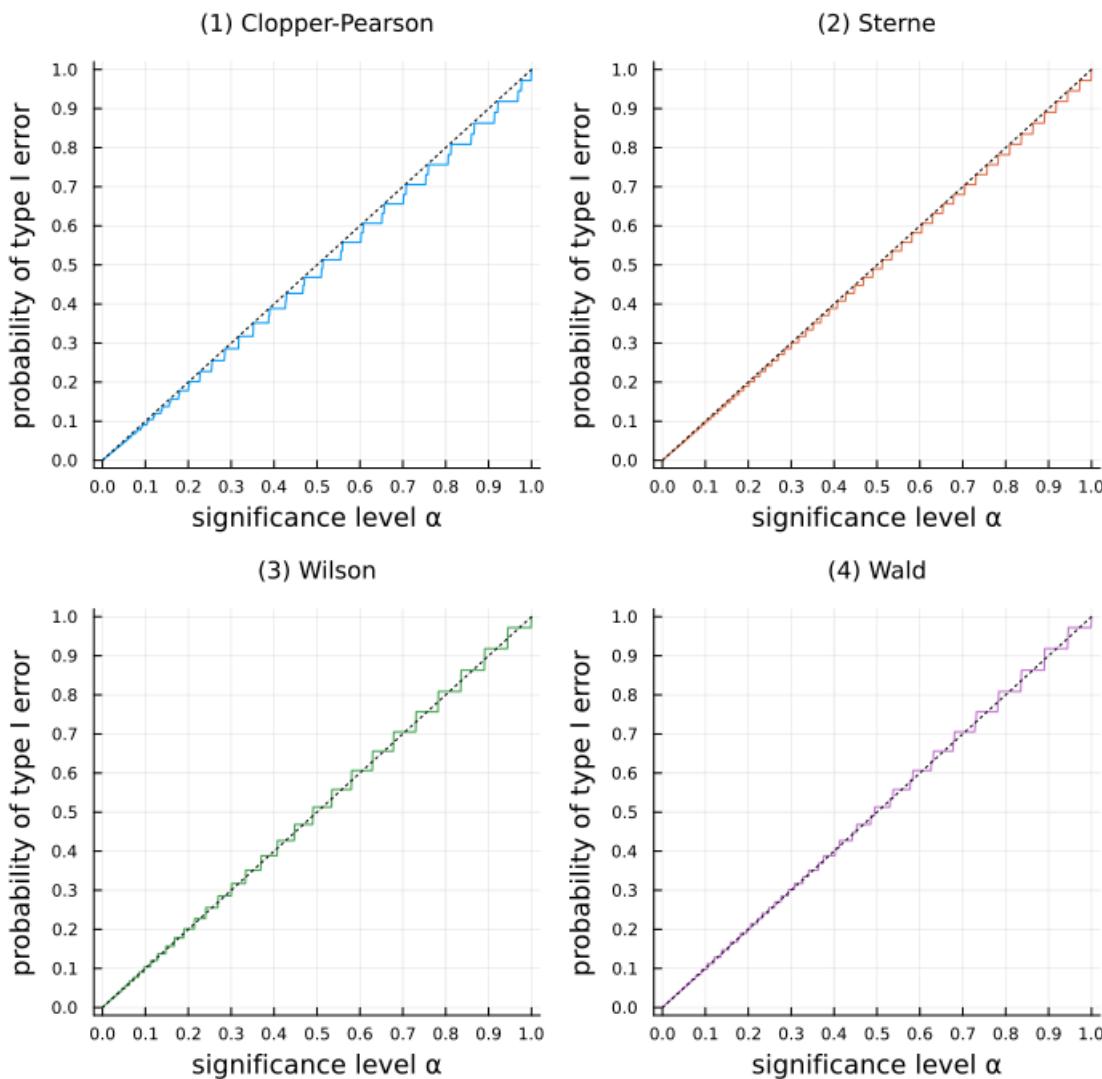
data: n = 100, p = $p_0 = 0.3$



In [20]: 1 plot_sim_binomtest(1000, 0.3)

Out[20]:

data: n = 1000, p = $p_0 = 0.3$



確かに二項分布モデルの4種のP値の帰無仮説下の二項分布内の累積分布函数は、確かに45度線を近似していることが分かる。このことはそのP値の分布が一様分布で近似されることを意味している。

6 信頼区間

6.1 信頼区間のP値もしくは検定を用いた定義

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。 $(\alpha = 5\% \text{ のとき}, \text{信頼度 } 1 - \alpha \text{ の信頼区間を } 95\% \text{ 信頼区間と呼ぶことが多い。})$

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 $\theta = \theta_0$ 下の統計モデルとデータの値 x との相性の良さの指標(P値)が有意水準 α 未満になることであった。そのような状況を

- 仮説 $\theta = \theta_0$ の下での統計モデルとデータの値 x の相性が悪い

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 との相性が閾値 α によって良いと判定されるパラメータの値全体の集合

であると言える。

まとめ: 信頼区間 = 検定で棄却されないモデルのパラメータ値全体の集合。

6.2 信頼区間の使い方

「閾値 α によって相性が良いと判定されること」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の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、現実における正しさの判断を保留する。
- 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。

まとめ: 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する。

7 信頼区間と検定の表裏一体性

7.1 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 $pvalue(x|\theta_0)$ 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 が与えられたとき、パラメータの数値 $\theta = \theta_0$ をP値 $pvalue(x|\theta_0)$ に対応させる函数

$$\theta_0 \mapsto pvalue(x|\theta_0)$$

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:

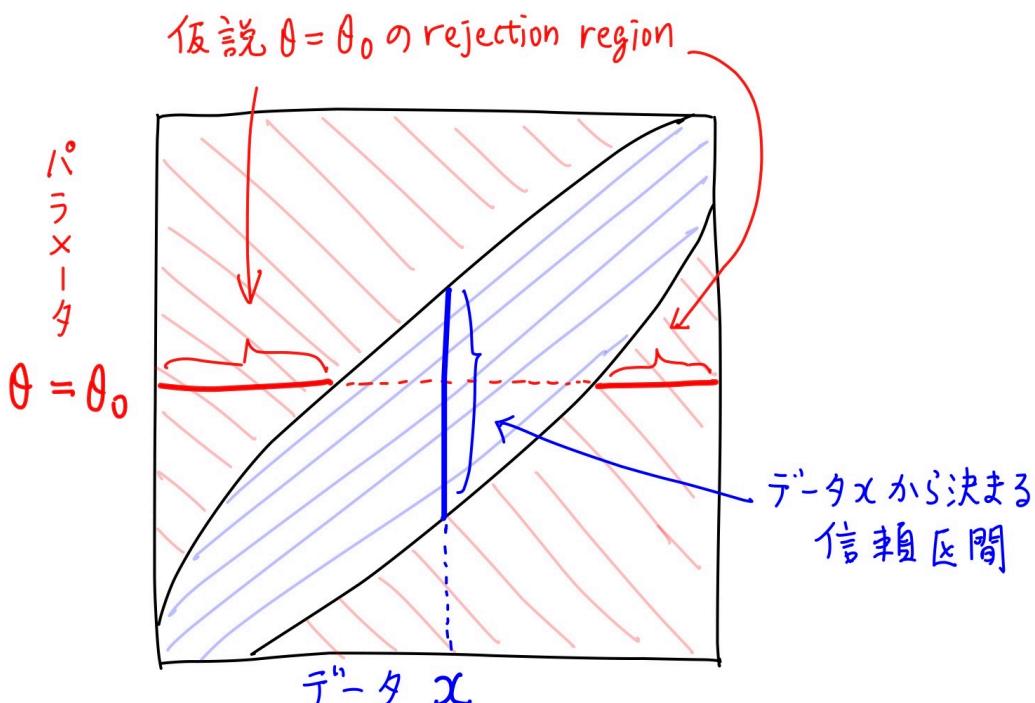
$$R_\alpha(\theta_0) = \{ x \mid pvalue(x|\theta_0) < \alpha \}.$$

信頼度 $1 - \alpha$ の信頼区間 $C_\alpha(x)$ は

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

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 $pvalue(x|\theta_0)$ の値が有意水準 α になる。



まとめ: 信頼区間と検定は表裏一体である。

7.2 信頼区間と検定の表裏一体性について書かれた教科書の例

以下の教科書にはこのノートが採用した信頼区間と検定の表裏一体性に関する考え方を書いてある。

- ・竹内啓, 数理統計学—データ解析の方法, 東洋経済新報社, 1963 (のp.103)
- ・小針覗宏, 確率・統計入門, 岩波書店, 1973 (のp.197)
- ・久保川達也, 現代数理統計学の基礎, 共立出版, 2017 (のp.169)
- ・竹村彰通, 新装改訂版 現代数理統計学, 学術図書出版社, 2020 (のp.202)

これらの教科書は有名である。

- ・<https://twitter.com/genkuroki/status/1531827825879445505> (<https://twitter.com/genkuroki/status/1531827825879445505>)

にその部分の引用がある。

まとめ: 信頼区間と検定が表裏一体であることは有名な教科書群に書いてある。

7.3 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 $pvalue(X|\theta_0) < \alpha$ となる確率は α もしくはその近似になるので、 $pvalue(X|\theta_0) \geq \alpha$ となる確率は $1 - \alpha$ もしくはその近似値になる。

そして、 $pvalue(X|\theta_0) \geq \alpha$ 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陷入っている疑いもある。

まとめ: 仮説 $\theta = \theta_0$ 下の統計モデル内部において、信頼区間がパラメータ値 $\theta = \theta_0$ を含む確率は信頼度 $1 - \alpha$ に等しい、もしくはそれに近い値になる。

これも証明するべきことではなく、「要請」だと考えた方がよい。

7.4 二項分布モデルでのP値函数の視覚化

文献: P値函数の使い方については次の文献が詳しい:

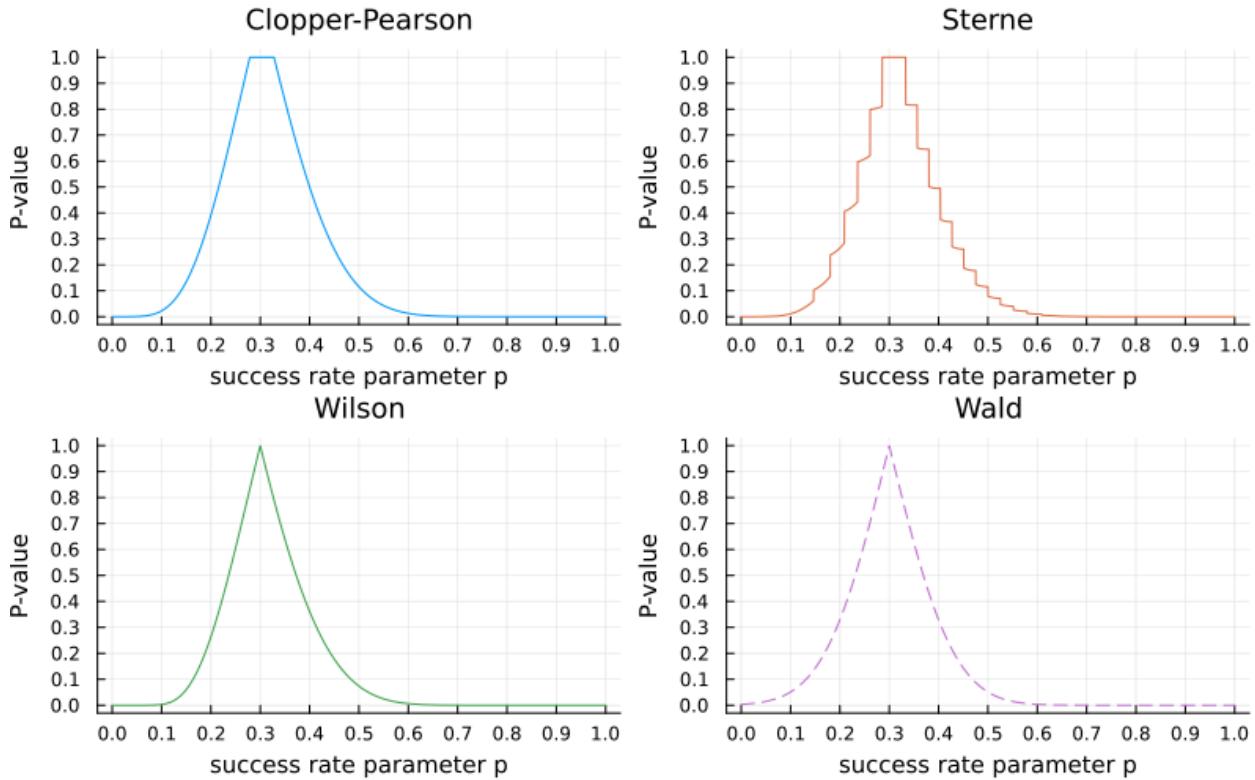
- Timothy L. Lash, Tyler J. VanderWeele, Sébastien Haneuse, and Kenneth J. Rothman.
Modern Epidemiology, 4th edition, 2020. [Google](https://www.google.com/search?q=Modern+Epidemiology+4th) (<https://www.google.com/search?q=Modern+Epidemiology+4th>)

第4版よりも古い版でもよい。

```
In [21]: 1 plot_binom_pvaluefunctions(; n = 20, k = 6)
```

Out[21]:

data: n = 20, k = 6



以上は「 $n = 20$ 回中 $k = 6$ 回成功」というデータの数値が与えられたときの、P値函数のグラフである。

P値が小さいほどデータの数値と成功確率パラメータ値が p の二項分布モデルの相性が悪いと考える。

上のグラフを見れば、P値函数による判定では、「 $n = 20$ 回中 $k = 6$ 回成功」というデータの数値に最も整合するパラメータ p の値は $p = 6/20 = 0.3$ になっており、そこから離れると相性の良さの指標(P値)の値が下がる。

どの場合も、「 $n = 20$ 回中 $k = 6$ 回成功」というデータの数値にあまりにも整合しないパラメータ p の値全体の集合の補集合は大雑把に 0.1 から 0.6 までの区間になることもわかる。これが信頼区間である。ただし、有意水準を決めずにグラフの見た目で判断してどんぶり勘定で区間を決めた。信頼区間の正式な定義はP値が有意水準という名の閾値 α 以上になるパラメータの範囲である。

信頼区間では「そこに含まれるか否か」(α 以上か否か)の情報しか残っていないが、P値函数の様子を直接見れば、パラメータを動かしたときの統計モデルとデータの数値の相性の良さ(compatibility)の度合いについても知ることができる。

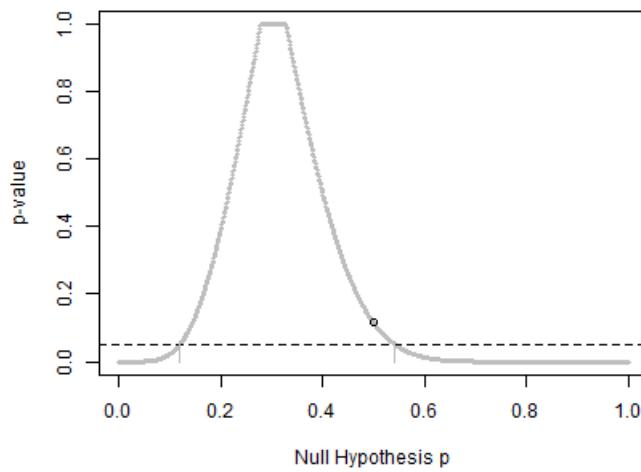
注意: 有意水準として $\alpha = 5\%$ がよく使われているが、そのことに科学的合理性はない。だから閾値 $\alpha = 5\%$ で計算した信頼区間を見て一喜一憂するのはバカげている。そのようにバカげたことをするくらいならば、閾値を決めずにP値函数の様子を直接眺めてどんぶり勘定で判断した方がましだと思われる。

注意: R言語では exactci パッケージを入れると、Clopper-Pearsonの信頼区間を与えるP値函数(`tsmethod = "central"`)と Sterneの信頼区間を与えるP値函数(`tsmethod = "minlik"`)を容易にプロットできる。

- <https://rdrr.io/cran/exactci/man/binom.exact.html> (<https://rdrr.io/cran/exactci/man/binom.exact.html>)

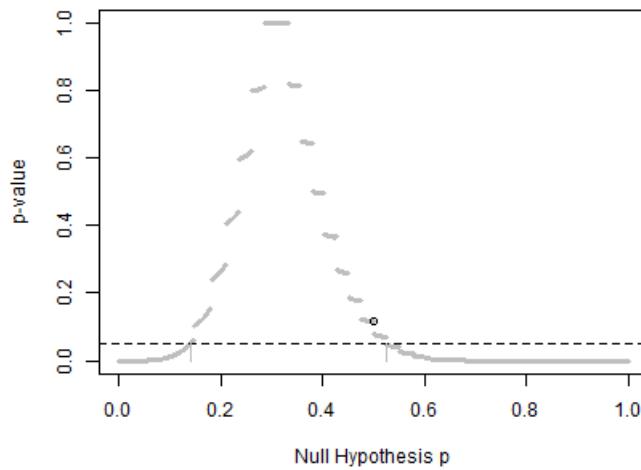
のExamplesに適当に `, plot=TRUE` を挿入してRunボタンを押してみよ。

```
julia> using RCall # requires the R language.  
julia> @library exactci # requires the package exactci of R.  
julia> binom_exact(6, 20; tsmethod = "central", plot = true)
```



RObject{VecSxp}

```
julia> binom_exact(6, 20; tsmethod = "minlik", plot = true)
```



RObject{VecSxp}

Exact two-sided binomial test (sum of minimum likelihood method)

```
data: 6L and 20L
number of successes = 6, number of trials = 20, p-value = 0.1153
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.1396 0.5254
sample estimates:
probability of success
0.3
```

In [22]: 1 `#binom_exact(6, 20; tsmethod = "central", plot = true)`

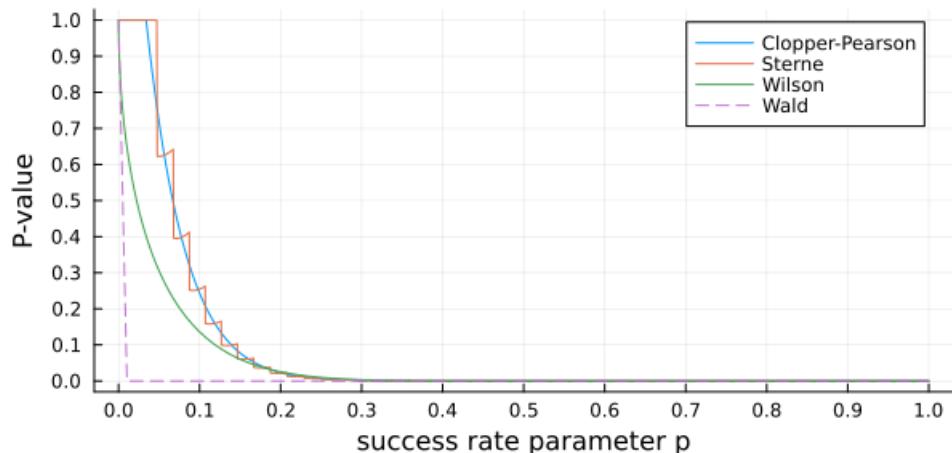
In [23]: 1 `#binom_exact(6, 20; tsmethod = "minlik", plot = true)`

```
In [24]: 1 gif_binomtest()
```

[Info: Saved animation to D:\OneDrive\work\Statistics\2022\images\pvaluefunction20.gif

Out[24]:

data: n = 20, k = 0

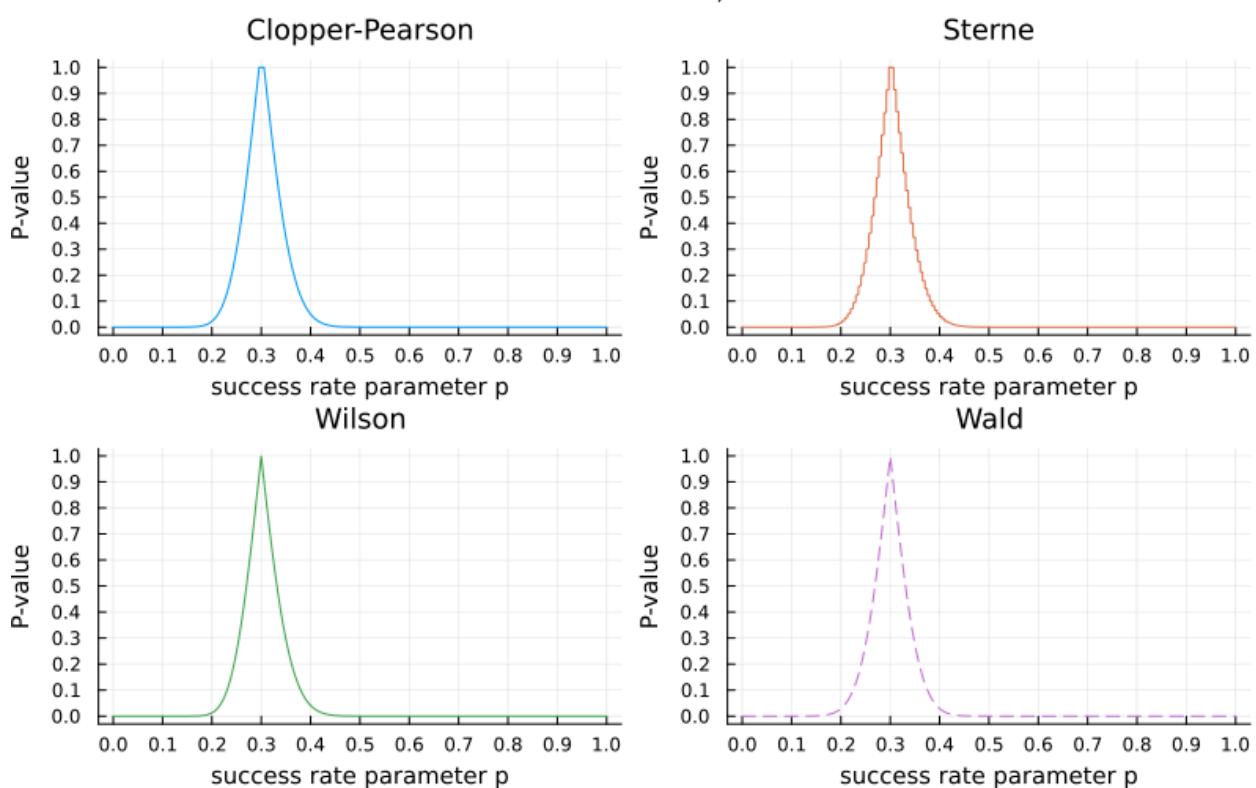


n を大きくするとP値函数の「幅」は狭くなる。

```
In [25]: 1 plot_binom_pvaluefunctions(; n = 100, k = 30)
```

Out[25]:

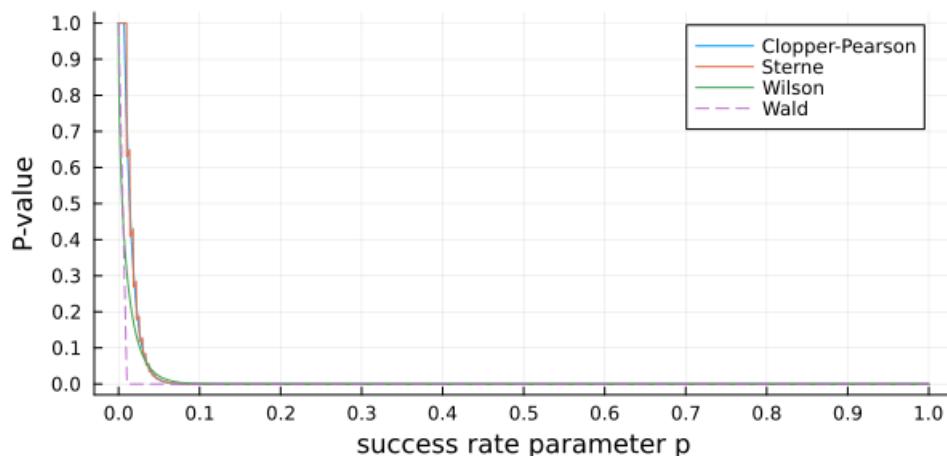
data: n = 100, k = 30



```
In [26]: 1 gif_binomtest(n = 100, fn = "images/pvaluefunction100.gif", fps=20)
```

[Info: Saved animation to D:\OneDrive\work\Statistics\2022\images\pvaluefunction100.gif

Out[26]:



PDFファイルではこの動画を見ることはできない。作成した動画は

- [\(https://github.com/genkuroki/Statistics/blob/master/2022/images/pvaluefunction20.gif\)](https://github.com/genkuroki/Statistics/blob/master/2022/images/pvaluefunction20.gif)
- [\(https://github.com/genkuroki/Statistics/blob/master/2022/images/pvaluefunction100.gif\)](https://github.com/genkuroki/Statistics/blob/master/2022/images/pvaluefunction100.gif)

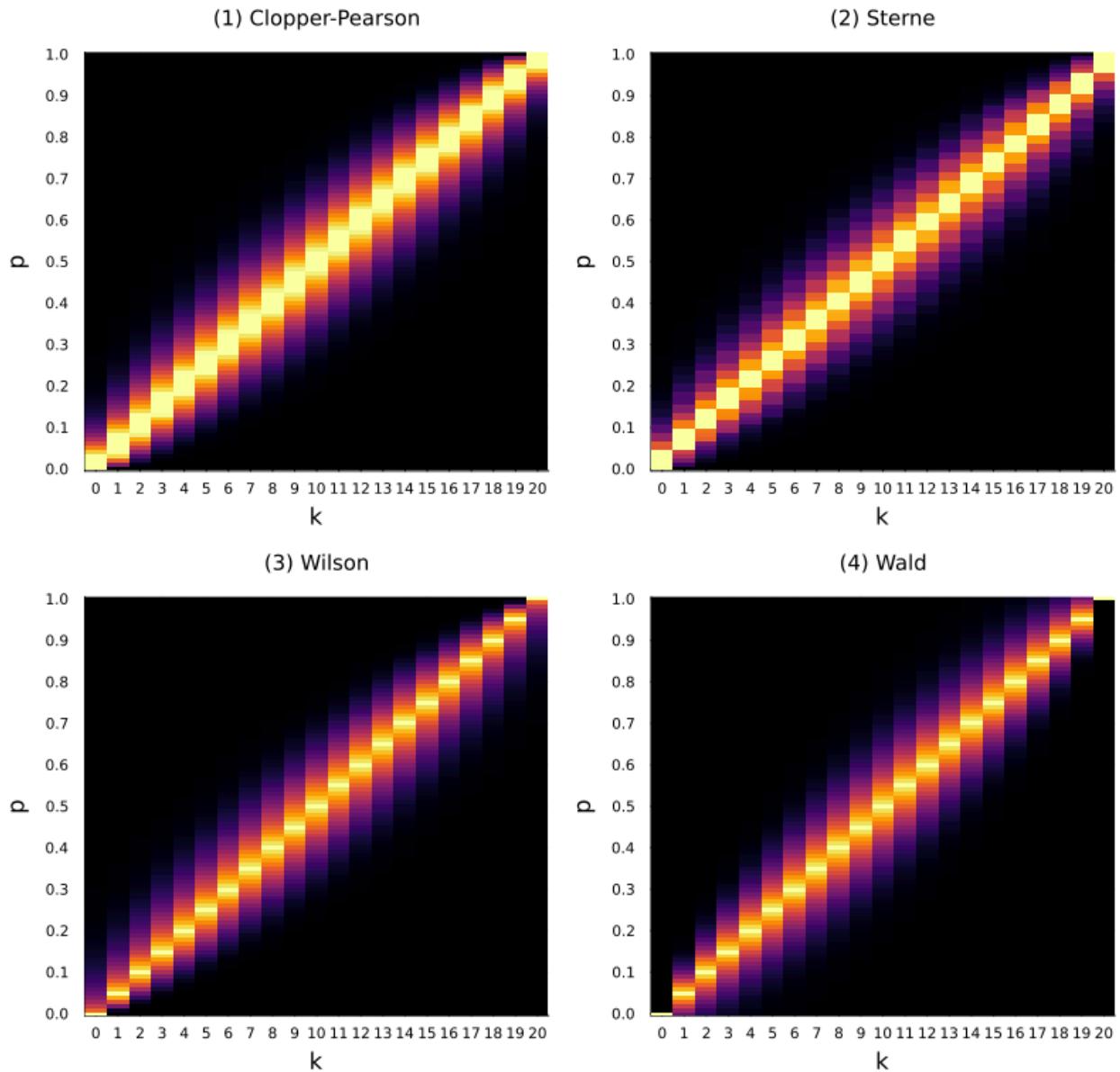
でも見ることができる。

pvalue(x|p) 達のヒートマップ (明るい所ほど値が大きい)

In [27]: 1 heatmap_binom_pvaluefunctions(; n = 20)

Out[27]:

data size: n = 20



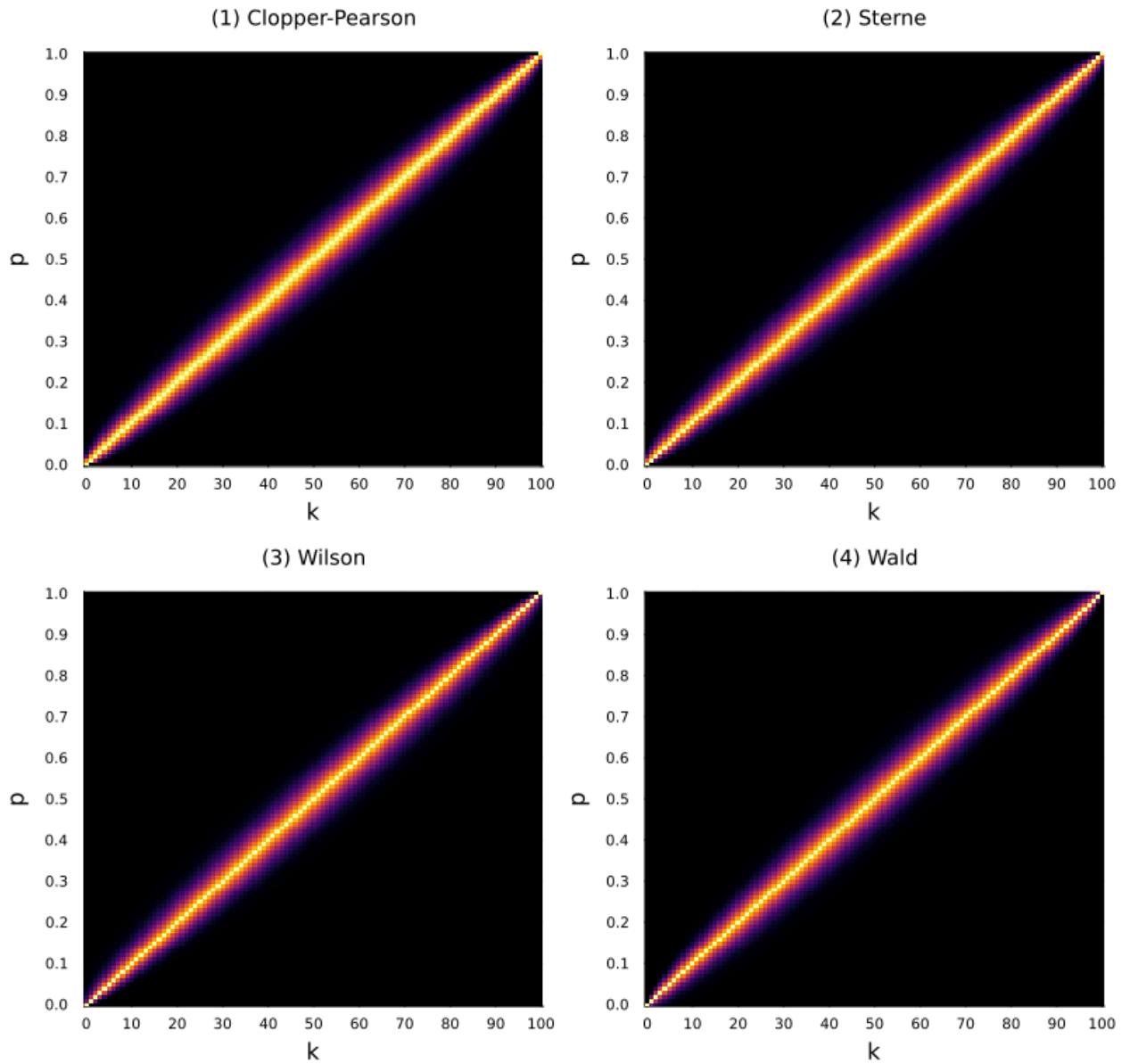
明るい部分ほどP値が大きい。P値が小さな部分はほぼ黒色になっている。そこでは「 n 回中 k 回成功」というデータの数値に成功回数パラメータ値が p の二項分布モデルが整合していないと考える。

例えば「 $n = 20$ 回中 $k = 6$ 回成功」というデータの数値が得られたとき $k = 6$ でのP値函数の「明るさ」を上のグラフで確認すると、 $p = 6/20 = 0.6$ で最も明るくなっている。そこから p が離れるほど暗くなることがわかる。P値による判定によれば、「 $n = 20$ 回中 $k = 6$ 回成功」というデータの数値に最も整合する二項分布モデルの成功確率パラメータ p の値は $p = 0.3$ であり、そこから p が離れるにつれて相性の良さは下がって行く。

```
In [28]: 1 heatmap_binom_pvaluefunctions(; n = 100, xtick = 0:10:100)
```

Out[28]:

data size: n = 100

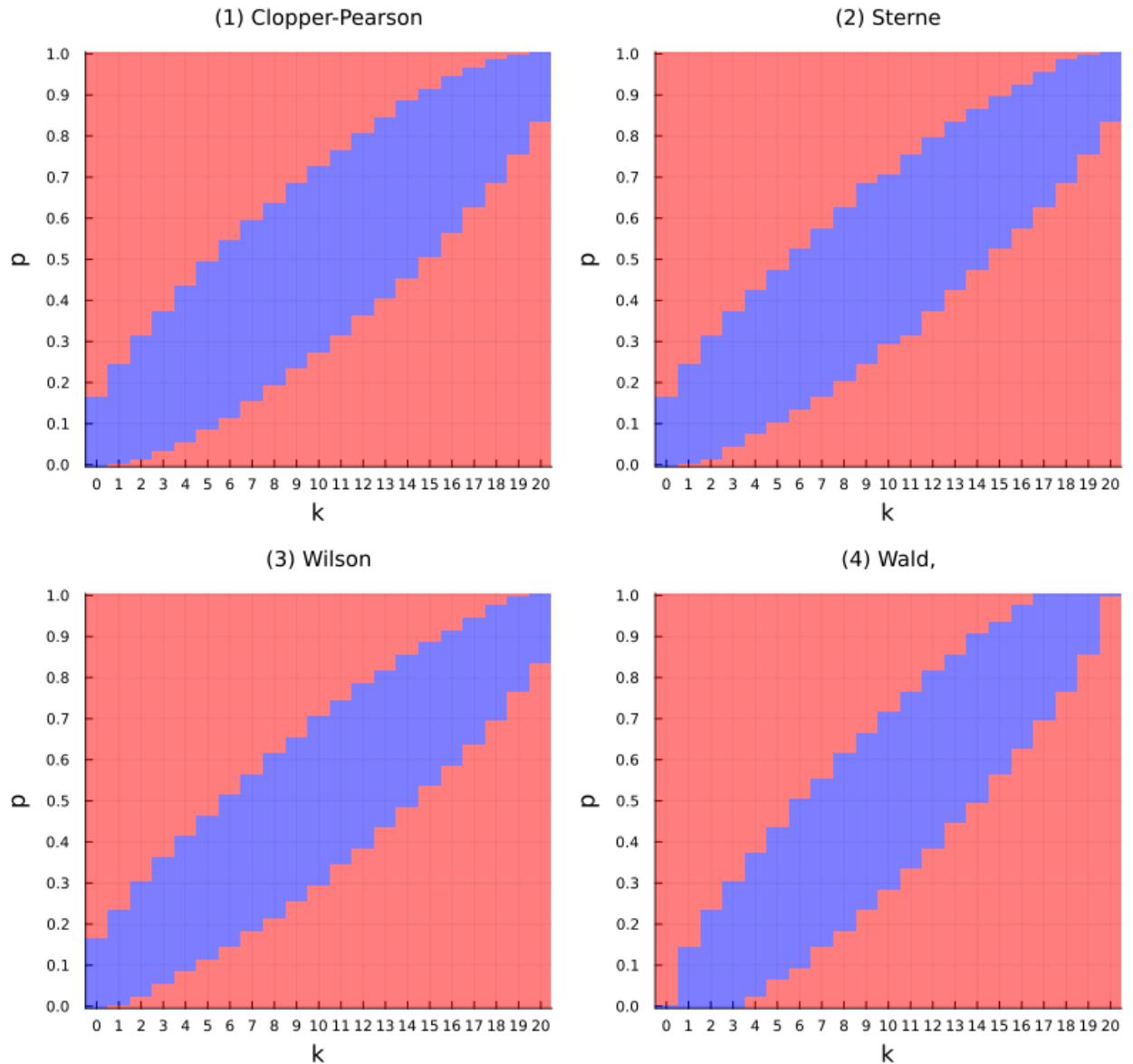


このように, n を大きくすると, 明るい部分が「細く」なる。これは, n を大きくすると, k を固定したときに得られるP値函数のグラフが「狭く」なることと同じ。

```
In [29]: 1 heatmap_binom_rejectionregions(; α = 0.05, n = 20)
```

Out[29]:

significane level: $\alpha = 0.05$, data size: $n = 20$

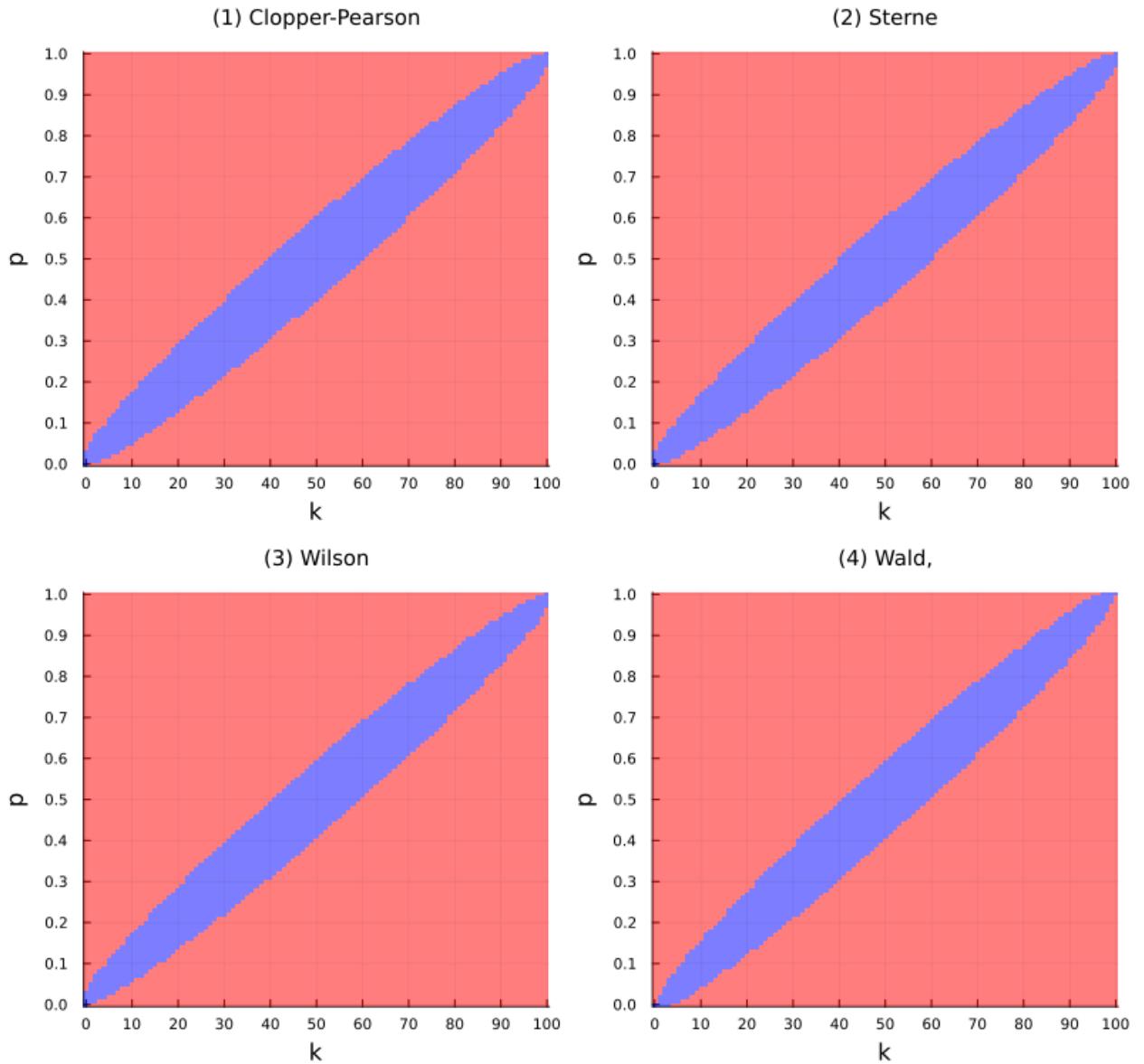


薄い赤の領域はP値が $\alpha = 5\%$ 未満になる部分であり、薄い青の領域はP値が $\alpha = 5\%$ 以上になる部分である。P値の定義の仕方に
よって結果は異なるが概ね似たような様子になっている。

```
In [30]: 1 heatmap_binom_rejectionregions(; α = 0.05, n = 100, xtick = 0:10:100)
```

Out[30]:

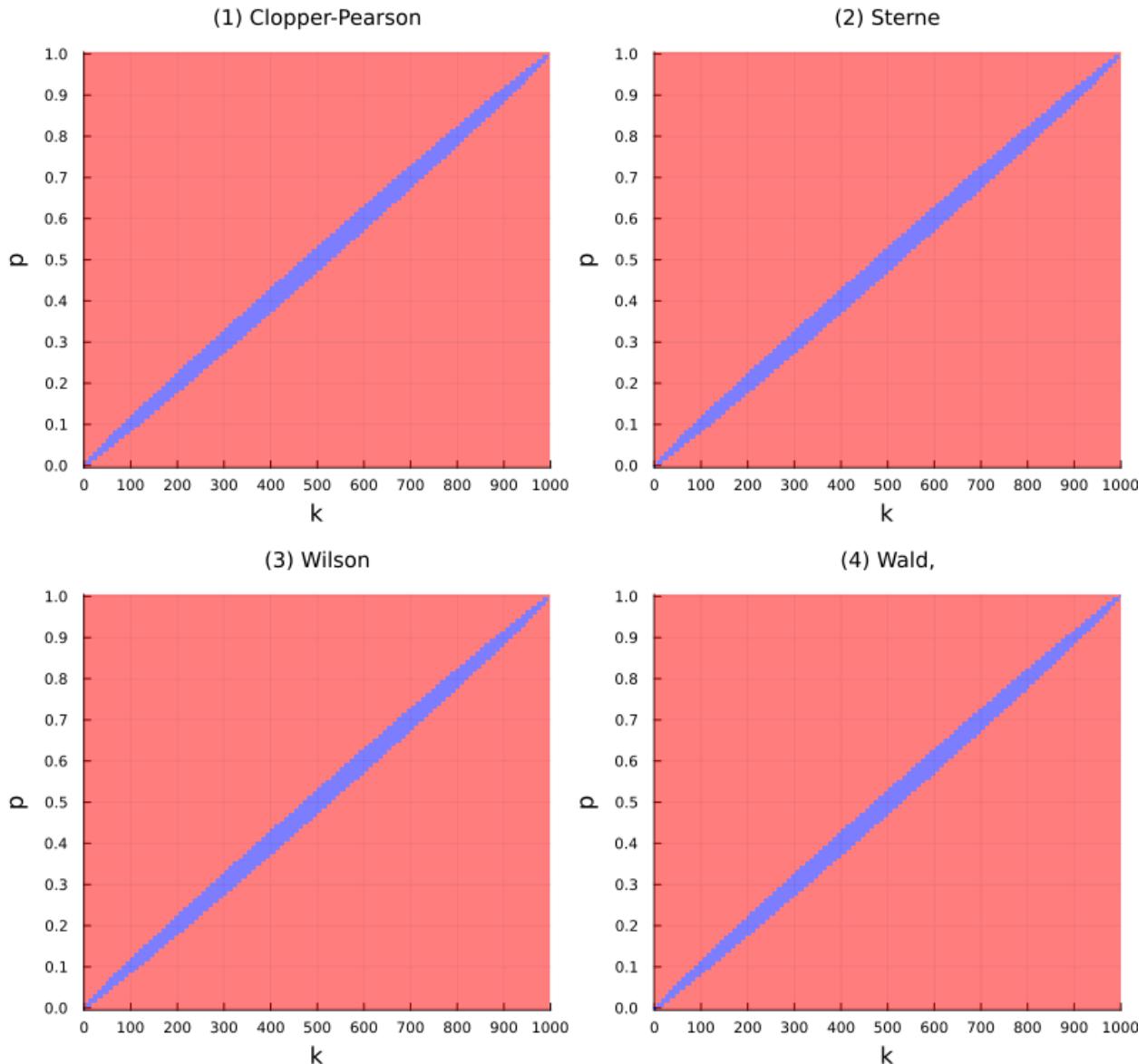
significane level: $\alpha = 0.05$, data size: $n = 100$



```
In [31]: 1 heatmap_binom_rejectionregions(; α = 0.05, n = 1000, xtick = 0:100:1000)
```

Out[31]:

significane level: $\alpha = 0.05$, data size: $n = 1000$



8 統計モデルやP値函数が「よい」かどうかの判断基準達

8.1 計算方法はシンプルな方がよい

このノートで扱っている4種のP値(Clopper-Pearson, Sterne, Wilson, Waldの信頼区間を与えるP値)では、計算効率を気にする実装を行うと、Sterneの信頼区間を与えるP値の実装がひどく複雑になってしまい、バグも発生し易い。

計算法がシンプルな方が実装のミスも少なく、計算効率もよいことが多い。

だから、たとえ別の方法が道具としての性能が勝っていたとしても、計算方法がシンプルな側を採用することは十分に検討に値する。

8.2 頑健な方がよい

統計モデルが現実のデータ生成法則をぴったり記述していると考えることは非現実的な場合が多い。

だから、統計モデルが現実のデータ生成法則からずれている場合について考えることも非常に重要になる。

統計モデルが現実のデータ生成法則からずれても統計分析の誤差が小さくなる傾向があるとき、その分析法は**頑健**(robust)であるという。

頑健な方が実践的な適用範囲が広がり、統計分析の失敗のリスクも減るので好ましい。

8.3 第一種の過誤の確率は有意水準に近い方がよい

第一種の過誤の確率は有意水準に近い方がよい。応用上特に重要なのは有意水準が $\alpha \leq 5\%$ と小さいところでの近似の精度である。

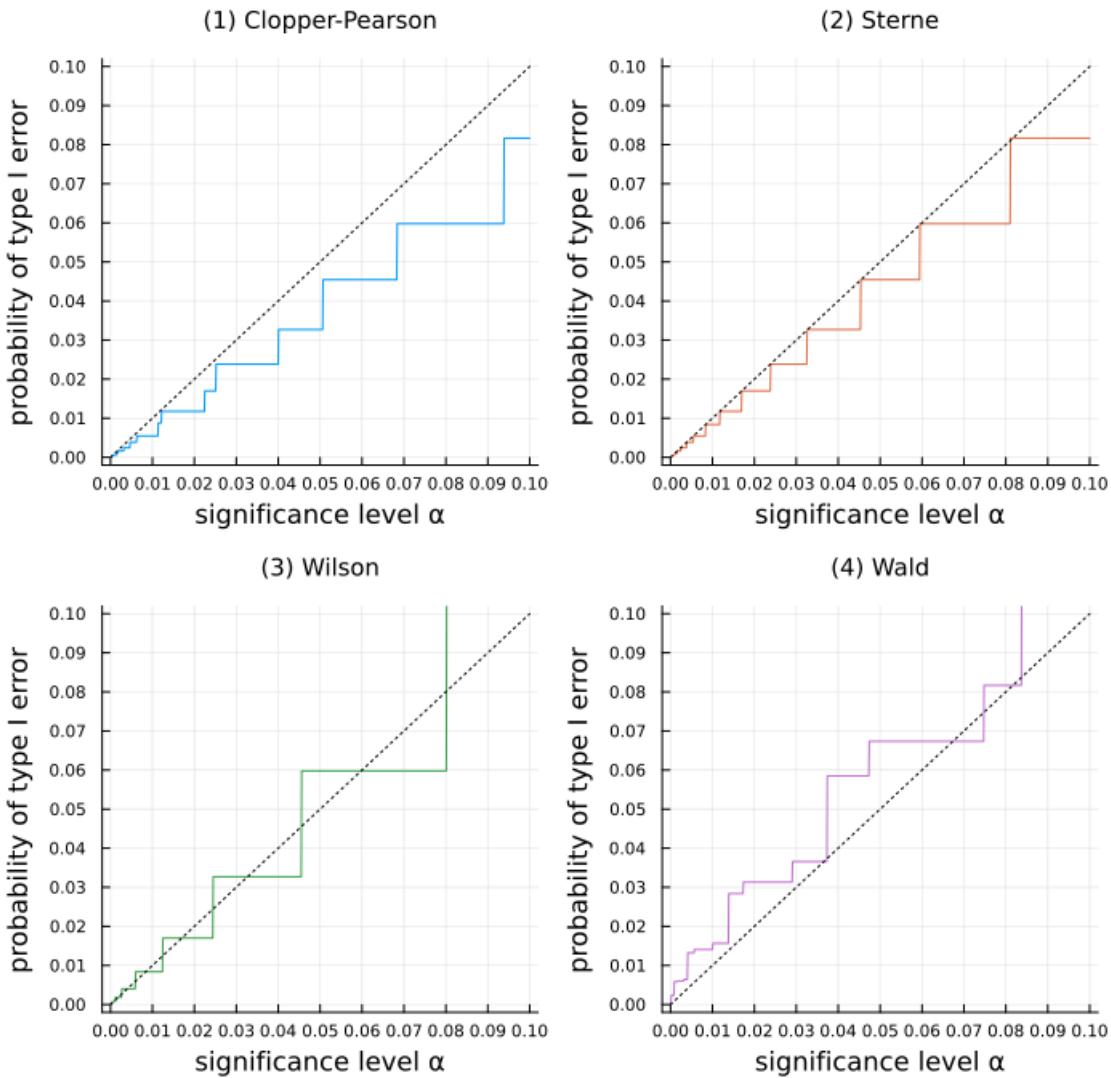
例えば、以下の二項分布モデルの4種のP値の例については、Clopper-Pearsonの信頼区間を与えるP値よりもSterneの信頼区間を与えるP値を使った方が誤差は小さくなっている。Waldの信頼区間を与えるP値よりもWilsonの信頼区間を与えるP値を使った方が誤差は小さくなっている。

In [32]: 1 Random.seed!(4649373)

Out[32]: TaskLocalRNG()

In [33]: 1 plot_sim_binomtest(100, 0.2; xmax=0.1, L=10^6)

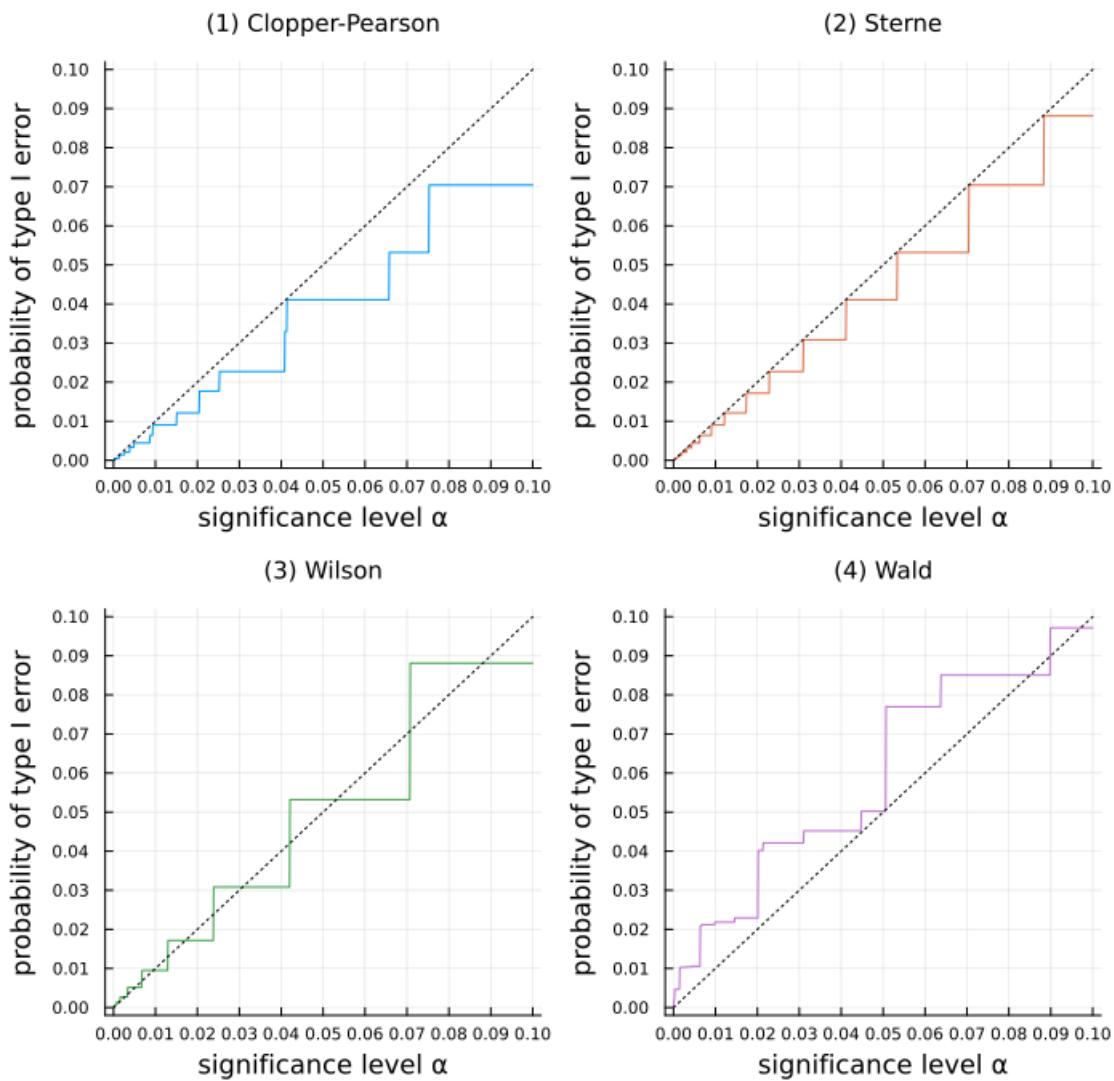
Out[33]: data: n = 100, p = p₀ = 0.2



```
In [34]: 1 plot_sim_binomtest(1000, 0.02; xmax=0.1, L=10^6)
```

Out[34]:

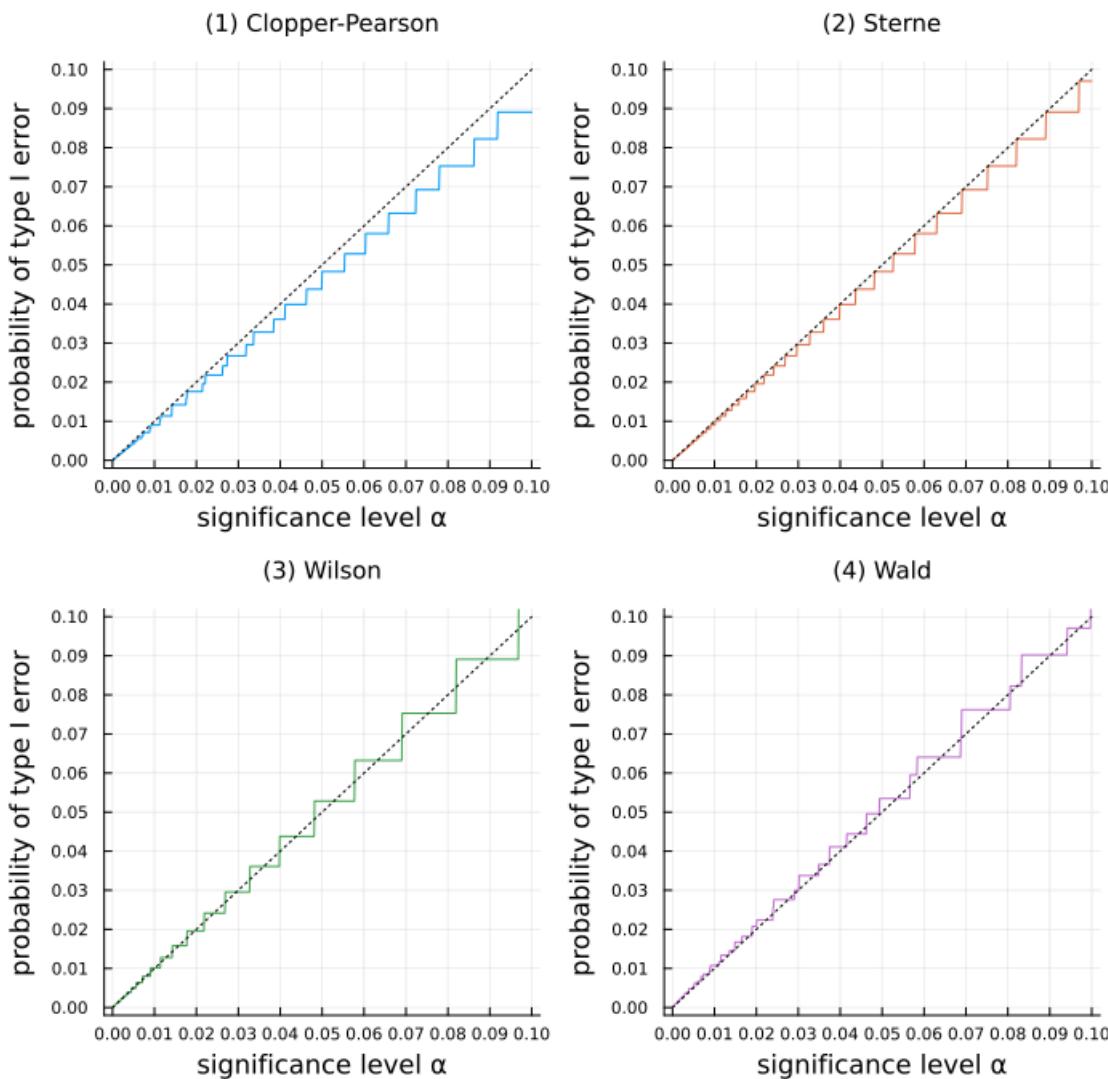
data: n = 1000, p = $p_0 = 0.02$



In [35]: 1 plot_sim_binomtest(1000, 0.2; xmax=0.1, L=10^6)

Out[35]:

data: n = 1000, p = p0 = 0.2



8.4 第一種の過誤の確率は有意水準以下である方がよい

「第一種の過誤の確率は有意水準以下である方がよい」という基準のことを **保守性** (conservativity)と呼ぶことがある。

上の例では, Clopper-Pearsonの信頼区間を与えるP値とSterneの信頼区間を与えるP値では, 第一種の過誤の確率が常に有意水準以下になり, **保守的** (conservative)である。しかし, Wilsonの信頼区間を与えるP値とWaldの信頼区間を与えるP値はその意味で保守的ではない(**リベラル** (liberal)ということがある)。

8.5 検出力は高い方がよい

大雑把に言うと, 仮説 $\theta = \theta_0$ のP値の **検出力** (power)はその仮説 $\theta = \theta_0$ と別の仮説 $\theta = \theta_1$ との分解能を意味する。

その文脈で別の仮説 $\theta = \theta_1$ を帰無仮説 $\theta = \theta_0$ の **対立仮説** (alternative hypothesis)と呼ぶ。

帰無仮説 $\theta = \theta_0$ のP値の対立仮説 $\theta = \theta_1 (\neq \theta_0)$ に関する **検出力** (power)は

- 対立仮説 $\theta = \theta_1$ 下の統計モデル内部で帰無仮説 $\theta = \theta_0$ のP値が有意水準 α 未満になる確率

と定義される。すなわち,

- X を対立仮説 $\theta = \theta_1$ 下の統計モデルに従う確率変数とし, データの数値を X としたとき仮説 $\theta = \theta_0$ が有意水準 α で棄却される確率

をそのように呼ぶ。

有意水準が等しいなら, 検出力は高い方がよい。

可能ならば任意の θ_1 についてそうであって欲しいがそのようにすることは一般には不可能である。

検出力については次の節も参照せよ。

まとめ: 検出力 = 対立仮説下の統計モデル内で帰無仮説が棄却される確率.

復習: 有意水準 ≈ 第一種の過誤の確率 = 帰無仮説下の統計モデル内で帰無仮説が棄却される確率.

9 Neyman-Pearsonの仮説検定

9.1 パラメータを持たない2つの統計モデルに関する仮説検定

データ x の生成のされ方に関するパラメータを持たない(もしくはすべてのパラメータの値が固定された)統計モデルが2つ与えられているとする。

簡単のために2つの統計モデルはそれぞれ確率密度函数 $p_0(x)$, $p_1(x)$ によって与えられているとする。

以下では統計モデル $p_0(x)$ を **帰無仮説** (null hypothesis)と呼び、統計モデル $p_1(x)$ を **対立仮説** (alternative hypothesis)と呼ぶ。(多くの場合に、帰無仮説は「薬Aには効果がない」のような否定したい仮説になり、対立仮説は例えば「薬Aには~という効果がある」のような正しいことを発見したい仮説になる。)

有意水準 $0 \leq \alpha \leq 1$ について、棄却領域と呼ばれる x の集合 R_α が与えられており、帰無仮説 $p_0(x)$ に従う確率変数 X_0 について、

$$P(X_0 \in R_\alpha) \leq \alpha$$

が成立していると仮定する。

このとき、棄却領域 R_α によって、**帰無仮説に関する有意水準 α の検定法** が与えられたと言い、データの数値 x が R_α に含まれるときに、帰無仮説 $p_0(x)$ は **棄却** されたと言う。

棄却領域 R_α で与えられた帰無仮説 $p_0(x)$ の検定法における対立仮説 $p_1(x)$ の **検出力** $\text{power}(R_\alpha)$ を、対立仮説 $p_1(x)$ に従う確率変数 X_1 を使って次のように定める:

$$\text{power}(R_\alpha) = P(X_1 \in R_\alpha).$$

対立仮説の検出力がより高い検定法は **より強力**(more powerful)であるという。

まとめ: 検出力 = power = 対立仮説下の統計モデル内で帰無仮説が棄却される確率.

復習: 第一種の過誤の確率 = 帰無仮説下の統計モデル内で帰無仮説が棄却される確率.

9.2 Neyman-Pearsonの補題

尤度比検定 (likelihood ratio test)と呼ばれる検定法を次の棄却領域 L_α によって定める:

$$L_\alpha = \left\{ x \mid \frac{p_0(x)}{p_1(x)} < c_\alpha \right\}.$$

ただし、定数 c_α は $P(X_0 \in L_\alpha) = \alpha$ をみたすものであるとする。

定理(Neyman-Pearsonの補題): 检定法 R_α で与えられた任意の検定法について

$$\text{power}(L_\alpha) \geq \text{power}(R_\alpha).$$

すなわち、尤度比検定は帰無仮説 $p_0(x)$ の検定法の中で対立仮説 $p_1(x)$ の検出力が最大である。

Neyman-Pearsonの補題: 帰無仮説と対立仮説がともにすべてのパラメータが固定された統計モデルであるときには、尤度比検定が最強力になる。

用語について補足説明: 対立仮説の下で帰無仮説が棄却されない確率を **第二種の過誤の確率** (type II error, βエラー, β-error)と呼ぶ。すなわち、

$$(第二種の過誤の確率) = 1 - (\text{検出力}).$$

仮説検定法では、第一種の過誤の確率 = 有意水準をある小さな値に定めたとき、第二種の過誤の確率を小さくしたい。

9.3 Neyman-Pearsonの補題の証明

X_0, X_1 はそれぞれ帰無仮説 $p_0(x)$ 、対立仮説 $p_1(x)$ に従う確率変数であるとする。このとき、

$$P(X_0 \in L_\alpha) = \alpha \geq P(X_1 \in R_\alpha).$$

L_α, R_α の補集合を L_α^c, R_α^c と書くことになると,

$$L_\alpha = (L_\alpha \cap R_\alpha) \cup (L_\alpha \cap R_\alpha^c), \quad R_\alpha = (L_\alpha \cap R_\alpha) \cup (L_\alpha^c \cap R_\alpha). \quad (*)$$

$P(X_0 \in L_\alpha) \geq P(X_0 \in R_\alpha)$ より,

$$P(X_0 \in L_\alpha \cap R_\alpha^c) \geq P(X_1 \in L_\alpha^c \cap R_\alpha). \quad (\#)$$

$x \in L_\alpha$ かつ $p_1(x) > p_0(x)/c_\alpha$ が同値であることと、すぐ上で示した(+)と、 $x \notin L_\alpha$ かつ $p_0(x)/c_\alpha \geq p_1(x)$ が同値であることを順番に使うと、

$$\begin{aligned} P(X_1 \in L_\alpha \cap R_\alpha^c) &= \int_{L_\alpha \cap R_\alpha^c} p_1(x) dx \geq \frac{1}{c_\alpha} \int_{L_\alpha \cap R_\alpha^c} p_0(x) dx \\ &= \frac{1}{c_\alpha} P(X_0 \in L_\alpha \cap R_\alpha^c) \geq \frac{1}{c_\alpha} P(X_0 \in L_\alpha^c \cap R_\alpha) \\ &= \frac{1}{c_\alpha} \int_{L_\alpha^c \cap R_\alpha} p_0(x) dx \geq \int_{L_\alpha^c \cap R_\alpha} p_1(x) dx = P(X_1 \in L_\alpha^c \cap R_\alpha). \end{aligned}$$

上と同様にして(*)より、

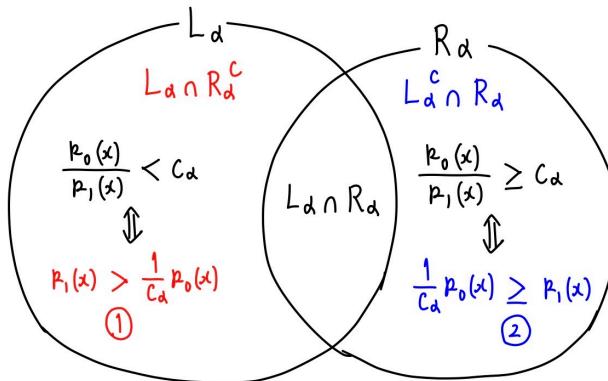
$$P(X_1 \in L_\alpha) - P(X_1 \in R_\alpha) = P(X_1 \in L_\alpha \cap R_\alpha^c) - P(X_1 \in L_\alpha^c \cap R_\alpha) \geq 0.$$

これで示すべきことが示された。

証明終

Neyman-Pearson Lemma

$$P(X_0 \in L_\alpha) = \alpha \geq P(X_0 \in R_\alpha) \Rightarrow P(X_0 \in L_\alpha \cap R_\alpha^c) \geq P(X_0 \in L_\alpha^c \cap R_\alpha) \quad (\#)$$



$$\begin{aligned} P(X_1 \in L_\alpha \cap R_\alpha^c) &= \int_{L_\alpha \cap R_\alpha^c} p_1(x) dx \stackrel{\textcircled{1}}{\geq} \frac{1}{c_\alpha} \int_{L_\alpha \cap R_\alpha^c} p_0(x) dx \\ &= \frac{1}{c_\alpha} P(X_0 \in L_\alpha \cap R_\alpha^c) \stackrel{(\#)}{\geq} \frac{1}{c_\alpha} P(X_0 \in L_\alpha^c \cap R_\alpha) \\ &= \frac{1}{c_\alpha} \int_{L_\alpha^c \cap R_\alpha} p_0(x) dx \stackrel{\textcircled{2}}{\geq} \int_{L_\alpha^c \cap R_\alpha} p_1(x) dx \\ &= P(X_1 \in L_\alpha^c \cap R_\alpha) \end{aligned}$$

$$P(X_1 \in L_\alpha \cap R_\alpha^c) \geq P(X_1 \in L_\alpha^c \cap R_\alpha) \Leftrightarrow \underbrace{P(X_1 \in L_\alpha) \geq P(X_1 \in R_\alpha)}, \quad \text{示したいこと。}$$

9.4 尤度比検定の例: 正規分布モデルの場合

$\mu_1 \in \mathbb{R}$ であるとし、 $p_0(x), p_1(x)$ が次の場合について考える:

$$p_0(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad p_1(x) = \frac{e^{-(x-\mu_1)^2/2}}{\sqrt{2\pi}}$$

このとき、

$$\frac{p_0(x)}{p_1(x)} = \exp(\mu_1^2 - 2\mu_1 x).$$

これがある値未満になるという条件は、

- $\mu_1 > 0$ の場合には、 x がある値 a より大きいという条件と同値になり、
- $\mu_1 < 0$ の場合には、 x がある値 a より小さいという条件と同値になる。

x がそのようになる標準正規分布における確率はそれぞれ

- $\mu_1 > 0$ の場合: $1 - \text{cdf}(\text{Normal}(0, 1), a)$,
- $\mu_1 < 0$ の場合: $\text{cdf}(\text{Normal}(0, 1), a)$

なので、これらが有意水準 α に等しくなる $a = a_\alpha$ はそれぞれ

- $\mu_1 > 0$ の場合: $a_\alpha = \text{quantile}(\text{Normal}(0, 1), 1 - \alpha)$,
- $\mu_1 < 0$ の場合: $a_\alpha = \text{quantile}(\text{Normal}(0, 1), \alpha)$

になる。ゆえに、帰無仮説 $p_0(x)$ と対立仮説 $p_1(x)$ に関する対数尤度比検定を与える棄却領域はそれぞれ次のようになる:

- $\mu_1 > 0$ の場合: $L_\alpha = \{x \mid x > a_\alpha\}$,
- $\mu_1 < 0$ の場合: $L_\alpha = \{x \mid x < a_\alpha\}$.

これは以下のように解釈される。

- 対立仮説の平均 μ_1 が帰無仮説の平均 0 より大きい場合には、データ x の値がある値より大きくなると、帰無仮説よりも対立仮説と整合的になり、帰無仮説が棄却される。
- 対立仮説の平均 μ_1 が帰無仮説の平均 0 より小さい場合には、データ x の値がある値より小さくなると、帰無仮説よりも対立仮説と整合的になり、帰無仮説が棄却される。

この例より、すべての μ_1 について、最強力になるような単一の検定法が存在しないことも分かる。

注意:しかし、**不偏** という非常に強い条件を満たす検定法の中ではすべての対立仮説パラメータ値 θ_1 について最強力になるような検定が存在することがある(一様最強力不偏検定(uniformly most powerful unbiased test, UMPU test))。20世紀の検定の理論はこのような方向で整備された。詳しくは次の教科書を参照せよ:

- Erich L. Lehmann, Joseph P. Romano, *Testing Statistical Hypotheses*, Third edition, 2005

しかし、一様最強力不偏検定(UMPU test)の理論は実践的にはそう役に立つわけではない。

特に、離散分布モデルを用いてUMPU検定を作るためには特定のデータによって帰無仮説が確率的に棄却される場合が出て来ることを受け入れる必要があり、通常の科学的研究でそれは受け入れらるる条件ではない。離散分布モデルによる実践的に使用されている検定法はどれもUMPU検定ではない。

とは言うものの、たとえ実践的ではないとしても、UMPU検定の話も数学的には結構面白い。興味がある人は、UMPU検定について自分で調べ、二項分布モデルなどの場合についてUMPU検定をコンピュータで実装してみるとよいだろう。

9.5 Neyman-Pearsonの仮説検定に関する極端な解釈の普及の問題

Neyman-Pearsonの仮説検定では、帰無仮説についての第一種の過誤の確率を有意水準 α の設定によって上からおさえ、対立仮説の検出力を可能な限り大きくすることによって、帰無仮説と対立仮説をデータの数値によって比較する方法である。

Neyman-Pearsonの仮説検定では、帰無仮説が棄却された場合には「帰無仮説よりも対立仮説の方が **相対的に妥当である**」と **暫定的に** 判断してもよいだろう。その判断は単に帰無仮説と対立仮説の相対的な比較の下での暫定的な判断に過ぎず、対立仮説の側が正しいと最終的に判断することとは異なる。

しかし、「帰無仮説が棄却されなかつた場合には帰無仮説を受容し、帰無仮説が棄却された場合には対立仮説の側が正しいと判断する」ことは、Neyman-Pearsonの仮説検定の本質的要請であるとみなす極端な解釈が広く普及している。

その解釈が極端過ぎることは、少なくとも、

- E. S. Pearson, *Statistical Concepts in the Relation to Reality*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 17, No. 2 (1955), pp. 204-207. [[link](https://www.jstor.org/stable/2983954)] [[pdf](https://errorstatistics.files.wordpress.com/2021/02/pearson_1955-stat-concepts-reality.pdf)]

のp.206からの以下の段落の引用を読めば明らかだと思われる:

Professor Fisher's next objection is to the use of such terms as the "acceptance" or "rejection" of a statistical hypothesis, and "errors of the first and second kinds". It may be readily agreed that in the first Neyman and Pearson paper of 1928, more space might have been given to discussing how the scientific worker's attitude of mind could be related to the formal structure of the mathematical probability theory that was introduced. Nevertheless it should be clear from the first paragraph of this paper that we were not speaking of the final acceptance or rejection of a scientific hypothesis on the basis of statistical analysis. We speak of accepting or rejecting a hypothesis with a "greater or less degree of confidence". Further, we were very far from suggesting that statistical methods should force an irreversible acceptance procedure upon the experimenter. Indeed, from the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is "a means of learning", for we remark: "the tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision". No doubt we could more aptly have said "his final or provisional decision"; even scientists, if they are employed in research departments by industry or government, may sometimes have to give a final decision.

翻訳:

Fisher教授の次の反対は、統計的仮説の「受容」または「棄却」、「第一種および第二種の過誤」といった用語の使用に対するものである。NeymanとPearsonによる1928年の最初の論文では、科学的な研究者の心構えが導入された数学的確率論の形式的構造にどのように関係付けられるか、について論じるためにより多くのスペースが設けることができたということについてはただちに同意できる。しかし、この論文の最初の段落から明らかなように、我々は科学的仮説の統計的分析に基く最終的受容や最終的棄却について語ってはいない。我々は、「信頼性の度合いの大小」によって仮説を受容するか棄却するかについて語っている。さらに、我々は、統計的手法が実験者に不可逆的な受容手順を強いることの示唆をまったくしていない。実際我々は、科学的探求において統計的検定は「学習の手段」であるというFisher教授の考えを最初から共有していた。その証拠に、我々は「検定そのものは最終的判断を与えないが、最終的決定のための道具として研究者にとって役に立つ。」と注意している。間違いなく我々はより適切に「研究者の最終的または暫定的な判断」と言えた。たとえ産業界や政府の研究部門に雇用されていて最終的決定を下さなければいけない科学者達であってもそうである。

このように、Neyman-PearsonのPearson氏は「Neuman-Pearsonの仮説検定では、検定の手続きによって仮説の受容と棄却に関する最終的判断を強いる」という解釈を明瞭に否定している。

検定の手続きは、科学的探求における「学習の手段」であり、最終的判断ではなく暫定的判断を与え、道具としては最終的決定にも役に立つと述べている。さらに、帰無仮説または対立仮説の絶対的な受容と棄却ではなく、信頼度の度合いの大小による相対的な判断に過ぎないことを説明している。

おそらく、「受容」(acceptance)と「棄却」(rejection)という1か0かの極端な判断を強いるニュアンスを与える用語を専門用語として採用してしまったことが悪かったのだろう。この「受容」「棄却」は現代の教科書でも使われている場合がある。Pearson氏にならって極端な意味に解釈しないように注意することが必要である。

なお、このノート群では「受容する」という言い方は避けて「棄却しない」という言い方を用いることにしている。(通常の意味で)「棄却しない」ことはただちに(通常の意味で)「受容する」ことを意味せず、「判断を保留する」ことが可能なことも意味している。

さらに次の解説も参照せよ:

- Deborah Mayo, Erich Lehmann's 100 Birthday: Neyman Pearson vs Fisher on P-values, Posted on November 19, 2017. [[link](https://errorstatistics.com/2017/11/19/erich-lehmans-100-birthday-neyman-pearson-vs-fisher-on-p-values/) (<https://errorstatistics.com/2017/11/19/erich-lehmans-100-birthday-neyman-pearson-vs-fisher-on-p-values/>)]

そこで引用されている次の論文では、Fisher流の検定とNeyman-Pearson流の検定の良い点を組み合わせて使うことを提案している:

- E. L. Lehmann, The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?, Journal of the American Statistical Association, Vol. 88, No. 424 (Dec., 1993), pp. 1242-1249. [[link](https://www.jstor.org/stable/2291263) (<https://www.jstor.org/stable/2291263>)], [[pdf](https://errorstatistics.files.wordpress.com/2013/11/lehmann_1-theory-or-2.pdf) (https://errorstatistics.files.wordpress.com/2013/11/lehmann_1-theory-or-2.pdf)]

非常に残念なことに、極端な解釈によって、Fisher流の検定とNeyman-Pearson流の検定の違いを強調し、対立を煽っているかのような解説を見ることがある。そのような解説では、Fisher流の検定とNeyman-Pearson流の検定の考え方を混ぜることは許さないという方針になりがちである。(これは上で引用したLehmann氏の論文とは正反対の態度である。)

そのようなことをしても科学的に建設的ではなく、現代における最良の視点から、科学的常識に沿った稳健で有用な解釈を普及させる方がよいと思われる。

補足: 上で引用した論文の著者のLehmann氏([Erich Leo Lehmann \(1917-2009\)](https://en.wikipedia.org/wiki/Erich_Leo_Lehmann) (https://en.wikipedia.org/wiki/Erich_Leo_Lehmann))はNeyman氏の弟子の有名な統計学者である。

補足: このノートでは「検定の手続きが1か0かの最終的判断を強いる」という極端な解釈を「統計学は科学のお墨付きを得るための手段ではない」という言い方でも明瞭に否定している。

補足: このノートで採用したP値を「仮説 $\theta = \theta_0$ 下の統計モデルとデータの数値 x の相性の良さ(compatibility)の指標」とみなす考え方方はFisherの有意性検定の考え方方に近いように思われる。そして、それと同時に、Neyman-Pearson的に対立仮説の検出力も重要な立場も採用している(その縛り抜きには無意味な検定法を排除できない)。

9.6 このノートの立場

検定の手続きは、例えば、「薬Aに効果はない」というような仮説をモデルのパラメータに関する仮説 $\theta = 0$ で代替し、「薬Aに効果はない」という仮説は疑わしい(もしくは誤りである)という結論を出したい状況で使われていることが多い。

そのとき「薬Aに効果はない」のような否定したい特別な仮説が帰無仮説と呼ばれることになり、仮説 $\theta = 0$ の 0 という値は特別な意味を持つものとして固定されていることが多い。

しかし、このノートでは、検定にかけられる仮説 $\theta = \theta_0$ を定めるパラメータ値 θ_0 は1つに固定されおらず、任意の値を取り得ることを強調し、可能なすべての θ_0 に関する検定の手続きによって信頼区間を定義している。

すなわち、このノートにおける信頼度 $1 - \alpha$ 信頼区間の定義は、仮説 $\theta = \theta_0$ がデータの数値 x による有意水準 α の検定の手続きで棄却されないような数値 θ_0 全体の集合だとしている。

このスタイルの利点は、信頼区間は、「薬Aに効果はない」という仮説だけではなく、任意の値 θ_0 に関する「薬Aの効果は θ_0 である」という仮説の検定結果の情報もすべて含んでいることが最初から明らかになることである。

信頼区間を見れば以下のような情報が得られる:

- a より小さな値を信頼区間は含まないので、薬Aの効果は概ね a 以上だと考えてよさそうだ.
- 信頼区間は 0 を含んでいるので、データの数値 x からは薬Aの効果の有無はわからなかった。(注意: 「効果はない」と判断するのは誤り。)
- a より大きな値を信頼区間は含まないので、薬Aの効果は概ね a 以下だと考えてよさそうだ.

このようなことが可能になったのは、検定にかける仮説 $\theta = \theta_0$ の θ_0 を自由に動かせる設定を採用したからである。

信頼区間は有意水準 α という閾値を設定することによって定義される。

しかし、 $\alpha = 5\%$ のような慣習化された有意水準をP値が下回るか否かに一喜一憂するというような愚行に警戒する必要が生じてしまった。

この人為的な閾値の設定に関わる問題について、このノートではP値函数について紹介することによって対処したつもりである。

データの数値 x が与えられたとき、パラメータ値 θ_0 に対して、仮説 $\theta = \theta_0$ のP値を対応させる函数をP値函数と呼ぶのであった。

信頼度 $1 - \alpha$ の信頼区間の定義は、P値函数の値が有意水準以上になるパラメータ値 θ_0 全体の集合なので、P値函数のグラフを描けば視覚的に信頼区間がどうなるかも確認できる。

すなわち、P値函数のグラフはすべての信頼度 $1 - \alpha$ に関する信頼区間の情報を持っている。

ゆえに、P値函数のグラフを描けば、有意水準 α を固定せずに、統計モデルのパラメータ値とデータの数値の相性の良さ(P値が小さいほど相性が悪い)を確認できる。

P値函数のグラフを利用すれば、有意水準を設定しなくても、そこから科学的に有益な情報を引き出すことが可能である。

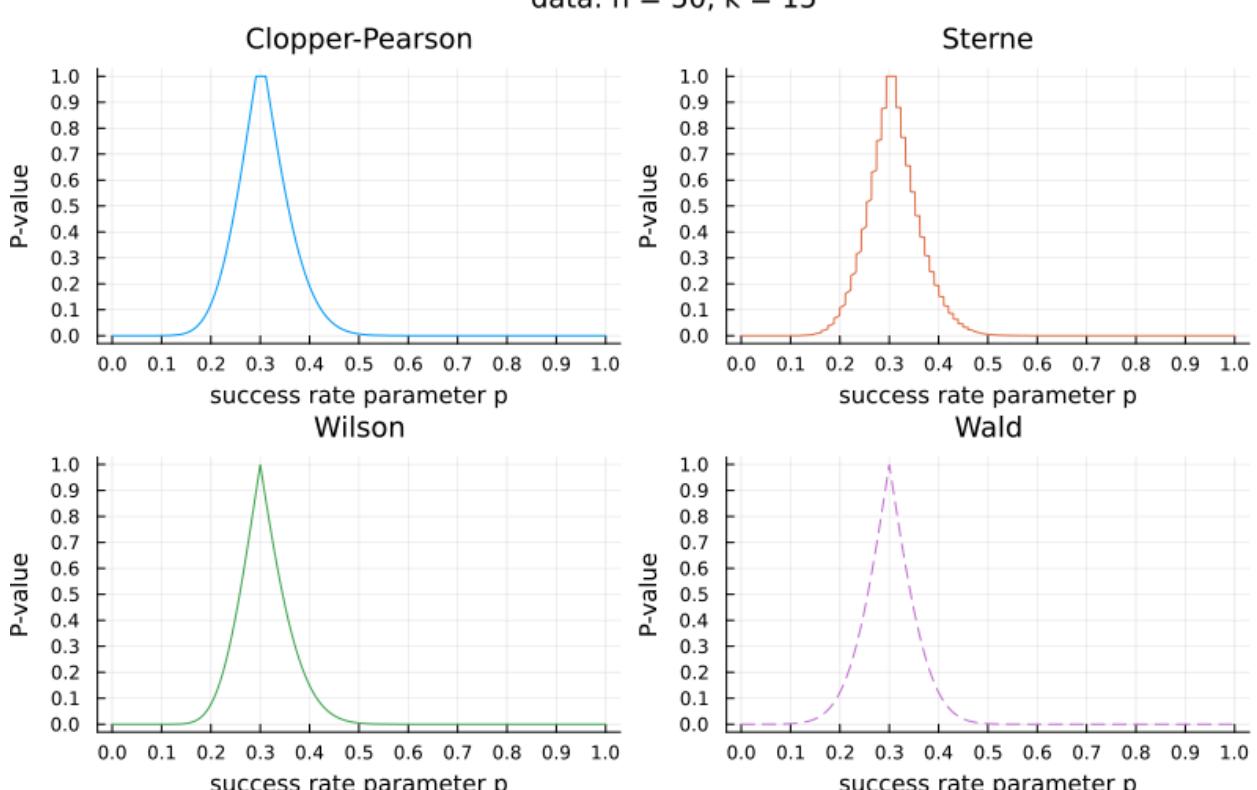
この考え方はRothman氏達の有名な疫学の教科書 *Modern Epidemiology* でも紹介されている。

P値函数のグラフは、ある特定の信頼度(例えば 95%)に関する信頼区間よりも、圧倒的に豊富な情報を持っている。まず、信頼度を自由に変えられる。さらに、P値函数が最大になるパラメータ値は、統計モデルがデータの数値に最もフィットするようなパラメータ値だと解釈される。実際、それは多くの場合に最尤推定値またはそれに近い値になっている。

P値函数がどういったものだったか忘れた人のために以下に二項分布モデルのP値函数のグラフを以下に掲載しておこう。P値函数のグラフは多くの場合に以下の図のように「とんがり帽子」のような形になる。そしてその「とんがり帽子」の幅が概ね信頼区間の

In [36]: 1 plot_binom_pvaluefunctions(; n = 50, k = 15)

Out[36]:



P値函数のグラフの例とR言語による書き方に関する解説が以下の場所にある:

- Denis Infanger, P-value functions: Tutorial using the pvaluefunctions package, 2021-11-30. [link (<https://cran.r-project.org/web/packages/pvaluefunctions/vignettes/pvaluefun.html>)]

- Plot a P-value Function from one or two Confidence Intervals. [[link](https://epijim.shinyapps.io/episheet_shiny/) (https://epijim.shinyapps.io/episheet_shiny/)]

以下のリンク先も面白い。

- https://twitter.com/ken_rothman/status/1258552231286063106
(https://twitter.com/ken_rothman/status/1258552231286063106)
- <https://twitter.com/MinatoNakazawa/status/656235447220224000>
(<https://twitter.com/MinatoNakazawa/status/656235447220224000>)
- <https://twitter.com/MinatoNakazawa/status/1202738229180416000>
(<https://twitter.com/MinatoNakazawa/status/1202738229180416000>)
- <https://twitter.com/MinatoNakazawa/status/1202893936181645312>
(<https://twitter.com/MinatoNakazawa/status/1202893936181645312>)

今後の目標: データ x とパラメータ値 θ_0 に対してP値を対応させる函数 $\text{op}\{\text{pvalue}\}(x|\theta_0)$ の例を構成して利用すること。

10 よくある誤解

10.1 P値の定義と使い方の復習

現実世界から得るデータ x の生成のされ方に関するパラメータ θ を持つ統計モデルが与えられているとき, データの数値 x に関する仮説 $\theta = \theta_0$ のP値は

- 仮説 $\theta = \theta_0$ 下の統計モデル内部でデータの数値 x 以上に極端な値が生成される確率もしくはその近似値と定義されるのであった。(「数値 x 以上に極端な値」の意味は別に定義しなければいけない。)

そして, そのようにして計算されるP値は, 仮説 $\theta = \theta_0$ 下の統計モデルとデータの数値の相性の良さ(comaptibility)の指標として使われ, 適当な閾値(有意水準)を設けて, P値が小さすぎる場合には「相性が悪い」と判定するのであった(これが検定の手続き)。

10.2 問題: P値に関するよくある誤解

データの数値 x に関する仮説 $\theta = \theta_0$ のP値を以下では単にP値と呼ぶこととする。

以下の主張の中で誤解であるものすべて挙げよ。

- (1) P値は現実世界において仮説 $\theta = \theta_0$ が正しい確率である。
- (2) 検定の手続きでは, P値が有意水準より小さいときには, 仮説 $\theta = \theta_0$ は科学的に疑わしいと考える。
- (3) 検定の手続きでは, P値が有意水準以上のときには, 仮説 $\theta = \theta_0$ は科学的に否定されたと考える。
- (4) P値が小さい結果ほど重要な結果である。

正しい主張を誤解だとみなさないように注意して欲しい。

この問題の解答例はこのノートの最後の方に掲載する。

10.3 信頼区間の定義の復習と使い方の復習

データの数値 x が与えられたとき, 有意水準 α の検定の手続きによって棄却されないパラメータ値 $\theta = \theta_0$ 全体の集合を, パラメータ θ に関する信頼度 $1-\alpha$ の信頼区間と呼ぶのであった。 $\alpha=5\%$ のときには 95% 信頼区間と呼ばれる。

10.4 問題: 信頼区間にに関するよく見る誤解

以下の主張が誤りもしくは杜撰である理由を説明せよ。

- (1) 95% 信頼区間の 95% は確率ではなく, 割合である。
- (2) 平均値の 95% 信頼区間にについて考える。このとき, 現実の母集団からの無作為抽出を繰り返して, 平均値の 95% 信頼区間を計算し直すとき, 現実の母集団の平均値を含む区間達の割合は 95% になる。

この問題の解答例はこのノートの最後に掲載する。

10.5 平均の信頼区間達の視覚化

後で別のノートで解説する予定の正規分布の標本分布モデルと $\$t\$$ 分布を用いて計算される平均の信頼区間を正規分布の標本達および正規分布以外の分布の標本達についてプロットしてみよう。

In [37]: 1 Random.seed!(4649373)

Out[37]: TaskLocalRNG()

10.5.1 正規分布の標本達から得られる平均の信頼区間達

標本(データ)を生成する分布が正規分布の場合は、信頼区間を計算するために用いる統計モデルで標本の生成のされ方がぴったり記述可能な場合になっている。

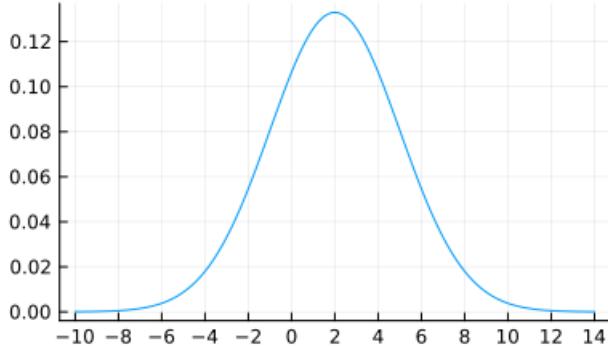
これは実践的な統計分析においては非現実的な想定である。

この場合には理論通りの結果がそのまま成立している。

In [38]: 1 dist = Normal(2, 3)
2 @show μ, σ = mean(dist), std(dist)
3 plot(dist, μ-4σ, μ+4σ; label="", title="\\$(distname(dist))",
4 xtick=-100:2:100)

(μ, σ) = (mean(dist), std(dist)) = (2.0, 3.0)

Out[38]: Normal(μ=2.0, σ=3.0)



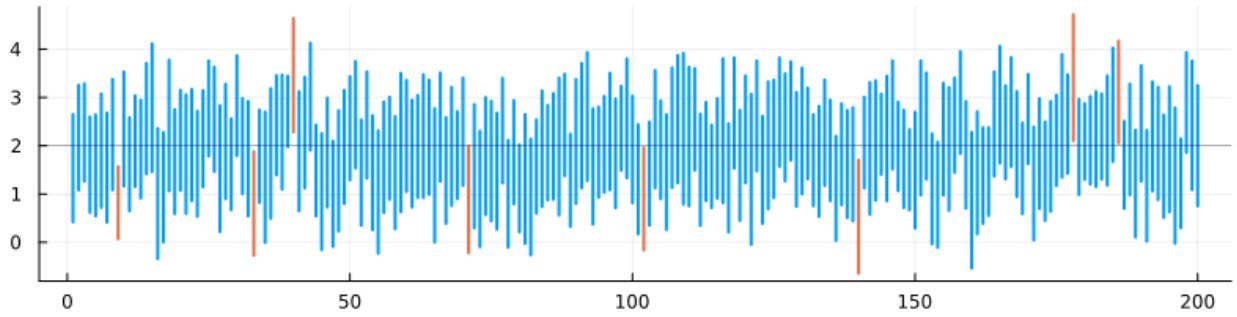
以下のセルでは正規分布 $\text{Normal}(2, 3)$ のサイズ $n=30$ の標本を200個ランダムに生成し、その各々について平均の 95% 信頼区間を計算してプロットしている。

橙色でプロットされている 95% 信頼区間は標本を生成した分布の平均値を含まない信頼区間である。

標本を生成した分布の平均値を含まない 95% 信頼区間の割合は 5% であって欲しい。

In [39]: 1 illustrate_confintmean(; dist, n = 30)

Out[39]: 95.0% confidence intervals for mean of size-30 samples of Normal(μ=2.0, σ=3.0)



In [40]: 1 sim_ttest(; dist, n = 30, L = 10^6)(0.05)

Out[40]: 0.050169

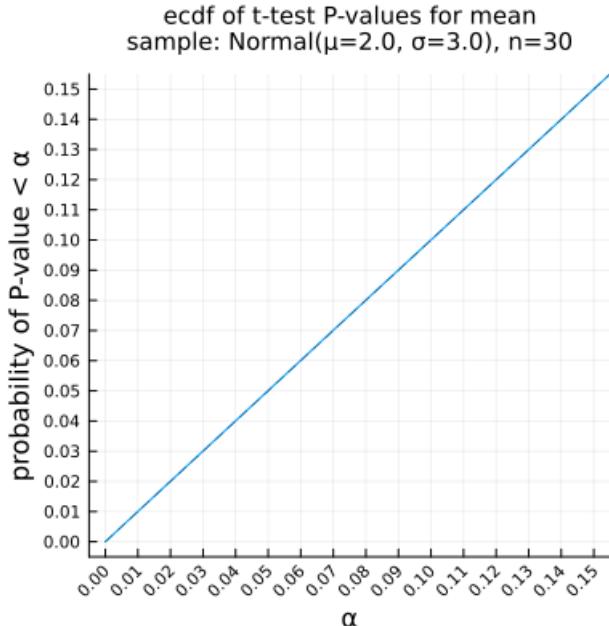
以下のセルでは正規分布 $\text{Normal}(2, 3)$ のサイズ $n=30$ の標本を100万個ランダムに生成し、信頼度 $1-\alpha$ 信頼区間に標本を生成した分布の平均値が含まれない確率(=割合)を計算している。

その割合は α であって欲しい。

この場合には実際にぴったり α になっている。

```
In [41]: 1 plot_sim_ttest(; dist, n = 30, L = 10^6)
```

Out[41]:



10.5.2 ガンマ分布の標本から得られる平均の信頼区間達

標本(データ)を生成する分布が左右非対称なガンマ分布 $\text{op}\{\text{Gamma}\}(3,4)$ の場合には、平均の信頼区間を構成するために使った正規分布とは異なる種類の分布で標本を生成しているので、誤差が生じることになる。

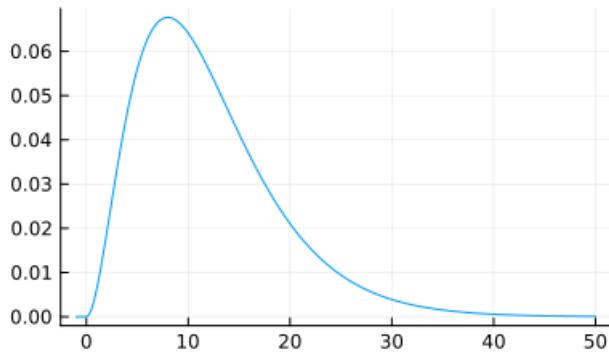
その誤差がどの程度であるかに注目して以下のグラフ達を眺めて欲しい。

```
In [42]: 1 dist = Gamma(3, 4)
2 @show mu, sigma = mean(dist), std(dist)
3 plot(dist, -1, 50; label="", title="\$distname(dist)\")
```

(μ, σ) = ($\text{mean}(\text{dist}), \text{std}(\text{dist})$) = (12.0, 6.928203230275509)

Out[42]:

Gamma($\alpha=3.0, \theta=4.0$)



以下のセルではガンマ分布 $\text{op}\{\text{Gamma}\}(3, 4)$ のサイズ $n=30$ の標本を200個ランダムに生成し、その各々について平均の 95% 信頼区間を計算してプロットしている。

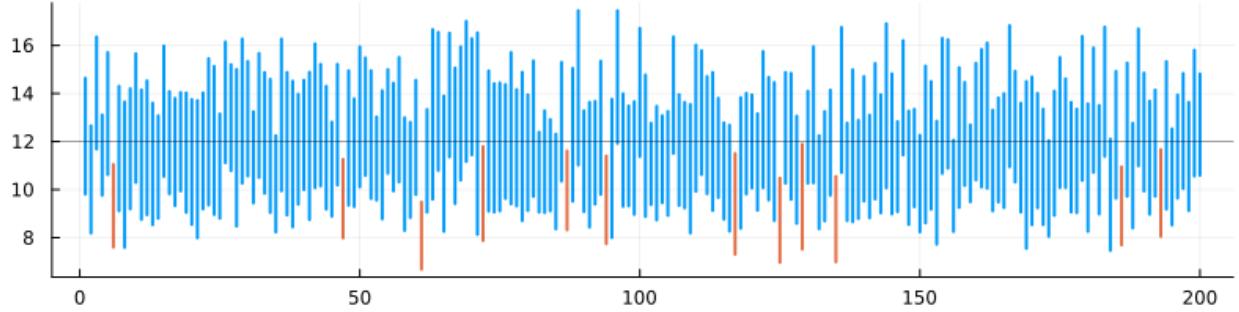
橙色でプロットされている 95% 信頼区間は標本を生成した分布の平均値を含まない信頼区間である。

標本を生成した分布の平均値を含まない 95% 信頼区間の割合は 5% であって欲しいが、この場合にはその割合は 5.8% 程度で 5% より少しだ大きい。

```
In [43]: 1 illustrate_confintmean(; dist, n = 30)
```

Out[43]:

95.0% confidence intervals for mean of size-30 samples of Gamma($\alpha=3.0, \theta=4.0$)



```
In [44]: 1 sim_ttest(; dist, n = 30, L = 10^6)(0.05)
```

Out[44]: 0.058244

以下のセルではガンマ分布 $\text{Gamma}(3, 4)$ のサイズ $n=30$ の標本を100万個ランダムに生成し、信頼度 $1-\alpha$ 信頼区間に標本を生成した分布の平均値が含まれない確率(=割合)を計算している。

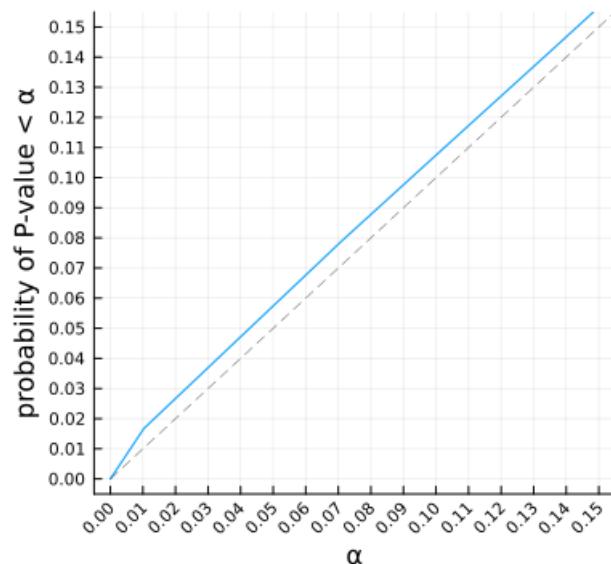
その割合は α であって欲しい。

この場合にその割合は α より少し大きな値になってしまっている。

```
In [45]: 1 plot_sim_ttest(; dist, n = 30, L = 10^6)
```

Out[45]:

ecdf of t-test P-values for mean
sample: Gamma($\alpha=3.0, \theta=4.0$), n=30

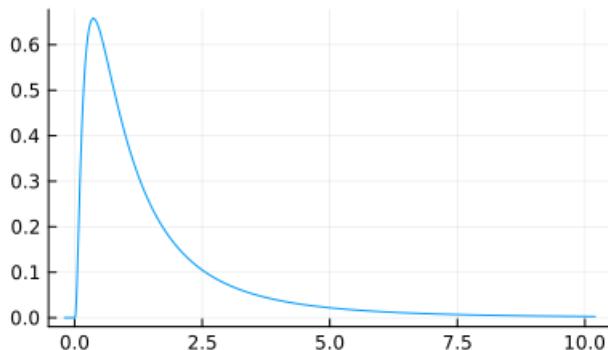


10.5.3 対数正規分布の標本達から得られる平均の信頼区間

標本(データ)を生成する分布が左右非対称でかつ非常に大きな値が生成され易い対数正規分布 $\text{LogNormal}(0,1)$ の場合には、非常に大きな誤差が生じることになる。

```
In [46]: 1 dist = LogNormal()
2 @show μ, σ = mean(dist), std(dist)
3 plot(dist, -0.2, 10.2; label="", title="$(distname(dist))")
(μ, σ) = (mean(dist), std(dist)) = (1.6487212707001282, 2.1611974158950877)
```

Out[46]: LogNormal($\mu=0.0, \sigma=1.0$)



以下のセルでは対数正規分布 $\text{LogNormal}(0,1)$ のサイズ $n=30$ の標本を200個ランダムに生成し、その各々について平均の 95% 信頼区間を計算してプロットしている。

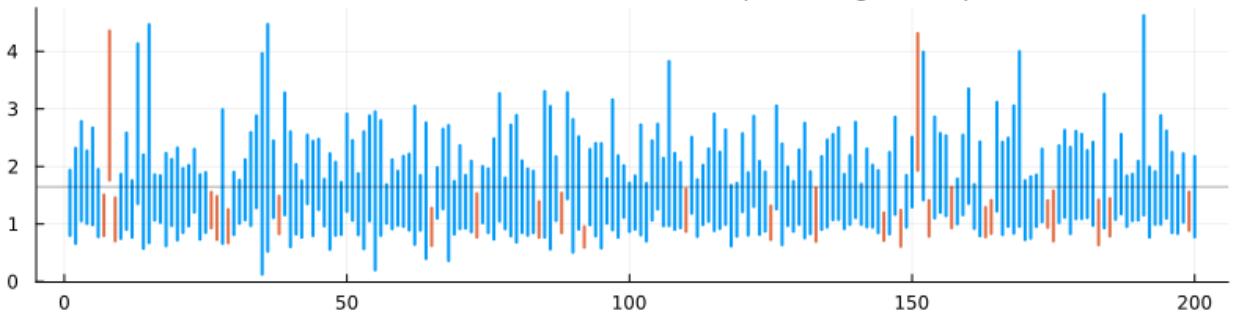
橙色でプロットされている 95% 信頼区間は標本を生成した分布の平均値を含まない信頼区間である。

標本を生成した分布の平均値を含まない 95% 信頼区間の割合は 5% であって欲しいが、この場合にはその割合は 12% 弱程度で 5% より非常に大きい。

正規分布の標本分布モデルと t 分布を使った平均に関する検定や信頼区間は小さな確率で極端な値が生じる分布(外れ値を持つ分布)の標本については誤差が大きくなり易いので注意を要する。

```
In [47]: 1 illustrate_confintmean(; dist, n = 30)
```

Out[47]: 95.0% confidence intervals for mean of size-30 samples of LogNormal($\mu=0.0, \sigma=1.0$)



```
In [48]: 1 sim_ttest(; dist, n = 30, L = 10^6)(0.05)
```

Out[48]: 0.117941

以下のセルでは対数正規分布 $\text{LogNormal}(0, 1)$ のサイズ $n=30$ の標本を100万個ランダムに生成し、信頼度 $1-\alpha$ 信頼区間に標本を生成した分布の平均値が含まれない確率(=割合)を計算している。

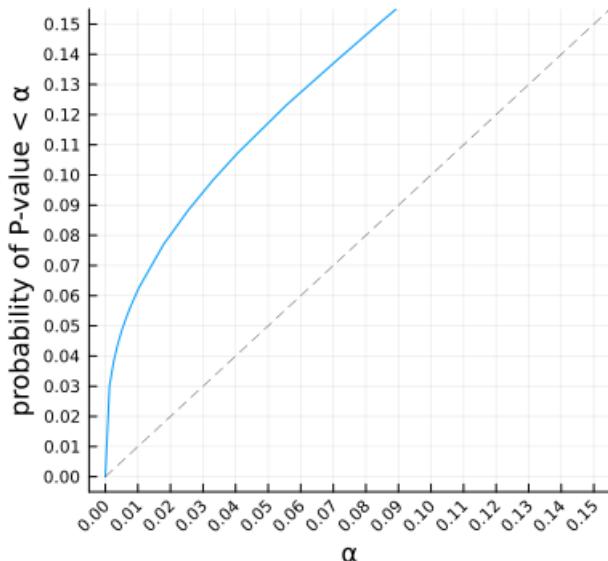
その割合は α であって欲しい。

この場合にその割合は α よりもずっと大きな値になってしまっている。

```
In [49]: 1 plot_sim_ttest(; dist, n = 30, L = 10^6)
```

Out[49]:

ecdf of t-test P-values for mean
sample: LogNormal($\mu=0.0, \sigma=1.0$), n=30

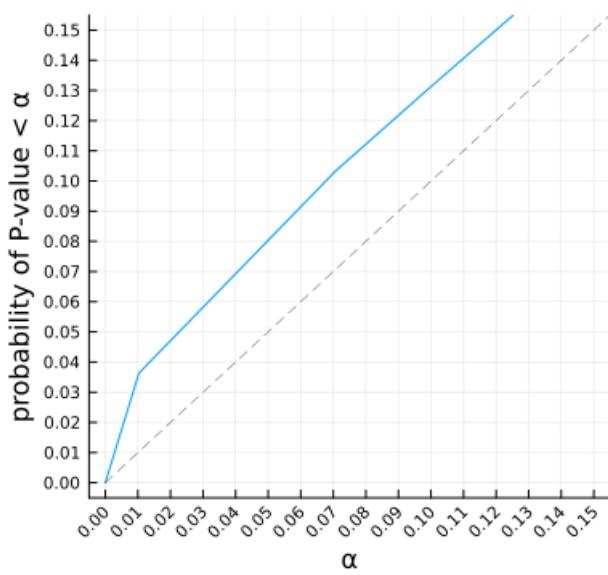


サンプルサイズ n を大きくすると誤差は小さくなる。

```
In [50]: 1 plot_sim_ttest(; dist, n = 100, L = 10^6)
```

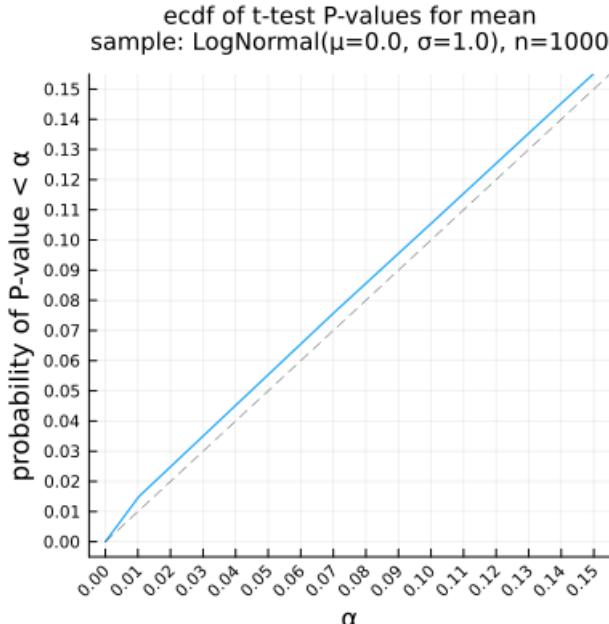
Out[50]:

ecdf of t-test P-values for mean
sample: LogNormal($\mu=0.0, \sigma=1.0$), n=100



```
In [51]: 1 plot_sim_ttest(; dist, n = 1000, L = 10^6)
```

Out[51]:



10.6 問題解答例: P値に関するよくある誤解

- (1) P値は現実世界において仮説 $\theta = \theta_0$ が正しい確率である。
- (2) 検定の手続きでは、P値が有意水準より小さいときには、仮説 $\theta = \theta_0$ は科学的に疑わしいと考える。
- (3) 検定の手続きでは、P値が有意水準以上のときには、仮説 $\theta = \theta_0$ は科学的に否定されたと考える。
- (4) P値が小さい結果ほど重要な結果である。

はすべて誤りである。各々についてコメントしておこう。

(1) P値の定義と全然違う。P値の定義は「仮説 $\theta = \theta_0$ 下の統計モデル内でデータの数値以上に極端な値が生じる確率またはその近似値」であった。

P値は「データの数値と仮説 $\theta = \theta_0$ 下の統計モデルの相性の良さ(compatibility)の指標」として使われる。統計モデル全体とデータの相性の良さ(compatibility)を見ているのであり、仮説 $\theta = \theta_0$ 単体との相性の良さを見ているのではない。

(2) 検定の手続きで、P値が有意水準より小さいときには、仮説 $\theta = \theta_0$ だけが疑わしいと考えるのではなく、統計モデルの前提の全体のどれかも疑わしいと考える必要がある。

(3) 検定の手続きで、P値が有意水準以上になった場合には、仮説 $\theta = \theta_0$ 下での統計モデルの現実における妥当性に関する判断を保留する(否定も肯定もしない)。

(4) 例えば、ある薬の効果について、巨大なサイズのデータによって非常に小さなP値が得られたとする。

しかし、存在することが確からしいその効果は実生活においてほとんど無視できるほど小さなものだったとする。

そのような場合にはP値が小さくても重要な結果が得られたとは言えない。

薬の効果について統計分析をする場合には、P値だけではなく、効果の大きさの指標にも注意を払う必要がある。

解答終

10.7 問題解答例: 信頼区間に関するよく見る誤解

- (1) 95% 信頼区間の 95% は確率ではなく、割合である。
 - (2) 平均値の 95% 信頼区間にについて考える。このとき、現実の母集団からの無作為抽出を繰り返して、平均値の 95% 信頼区間を計算し直すとき、現実の母集団の平均の真の値を含む区間達の割合は 95% になる。
- はどちらも誤りもしくは杜撰な考え方である。
- (1) 仮説 $\theta = \theta_0$ 下の統計モデル内部で生成された仮想的なデータの数値(確率変数になる)に関するパラメータ θ に関する 95% 信頼区間に θ_0 が含まれる確率は 95% またはそれに近い値になる。
- このように 95% 信頼区間の 95% は統計モデル内部での確率になっていると解釈される。ゆえに「確率ではない」と言ってしまうと自明に誤りになる。

ただし, 確率的に揺らぐのはデータを生成するために使われたパラメータ値 θ_0 ではなく, そのパラメータ値を使って生成されたデータの値とそれを使って計算される信頼区間である. 固定されている値 θ_0 が確率的に様々に変化する 95% 信頼区間に含まれる確率が 95% またはそれに近い値になる.

(2) 現実の母集団の未知の平均値を μ_{real} と書くことにする. もしも平均値の信頼区間を定義するために使った統計モデルで平均値パラメータを $\mu = \mu_{\text{real}}$ とおくことによって得られる確率分布が現実の母集団の分布をよく近似していれば、問題文の(2)の主張は確かに正しい.

しかし, 数学的フィクションである統計モデルが現実の母集団の分布をよく近似しているという仮定が成立しているは限らない. 成立していない場合には(2)の主張も成立しているとは限らない.

この問題中の(1),(2)の誤解の背景には現実とモデルの混同が隠れていると考えられる.

解答終