検定と信頼区間: 平均の比較

- 黒木玄

- 2022-06-16〜2022-06-16

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)

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 : isinf(y) ? zero(y) : x/y

x ≲ y = x < y || x ≈ y

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

distname(dist::Distribution) = replace(string(dist), r"{.*}" => "")
myskewness(dist) = skewness(dist)
mykurtosis(dist) = kurtosis(dist)
function standardized_moment(dist::ContinuousUnivariateDistribution, m)
    μ, σ = mean(dist), std(dist)
    quadgk(x -> (x - μ)^m * pdf(dist, x), extrema(dist)...)[1] / σ^m
end
myskewness(dist::MixtureModel{Univariate, Continuous}) =
    standardized_moment(dist, 3)
mykurtosis(dist::MixtureModel{Univariate, Continuous}) =
    standardized_moment(dist, 4) - 3
```

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

1 平均の差に関するP値と信頼区間

1.1 平均の差に関するP値と信頼区間を使って行いたいこと

以下のようなことを行いたい.

(1) 日本にいる12歳の男子と女子をそれぞれ m 人と n 人無作為抽出して, 身長を測って得た数値のデータをそれぞれ $x_1, \ldots, x_m$, $y_1, \ldots, y_n$ と書くことにする. そのようなデータを用いて日本にいる12歳の男子と女子の平均身長の差がどれだけあるか(もしくはないか)を推定したい.

(2) 薬Xの効き目を調べるために, 同一の母集団から $m+n$ 人を無作為に選んで, ランダムに選んだそのうち m 人は薬Xを与え, 残りの n 人にはプラセボ(偽薬)を与えた. そして, $m+n$ 人について治療効果を表す指標の数値を測定し, 薬Xを与えた m 人分の数値は $x_1, ldots, x_m$ でプラセボを与えた n 人分の数値は $y_1, \ldots, y_n$ であったとする. そのデータから, 薬Xを与えた場合の治療効果を表す指標の平均値と与えなかった場合の治療効果を表す指標の平均値の差がどうなっているかについて推定したい.

目標は2つの群の平均の差の信頼区間の構成である.

そのためには, 検定と信頼区間の表裏一体性より, P値を適切に定義すればよい.

1.2 平均の差の検定で使用されるP値の定義

データ: m 個の実数値 $x_1, \ldots, x_m$ と n 個の実数値 $y_1, \ldots, y_n$.

x_i 達と y_i 達の標本平均と不偏分散を以下のように書くことにする:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2,$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{m}(y_i - \bar{y})^2.$$

統計モデル: 平均 μ_x と分散 σ_x^2 を持つ確率分布 D_x のサイズ m の標本分布 D_x^m と平均 μ_y と分散 σ_y^2 を持つ確率分布 D_y のサイズ n の標本分布 D_y^n の直積分布 $D_x^m \times D_y^n$ を統計モデルとして採用する.

以下では簡単のため D_x も D_y も連続分布であると仮定し, それぞれの確率密度函数を $p_x(x), p_y(y)$ と書くことにする. このとき, 統計モデル $D_x^m \times D_y^n$ の確率密度函数は

$$p(x_1, \ldots, x_m, y_1, \ldots, y_n) = p_x(x_1) \cdots p_x(x_m) \cdot p_y(y_1) \cdots p_y(y_n)$$

になる. この確率分布に従う確率変数達(独立になる)を $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ と書き, X_i 達と Y_i 達の標本平均と不偏分散を以下のように書くことにする:

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2,$$
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

検定したい仮説: $\mu_x - \mu_y = \Delta$ (Δ は具体的な数値).

中心極限定理: モデル内確率変数としての2つの標本平均達の分布について, 中心極限定理による正規分布近似が使えると仮定する.

$\bar{X}, \bar{Y}$ の平均(期待値)と分散は以下のようになる:

$$E[\bar{X}] = \mu_x, \quad E[\bar{Y}] = \mu_y, \quad \mathrm{var}(\bar{X}) = \frac{\sigma_x^2}{m}, \quad \mathrm{var}(\bar{Y}) = \frac{\sigma_y^2}{n}.$$

さらに, $\bar{X}$ と $\bar{Y}$ が確率変数として独立であることより,

$$E[\bar{X} - \bar{Y}] = \mu_x - \mu_y, \quad \mathrm{var}(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}.$$

さらに, 中心極限定理より, 次の近似が使える:

$$\bar{X} - \bar{Y} \sim \mathrm{Normal}\left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}\right), \quad \text{approximately.}$$

すなわち,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} \sim \mathrm{Normal}(0, 1), \quad \text{approximately.}$$

大数の法則: モデル内確率変数としての不偏分散 S_x^2, S_y^2 でモデルの分散 σ_x^2, σ_y^2 がよく近似されていると仮定する.

このとき, 次の近似が使える.

$$T := \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}} \sim \mathrm{Normal}(0, 1), \quad \text{approximately.}$$

P値の定義: これを用いて, 具体的に与えられた数値 Δ に関する仮説「$\mu_x - \mu_y = \Delta$」のP値を以下のように定義する. まず, データの数値の t 値 $t = t(\Delta)$ を次のように定義する:

$$t = t(\Delta) = \frac{(\bar{x} - \bar{y}) - \Delta}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}.$$

仮説「$\mu_x - \mu_y = \Delta$」のP値を, その仮説下のモデル内の確率変数としての t 値 T の値の絶対値がデータから計算した t 値 $t = t(\Delta)$ の絶対値以上になる確率の近似値として定義する:

$$\mathrm{pvalue}_{\mathrm{Normal}}(\bar{x}, \bar{y}, s_x^2, s_y^2 | m, n, \mu_x - \mu_y = \Delta) = 2(1 - \mathrm{cdf}(\mathrm{Normal}(0, 1), |t(\Delta)|)).$$

しかし, 実際に使用されるのは次に定義する t 分布を使って定義されたP値の方である.

t分布を使って補正されたP値の定義: 上のP値の t 分布版を定義しよう.

天下り的になってしまうが, 自由度 ν を次のように定義する(これの導出は別の節で行う):

$$\nu = \frac{\left(\frac{s_x^2}{m} + \frac{s_y^2}{n}\right)^2}{\frac{s_x^4}{m^2(n-1)} + \frac{s_y^4}{n^2(m-1)}}.$$

仮に $s_x^2 = s_y^2$, $m = n$ だとすると, $\nu = 2n - 2$ となる.

一般にこの ν は整数にならないがそのまま用いる.

この ν を用いて t 分布を使って計算されるP値を次のように定める:

$$\mathrm{pvalue}_{\mathrm{Welch}}(\bar{x}, \bar{y}, s_x^2, s_y^2 | m, n, \mu_x - \mu_y = \Delta) = 2(1 - \mathrm{cdf}(\mathrm{TDist}(\nu), |t(\Delta)|)).$$

以下ではこれを使うことにする. このP値は **Welchのt検定** と呼ばれる検定のP値である.

ν の定義を覚える必要はない. m, n が大きいならば, ν も大きくなり, 自由度 ν の t 分布 $\mathrm{TDist}(\nu)$ は標準正規分布 $\mathrm{Normal}(0, 1)$ でよく近似されるようになるので, ν の値がどうであるかを実質的に気にする必要がなくなることにも注意せよ. この t 分布による補正が有効なのは特別な場合に限るが, 有効でない場合も害はないのでこちらの方を使うことにする.

信頼区間: Welchの t 検定のP値から定まる信頼区間は以下のように書ける.

まず, 自由度 ν の t 分布において $t_{\nu,\alpha/2}$ 以上になる確率は $\alpha/2$ になると仮定する:

$$t_{\nu,\alpha/2} = \mathrm{quantile}(\mathrm{TDist}(\nu), 1 - \alpha/2).$$

このとき, 平均の差 $\mu_x - \mu_y$ の信頼度 $1 - \alpha$ の信頼区間が次のように定義される:

$$\mathrm{confint}_{\mathrm{Welch}}(\bar{x}, \bar{y}, s_x^2, s_y^2 | m, n, \alpha) = \left[\bar{X} - \bar{Y} - t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}},\ \bar{X} - \bar{Y} + t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}\right].$$

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

```
1 https://www.e-stat.go.jp/dbview?sid=0003224177
```