検定と信頼区間: 比率の比較

- 黒木玄
- 2022-06-14〜2022-06-15

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 Base.Threads
using BenchmarkTools
using DataFrames
using Distributions
using LinearAlgebra
using Memoization
using Printf
using QuadGK
using RCall
using Random
Random.seed!(4649373)
using Roots
using SpecialFunctions
using StaticArrays
using StatsBase
using StatsFuns
using StatsPlots
default(fmt = :png, size = (400, 250),
    titlefontsize = 10, plot_titlefontsize = 12)
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=false)))
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)

In [4]:
```
oddsratiohat(a, b, c, d) = safediv(a*d, b*c)
stderr_logoddsratiohat(a, b, c, d) = √(1/a + 1/b + 1/c + 1/d)

function pvalue_or_wald(a, b, c, d; ω=1)
    logORhat = log(oddsratiohat(a, b, c, d))
    SElogORhat = stderr_logoddsratiohat(a, b, c, d)
    2ccdf(Normal(0, 1), abs(logORhat - log(ω))/SElogORhat)
end

function confint_or_wald(a, b, c, d; α=0.05)
    z = quantile(Normal(), 1-α/2)
    ORhat = oddsratiohat(a, b, c, d)
    SElogORhat = stderr_logoddsratiohat(a, b, c, d)
    [exp(-z*SElogORhat)*ORhat, exp(z*SElogORhat)*ORhat]
end
```

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

In [5]:
```
riskratiohat(a, b, c, d) = safediv(a*(c+d), (a+b)*c)
stderr_logriskratiohat(a, b, c, d) = √(1/a - 1/(a+b) + 1/c - 1/(c+d))

function pvalue_rr_wald(a, b, c, d; ρ=1)
    logRRhat = log(riskratiohat(a, b, c, d))
    SElogRRhat = stderr_logriskratiohat(a, b, c, d)
    2ccdf(Normal(0, 1), abs(logRRhat - log(ρ))/SElogRRhat)
end

function confint_rr_wald(a, b, c, d; α=0.05)
    z = quantile(Normal(), 1-α/2)
    RRhat = riskratiohat(a, b, c, d)
    SElogRRhat = stderr_logriskratiohat(a, b, c, d)
    [exp(-z*SElogRRhat)*RRhat, exp(z*SElogRRhat)*RRhat]
end
```

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

In [6]:
```
function delta(a, b, c, d; ω=1)
    A, B, C = 1-ω, a+d+ω*(b+c), a*d-ω*b*c
    isinf(ω) ? typeof(ω)(-min(b, c)) : safediv(2C, B + √(B^2 - 4A*C))
end

# correction = 0.5 は連続性補正を与える.
function _chisqstat_or(a, b, c, d, δ; correction=0.0)
    ã, b̃, c̃, d̃ = a-δ, b+δ, c+δ, d-δ
    safemul(max(0, abs(δ)-correction)^2, 1/ã + 1/b̃ + 1/c̃ + 1/d̃)
end

function chisqstat_or(a, b, c, d; ω=1, correction=0.0)
    δ = delta(a, b, c, d; ω)
    _chisqstat_or(a, b, c, d, δ; correction)
end

function pvalue_or_pearson(a, b, c, d; ω=1, correction=0.0)
    χ² = chisqstat_or(a, b, c, d; ω, correction)
    ccdf(Chisq(1), χ²)
end

function confint_or_pearson(a, b, c, d; α=0.05, correction=0.0)
    ω_L, ω_U = confint_or_wald(a, b, c, d; α)
    f(ω) = logit(pvalue_or_pearson(a, b, c, d; ω, correction)) - logit(α)
    [find_zero(f, ω_L), find_zero(f, ω_U)]
end
```

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

In [7]:
```
# Cornfeild (1956), p.139, (4.2) と比較せよ.
a, b, c, d = 3, 11, 60, 32
@show confint_or_wald(a, b, c, d)
@show confint_or_pearson(a, b, c, d)
@show confint_or_pearson(a, b, c, d; correction=0.5);
```

```
confint_or_wald(a, b, c, d) = [0.037829803607039435, 0.5592686923030045]
confint_or_pearson(a, b, c, d) = [0.04084291410483445, 0.5253472140108638]
confint_or_pearson(a, b, c, d; correction = 0.5) = [0.02962955731952912, 0.6228119542455193]
```

In [8]:

```
function Delta(a, b, c, d; ρ=1)
    m, n = a+b, c+d
    A, B, C = ρ-1, n-a+ρ*(m-c), a*n-ρ*m*c
    isinf(ρ) ? typeof(ω)(-c) : safediv(2C, B + √(B^2 - 4A*C))
end

function _chisqstat_rr(a, b, c, d, Δ)
    m, n = a+b, c+d
    safemul(Δ^2, safediv(b, m*(a-Δ)) + safediv(d, n*(c+Δ)))
end

function chisqstat_rr(a, b, c, d; ρ=1)
    Δ = Delta(a, b, c, d; ρ)
    _chisqstat_rr(a, b, c, d, Δ)
end

function pvalue_rr_pearson(a, b, c, d; ρ=1)
    χ² = chisqstat_rr(a, b, c, d; ρ)
    ccdf(Chisq(1), χ²)
end

function confint_rr_pearson(a, b, c, d; α=0.05)
    ρ_L, ρ_U = confint_rr_wald(a, b, c, d; α)
    f(ρ) = logit(pvalue_rr_pearson(a, b, c, d; ρ)) - logit(α)
    [find_zero(f, ρ_L), find_zero(f, ρ_U)]
end
```

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

In [9]:

```
_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

function pvalue_or_sterne(a, b, c, d; ω=1)
    pvalue_sterne(FisherNoncentralHypergeometric(a+b, c+d, a+c, ω), a)
end

function confint_or_sterne(a, b, c, d; α = 0.05)
    ω_L, ω_U = confint_or_wald(a, b, c, d; α = α/10)
    ps = find_zeros(ω_L, ω_U) do ω
        logit(pvalue_or_sterne(a, b, c, d; ω)) - logit(α)
    end
    # 次の行は稀に区間にならない場合への対策
    [first(ps), last(ps)]
end
```

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

In [10]:
```
function pvalue_or_clopper_pearson(a, b, c, d; ω=1)
    fnch = FisherNoncentralHypergeometric(a+b, c+d, a+c, ω)
    min(1, 2cdf(fnch, a), 2ccdf(fnch, a-1))
end

function confint_or_clopper_pearson(a, b, c, d; α = 0.05)
    ω_L, ω_U = confint_or_wald(a, b, c, d; α = α/10)
    find_zeros(ω_L, ω_U) do ω
        logit(pvalue_or_clopper_pearson(a, b, c, d; ω)) - logit(α)
    end
end
```

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

In [11]:
```
a, b, c, d = 49, 965, 26, 854
pvalue_or_sterne(a, b, c, d), pvalue_or_clopper_pearson(a, b, c, d)
```

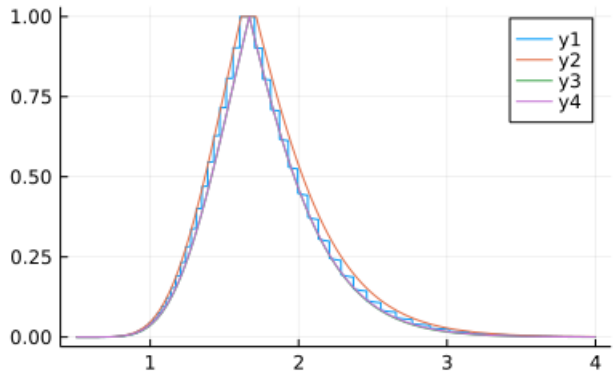
Out[11]:
```
(0.04407901464506375, 0.04708076157076113)
```

In [12]:
```
a, b, c, d = 49, 965, 26, 854
[
    confint_or_sterne(a, b, c, d),
    confint_or_clopper_pearson(a, b, c, d)
]
```

Out[12]:
```
2-element Vector{Vector{Float64}}:
 [1.0202716568785046, 2.7656505208089848]
 [1.0061031380773773, 2.8212565352058787]
```

In [13]:
```
plot(ω -> pvalue_or_sterne(a, b, c, d; ω), 0.5, 4)
plot!(ω -> pvalue_or_clopper_pearson(a, b, c, d; ω), 0.5, 4)
plot!(ω -> pvalue_or_pearson(a, b, c, d; ω), 0.5, 4)
plot!(ω -> pvalue_or_wald(a, b, c, d; ω), 0.5, 4)
```

Out[13]:



1 比率の比較に関するP値と信頼区間

1.1 比率の比較に関するP値と信頼区間を使って行いたいこと

(1) ウェブページのデザインAとBではどちらの側がどれだけ商品が売れ易いかを知りたい.

(2) 薬Xを処方した側ではしなかった側よりも快復者の割合がどれだけ増えるかを知りたい.

例えば, (1)の場合に次のようなデータが得られたとする:

	商品を購入した	購入しなかった	合計
デザインA	49	965	1014
デザインB	26	854	880
合計	75	1819	1894

このデータの数値は, デザインAの側にアクセスした1014人中の49人が商品を購入し, デザインBの側にアクセスした880人中の26人が商品を購入したことを表している.

このデータの数値からの印象では, デザインAの方がデザインBの方が商品の購入確率が高いように見える. 実際, デザインAでの購入者割合は $49/1014 \approx 4.8\%$ 程度で, デザインBでの購入者割合の $26/880 \approx 3.0\%$ より大きい.

しかし, データの確率的揺らぎのせいで偶然に, デザインAの側がよく売れるように見えるデータの数値が得られただけなのかもしれない. (実際には他にも様々な原因で偏ったデータが得られる場合がある.)

もしもそうならば, 上のデータに基くウェブページのデザイン採用に関する意思決定は無駄に終わることになってしまう. だから, そのようなリスクの程度を見積もる必要がある. そのための道具がP値や信頼区間である.

このようなデータに適当な統計モデルを適用して, デザインAとデザインBでの購入確率の違いを信頼区間で表したい.

そのためには以下のようにすればよい:

1. 統計モデルの設定.
2. 購入確率の違いを表す指標を導入.
3. その指標の値が○○であるという仮説のP値を定義する.
4. そのP値を使って信頼区間を定義する.
5. 以上によって得られたP値函数や信頼区間を利用する.

1.2 2×2の分割表型データとその2つの二項分布モデル

このノートでは次の形の 2×2 の分割表型のデータを扱う:

	$Y=1$	$Y=0$	
$X=1$	a	b	$m=a+b$
$X=0$	c	d	$n=c+d$
	$r=a+c$	$d=b+d$	$N=a+b+c+d$

ここで a, b, c, d は 0 以上の整数である.

このノートでは2×2の分割表型データに関する以下のような設定の統計モデルを扱う.

(1) 横方向の合計 $m=a+b$, $n=c+d$ (ゆえに全体の合計 $N=m+n$) は固定されていると仮定する.

(2) a は成功確率パラメータ p の二項分布に従っており, c は成功確率パラメータ q の二項分布に従っているとし, a, c は独立であると仮定する. p, q を比率と呼ぶこともある.

このとき, a, b, c, d の同時確率質量函数は次のように表される:

$$P(a,b,c,d|m,n,p,q) = \binom{m}{a} p^a (1-p)^b \binom{n}{c} q^c (1-q)^d$$
$$(a,b,c,d \in \mathbb{Z}_{\geq 0},\ a+b=m,\ c+d=n).$$

これを **2つの二項分布モデル** と呼ぶ.

(3) このモデルは上の分割表において, 以下が成立しているという設定になっている:

- $X=1$ のときには $Y=1$ となる確率は p である.
- $X=0$ のときには $Y=1$ となる確率は q である.

そして, $X=1$ となる人達を m 人選び, $X=0$ となる人を n 人選んで Y がどうなるかを調べた.

X, Y として以下のような場合を想定している:

- $X=1,0$ はウェブページのデザインがA,Bのどちらであるかを意味し, $Y=1,0$ は商品を購入したかしないかを意味する.
- $X=1,0$ は薬Xを処方したか否かを意味し, $Y=1,0$ は快復したか否かを意味する.

注意: このモデルの採用が妥当であるかどうかはデータの数値とは別の情報を使って判断しなければいけない. 例えば, 比較したい母集団とは異なる偏った母集団からの無作為抽出になってしまっている疑いが強い場合には, モデルを変更するか, データの取得の仕方を変更するか, どちらかの処置が必要になる.

注意: 分割表の縦方向と横方向の役割を逆転させた場合も数学的には同様である. ただし, データの分析結果の解釈では立場を交換したことによって違いが出るので注意すること.

注意: 以上で説明した2つの二項分布モデル以外に, 多項分布(四項分布)モデルや4つのPoisson分布モデルも考えられるが, 以下では省略する. 実は統計モデルをそのように変更しても以下で説明するP値や信頼区間の計算の仕方はどれも同じになることが知られている. 2×2の分割表に関する統計モデルについては, 「条件付き確率分布, 尤度, 推定, 記述統計」のノート (https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/06%20Conditional%20distribution%2C%20likelihood%2C%20estin の「2×2の分割表の分布」の節を参照せよ.

1.3 比率の違いを表す2つの指標: オッズ比パラメータ OR とリスク比パラメータ RR

比率 $0 < p < 1$ に関する

$$u = \frac{p}{1-p}$$

を **オッズ**(odds)と呼び, 2つの比率 $0 < p < 1$, $0 < q < 1$ に関する

$$\mathrm{OR} = \frac{p/(1-p)}{q/(1-q)} = \frac{p(1-q)}{(1-p)q}$$

を **オッズ比**(odds ratio, OR)と呼び, これの対数 β を **対数オッズ比** (log odds ratio)と呼ぶ:

$$\log \mathrm{OR} = \log \frac{p(1-q)}{(1-p)q} = \log \frac{p}{1-p} - \log \frac{q}{1-q}.$$

さらにこのノートでは, 2つの比率 $0 < p < 1$, $0 < q < 1$ に関する

$$\mathrm{RR} = \frac{p}{q}$$

を **リスク比**(risk ratio, RR)と呼び, これの対数 γ を **対数リスク比** (log risk ratio)と呼ぶ:

$$\log \mathrm{RR} = \log \frac{p}{q} = \log p - \log q.$$

もしも, p, q の両方が 0 に近ければ, $1-p \approx 1$, $1-q \approx 1$ という近似を使えるので, リスク比はオッズ比で近似される:

$$1-p \approx 1,\ 1-q \approx 1 \implies \mathrm{RR} \approx \mathrm{OR}.$$

比率 p と q が等しいという条件はオッズ比やリスク比を使って以下のように言い換えられる:

$$p = q \iff \mathrm{OR} = 1 \iff \mathrm{RR} = 1.$$

だから, p と q が等しいという仮説は $\mathrm{OR} = 1$ や $\mathrm{RR} = 1$ のように表現できる. (もちろん, それらの対数版である $\log \mathrm{OR} = 0$ や $\log \mathrm{RR} = 0$ を採用してもよい.)

そして, オッズ比 OR とリスク比 RR が 1 からどれだけどのように離れているかは, 2つの比率 p, q の違いを表す指標として使える. (もちろん, それらの対数版である $\log \mathrm{OR}$ や $\log \mathrm{RR}$ が 0 からどれだけどのように 離れているかを指標として採用してもよい.)

注意: 他にも **比率の差** $p - q$ も重要な指標だが, このノートでは面倒だという理由で扱わないことにする.

注意: オッズは賭け事ではよく使われる用語である. 例えばある勝負で A が勝つことに賭けるときに

$$\text{勝つ確率} : \text{負ける確率} = 1 : 4$$

だと思っているなら, あなたはオッズは「1 対 4」だと思っていることになる.

注意: 例えば確率 p, q が「死亡確率」ならば「リスク」という呼び方は適切になる. ここではそうでない場合も p, q を「リスク」と呼んでしまい, p/q を「リスク比」と呼ぶことにする. 具体的な応用先で, この呼び方に違和感を感じる場合には適宜別の呼び方をするようにして欲しい.

注意: このノートでは主にオッズ比を比率の違いを表す指標として利用する. その理由は. 応用上の理由ではなく, 単に数学的に扱いやすいからである. そして, 以下の注意で説明するように, オッズ比を考えることは, その対数を考えることによって, ロジスティック回帰の話とも関係付けることができる.

注意: オッズ $u = p/(1-p)$ の対数を x と書き, **対数オッズ**(log odds)と呼ぶ. このとき, 以下の公式によって $0 < p < 1$, $0 < u < \infty$, $-\infty < x < \infty$ が一対一に対応する:

$$\begin{aligned}
p &= \frac{u}{1+u} = \frac{1}{1+e^{-x}} = \mathrm{logistic}(x), \\
u &= \frac{p}{1-p} = \exp(x), \\
x &= \log(u) = \log \frac{p}{1-p} = \mathrm{logit}(p).
\end{aligned}$$

x を p に対応させる函数は **ロジスティック函数** と呼ばれ, その逆函数は **ロジット函数** と呼ばれる.

注意: 対数オッズ比を $\beta = \log \mathrm{OR}$ と書き, さらに q の対数を $\alpha = \log q$ と書くと, p, q の対数オッズはそれぞれ次のように表される:

$$\mathrm{logit}(p) = \log \frac{p}{1-p} = \alpha + \beta, \quad \mathrm{logit}(q) = \log \frac{q}{1-p} = \alpha.$$

これらは次と同値である:

$$p = \mathrm{logistic}(\alpha + \beta), \quad q = \mathrm{logistic}(\alpha).$$

モデルの2つの比率パラメータ p, q をこのように表して, データの数値から α, β の値を推定することを **ロジスティック回帰** (logisti regression)と呼ぶ. (実際にはもっと一般の場合もロジスティック回帰と呼ぶ.)

注意: 対数リスク比を $\beta = \log \mathrm{RR} = \log p - \log q$ と書き, q の対数を $\alpha = \log q$ と書くと, p, q の対数はそれぞれ次のように表される:

$$\log p = \alpha + \beta, \quad \log q = \alpha.$$

これは次と同値である:

$$p = \exp(\alpha + \beta), \quad q = \exp(\alpha).$$

これと上の違いは logistic と exp の違いになっている.

2 Wald版のオッズ比に関するP値と信頼区間

2.1 Wald版のオッズ比に関するP値と信頼区間の定義

検定したい仮説: 検定したい仮説は

$$\text{オッズ比パラメータは } \mathrm{OR} = \omega \text{ である.}$$

であるとする. ここで $\omega > 0$ は具体的な正の実数である.

正規分布近似: 仮に a, b, c, d が2つの二項分布モデルに従う確率変数ならば, それらの **オッズ比**

$$\widehat{\mathrm{OR}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

の対数 $\log \widehat{\mathrm{OR}}$ (確率変数としての対数オッズ比)は, 平均がモデルのパラメータの対数オッズ比

$$\log \mathrm{OR} = \log \frac{p(1-q)}{(1-p)q}$$

で, 分散が

$$\widehat{\mathrm{SE}}^2_{\log \widehat{\mathrm{OR}}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

であるような正規分布に近似的に従うことを示せる. ゆえに,

$$\frac{\log \widehat{\mathrm{OR}} - \log \mathrm{OR}}{\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}} \sim \mathrm{Normal}(0, 1), \text{ approximately.}$$

(「大数の法則と中心極限定理」のノート (https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/05%20Central%20limit%20theorem.ipynb)で解説したデルタ法を使えば示せる. この点については後の節で説明する.)

P値の構成法: 上の近似を使えば, 「オッズ比パラメータは $\mathrm{OR} = \omega$ である」という仮説のP値を次のように定めることができる:

$$\mathrm{pvalue}_{\mathrm{Wald}}(a, b, c, d | \, \mathrm{OR} = \omega) = 2\left(1 - \mathrm{cdf}\left(\mathrm{Normal}(0,1), \frac{\left|\log \widehat{\mathrm{OR}} - \log \omega\right|}{\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}}\right)\right)$$

ただし, $\log \widehat{\mathrm{OR}}$ と $\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}$ はデータの数値 a, b, c, d から上で示した式で計算された値であるとする.

対応する信頼区間: このP値の定義に対応するオッズ比パラメータ OR に関する信頼度 $1 - \alpha$ の信頼区間は次のようになる:

$$\mathrm{confint}^{\log \mathrm{OR}}_{\mathrm{Wald}}(a, b, c, d | \alpha) = \left[\log \widehat{\mathrm{OR}} - z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}, \ \log \widehat{\mathrm{OR}} + z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}\right]$$

対応するオッズ比パラメータ OR の信頼区間は次のようになる:

$$\mathrm{confint}^{\mathrm{OR}}_{\mathrm{Wald}}(a, b, c, d | \alpha) = \left[\exp\left(-z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}\right)\widehat{\mathrm{OR}}, \ \exp\left(z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}\right)\widehat{\mathrm{OR}}\right]$$

ここで, $z_{\alpha/2} = \mathrm{quantile}(\mathrm{Normal}(0,1), 1 - \alpha/2)$ である.

2.2 Wald版のオッズ比に関するP値と信頼区間の計算例

データが次の場合のWald版の仮説「オッズ比パラメータは $\mathrm{OR} = 1$ である」のP値とオッズ比パラメータ OR の 95% 信頼区間を計算してみよう:

	商品を購入した	購入しなかった	合計
デザインA	49	965	1014
デザインB	26	854	880
合計	75	1819	1894

結果は次のようになる:

- (Wald版の仮説「オッズ比パラメータは $\mathrm{OR} = 1$ である」のP値) ≈ 3.847%
- (Wald版のオッズ比パラメータ OR の 95% 信頼区間) ≈ [1.0275, 2.7072]

2.2.1 WolframAlphaによるWald版のP値と信頼区間の計算の仕方

`sqrt(1/a+1/b+1/c+1/d) where a=49.0, b=965, c=26, d=854` → 実行 (https://www.wolframalpha.com/input?i=sqrt%281%2Fa%2B1%2Fb%2B1%2Fc%2B1%2Fd%29+where+a%3D49.0%2C+b%3D965%2C+c%3D26%2C+d%3D854) → 0.247137

`log(a*d/(b*c))/0.247137 where a=49.0, b=965, c=26, d=854` → 実行 (https://www.wolframalpha.com/input?i=log%28a*d%2F%28b*c%29%29%2F0.247137+where+a%3D49.0%2C+b%3D965%2C+c%3D26%2C+d%3D854) → 2.06981

`2(1 - cdf(NormalDistrubution(0,1), 2.06981))` → 実行 (https://www.wolframalpha.com/input?i=2%281+-+cdf%28NormalDistrubution%280%2C1%29%2C+2.06981%29%29) → 0.0384701 (P値)

`quantile(NormalDistribution(0,1), 0.975))` → 実行 (https://www.wolframalpha.com/input?i=quantile%28NormalDistribution%280%2C1%29%2C+0.975%29) → 1.95996

`{exp(-0.247137z)*a*d/(b*c), exp(0.247137z)*a*d/(b*c)} where a=49.0, b=965, c=26, d=854, z=1.95996` → 実行 (https://www.wolframalpha.com/input?i=%7Bexp%28-0.247137z%29*a*d%2F%28b*c%29%2C+exp%280.247137z%29*a*d%2F%28b*c%29%7D+where+a%3D49.0%2C+b%3D965%2C → {1.02752, 2.70717} (95%信頼区間)

2.2.2 Julia言語によるWald版のP値と信頼区間の計算の仕方(1)

素朴に定義通りにコードを入力すれば計算できる.

In [14]:
```
a, b, c, d = 49, 965, 26, 854
@show a, b, c, d
ω = 1.0
@show ω
@show ORhat = a*d/(b*c)
@show SEhat = √(1/a + 1/b + 1/c + 1/d)
@show pvalue = 2ccdf(Normal(), abs(log(ORhat) - log(ω))/SEhat)
α = 0.05
@show α
@show z = quantile(Normal(), 1-α/2)
@show confint = [exp(-z*SEhat)*ORhat, exp(z*SEhat)*ORhat]
;
```

```
(a, b, c, d) = (49, 965, 26, 854)
ω = 1.0
ORhat = (a * d) / (b * c) = 1.6678357911518533
SEhat = √(1 / a + 1 / b + 1 / c + 1 / d) = 0.247137474584997
pvalue = 2 * ccdf(Normal(), abs(log(ORhat) - log(ω)) / SEhat) = 0.03847043006362489
α = 0.05
z = quantile(Normal(), 1 - α / 2) = 1.9599639845400576
confint = [exp(-z * SEhat) * ORhat, exp(z * SEhat) * ORhat] = [1.0275181208761617, 2.7071797272783877]
```

2.2.3 Julia言語によるWald版のP値と信頼区間の計算の仕方(2)

このノートブックの最初の方で定義した函数を使って計算.

In [15]:
```
a, b, c, d = 49, 965, 26, 854
@show a, b, c, d
@show pvalue_or_wald(a, b, c, d; ω=1)
@show confint_or_wald(a, b, c, d; α=0.05);
```

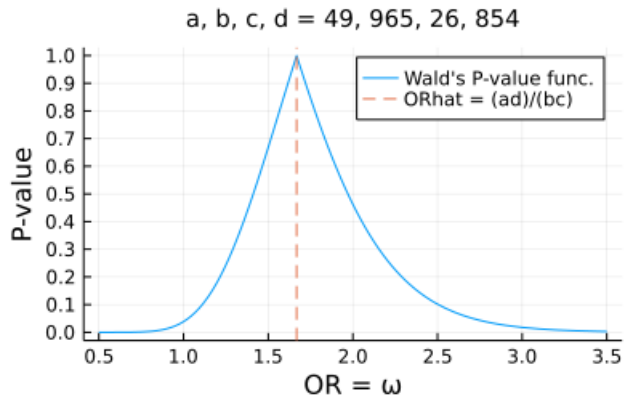
```
(a, b, c, d) = (49, 965, 26, 854)
pvalue_or_wald(a, b, c, d; ω = 1) = 0.03847043006362489
confint_or_wald(a, b, c, d; α = 0.05) = [1.0275181208761617, 2.7071797272783877]
```

2.2.4 Julia言語によるWald版のP値函数の視覚化

P値函数 $\omega \mapsto \mathrm{pvalue}_{\mathrm{Wald}}(a, b, c, d | \mathrm{OR} = \omega)$ のグラフを描いてみよう.

In [16]:
```
a, b, c, d = 49, 965, 26, 854
plot(ω -> pvalue_or_wald(a,b,c,d; ω), 0.5, 3.5; label="Wald's P-value func.")
vline!([oddsratiohat(a,b,c,d)]; label="ORhat = (ad)/(bc)", ls=:dash)
plot!(; xguide="OR = ω", yguide="P-value")
plot!(; ytick=0:0.1:1)
title!("a, b, c, d = $a, $b, $c, $d")
```

Out[16]:



このようにWald版P値函数は, $\omega = \widehat{\mathrm{OR}} = (ad)/(bc)$ で最大値の 1 になり, 「とんがり帽子」型のグラフになる.

2.2.5 R言語での計算の仕方

定義通りにコードを入力すれば計算できる.

```
a = 49
b = 965
c = 26
d = 854
omega = 1
ORhat = a*d/(b*c)
SEhat = sqrt(1/a + 1/b + 1/c + 1/d)
p.value = 2*(1 - pnorm(abs(log(ORhat) - log(omega))/SEhat))
alpha = 0.05
z = qnorm(1 - alpha/2)
conf.int = c(exp(-z*SEhat)*ORhat, exp(z*SEhat)*ORhat)

cat("data: a b c d = ", a, b, c, d, "\n")
cat("null hypothesis: OR = ", omega, "\n")
cat("ORhat = ", ORhat, "\n")
cat("SEhat = ", SEhat, "\n")
cat("p.value = ", p.value, "\n")
cat("conf.level = ", 1 - alpha, "\n")
cat("conf.int = ", conf.int, "\n")
```

In [17]:
```
R"""
a = 49
b = 965
c = 26
d = 854
omega = 1
ORhat = a*d/(b*c)
SEhat = sqrt(1/a + 1/b + 1/c + 1/d)
p.value = 2*(1 - pnorm(abs(log(ORhat) - log(omega))/SEhat))
alpha = 0.05
z = qnorm(1 - alpha/2)
conf.int = c(exp(-z*SEhat)*ORhat, exp(z*SEhat)*ORhat)

cat("data: a b c d = ", a, b, c, d, "\n")
cat("null hypothesis: OR = ", omega, "\n")
cat("ORhat = ", ORhat, "\n")
cat("SEhat = ", SEhat, "\n")
cat("p.value = ", p.value, "\n")
cat("conf.level = ", 1 - alpha, "\n")
cat("conf.int = ", conf.int, "\n")
""";
```

```
data: a b c d =  49 965 26 854
null hypothesis: OR =  1
ORhat =  1.667836
SEhat =  0.2471375
p.value =  0.03847043
conf.level =  0.95
conf.int =  1.027518 2.70718
```

95% 信頼区間はepiR パッケージを使えば

```
A = matrix(c(49, 965, 26, 854), 2, 2, byrow=T)
result = epiR::epi.2by2(A, digits=4, conf.level=0.95)
```

で計算できる. 色々表示されるが,

```
Odds ratio                                     1.6678 (1.0275, 2.7072)
```

の行の括弧の内側にWald版の信頼区間が表示されている.

In [18]:
```
R"""
A = matrix(c(49, 965, 26, 854), 2, 2, byrow=T)
result = epiR::epi.2by2(A, digits=4, conf.level=0.95)
"""
```

Out[18]:
```
RObject{VecSxp}
             Outcome +    Outcome -      Total        Inc risk *        Odds
Exposed +           49          965       1014             4.83      0.0508
Exposed -           26          854        880             2.95      0.0304
Total               75         1819       1894             3.96      0.0412

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 1.6356 (1.0254, 2.6088)
Odds ratio                                     1.6678 (1.0275, 2.7072)
Attrib risk in the exposed *                   1.8778 (0.1475, 3.6081)
Attrib fraction in the exposed (%)            38.8590 (2.4779, 61.6679)
Attrib risk in the population *                1.0053 (-0.4170, 2.4276)
Attrib fraction in the population (%)         25.3879 (-1.2434, 45.0140)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 4.368 Pr>chi2 = 0.037
Fisher exact test that OR = 1: Pr>chi2 = 0.044
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

2.3 確率変数としての対数オッズ比の分散の近似値の導出

この節では a, b, c, d は2つの二項分布モデルに従う確率変数であるとし,

$$\hat{p} = \frac{a}{a+b} = \frac{a}{m}, \quad \hat{q} = \frac{c}{c+d} = \frac{c}{n}$$

とおく. これの対数オッズ比

$$\log \widehat{\mathrm{OR}} = \log \frac{a/b}{c/d} = \log \frac{\hat{p}/(1-\hat{p})}{\hat{q}(1-\hat{q})} = \log \frac{\hat{p}}{1-\hat{p}} + \log \frac{\hat{q}}{1-\hat{q}}$$

の分散の近似値を求めたい. $a \sim \mathrm{Binomial}(m, p)$, $c \sim \mathrm{Binomial}(n, q)$ なので, 二項分布の中心極限定理より,

$$\hat{p} \sim \mathrm{Normal}\left(p, \sqrt{\frac{p(1-p)}{m}}\right), \quad \hat{q} \sim \mathrm{Normal}\left(q, \sqrt{\frac{q(1-q)}{n}}\right), \quad \text{approximately.}$$

a, c は独立なので $\hat{p}, \hat{q}$ も独立になる.

$f(t) = \log(t/(1-t)) = \log(1/(1-t) - 1)$ とおくと, $f'(t) = ((1-t)/t)(1/(1-t)^2) = 1/(t(1-t))$ なので,

$$\log \frac{\hat{p}}{1-\hat{p}} = \frac{p}{1-p} + \frac{\hat{p}-p}{p(1-p)} + O((\hat{p}-p)^2).$$

これより, 次の近似が得られる:

$$\log \frac{\hat{p}}{1-\hat{p}} \sim \mathrm{Normal}\left(\log \frac{p}{1-p}, \frac{1}{\sqrt{mp(1-p)}}\right), \quad \text{approximately.}$$

同様にして,

$$\log \frac{\hat{q}}{1-\hat{q}} \sim \mathrm{Normal}\left(\log \frac{q}{1-q}, \frac{1}{\sqrt{nq(1-q)}}\right), \quad \text{approximately.}$$

一般に X, Y が独立な確率変数で $X \sim \mathrm{Normal}(\mu_X, \sigma_X)$, $Y \sim \mathrm{Normal}(\mu_Y, \sigma_Y)$ のとき,
$X + Y \sim \mathrm{Normal}\left(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right)$ となる. そして,

$$\begin{aligned}\left(\frac{1}{\sqrt{mp(1-p)}}\right)^2 + \left(\frac{1}{\sqrt{nq(1-q)}}\right)^2 &= \frac{1}{mp(1-p)} + \frac{1}{nq(1-q)} \\ &= \frac{1}{mp} + \frac{1}{m(1-p)} + \frac{1}{nq} + \frac{1}{n(1-q)}\end{aligned}$$

なので,

$$\mathrm{SE}_{\log \widehat{\mathrm{OR}}} = \sqrt{\frac{1}{mp} + \frac{1}{m(1-p)} + \frac{1}{nq} + \frac{1}{n(1-q)}}$$

とおいて, $\log \widehat{\mathrm{OR}}$ の定義と

$$\log \mathrm{OR} = \log \frac{p/(1-p)}{q/(1-q)} = \log \frac{p}{1-p} + \log \frac{q}{1-q}$$

を使うと,

$$\log \widehat{\mathrm{OR}} \sim \mathrm{Normal}\left(\log \mathrm{OR}, \mathrm{SE}_{\log \widehat{\mathrm{OR}}}\right), \quad \text{approximately}$$

となることがわかる.

もしも $mp, m(1-p), nq, n(1-q)$ のそれぞれを a, b, c, d で近似できるならば, SE の中の $mp, m(1-p), nq, n(1-q)$ のそれぞれを a, b, c, d で置き換えて,

$$\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

を作って,

$$\log \widehat{\mathrm{OR}} \sim \mathrm{Normal}\left(\log \mathrm{OR}, \widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}\right), \quad \text{approximately}$$

という近似が得られる. この近似のもとで

$$\frac{\log \widehat{\mathrm{OR}} - \log \mathrm{OR}}{\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}} \sim \mathrm{Normal}(0, 1), \quad \text{approximately.}$$

注意: 以上のような議論ではハット付きの $\widehat{\mathrm{OR}}$ とハットが付いていない OR の区別が曖昧になると理解できなくなる. ハットが付いていない $\mathrm{OR} = (p/(1-p))/(q/(1-q))$ はモデルのパラメータ p, q で書かれた量であり, ハットが付いている $\widehat{\mathrm{OR}} = (a/b)/(c/d)$ はモデルに従う確率変数 a, b, c, d で書かれた量になっていることに注意すれば区別がし易くなるだろう. 実際にその方法で $\mathrm{SE}_{\log \widehat{\mathrm{OR}}}$ と $\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{OR}}}$ の区別を容易にできることを確認してみよ.

注意: ハット付きの記号はデータの数値で表された同様の量にもそのまま流用されるので, 文脈によって, 確率変数になっているかと, データの数値から計算された数値になっているかを区別しなければいけない. この問題はモデル内の確率変数とデータの数値から計算さらた数値を記号的にも区別できるようにすれば解決するのだが, 実際にやってみればわかるように, 記号が煩雑過ぎるという点でつらくなる場合が多数出て来る. しかし, 初学者にとって「文脈を利用して正しく解釈すること」は難しい場合があるので, 実際に難しいと感じる人は, 文脈に頼らずに記号を見ただけで区別できるように書かれたノートを **自分で** 作成してみるとよいだろう. (おそらくそれを本当に実行した人は結果的に文脈を利用して正しく解釈することが楽にできるようになるだろう.)

2.4 問題: 対数リスク比の場合の正規分布近似

a, b, c, d は2つの二項分布モデルに従う確率変数であるとし, $\hat{p} = a/m$, $\hat{q} = c/n$ とおき, それらに関する確率変数としての対数リスク比を次のように定める:

$$\log \widehat{\mathrm{RR}} = \log \frac{\hat{p}}{\hat{q}} = \log \hat{p} - \log \hat{q}.$$

このとき,

$$\log \mathrm{RR} = \log \frac{p}{q}, \quad \widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}} = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

とおくと, 二項分布の中心極限定理を使って, 次の近似が得られることを示せ:

$$\log \widehat{\mathrm{RR}} \sim \mathrm{Normal}\left(\log \mathrm{RR}, \widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}\right), \quad \text{approximately}.$$

解答例: $f(t) = \log t$ とおくと, $f'(t) = 1/t$ なので,

$$\log \hat{p} = \log p + \frac{\hat{p} - p}{p} + O((\hat{p} - p)^2).$$

これと二項分布の中心極限定理 $\hat{p} \sim \mathrm{Normal}\left(p, \sqrt{p(1-p)/m}\right)$ をから次の近似を得る:

$$\log \hat{p} \sim \mathrm{Normal}\left(\log p, \sqrt{\frac{1-p}{mp}}\right), \quad \text{approximately}.$$

同様にして,

$$\log \hat{q} \sim \mathrm{Normal}\left(\log q, \sqrt{\frac{1-q}{nq}}\right), \quad \text{approximately}.$$

ゆえに, $\mathrm{SE}_{\log \widehat{\mathrm{RR}}}$ を

$$\begin{aligned}\mathrm{SE}_{\log \widehat{\mathrm{RR}}} &= \sqrt{\frac{1-p}{mp} + \frac{1-q}{nq}} \\ &= \sqrt{\frac{1}{mp} - \frac{1}{m} + \frac{1}{nq} - \frac{1}{n}}\end{aligned}$$

と定めると,

$$\log \widehat{\mathrm{RR}} \sim \mathrm{Normal}\left(\log \mathrm{RR}, \mathrm{SE}_{\log \widehat{\mathrm{RR}}}\right), \quad \text{approximately}.$$

ゆえにもしも, mp, nq のそれぞれを a, c で近似できるならば, $a + b = m$, $c + d = n$ も使って

$$\log \widehat{\mathrm{RR}} \sim \mathrm{Normal}\left(\log \mathrm{RR}, \widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}\right), \quad \text{approximately}.$$

が得られる.

解答終

2.5 問題: Wald版のリスク比に関するP値と信頼区間の定義

上の問題の結果を用いてWald版のリスク比に関するP値と信頼区間を適切に定義せよ.

注意: 数学的にややこしい事柄を理解したい場合には「定義だけは最初に覚えなければいけない」のように考えると失敗することが多い. 適切な定義を自分で作れるような考え方をした方がよい. この問題を適切に解けた人は今度は自力でP値と信頼区間を適切に定義して自由自在に使えるようになるだろう.

解答例: 現実世界における調査によって2×2の分割表のデータの数値 a, b, c, d が得られているとする. そのデータの数値の対数リスク比を次のように定める:

$$\log \widehat{\mathrm{RR}} = \frac{a/(a+b)}{c/(c+d)}.$$

仮説「リスク比パラメータは $\mathrm{RR} = \rho$ である」の下での2つの二項分布モデル内でランダムに生成されたデータの対数リスク比が現実世界で得たデータの対数リスク比以上に $\mathrm{RR} = \rho$ から離れる確率の近似値として, 「リスク比パラメータは $\mathrm{RR} = \rho$ である」という仮説のP値を定める:

$$\mathrm{pvalue}_{\mathrm{Wald}}(a, b, c, d | \, \mathrm{RR} = \rho) = 2\left(1 - \mathrm{cdf}\left(\mathrm{Normal}(0,1), \frac{\left|\log \widehat{\mathrm{RR}} - \log \rho\right|}{\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}}\right)\right).$$

ここで1つ前の問題の結果と記号

$$\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}} = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

を使った. このP値に定義に対応する対数リスク比パラメータ $\log \mathrm{RR}$ の信頼度 $1-\alpha$ の信頼区間は

$$\mathrm{confint}_{\mathrm{Wald}}^{\log \mathrm{RR}}(a, b, c, d | \alpha) = \left[\log \widehat{\mathrm{RR}} - z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}, \ \log \widehat{\mathrm{RR}} + z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}\right]$$

になり, リスク比パラメータ RR の信頼度 $1-\alpha$ の信頼区間は

$$\mathrm{confint}_{\mathrm{Wald}}^{\mathrm{RR}}(a, b, c, d | \alpha) = \left[\exp\left(-z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}\right)\widehat{\mathrm{RR}}, \ \exp\left(z_{\alpha/2}\widehat{\mathrm{SE}}_{\log \widehat{\mathrm{RR}}}\right)\widehat{\mathrm{RR}}\right]$$

になる. ここで $z_{\alpha/2} = \mathrm{cdf}(\mathrm{Normal}(0,1), 1-\alpha/2)$.

2.6 問題: Wald版のリスク比に関するP値と信頼区間の計算例

データが次の場合のWald版の仮説「リスク比パラメータは $\mathrm{RR} = 1$ である」のP値とリスク比パラメータ RR の 95% 信頼区間を計算してみよ:

	商品を購入した	購入しなかった	合計
デザインA	49	965	1014
デザインB	26	854	880
合計	75	1819	1894

2.6.1 Julia言語によるWald版のリスク比に関するP値と信頼区間の計算例

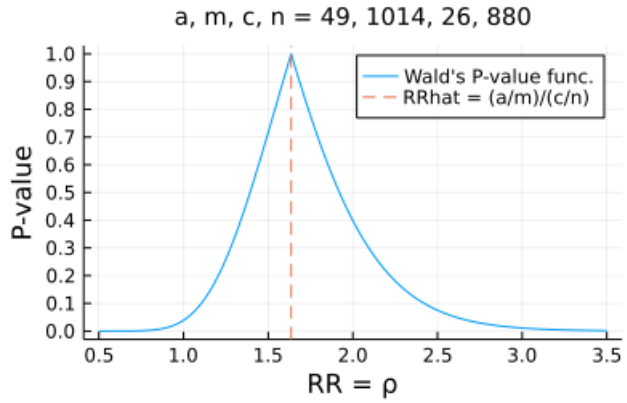
In [19]:
```
a, b, c, d = 49, 965, 26, 854
@show riskratiohat(a, b, c, d)
@show pvalue_rr_wald(a, b, c, d; ρ=1)
@show confint_rr_wald(a, b, c, d; α = 0.05);
```

```
riskratiohat(a, b, c, d) = 1.635563647397967
pvalue_rr_wald(a, b, c, d; ρ = 1) = 0.038895117565198624
confint_rr_wald(a, b, c, d; α = 0.05) = [1.025408984412048, 2.6087819449169145]
```

2.6.2 Julia言語によるWald版のリスク比に関するP値函数の視覚化

In [20]:
```
a, b, c, d = 49, 965, 26, 854
plot(ρ -> pvalue_rr_wald(a,b,c,d; ρ), 0.5, 3.5; label="Wald's P-value func.")
vline!([riskratiohat(a,b,c,d)]; label="RRhat = (a/m)/(c/n)", ls=:dash)
plot!(; xguide="RR = ρ", yguide="P-value")
plot!(; ytick=0:0.1:1)
title!("a, m, c, n = $a, $(a+b), $c, $(c+d)")
```

Out[20]:



2.6.3 Wald版のリスク比に関するP値と信頼区間の計算例の問題の解答例

- (Wald版の仮説「リスク比パラメータは $\mathrm{RR} = 1$ である」のP値) ≈ 3.890%
- (Wald版のリスク比パラメータ RR の 95% 信頼区間) ≈ [1.0254, 2.6088]

注意: この場合にはリスク比がオッズ比で近似される場合になっている.

In [21]:
```
a, b, c, d = 49, 965, 26, 854
@show riskratiohat(a, b, c, d)
@show oddsratiohat(a, b, c, d)
@show pvalue_rr_wald(a, b, c, d; ρ=1)
@show pvalue_or_wald(a, b, c, d; ω=1)
@show confint_rr_wald(a, b, c, d; α = 0.05)
@show confint_or_wald(a, b, c, d; α = 0.05);
```

```
riskratiohat(a, b, c, d) = 1.635563647397967
oddsratiohat(a, b, c, d) = 1.6678357911518533
pvalue_rr_wald(a, b, c, d; ρ = 1) = 0.038895117565198624
pvalue_or_wald(a, b, c, d; ω = 1) = 0.03847043006362489
confint_rr_wald(a, b, c, d; α = 0.05) = [1.025408984412048, 2.6087819449169145]
confint_or_wald(a, b, c, d; α = 0.05) = [1.0275181208761617, 2.7071797272783877]
```

3 Pearsonのχ²検定版のオッズ比に関するP値と信頼区間

3.1 Pearsonのχ²検定版のオッズ比に関するP値と信頼区間の定義

天下り的になってしまい非常に申し訳ないのだが, オッズ比に関するP値のPearsonのχ²検定版を以下のように定義する. 特にP値の構成法の(1)の段階がひどく天下り的である. しかし, それ以外の部分については計算の筋道を詳しく書いておくので, 自分の手で計算して確認して欲しい.

そして, さらに, それだけだとどういうことなのか分かりにくいので, 後の方で示す具体的な数値の計算例と以下で説明する各ステップの対応を繰り返し確認して欲しい.

検定したい仮説: 検定したい仮説は

$$\text{オッズ比パラメータは } \mathrm{OR} = \omega \text{ である.}$$

であるとする. ここで $\omega > 0$ は具体的な正の実数である.

仮説が $\mathrm{OR} = 1$ である場合に, 以下で定義するP値はちょうど **独立性に関するPearsonのχ²検定** で使うP値になっている.

P値の構成法: オッズ比パラメータの値に関する仮説のP値のPearsonのχ²検定版の構成法は以下の通り.

(1) 分割表のデータの数値 a, b, c, d に対して, 次を満たす $\delta = \delta(a, b, c, d|\omega)$ を求める:

$$\frac{(a-\delta)(d-\delta)}{(b+\delta)(c+\delta)} = \omega, \quad -\min(b, c) \leq \delta \leq \min(a, d).$$

これの左側の等式は

$$A = 1 - \omega, \quad B = a + d + \omega(b + c), \quad C = ad - \omega bc$$

とおくと, 次の二次方程式に書き直される:

$$A\delta^2 - B\delta + C = 0.$$

これの解で $-\min(b, c) \leq \delta \leq \min(a, d)$ を満たす側が欲しい $\delta = \delta(a, b, c, d|\omega)$ である. それは次のように表される:

$$\delta = \delta(a, b, c, d|\omega) = \frac{2C}{B + \sqrt{B^2 - 4AC}}.$$

この表示で浮動小数点数の計算をした方が安全でかつ誤差も小さくなり易い.

特に $\omega = 1$ の場合(独立性検定の場合)には, δ を決める方程式は $-(a + b + c + d)\delta + ad - bc = 0$ になるので,

$$\omega = 1 \implies \delta = \delta(a, b, c, d|1) = \frac{ad - bc}{a + b + c + d} = \frac{ad - bc}{N}.$$

(2) $\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}$ を次のように定める:

$$\tilde{a} = a - \delta, \quad \tilde{b} = b + \delta, \quad \tilde{c} = c + \delta, \quad \tilde{d} = d - \delta.$$

これが実は仮説 $\mathrm{OR} = 1$ 下での期待値の最尤推定値になっていることを後の方の節で説明する.

特に $\omega = 1$ の場合(独立性検定の場合)には次のようになっている:

$$\omega = 1 \implies \begin{cases} \tilde{a} = \dfrac{(a+b)(a+c)}{N} = \dfrac{ms}{N}, & \tilde{b} = \dfrac{(a+b)(b+d)}{N} = \dfrac{mr}{N}, \\ \tilde{c} = \dfrac{(c+d)(a+c)}{N} = \dfrac{ns}{N}, & \tilde{d} = \dfrac{(c+d)(b+d)}{N} = \dfrac{nr}{N}. \end{cases}$$

(3) データの数値に対応するPearsonのχ²統計量 $\chi^2 = \chi^2(a, b, c, d|\delta)$ の値を次のように定める:

$$\begin{aligned} \chi^2 = \chi^2(a, b, c, d|\delta) &= \delta^2 \left(\frac{1}{a-\delta} + \frac{1}{b+\delta} + \frac{1}{c+\delta} + \frac{1}{d-\delta} \right) \\ &= \frac{(a-\tilde{a})^2}{\tilde{a}} + \frac{(b-\tilde{b})^2}{\tilde{b}} + \frac{(c-\tilde{c})^2}{\tilde{c}} + \frac{(d-\tilde{d})^2}{\tilde{d}}. \end{aligned}$$

ここで, $\delta = \delta(a, b, c, d|\omega)$ は a, b, c, d, ω のみを使って計算できる数値なので, この χ^2 = \chi^2(a,b,c,d|\delta)もa,b,c,d,\omega$ のみを使って計算できる数値になっている.

上の χ^2 に関する後者の表示はPearsonのχ²統計量の一般的な表示

$$\chi^2 = \sum \frac{(\text{(観測値)} - \text{(帰無仮説下のモデル内での期待値)})^2}{\text{(帰無仮説下のモデル内での期待値)}}$$

の特別な場合になっていることに注意せよ. ただし, 「帰無仮説下のモデル内での期待値」は最尤推定値として求めたものになっている.

特に $\omega = 1$ の場合(独立性検定の場合)には次のようになっている:

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{N(ad - bc)^2}{mnrs}$$

これは独立性に関するPearsonのχ²統計量の公式として有名である.

(4) 仮説 $\mathrm{OR} = \omega$ の下での統計モデル内で同様に定義されたPearsonのχ²統計量の値がデータから計算した χ^2 の値以上になる確率の近似値としてP値を定義したい. (一般にP値は, 統計モデル内でデータの数値以上に極端な場合が生じる確率またはその近似値として定義されるのであった.)

仮に a, b, c, d が現実から得たデータの数値ではなく, 仮説 $\mathrm{OR} = \omega$ の下での統計モデル内でランダムに生成された仮想的なデータであるとき, (3)のように定義された χ^2 は a, b, c, d が十分に大きなとき(実はそう大きくなくてもよい), 近似的に自由度 1 のχ²分布に従うことを示せる.

このことを使って, 仮説 $\mathrm{OR} = \omega$ のP値を次のように定義する:

$$\mathrm{pvalue}_{\mathrm{Pearson}}(a, b, c, d|\ \mathrm{OR} = \omega) = 1 - \mathrm{cdf}(\mathrm{Chisq}(1), \chi^2).$$

仮説 $\mathrm{OR} = \omega$ に関するデータの数値 a, b, c, d に関するPearsonのχ²統計量の値 $\chi^2 = \chi^2(a, b, c, d|\delta(a, b, c, d|\omega)$ は a, b, c, d, ω だけで計算される数値なので, これでP値が定義されている.

対応する信頼区間: このP値の定義に対応するオッズ比パラメータ OR の信頼度 $1 - \alpha$ の信頼区間は次のように定義される:

$$\mathrm{confint}^{\mathrm{OR}}_{\mathrm{Pearson}}(a, b, c, d|\alpha) = \{\, \omega > 0 \mid \mathrm{pvalue}_{\mathrm{Pearson}}(a, b, c, d|\ \mathrm{OR} = \omega) \geq \alpha \,\}.$$

この信頼区間を計算するために使えるシンプルな公式はないように思われる.

信頼区間の計算法1: ω に関する次の方程式の解を2つ求めれば, その2つがオッズ比パラメータ OR の信頼区間の両端の値になる:

$$\mathrm{pvalue}_{\mathrm{Pearson}}(a, b, c, d|\ \mathrm{OR} = \omega) = \alpha.$$

信頼区間の計算法2: まず, $\chi^2_{\nu,\alpha}$ を

$$\chi^2_{\nu,\alpha} = \mathrm{quantile}(\mathrm{Chisq}(\nu), 1 - \alpha)$$

と定める. そして, $-\min(b, c) < \delta < \min(a, d)$ と次を満たす δ を2つ求める:

$$\chi^2(a, b, c, d|\delta) = \delta^2 \left(\frac{1}{a - \delta} + \frac{1}{b + \delta} + \frac{1}{c + \delta} + \frac{1}{d - \delta} \right) = \chi^2_{1,\alpha}$$

そして, それらの δ に対して, ω を

$$\omega = \frac{(a - \delta)(d - \delta)}{(b + \delta)(c + \delta)}$$

によって対応させれば, 信頼区間の両端の値が得られる.

文献: この節の構成は次の論文に書いてある方法の連続補正無し版になっている:

- Jerome Cornfield, A Statistical Problem Arising from Retrospective Studies, Berkeley Symposium on Mathematical Statistics and Probability, 1956: 135-148 (1956) link (https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings%20of%20the%20Third%20Berkeley%20Symposium%20on%20Mathematical%20Statistics%20and%20

3.2 Pearsonのχ²検定版のオッズ比に関するP値と信頼区間の計算例

データが次の場合のPearsonのχ²検定版仮説「オッズ比パラメータは $\mathrm{OR} = 1$ である」のP値とオッズ比 OR の 95% 信頼区間を計算してみよう:

	商品を購入した	購入しなかった	合計
デザインA	49	965	1014
デザインB	26	854	880
合計	75	1819	1894

結果は次のようになる:

- (Pearsonのχ²検定版の仮説「オッズ比パラメータは $\mathrm{OR} = 1$ である」のP値) ≈ 3.661%
- (Pearsonのχ²検定版のオッズ比パラメータ OR の 95% 信頼区間) ≈ [1.0318, 2.6957]

信頼区間ついてはJulia言語版の計算例のみを示す. WolframAlphaでこの信頼区間を求めることはかなり面倒である.

3.2.1 WolframAlphaによるPearsonのχ²検定のP値の計算の仕方

$\omega = 1$ の場合には, Pearsonのχ²統計量は,

$$\chi^2 = \frac{(a + b + c + d)(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

となるのであった. これを用いてP値を計算してみよう.

`(a+b+c+d)(ad-bc)^2/((a+b)(c+d)(a+c)(b+d)) where a=49.0, b=965, c=26, d=854` → 実行 (https://www.wolframalpha.com/input?i=%28a%2Bb%2Bc%2Bd%29%28ad-bc%29%5E2%2F%28%28a%2Bb%29%28c%2Bd%29%28a%2Bc%29%28b%2Bd%29%29+where+a%3D49.0%2C+b%3D965%2C- → 4.36824 (独立性に関するPearsonのχ²統計量)

`1 - cdf(ChisqDistribution(1), 4.36824)` → 実行 (https://www.wolframalpha.com/input?i=1+-+cdf%28ChisqDistribution%281%29%2C+4.36824%29) → 0.0366148 (P値)

3.2.2 Julia言語によるPearsonのχ²検定版のオッズ比に関するP値と信頼区間の計算の仕方(1)

$\omega = 1$ の場合のPearsonのχ²検定版のオッズ比に関するP値は, 独立性に関するPearsonのχ²検定のP値そのものになる. それを上でやったのと同じ方法で計算する.

信頼区間の側は函数の零点を見つけてくれる函数を使って求めてみよう.

In [22]:
```
a, b, c, d = 49, 965, 26, 854
@show a, b, c, d
χ² = (a+b+c+d)*(a*d-b*c)^2/((a+b)*(c+d)*(a+c)*(b+d))
@show χ²
@show pvalue = ccdf(Chisq(1), χ²)
α = 0.05
@show α
confint = find_zeros(0.5, 3.0) do ω
    A, B, C = 1-ω, a+d+ω*(b+c), a*d-ω*b*c
    δ = 2C/(B + √(B^2 - 4A*C))
    χ² = δ^2 * (1/(a-δ) + 1/(b+δ) + 1/(c+δ) + 1/(d-δ))
    ccdf(Chisq(1), χ²) - α
end
@show confint;
```

```
(a, b, c, d) = (49, 965, 26, 854)
χ² = 4.368235947209686
pvalue = ccdf(Chisq(1), χ²) = 0.03661487827607113
α = 0.05
confint = [1.031778110800563, 2.6957054358850328]
```

`f(x, y) do z ... end` 構文については

- https://docs.julialang.org/en/v1/manual/functions/#Do-Block-Syntax-for-Function-Arguments (https://docs.julialang.org/en/v1/manual/functions/#Do-Block-Syntax-for-Function-Arguments)

を参照せよ. `find_zeros` 函数の使い方については,

- https://juliamath.github.io/Roots.jl/stable/#Basic-usage (https://juliamath.github.io/Roots.jl/stable/#Basic-usage)

を参照せよ.

3.2.3 Julia言語によるPearsonのχ²検定版のオッズ比に関するP値と信頼区間の計算の仕方(2)

このノートブックの最初の方で定義した函数を使って計算.

In [23]:
```
a, b, c, d = 49, 965, 26, 854
@show a, b, c, d
@show chisqstat_or(a, b, c, d; ω=1)
@show pvalue_or_pearson(a, b, c, d; ω=1)
@show confint_or_pearson(a, b, c, d; α=0.05);
```

```
(a, b, c, d) = (49, 965, 26, 854)
chisqstat_or(a, b, c, d; ω = 1) = 4.368235947209686
pvalue_or_pearson(a, b, c, d; ω = 1) = 0.03661487827607113
confint_or_pearson(a, b, c, d; α = 0.05) = [1.031778110800563, 2.6957054358850328]
```

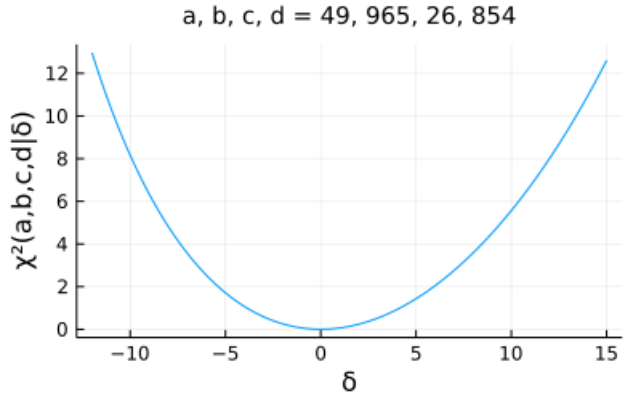
3.2.4 Julia言語によるPearsonのχ²検定版のオッズ比に関するP値の視覚化

まず, 次の函数のグラフを描いてみよう:

$$\delta \mapsto \chi^2(a,b,c,d|\delta) = \delta^2 \left(\frac{1}{a-\delta} + \frac{1}{b+\delta} + \frac{1}{c+\delta} + \frac{1}{d-\delta} \right).$$

In [24]:
```
a, b, c, d = 49, 965, 26, 854
plot(δ -> _chisqstat_or(a,b,c,d,δ), -12, 15; label="")
plot!(; xguide="δ", yguide="χ²(a,b,c,d|δ)")
title!("a, b, c, d = $a, $b, $c, $d")
```

Out[24]:

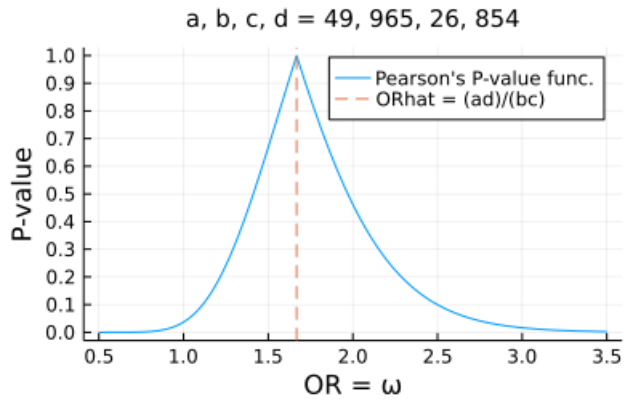


このように $\delta \mapsto \chi^2(a,b,c,d|\delta)$ は $\delta = 0$ で最小値 0 になるU型のグラフになる. ゆえに, δ に関する方程式 $\chi^2(a,b,c,d|\delta) = \chi^2_{1,\alpha}$ の $-\min(b,c) < \delta < \min(a,d)$ を満たす解は負値と正値の2つになる.

次に, P値函数 $\omega \mapsto \mathrm{pvalue}(a,b,c,d|\,\mathrm{OR} = \omega)$ のグラフを描いてみよう.

In [25]:
```
a, b, c, d = 49, 965, 26, 854
plot(ω -> pvalue_or_pearson(a,b,c,d;ω), 0.5, 3.5; label="Pearson's P-value func.")
vline!([oddsratiohat(a,b,c,d)]; label="ORhat = (ad)/(bc)", ls=:dash)
plot!(; xguide="OR = ω", yguide="P-value")
plot!(; ytick=0:0.1:1)
title!("a, b, c, d = $a, $b, $c, $d")
```

Out[25]:

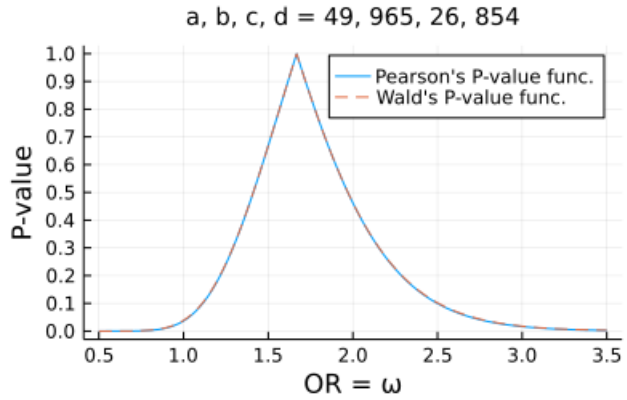


このようにP値函数 $\omega \mapsto \mathrm{pvalue}(a,b,c,d|\,\mathrm{OR} = \omega)$ は $\omega = \widehat{\mathrm{OR}} = (ad)/(bc)$ で最大値 1 になり, 「とんがり帽子」の形のグラフになる.

Wald版のP値函数のグラフと重ねてプロットすると次のようになる.

In [26]:
```
a, b, c, d = 49, 965, 26, 854
plot(ω -> pvalue_or_pearson(a,b,c,d; ω), 0.5, 3.5; label="Pearson's P-value func.")
plot!(ω -> pvalue_or_wald(a,b,c,d; ω), 0.5, 3.5; label="Wald's P-value func.", ls=:dash)
plot!(; xguide="OR = ω", yguide="P-value")
plot!(; ytick=0:0.1:1)
title!("a, b, c, d = $a, $b, $c, $d")
```

Out[26]:



この場合には a, b, c, d が十分に大きいので, ほぼぴったり重なっている!

このことから, a, b, c, d が十分に大きな場合には, 計算法が複雑なPearsonのχ²検定版のP値と信頼区間を使う必要がないことがわかる. そのような場合にはWald型のP値と信頼区間で実用的には十分である.

P値と 95% 信頼区間も以下のようにほぼ同じになっている:

- P値の比較:
  - (Pearsonのχ²検定版の仮説「オッズ比パラメータは $\mathrm{OR} = 1$ である」のP値) ≈ 3.7%
  - (Wald版の仮説「オッズ比パラメータは $\mathrm{OR} = 1$ である」のP値) ≈ 3.8%
- 信頼区間の比較:
  - (Pearsonのχ²検定版のオッズ比パラメータ OR の 95% 信頼区間) ≈ [1.03, 2.70]
  - (Wald版のオッズ比パラメータ OR の 95% 信頼区間) ≈ [1.03, 2.71]

数学的注意: 数学の世界では, 定義が全く異なる2つの量がある条件のもとで近似的によく一致するというようなことが起こる. 異なる動機や思想のもとで定義された2つの量が現実世界での応用では「同じ」とみなしてよいだけ近似的によく一致することがある. そのような場合に元の動機や思想にこだわって解釈しようとすることは誤りである.

注意: a, b, c, d が小さい場合には違いが大きくなる場合がある.

In [27]:
```
a, b, c, d = 4, 1, 1, 5
@show pvalue_or_pearson(a,b,c,d)
@show pvalue_or_wald(a,b,c,d)
println()
@show confint_or_pearson(a,b,c,d)
@show confint_or_wald(a,b,c,d);
```

```
pvalue_or_pearson(a, b, c, d) = 0.03568232934330229
pvalue_or_wald(a, b, c, d) = 0.05563253253702121

confint_or_pearson(a, b, c, d) = [1.1871236326245318, 337.7203835779612]
confint_or_wald(a, b, c, d) = [0.930440993688474, 429.9036722514896]
```

このように a, b, c, d が小さい場合に信頼区間の幅はものすごく大きくなる.

3.2.5 R言語によるPearsonのχ²検定のP値の計算の仕方

独立性に関するPearsonのχ²検定のP値は次のようにして計算できる.

```
A = matrix(c(49, 965, 26, 854), 2, 2, byrow=T)
result = chisq.test(A, correct=F)
```

In [28]:
```
R"""
A = matrix(c(49, 965, 26, 854), 2, 2, byrow=T)
result = chisq.test(A, correct=F)
"""
```

Out[28]:
```
RObject{VecSxp}

        Pearson's Chi-squared test

data:  A
X-squared = 4.3682, df = 1, p-value = 0.03661
```

3.3 「オッズ比パラメータはOR=ωである」という仮説下での最尤法との関係

データの数値 a, b, c, d が与えらていると仮定し, $a+b=m$, $c+d=n$ であるとする.

3.3.1 2つの二項分布モデル全体での最尤法

2つの二項分布モデルの対数尤度函数は

$$\begin{aligned}(p,q) \mapsto L &= \log P(a,b,c,d|m,n,p,q) \\ &= \log\left(\binom{m}{a}p^a(1-p)^b\binom{n}{c}q^c(1-q)^d\right) \\ &= a\log p + b\log(1-p) + c\log q + d\log(1-q) + (\text{constant w.r.t. } p,q).\end{aligned}$$

なので,

$$\frac{\partial L}{\partial p} = \frac{a}{p} - \frac{b}{1-p} = \frac{a-(a+b)p}{p(1-p)}.$$

これより, $\partial L/\partial p = 0$ と $p = \hat{p} := a/(a+b) = a/m$ は同値になる. 同様にして, $\partial L/\partial q = 0$ と $q = \hat{q} := c/(c+d) = c/n$ は同値になる. このことより, 対数尤度 L を最大化する p, q は $\hat{p} = a/m$, $\hat{q} = c/n$ になることがわかる.

$p = \hat{p} = a/m$, $q = \hat{q} = c/n$ のときの2つの二項分布モデルでの期待値は

$$m\hat{p} = a, \quad m(1-\hat{p}) = b, \quad n\hat{q} = c, \quad n(1-\hat{q}) = d$$

になる. これはパラメータの動く範囲に制限を付けない2つの二項分布モデルの場合には, パラメータに最尤推定値を代入して得られる確率分布(**予測分布** という)の期待値が, データの値 a, b, c, d そのものになることを意味している.

注意: この性質は, 4つのPoisson分布モデルでも四項分布モデルでも成立している.

3.3.2 条件OR=ωで制限した2つの二項分布モデルでの最尤法

$\omega > 0$ を任意に取って固定する.

条件 $\mathrm{OR} = (p/(1-p))/(q/(1-q)) = \omega$ という条件の下で対数尤度を最大化する問題をLagrangeの未定乗数法を使って解こう. 函数 $(p,q,\delta) \mapsto M$ を次のように定める:

$$\begin{aligned}M &= L - \delta(\log \mathrm{OR} - \log\omega) \\ &= a\log p + b\log(1-p) + c\log q + d\log(1-q) \\ &\quad - \delta(\log p - \log(1-p) - \log q + \log(1-q) - \log\omega) + \text{const.}\end{aligned}$$

このとき,

$$\begin{aligned}\frac{\partial M}{\partial p} &= \frac{a}{p} - \frac{b}{1-p} - \frac{\delta}{p} - \frac{\delta}{1-p} = \frac{a-\delta}{p} - \frac{b+\delta}{1-p}, \\ \frac{\partial M}{\partial q} &= \frac{c}{q} - \frac{d}{1-q} + \frac{\delta}{q} + \frac{\delta}{1-q} = \frac{c+\delta}{q} - \frac{d-\delta}{1-q}, \\ -\frac{\partial M}{\partial \delta} &= \log \mathrm{OR} - \log\omega = \log\frac{p/(1-p)}{q/(1-q)} - \log\omega.\end{aligned}$$

これらがすべて 0 になることと次が成立することは同値である:

$$\frac{p}{1-p} = \frac{a-\delta}{b+\delta}, \quad \frac{q}{1-q} = \frac{c+\delta}{d-\delta}, \quad \mathrm{OR} = \frac{(a-\delta)(d-\delta)}{(b+\delta)(c+\delta)} = \omega.$$

ゆえに

$$\frac{(a-\delta)(d-\delta)}{(b+\delta)(c+\delta)} = \omega, \quad -\min(b,c) < \delta < \min(a,d)$$

を満たすように δ を取り, $\tilde{p}, \tilde{q}$ を

$$\frac{\tilde{p}}{1-\tilde{p}} = \frac{a-\delta}{b+\delta}, \quad \frac{\tilde{q}}{1-\tilde{q}} = \frac{c+\delta}{d-\delta}$$

を満たすように

$$\tilde{p} = \frac{\tilde{a}}{\tilde{a}+\tilde{b}} = \frac{\tilde{a}}{m}, \quad \tilde{q} = \frac{\tilde{c}}{\tilde{c}+\tilde{d}} = \frac{\tilde{c}}{n}$$

と定めれば, $p = \tilde{p}$, $q = \tilde{q}$ が条件 $\mathrm{OR} = \omega$ の下での最尤推定値になる.

そのとき, モデルの期待値は

$$m\tilde{p} = a - \delta, \quad m(1-\tilde{p}) = b + \delta, \quad n\tilde{q} = c + \delta, \quad n(1-\tilde{q}) = d - \delta$$

になる. **これでPearsonのχ²検定版のオッズ比に関するP値の構成法の(1)のステップが本質的に条件 $\mathrm{OR} = \omega$ の下での最尤法そのものであったことがわかった.**

3.3.3 Wilks's theoremの適用

Wilks' theorem (https://en.wikipedia.org/wiki/Wilks%27_theorem)について説明しよう.

適当な緩い条件のもとで, パラメータ θ, η を持つ統計モデル $D(\theta,\eta)^n$ (ただし $\theta = (\theta_1, \ldots, \theta_\nu)$ は ν 個の独立なパラメータ達であるとする)について以下が成立している.

$X = (X_1, \ldots, X_n)$ は分布 $D(\theta_0, \eta_0)^n$ 従う確率変数であると仮定する.

パラメータの動ける範囲に制限を付けないモデル $D(\theta,\eta)^n$ における確率変数としてのデータ X から得られる最尤推定量を $\hat{\theta}, \hat{\eta}$ と表す.

パラメータの動ける範囲に $\theta = \theta_0$ という制限を付けたモデル $D(\theta,\eta)^n$ における確率変数としてのデータ X から得られる最尤推定量を $\tilde{\theta} = \theta_0, \tilde{\eta}$ と表す.

このとき, n が大きければ, 対数尤度比は自由度 ν のχ²分布に近似的に従う:

$$2\log\frac{(X \text{ に関する } D(\hat{\theta},\hat{\eta})^n \text{ の尤度})}{(X \text{ に関する } D(\theta_0,\tilde{\eta})^n \text{ の尤度})} \sim \mathrm{Chisq}(\nu), \quad \text{approximately}$$

これがWilksの定理である. 要するにパラメータの動ける範囲を ν 次元下げた場合の最尤法を考えると, 自然に自由度 ν のχ²分布に近似的に従う確率変数が得られるという仕組みになっている.

上の結果における左辺の対数尤度比は $n \to \infty$ で漸近的に同値な別の確率変数に置き換えられる.

例えば, 2つの二項分布モデルで $\nu = 1$, $\theta = \mathrm{OR}$, $\theta_0 = \omega$, $\eta = q$ のとき, 対数尤度比は分子分母が大幅にキャンセルして次のようにシンプルな式になる:

$$2\log\frac{(a/m)^a(b/m)^b(c/n)^c(d/n)^d}{(\tilde{a}/m)^a(\tilde{b}/m)^b(\tilde{c}/n)^c(\tilde{d}/n)^d} = 2a\log\frac{a}{\tilde{a}} + 2b\log\frac{b}{\tilde{b}} + 2c\log\frac{c}{\tilde{c}} + 2d\log\frac{d}{\tilde{d}}.$$

これに, $f(x) = 2x\log(x/\tilde{x})$ のとき, $f'(x) = 2\log(x/\tilde{x}) + 2$, $f''(x) = 2/x$ より,

$$2x\log\frac{x}{\tilde{x}} = 2(x-\tilde{x}) + \frac{(x-\tilde{x})^2}{\tilde{x}} + O((x-\tilde{x})^2)$$

となることを適用すると, $a+b+c+d = \tilde{a}+\tilde{b}+\tilde{c}+\tilde{d}$ より,

$$\begin{aligned} &2a\log\frac{a}{\tilde{a}} + 2b\log\frac{b}{\tilde{b}} + 2c\log\frac{c}{\tilde{c}} + 2d\log\frac{d}{\tilde{d}} \\ &\approx \frac{(a-\tilde{a})^2}{\tilde{a}} + \frac{(b-\tilde{b})^2}{\tilde{b}} + \frac{(c-\tilde{c})^2}{\tilde{c}} + \frac{(d-\tilde{d})^2}{\tilde{d}} = (\text{Pearsonの}\chi^2\text{統計量}) \end{aligned}$$

とPearsonのχ²統計量が得られ,

$$(\text{Pearsonの}\chi^2\text{統計量}) \sim \mathrm{Chisq}(1), \quad \text{approximately}$$

が得られる. これがPearsonのχ²検定の基礎になる.

注意: 以上の話を少し一般化するとPearsonのχ²統計量の一般化である **スコア統計量** を使ったχ²検定の構成法が得られる.

3.4 問題: 「リスク比パラメータはRR=ρである」という仮説下での最尤法

以下示せ.

$\mathrm{RR} = \rho$ という条件の下での, データの数値 a, b, c, d に関する2つの二項分布モデルの最尤推定値 $\tilde{p}, \tilde{q}$ が

$$\tilde{p} = \frac{a - \Delta}{a - \Delta + b}, \quad \tilde{q} = \frac{c + \Delta}{c + \Delta + d}$$

になることを示せ. ただし Δ は次の条件で定める:

$$\frac{(a - \Delta)(c + \Delta + d)}{(a - \Delta + b)(c + \Delta)} = \rho, \quad -c < \Delta < a.$$

注意: パラメータの値をその最尤推定値にしたとき, 2つの二項分布モデルの期待値は次のようになる:

$$\begin{aligned} m\tilde{p} &= \frac{m}{m - \Delta}(a - \Delta), \quad m(1 - \tilde{p}) = \frac{m}{m - \Delta} b, \\ n\tilde{q} &= \frac{n}{n + \Delta}(c + \Delta), \quad m(1 - \tilde{q}) = \frac{n}{n + \Delta} d. \end{aligned}$$

解答例: Lagrangeの未定乗数法をオッズ比を扱った場合と同様に用いる.

函数 $(p, q, \Delta) \mapsto M$ を次のように定める:

$$\begin{aligned} M &= \log P(a, b, c, d | m, n, p, q) - \Delta(\log \mathrm{RR} - \log \rho) \\ &= a \log p + b \log(1 - p) + c \log q + d \log(1 - q) \\ &\quad - \Delta(\log p - \log q - \log \rho) + \mathrm{const.} \end{aligned}$$

このとき,

$$\begin{aligned} \frac{\partial M}{\partial p} &= \frac{a}{p} - \frac{b}{1 - p} - \frac{\Delta}{p} = \frac{a - \Delta}{p} - \frac{b}{1 - p}, \\ \frac{\partial M}{\partial q} &= \frac{c}{q} - \frac{d}{1 - q} + \frac{\Delta}{q} = \frac{c + \delta}{q} - \frac{d}{1 - q}, \\ -\frac{\partial M}{\partial \Delta} &= \log \mathrm{RR} - \log \omega = \log \frac{p}{q} - \log \omega. \end{aligned}$$

これらがすべて 0 になることと次が成立することは同値である:

$$\frac{p}{1 - p} = \frac{a - \Delta}{b}, \quad \frac{q}{1 - q} = \frac{c + \Delta}{d}, \quad \mathrm{RR} = \frac{p}{q} = \rho.$$

さらに

$$\frac{p}{1 - p} = \frac{a - \Delta}{b}, \quad \frac{q}{1 - q} = \frac{c + \Delta}{d}$$

と

$$p = \tilde{p} := \frac{a - \Delta}{a - \Delta + b}, \quad q = \tilde{q} := \frac{c + \Delta}{c + \Delta + d}$$

は同値であり, これが成立しているとき, $\mathrm{RR} = p/q = \rho$ は

$$\frac{(a - \Delta)(c + \Delta + d)}{(a - \Delta + b)(c + \Delta)} = \rho$$

と同値である. これと $-c < \Delta < a$ を満たす Δ を取って, すぐ上の式で $\tilde{p}, \tilde{q}$ を定義すると, $p = \tilde{p}$, $q = \tilde{q}$ が $\mathrm{RR} = \rho$ という条件の下での最尤推定量になる.

解答終

3.5 問題: Pearsonのχ²検定版のリスク比に関するP値と信頼区間の定義

上の問題の結果を用いてPearsonのχ²検定版のリスク比に関するP値と信頼区間を適切に定義せよ.

注意: 適切な定義を自分で考えることは物事を楽に理解するために必要なことである.

解答例: Δ を

$$\frac{(a-\Delta)(n+\Delta)}{(m-\Delta)(c+\Delta)} = \rho, \quad -c < \Delta < a.$$

という条件で定める. このような Δ は

$$A = \rho - 1, \quad B = n - a + \rho(m-c), \quad C = an - \rho mc$$

と置き,

$$\Delta = \frac{2C}{B + \sqrt{B^2 - 4AC}}$$

とおけば求まる. さらに, $\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}$ を次のように定める:

$$\begin{aligned} \tilde{a} &= \frac{m}{m-\Delta}(a-\Delta), \quad \tilde{b} = \frac{m}{m-\Delta}b, \\ \tilde{c} &= \frac{n}{n+\Delta}(c+\Delta), \quad \tilde{d} = \frac{n}{n+\Delta}d. \end{aligned}$$

このとき, $(\tilde{a}/m)/(\tilde{c}/n) = \rho$ となっていることに注意せよ.

Pearsonのχ²統計量 χ^2 を次のように定める:

$$\chi^2 = \frac{(a-\tilde{a})^2}{\tilde{a}} + \frac{(b-\tilde{b})^2}{\tilde{b}} + \frac{(c-\tilde{c})^2}{\tilde{c}} + \frac{(d-\tilde{d})^2}{\tilde{d}}.$$

上の定義のもとで,

$$\begin{aligned} a - \tilde{a} &= \frac{b\Delta}{m-\Delta}, \quad b - \tilde{b} = \frac{-b\Delta}{m-\Delta}, \\ c - \tilde{c} &= \frac{-d\Delta}{n+\Delta}, \quad d - \tilde{d} = \frac{d\Delta}{n+\Delta} \end{aligned}$$

でかつ

$$\frac{1}{\tilde{a}} + \frac{1}{\tilde{b}} = \frac{(m-\Delta)^2}{m(a-\Delta)b}, \quad \frac{1}{\tilde{c}} + \frac{1}{\tilde{d}} = \frac{(n+\Delta)^2}{n(c+\Delta)d}$$

なので,

$$\frac{(a-\tilde{a})^2}{\tilde{a}} + \frac{(b-\tilde{b})^2}{\tilde{b}} = \frac{b\Delta^2}{m(a-\Delta)}, \quad \frac{(c-\tilde{a})^2}{\tilde{c}} + \frac{(d-\tilde{b})^2}{\tilde{d}} = \frac{d\Delta^2}{n(c+\Delta)}$$

となることより,

$$\chi^2 = \Delta^2 \left(\frac{b}{m(a-\Delta)} + \frac{d}{n(c+\Delta)} \right).$$

仮説「リスク比パラメータは $\mathrm{RR} = \rho$ である」のP値を次のように定めることができる:

$$\mathrm{pvalue}_{\mathrm{Pearson}}(a, b, c, d | \, \mathrm{RR} = \rho) = 1 - \mathrm{cdf}(\mathrm{Chisq}(1), \chi^2).$$

これに対応するリスク比パラメータ RR の信頼度 $1-\alpha$ の信頼区間は

$$\mathrm{confint}^{\mathrm{RR}}_{\mathrm{Pearson}}(a, b, c, d | \alpha) = \{ \rho > 0 \mid \mathrm{pvalue}_{\mathrm{Pearson}}(a, b, c, d | \, \mathrm{RR} = \rho) \geq \alpha \}.$$

になる.

解答終

3.6 問題: Pearsonのχ²検定版のリスク比に関するP値と信頼区間の計算例

データが次の場合のPearsonのχ²検定版の仮説「リスク比パラメータは $\mathrm{RR} = 1$ である」のP値とリスク比パラメータ RR の 95% 信頼区間を計算してみよ:

	商品を購入した	購入しなかった	合計
デザインA	49	965	1014
デザインB	26	854	880
合計	75	1819	1894

3.6.1 Julia言語によるPearsonのχ²検定版のリスク比に関するP値と信頼区間の計算例

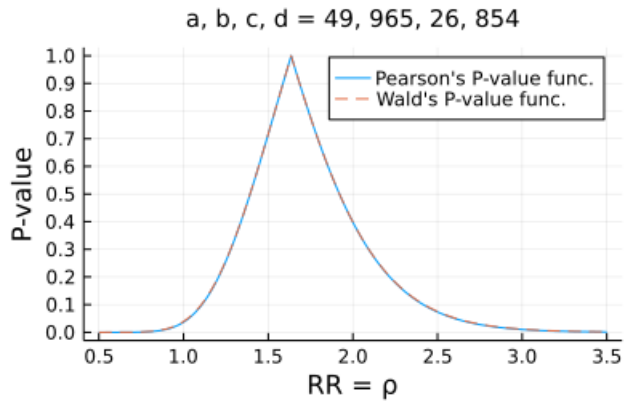
In [29]:
```
a, b, c, d = 49, 965, 26, 854
@show oddsratiohat(a, b, c, d)
@show riskratiohat(a, b, c, d)
@show pvalue_rr_wald(a, b, c, d)
@show pvalue_or_wald(a, b, c, d)
@show pvalue_rr_pearson(a, b, c, d)
@show pvalue_or_pearson(a, b, c, d)
@show confint_rr_wald(a, b, c, d)
@show confint_or_wald(a, b, c, d)
@show confint_rr_pearson(a, b, c, d)
@show confint_or_pearson(a, b, c, d);
```

```
oddsratiohat(a, b, c, d) = 1.6678357911518533
riskratiohat(a, b, c, d) = 1.635563647397967
pvalue_rr_wald(a, b, c, d) = 0.038895117565198624
pvalue_or_wald(a, b, c, d) = 0.03847043006362489
pvalue_rr_pearson(a, b, c, d) = 0.03661487827607109
pvalue_or_pearson(a, b, c, d) = 0.03661487827607113
confint_rr_wald(a, b, c, d) = [1.025408984412048, 2.6087819449169145]
confint_or_wald(a, b, c, d) = [1.0275181208761617, 2.7071797272783877]
confint_rr_pearson(a, b, c, d) = [1.03047149323473, 2.5998459013273476]
confint_or_pearson(a, b, c, d) = [1.031778110800563, 2.6957054358850328]
```

3.6.2 Julia言語によるPearsonのχ²検定版のリスク比に関するP値函数の視覚化

In [30]:
```
a, b, c, d = 49, 965, 26, 854
plot(ρ -> pvalue_rr_pearson(a,b,c,d; ρ), 0.5, 3.5; label="Pearson's P-value func.")
plot!(ρ -> pvalue_rr_wald(a,b,c,d; ρ), 0.5, 3.5; label="Wald's P-value func.", ls=:dash)
plot!(; xguide="RR = ρ", yguide="P-value")
plot!(; ytick=0:0.1:1)
title!("a, b, c, d = $a, $b, $c, $d")
```

Out[30]:



この場合には a, b, c, d が十分に大きいので, ほぼぴったり重なっている!

3.6.3 Pearsonのχ²検定版のリスク比に関するP値と信頼区間の計算の解答例

- (Pearsonのχ²検定版の仮説「リスク比パラメータは $\mathrm{RR} = 1$ である」のP値) ≈ 3.661%
- (Pearsonのχ²検定版のリスク比パラメータ RR の 95% 信頼区間) ≈ [1.0305, 2.5998]

3.7 ニューサンスパラメータの問題に関する注意

我々が想定している統計モデルには p と q という2つのパラメータが含まれている. そこに仮説 $\mathrm{OR} = \omega$ によって制限を課しても, 独立なパラメータの個数は1つしか減らず, たとえば q の側を自由に動けるパラメータとして採用し, 仮説 $\mathrm{OR} = \omega$ によって p の値は q の値から決まると考えることができる.

だから, P値を定義するために必要な「仮説 $\mathrm{OR}=\omega$ の下で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」は, パラメータ q の値を決めるごとに別々に決まる値になってしまう. このような状況のとき, q はニューサンスパラメータ(nuisance parameter, 局外パラメータ, 撹乱パラメータ, 迷惑パラメータ)であるという.

しかし, 「データの数値以上に極端な値」の意味を, Wald的な正規分布近似を使ったり, Pearsonのχ²統計量を使ったりして定義すれば, 「仮説 $\mathrm{OR}=\omega$ の下で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」が, a,b,c,d が十分に大きなとき, 近似的にパラメータ q の値によらなくなることを示せる.

以上で紹介したP値の定義にはこのような非常に巧妙な方法が使われている.

ニューサンスパラメータの問題への対処の仕方にはそれら以外にも, 以下の節で説明する条件付き確率分布を利用する方法がある. その方法による検定は **Fisher検定** と呼ばれている.

4 Fisher検定版のオッズ比に関するP値と信頼区間

4.1 条件付き確率分布を利用するアイデア

2つの二項分布モデルの確率質量函数は

$$\begin{gathered}P(a,b,c,d|m,n,p,q)=\binom{m}{a}p^a(1-p)^b\binom{n}{c}q^c(1-q)^d\\(a,b,c,d\in\mathbb{Z}_{\geq 0},\ a+b=m,\ c+d=n).\end{gathered}$$

と書けるのであった. この分布を条件 $a+c=r$ (そのとき $b+d=m+n-r$)という条件で制限して得られる条件付き確率分布を求めてみよう. パラメータのオッズ比 OR を

$$\mathrm{OR}=\frac{p(1-q)}{(1-p)q}$$

と定めると, 条件 $a+b=m$, $c+d=n$, $a+c=r$ の下で,

$$\begin{array}{llll}a=a, & b=m-a, & c=r-a, & d=n-r+a,\\ a\geq 0, & a\leq m, & a\leq r, & a\geq r-n\end{array}$$

が成立しているので, 上の確率質量函数は

$$P(a,b,c,d|m,n,p,q)=\binom{m}{a}\binom{n}{r-a}(1-p)^m q^r(1-q)^{n-r}(\mathrm{OR})^a\propto\binom{m}{a}\binom{n}{r-a}(\mathrm{OR})^a$$

従って,

$$Z(m,n,r,\mathrm{OR})=\sum_a\binom{m}{a}\binom{n}{r-a}(\mathrm{OR})^a$$

とおいて, 確率質量函数 $P(a|m,n,r,\mathrm{OR})$ を

$$P(a|m,n,r,\mathrm{OR})=\frac{1}{Z(m,n,r,\mathrm{OR})}\binom{m}{a}\binom{n}{r-a}(\mathrm{OR})^a\quad(\max(0,r-n)\leq a\leq\min(m,r))$$

と定義すると, この確率質量函数で定義される確率分布が条件 $a+c=r$ で2つの二項分布モデルを制限して得られる条件付き確率分布になる. この確率分布を **Fisherの非心超幾何分布** (Fisher's noncentral hypergeometric distribution)と呼び,

$$\mathrm{FisherNoncentralHypergeometric}(m,n,r,\mathrm{OR})$$

と表すことにする.

このように, 条件 $a+c=r$ によって条件付き確率分布に移れば, モデルのパラメータはオッズ比パラメータ OR だけになる. ニューサンスパラメータ(nuisance parameter, 局外パラメータ, 撹乱パラメータ, 迷惑パラメータ)がなくなるのである!

だから, 2×2の分割表のデータの数値 a,b,c,d が得られたときに, $a+b=m$, $c+d=n$, $a+c=r$, $b+d=m+n-r$ という縦と横方向のすべての合計(すべてのマージン)の値を固定することによって得られる条件付き確率分布(Fisherの非心超幾何分布)に移って, 「データの数値以上に極端な値が生じる確率」を正確に求めることによってP値が定義できることがわかる. P値の定義の方法は二項分布モデルの場合にならえばよい.

この方法には利点と欠点がある.

- **利点:** Wald版やPearsonのχ²検定版と違って, 確率の計算を近似計算にする必要がなくなる. それによって, 第一種の過誤が起こる確率を確実に有意水準以下にできる.

- **欠点:** 有限離散分布では分布によって生じる値の個数が減ると, 確率を正確に計算する方法で定義されたP値について, 第一種の過誤の確率が有意水準よりも真に小さくなってしまうことが普通になる. しかもそれを条件付き確率分布に移ることによって行うと, 第一種の過誤の確率が有意水準よりも大幅に小さくなってしまう場合が出て来る.

第一種の過誤の確率を確実に有意水準以下にしたい場合にはFisher検定は優れた方法になるが, 検出力はその分だけ下がってしまうという欠点がある. 利点と欠点を理解して, 自分の目的に合わせて適切な道具を使えばよいだろう.

注意: Fisherの非心超幾何分布を $\mathrm{OR} = 1$ に特殊化すると, 超幾何分布が得られる:

$$P(a|m,n,r,1) = \binom{m+n}{r}^{-1}\binom{m}{a}\binom{n}{r-a} \quad (\max(0,r-n) \leq a \leq \min(m,r)).$$

この確率は, 赤い玉が m 個, 白い玉が n 個入っている壺から, r 個の玉を無作為に取り出すときに(非復元抽出, r 個をまとめて一度に取り出す様子を想像すればよい), 取り出した r 個の玉の中に含まれる赤い玉の個数が a 個になる確率だと解釈される.

4.2 Sterne型のFisher検定のP値とそれに付随する信頼区間の定義

前節の記号の下で, データの数値 a,b,c,d に対して, $m = a+b$, $n = c+d$, $r = a+c$ とおき, 仮説「オッズ比パラメータは $\mathrm{OR} = \omega$ である」のSterne型のP値を次のように定める:

$$\mathrm{pvalue}_{\mathrm{Sterne}}(a,b,c,d|\,\mathrm{OR} = \omega) = \sum_{P(i|m,n,r,\omega) \leq P(a|m,n,r,\omega)} P(i|m,n,r,\omega).$$

これに対応する信頼度 $1-\alpha$ の信頼区間を

$$\mathrm{confint}^{\mathrm{OR}}_{\mathrm{Sterne}}(a,b,c,d|\alpha) = \{\,\omega > 0 \mid \mathrm{pvalue}_{\mathrm{Sterne}}(a,b,c,d|\,\mathrm{OR} = \omega) \geq \alpha\,\}$$

によって定める. もしくは, 実用的には, これの上限と下限に挟まれた区間として, $\mathrm{confint}_{\mathrm{Sterne}}(a,b,c,d|\,\mathrm{OR} = \omega)$ を定義し直したものを使用する. (このノートでの実装ではそうなっている.)

4.3 Clopper-Pearson型のFisher検定のP値とそれに付随する信頼区間

さらに続けて, Clopper-Pearson型のP値を次のように定める:

$$\mathrm{pvalue}_{\mathrm{CP}}(a,b,c,d|\,\mathrm{OR} = \omega) = \min\begin{pmatrix} 1 \\ 2\,\mathrm{cdf}(\mathrm{FisherNoncentralHypergeometric}(a+b,c+d,a+c,\omega),a) \\ 2(1-\mathrm{cdf}(\mathrm{FisherNoncentralHypergeometric}(a+b,c+d,a+c,\omega),a-1)) \end{pmatrix}$$

これに対応する信頼度 $1-\alpha$ の信頼区間を

$$\mathrm{confint}^{\mathrm{OR}}_{\mathrm{CP}}(a,b,c,d|\alpha) = \{\,\omega > 0 \mid \mathrm{pvalue}_{\mathrm{CP}}(a,b,c,d|\,\mathrm{OR} = \omega) \geq \alpha\,\}$$

によって定める.

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```