

誤用を避けるための注意

- ・ 黒木玄
- ・ 2022-07-01～2022-07-11, 2022-08-24, 2023-07-12, 2024-01-06, 2024-01-13, 2024-07-03, 2025-05-19, 2025-07-02

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

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

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

このノートに書いてある式を文字通りにそのまま読んで正しいと思ってしまうとひどい目に会う可能性が高い。しかし、数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/2022/13%20How%20to%20avoid%20misuse.ipynb)

(<https://colab.research.google.com/github/genkuroki/Statistics/blob/master/2022/13%20How%20to%20avoid%20misuse.ipynb>).

目次

- ▼ [1 データ全体の様子を適切なグラフ作画によって確認せよ！](#)
 - [1.1 Anscombeの例](#)
 - [1.2 問題: Anscombeの例](#)
 - [1.3 データサウルス](#)
 - [1.4 問題: データサウルス](#)
 - [1.5 箱ひげ図は十分に大きなサイズの単峰型のデータにのみ使用すること](#)
 - [1.6 不適切な箱ひげ図の例](#)
 - [1.7 箱ひげ図などの正しい使い方](#)
 - [1.8 ゴリラ](#)
 - [1.9 問題: ゴリラ](#)
- ▼ [2 印象操作のためにグラフを利用しない](#)
 - [2.1 3次元円グラフは絶対に避ける](#)
- ▼ [3 無作為抽出の失敗に注意を払う](#)
 - [3.1 1936年のアメリカ大統領選挙の結果の予測の失敗](#)
 - [3.2 1948年のアメリカ大統領選挙の結果の予測の失敗](#)
 - [3.3 選挙の予測に関する参考文献](#)
- ▼ [4 2つのグループを比較するときの注意](#)
 - [4.1 データは偏っていることが多い](#)
 - [4.2 2つのグループを比較するときの注意に関する参考資料](#)
- ▼ [5 P値や信頼区間の誤用](#)
 - [5.1 P値の定義](#)
 - [5.2 P値の正しい使い方](#)
 - [5.3 復習: 累積分布函数 cdf と分位点函数 quantile の定義](#)
 - [5.4 信頼区間の定義](#)
 - [5.5 信頼区間の正しい解釈の仕方](#)
 - [5.6 P値の誤用の例](#)
 - [5.6.1 P値は仮説が正しい確率ではない](#)
 - [5.6.2 単独のP値のみを使って重要な決定をしてはいけない](#)
 - [5.6.3 扱う仮説達の中で正しいものの割合も重要](#)
 - [5.6.4 実際の検出力が低くなつたせいで再現率が低くなる場合がある](#)
 - [5.6.5 すべてを正直に報告しなければいけない](#)
 - [5.7 P値や信頼区間の誤用に関する参考資料](#)

In [1]:

```
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（高速化のために用意）"""
6 _packages_added = [info.name for (uuid, info) in Pkg.dependencies() if info.is_direct_dep]
7
8 """_packages_added内にないパッケージをPkg.addする"""
9 add_pkg_if_not_added_yet(pkg) = if !(pkg in _packages_added)
10    println(stderr, "# $(pkg).jl is not added yet, so let's add it.")
11    Pkg.add(pkg)
12 end
13
14 """expr::Exprからusing内の`.`を含まないモジュール名を抽出"""
15 function find_using_pkgs(expr::Expr)
16    pkgs = String[]
17    function traverse(expr::Expr)
18        if expr.head == :using
19            for arg in expr.args
20                if arg.head == .. && length(arg.args) == 1
21                    push!(pkgs, string(arg.args[1]))
22                elseif arg.head == :( :) && length(arg.args[1].args) == 1
23                    push!(pkgs, string(arg.args[1].args[1]))
24                end
25            end
26        else
27            for arg in expr.args arg isa Expr && traverse(arg) end
28        end
29    end
30    traverse(expr)
31    pkgs
32 end
33
34 """必要そうなPkg.addを追加するマクロ"""
35 macro autoadd(expr)
36     pkgs = find_using_pkgs(expr)
37     :(add_pkg_if_not_added_yet.($pkgs); $expr)
38 end
39
40 isdir("images") || mkdir("images")
41 ENV["LINES"], ENV["COLUMNS"] = 100, 100
42 using Base.Threads
43 using LinearAlgebra
44 using Printf
45 using Random
46 Random.seed!(4649373)
47
48 @autoadd begin
49 #using BenchmarkTools
50 using CSV
51 using DataFrames
52 using Distributions
53 using HTTP
54 #using Memoization
55 using QuadGK
56 #using RCall
57 using RDatasets
58 #using Roots
59 #using SpecialFunctions
60 #using StaticArrays
61 #using StatsBase
62 #using StatsFuns
63 using StatsPlots
64 default(fmt = :png, size = (400, 250),
65          titlefontsize = 10, plot_titlefontsize = 12)
66 #using SymPy
67 end
```

```
In [2]: 1 #=
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(latex(x), mode="inline", fold_short_frac=false)
6         out = replace(out, r"\frac{" ⇒ "\dfrac{")
7         print(io, string(out))
8     end
9 end
10 =#
```

```
In [3]: 1 safemul(x, y) = x == 0 ? x : isinf(x) ? typeof(x)(Inf) : x*y
2 safediv(x, y) = x == 0 ? x : isinf(y) ? zero(y) : 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}) =
17     standardized_moment(dist, 3)
18 mykurtosis(dist::MixtureModel{Univariate, Continuous}) =
19     standardized_moment(dist, 4) - 3
```

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

```
In [4]: 1 function plot_ols(x, y;
2             ms=4, lw=1.5, alpha=0.8, xguide="", yguide="",
3             xlim=nothing, ylim=nothing, kwargs...)
4     X = x .^ (0:1)'
5     ā, ī = X \ y
6
7     if isnothing(xlim)
8         a, b = extrema(x)
9         xlim = (a - 0.1(b-a), b + 0.1(b-a))
10    end
11    if isnothing(ylim)
12        c, d = extrema(y)
13        ylim = (c - 0.1(d-c), d + 0.1(d-c))
14    end
15
16    scatter(x, y; label="", msw=0, ms, alpha)
17    plot!(x → ā + ī*x, xlim..., label="", lw)
18    plot!(; xguide, yguide, xlim, ylim)
19    plot!(; kwargs...)
20 end
```

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

```
In [5]: 1 #using RDatasets
2 anscombe = dataset("datasets", "anscombe")
3 xlim, ylim = (3, 20), (2, 14)
4 xguide, yguide = "x", "y"
5 PP = []
6 for i in 1:4
7     x, y = anscombe[!, "X$i"], anscombe[!, "Y$i"]
8     P = plot_ols(x, y; xguide, yguide, xlim, ylim, title="Anscombe $i")
9     push!(PP, P)
10 end
11 Anscombe_quartet = plot(PP...; size=(500, 450));
```

In [6]:

```
1 # R の側にdatasaurRusをインストールしておき、using RCallしておくこと
2 #datasaurus = rcopy(R"datasaurRus::datasaurus_dozen")
3
4 # インターネット上からダウンロード
5 ## See https://www.openintro.org/data/index.php?data=datasaurus
6 url_datasaurus = "https://www.openintro.org/data/csv/datasaurus.csv"
7 ## See https://github.com/rfordatascience/tidytuesday/tree/main/data/2020/2020-10-13
8 #url_datasaurus = "https://github.com/rfordatascience/tidytuesday/raw/refs/heads/main/data/
9 datasaurus = CSV.read(IOBuffer(HTTP.get(url_datasaurus).body), DataFrame)
10
11 datanames = unique(datasaurus.dataset)
12 xlim = ylim = (-5, 105)
13 xguide = yguide = ""
14 PP = []
15 for name in datanames
16     x = datasaurus.x[datasaurus.dataset .== name]
17     y = datasaurus.y[datasaurus.dataset .== name]
18     P = plot_ols(x, y; xlim, ylim, title=name, ms=3)
19     push!(PP, P)
20 end
21 layout = @layout[ _ a _ ; b c d ; e f g ; h i j ; k l m ]
22 Datasaurus_dozen = plot(PP...; layout, size=(800, 1000));
```


In [7]:

```

1  function pvalue_clopper_pearson(dist::DiscreteUnivariateDistribution, x)
2      min(1, 2cdf(dist, x), 2ccdf(dist, x-1))
3  end
4  pvalue_clopper_pearson(n, k, p) = pvalue_clopper_pearson(Binomial(n, p), k)
5
6  _pdf_le(x, (dist, y)) = pdf(dist, x) ≈ y
7
8  function _search_boundary(f, x0, Δx, param)
9      x = x0
10     if f(x, param)
11         while f(x - Δx, param) x -= Δx end
12     else
13         x += Δx
14         while !f(x, param) x += Δx end
15     end
16     x
17 end
18
19 function pvalue_sterne(dist::DiscreteUnivariateDistribution, x)
20     Px = pdf(dist, x)
21     Px == 0 && return Px
22     Px == 1 && return Px
23     m = mode(dist)
24     Px ≈ pdf(dist, m) && return one(Px)
25     if x < m
26         y = _search_boundary(_pdf_le, 2m - x, 1, (dist, Px))
27         cdf(dist, x) + ccdf(dist, y-1)
28     else # x > m
29         y = _search_boundary(_pdf_le, 2m - x, -1, (dist, Px))
30         cdf(dist, y) + ccdf(dist, x-1)
31     end
32 end
33
34 function pvalue_sterne(n, k, p)
35     pvalue_sterne(Binomial(n, p), k)
36 end
37
38 function pvalue_wilson(n, k, p)
39     p̂ = k/n
40     SE = √(p*(1-p)/n)
41     2ccdf(Normal(), abs(p̂ - p)/SE)
42 end
43
44 function pvalue_wald(n, k, p)
45     p̂ = k/n
46     SEhat = √(p̂*(1-p̂)/n)
47     2ccdf(Normal(), abs(p̂ - p)/SEhat)
48 end
49
50 function illustrate_binomial_pvalues()
51     n = 20
52     p = 0.25
53     k = 9
54
55     p̂ = k/n
56     σ² = n*p̂*(1-p̂)
57
58     null = Binomial(n, p)
59     μ, σ² = mean(null), var(null)
60     nullname = distname(null)
61
62     f(x) = mypdf(null, x)
63     xlim = (-2, n+1)
64     xlimk = (k-0.5, n+0.5)
65
66     pval = @sprintf "%.03f" pvalue_clopper_pearson(n, k, p)
67     cval = @sprintf "%.03f" ccdf(null, k-1)
68     P1 = plot(title="(1) Clopper-Pearson")
69     plot!(f, xlim...; label="", c=1)
70     plot!(f, xlimk...; label="", fillrange=0, c=1, fc=:red, fa=0.5)
71     vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
72     annotate!(k+5, 4f(k), text("P-value = $pval", 10, :red))
73     plot!([k+7, k+7], [2.4f(k), 3.6f(k)]; arrow=true, c=:red, label="")
74     annotate!(k+6, 3f(k), text("2x", 10, :red, :right))
75     annotate!(k+5, 2f(k), text("one-tailed = $cval", 10, :red))
76     plot!([k+7, k+1], [1.5f(k), 1.5f(k+1)]; arrow=true, c=:red, label="")
77     plot!(; xtick=0:n, tickfontsize=7)
78
79     pval = @sprintf "%.03f" pvalue_sterne(n, k, p)
80     j = _search_boundary(_pdf_le, 2mode(null)-k, -1, (null, pdf(null, k)))

```

```

81    xlimj = (-0.5, j+0.5)
82    P2 = plot(title="(2) Sterne")
83    plot!(f, xlim...; label="", c=1)
84    plot!(f, xlimk...; label="", fillrange=0, c=1, fc=:red, fa=0.5)
85    plot!(f, xlimj...; label="", fillrange=0, c=1, fc=:red, fa=0.5)
86    vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
87    plot!([k-0.5, j+0.5], [f(k), f(k)]; label="", c=2, ls=:dash)
88    annotate!(k+5, 3f(k), text("P-value = $pval", 10, :red))
89    plot!([k+5, k+1], [2.5f(k), 0.8f(k)]; arrow=true, c=:red, label="")
90    plot!([k+5, j+1], [2.5f(k), 0.8f(j)]; arrow=true, c=:red, label="")
91    plot!(; xtick=0:n, tickfontsize=7)
92
93    pval = @sprintf "%.03f" pvalue_wilson(n, k, p)
94    normal = Normal(μ, √σ²)
95    g(x) = pdf(normal, x)
96    P3 = plot(title="(3) Wilson")
97    plot!(f, xlim...; label="", c=1)
98    vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
99    plot!(g, xlim...; ls=:dash, label="Normal(μ=$μ, σ=√($σ²))", c=2)
100   plot!(g, k, n; label="", fillrange=0, c=2, fc=:red, fa=0.5)
101   plot!(g, -2, 2μ-k; label="", fillrange=0, c=2, fc=:red, fa=0.5)
102   annotate!(k+5, 3f(k), text("P-value = $pval", 10, :red))
103   plot!([k+5, k+0.5], [2.5f(k), 0.8g(k)]; arrow=true, c=:red, label="")
104   plot!([k+5, j+0.2], [2.5f(k), 0.7g(j)]; arrow=true, c=:red, label="")
105   plot!(; xtick=0:n, tickfontsize=7)
106
107   pval = @sprintf "%.03f" pvalue_wald(n, k, p)
108   normal = Normal(μ, √σ²)
109   h(x) = pdf(normal, x)
110   P4 = plot(title="(4) Wald")
111   plot!(f, xlim...; label="", c=1)
112   vline!([μ]; label="μ=$μ", c=1, lw=0.5, ls=:dash)
113   plot!(h, xlim...; ls=:dash, label="Normal(μ=$μ, σ=√($σ²))", c=2)
114   plot!(h, k, n; label="", fillrange=0, c=2, fc=:red, fa=0.5)
115   plot!(h, -2, 2μ-k; label="", fillrange=0, c=2, fc=:red, fa=0.5)
116   annotate!(k+5, 3f(k), text("P-value = $pval", 10, :red))
117   plot!([k+5, k+0.5], [2.5f(k), 0.8g(k)]; arrow=true, c=:red, label="")
118   plot!([k+5, j+0.2], [2.5f(k), 0.7g(j)]; arrow=true, c=:red, label="")
119   plot!(; xtick=0:n, tickfontsize=7)
120
121   plot(P1, P2, P3, P4; size=(800, 520), layout=(2, 2))
122   plot!(; plot_title="model: $nullname, data: n=$n, k=$k")
123 end
124
125 #illustrate_binomial_pvalues()

```

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

```

In [8]: 
1 oddsratiohat(a, b, c, d) = safediv(a*d, b*c)
2 stderr_logoddsratiohat(a, b, c, d) = √(1/a + 1/b + 1/c + 1/d)
3
4 function pvalue_or_wald(a, b, c, d; ω=1)
5     logORhat = log(oddsratiohat(a, b, c, d))
6     SEhat_logORhat = stderr_logoddsratiohat(a, b, c, d)
7     2ccdf(Normal(0, 1), safediv(abs(logORhat - log(ω)), SEhat_logORhat))
8 end
9
10 function confint_or_wald(a, b, c, d; α=0.05)
11     z = quantile(Normal(), 1-α/2)
12     ORhat = oddsratiohat(a, b, c, d)
13     SEhat_logORhat = stderr_logoddsratiohat(a, b, c, d)
14     [safemul(exp(-z*SEhat_logORhat), ORhat), safemul(exp(z*SEhat_logORhat), ORhat)]
15 end

```

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

```
In [9]: 1 riskratiohat(a, b, c, d) = safediv(a*(c+d), (a+b)*c)
2 stderr_logriskratiohat(a, b, c, d) = sqrt(1/a - 1/(a+b) + 1/c - 1/(c+d))
3
4 function pvalue_rr_wald(a, b, c, d; ρ=1)
5   (a+b==0 || c+d==0) && return 1.0
6   logRRhat = log(riskratiohat(a, b, c, d))
7   SEhat_logRRhat = stderr_logriskratiohat(a, b, c, d)
8   2ccdf(Normal(0, 1), safediv(abs(logRRhat - log(ρ)), SEhat_logRRhat))
9 end
10
11 function confint_rr_wald(a, b, c, d; α=0.05)
12   (a+b==0 || c+d==0) && return [0, Inf]
13   z = quantile(Normal(), 1-α/2)
14   RRhat = riskratiohat(a, b, c, d)
15   SEhat_logRRhat = stderr_logriskratiohat(a, b, c, d)
16   [safemul(exp(-z*SEhat_logRRhat), RRhat), safemul(exp(z*SEhat_logRRhat), RRhat)]
17 end
```

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

```
In [10]: 1 function logtick(; xlim=(0.03, 30))
2   xmin, xmax = xlim
3   a = floor(Int, log10(xmin))
4   b = ceil(Int, log10(xmax))
5   nums = [1, 2, 3, 4, 5, 6, 7, 8, 9]
6   mask = Bool[1, 1, 0, 0, 1, 0, 0, 0, 0]
7
8   logtick = foldl(vcat, ([10.0^k*x for x in nums if xmin ≤ 10.0^k*x ≤ xmax] for k in a:b)
9   logticklabel_a = foldl(vcat,
10     ([mask[i] ? string(round(10.0^k*x; digits=-k)) : ""
11       for (i, x) in enumerate(nums) if xmin ≤ 10.0^k*x ≤ xmax]
12       for k in a:-1)])
13   logticklabel_b = foldl(vcat,
14     ([mask[i] ? string(10^k*x) : ""
15       for (i, x) in enumerate(nums) if xmin ≤ 10.0^k*x ≤ xmax]
16       for k in 0:b)))
17   logticklabel = vcat(logticklabel_a, logticklabel_b)
18   (logtick, logticklabel)
19 end
20
21 #logtick()
```

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

In [11]:

```
1 risk_ratio_example1 = plot()
2
3 xlim = (0.1, 1.5)
4
5 k=250
6 a, b, c, d = 10000-k, 10000+k, 10000, 10000
7 RRhat = riskratiohat(a, b, c, d)
8 pval = pvalue_rr_wald(a, b, c, d)
9 ci = confint_rr_wald(a, b, c, d)
10 label = """
11 ----- A -----
12 Data: $a $b
13      $c $d
14 RRhat: $(round(RRhat; digits=3))
15 95% CI: $(round.(ci; digits=3))
16 RR=1 P-val.: $(round(100pval; digits=2))%
17 """
18 plot!(ρ → pvalue_rr_wald(a, b, c, d; ρ), xlim...; label)
19
20 a, b, c, d = 4, 16, 10, 10
21 RRhat = riskratiohat(a, b, c, d)
22 pval = pvalue_rr_wald(a, b, c, d)
23 ci = confint_rr_wald(a, b, c, d)
24 label = """
25 ----- B -----
26 Data: $a $b
27      $c $d
28 RRhat: $(round(RRhat; digits=3))
29 95% CI: $(round.(ci; digits=3))
30 RR=1 P-val.: $(round(100pval; digits=2))%
31 """
32 plot!(ρ → pvalue_rr_wald(a, b, c, d; ρ), xlim...; label, ls=:dash)
33
34 plot!(xscale=:log10, xtick=logtick(; xlim), ytick=0:0.1:1)
35 #plot!(y tick=0:0.1:1)
36 vline!([1]; label="", c=:red, lw=0.8)
37 plot!(xguide="RR (risk ratio)", yguide="P-value")
38 plot!(size=(800, 300), legend=:outertopright)
39 plot!(leftmargin=4Plots.mm, bottommargin=4Plots.mm)
40
41 risk_ratio_example1;
```

In [12]:

```
1 risk_ratio_example2 = plot()
2
3 xlim = (0.8, 3.5)
4
5 k=250
6 a, b, c, d = 10000+k, 10000-k, 10000, 10000
7 RRhat = riskratiohat(a, b, c, d)
8 pval = pvalue_rr_wald(a, b, c, d)
9 ci = confint_rr_wald(a, b, c, d)
10 label = """
11 ----- A -----
12 Data: $a $b
13      $c $d
14 RRhat: $(round(RRhat; digits=3))
15 95% CI: $(round.(ci; digits=3))
16 RR=1 P-val.: $(round(100pval; digits=2))%
17 """
18 plot!(ρ → pvalue_rr_wald(a, b, c, d; ρ), xlim...; label)
19
20 a, b, c, d = 16, 4, 10, 10
21 RRhat = riskratiohat(a, b, c, d)
22 pval = pvalue_rr_wald(a, b, c, d)
23 ci = confint_rr_wald(a, b, c, d)
24 label = """
25 ----- B -----
26 Data: $a $b
27      $c $d
28 RRhat: $(round(RRhat; digits=3))
29 95% CI: $(round.(ci; digits=3))
30 RR=1 P-val.: $(round(100pval; digits=2))%
31 """
32 plot!(ρ → pvalue_rr_wald(a, b, c, d; ρ), xlim...; label, ls=:dash)
33
34 plot!(xscale=:log10, xtick=logtick(; xlim), ytick=0:0.1:1)
35 #plot!(ytick=0:0.1:1)
36 vline!([1]; label="", c=:red, lw=0.8)
37 plot!(xguide="RR (risk ratio)", yguide="P-value")
38 plot!(size=(800, 300), legend=:outertopright)
39 plot!(leftmargin=4Plots.mm, bottommargin=4Plots.mm)
40
41 risk_ratio_example2;
```

1 データ全体の様子を適切なグラフ作画によって確認せよ！

1.1 Anscombeの例

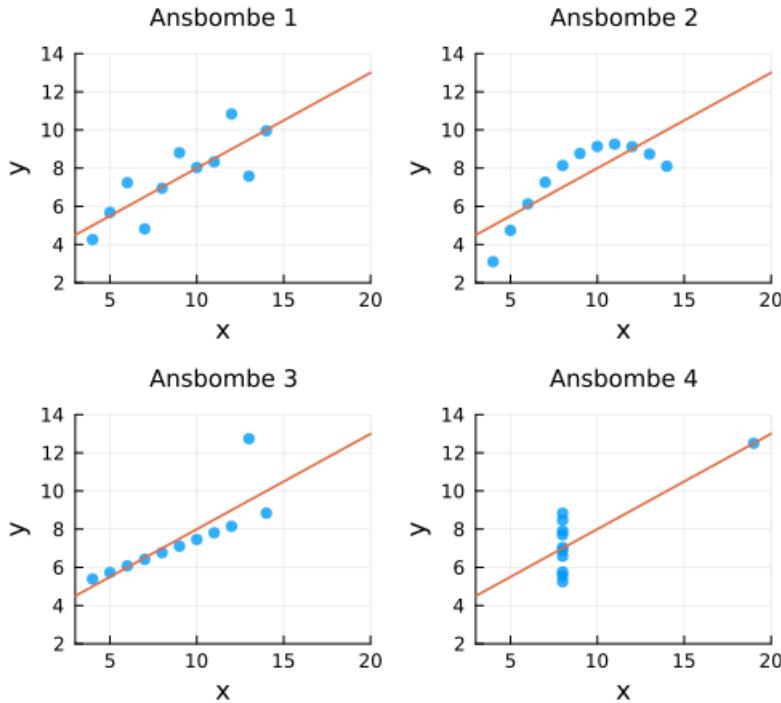
Anscombeの例 (https://en.wikipedia.org/wiki/Anscombe%27s_quartet)については

- ・「標本分布について」のノート
(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/04%20Distribution%20of%20samples.ipynb>)
- ・Anscombeの例 - Wikipedia
(<https://ja.wikipedia.org/wiki/%E3%82%A2%E3%83%B3%E3%82%B9%E3%82%B3%E3%83%A0%E3%81%AE%E4%BE%8B>)

の必修問題を参照せよ。

In [13]: 1 Anscombe_quartet

Out[13]:



4つのデータの標本平均 \bar{x} , \bar{y} と不偏共分散 s_x^2 , s_y^2 と不偏共分散 s_{xy} はほぼ一致し、その結果、回帰直線 $y = \hat{\alpha} + \hat{\beta}x$ もほぼ一致する。

このようなデータの場合にコンピュータで回帰直線だけを計算してしまうと、不適切な回帰分析をしてしまう可能性がある。

最小二乗法によって回帰直線を求めることが適切なのはAnscombe 1の形のデータだけである。

現実世界で得られるリアルなデータでも、Anscombe 2, 3, 4 のような様子をしている場合がある。

Anscombe 2は直線で近似するには適さない場合である。

何らかの事故によって一度だけ測定ミスが起こるとAnscombe 3のようなデータができあがる。

特別に大きな x の値を持つものが1つだけ含まれているデータはAnscombe 4に似た形になる。

これらのデータは平均、分散、共分散のような要約統計量だけを見ても区別をできない。

1.2 問題: Anscombeの例

[「標本分布について」のノート](#)

(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/04%20Distribution%20of%20samples.ipynb>) にあるAnscombeの例に関する必修問題をまだ解いていない人は解いてみよ。

1.3 データサウルス

データサウルスの例については

- 「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%C2%20likelihood%C2%20inference.ipynb>)

の最後の方の問題で紹介した。データサウルスについては

- <https://www.research.autodesk.com/publications/same-stats-different-graphs/>
(<https://www.research.autodesk.com/publications/same-stats-different-graphs/>)

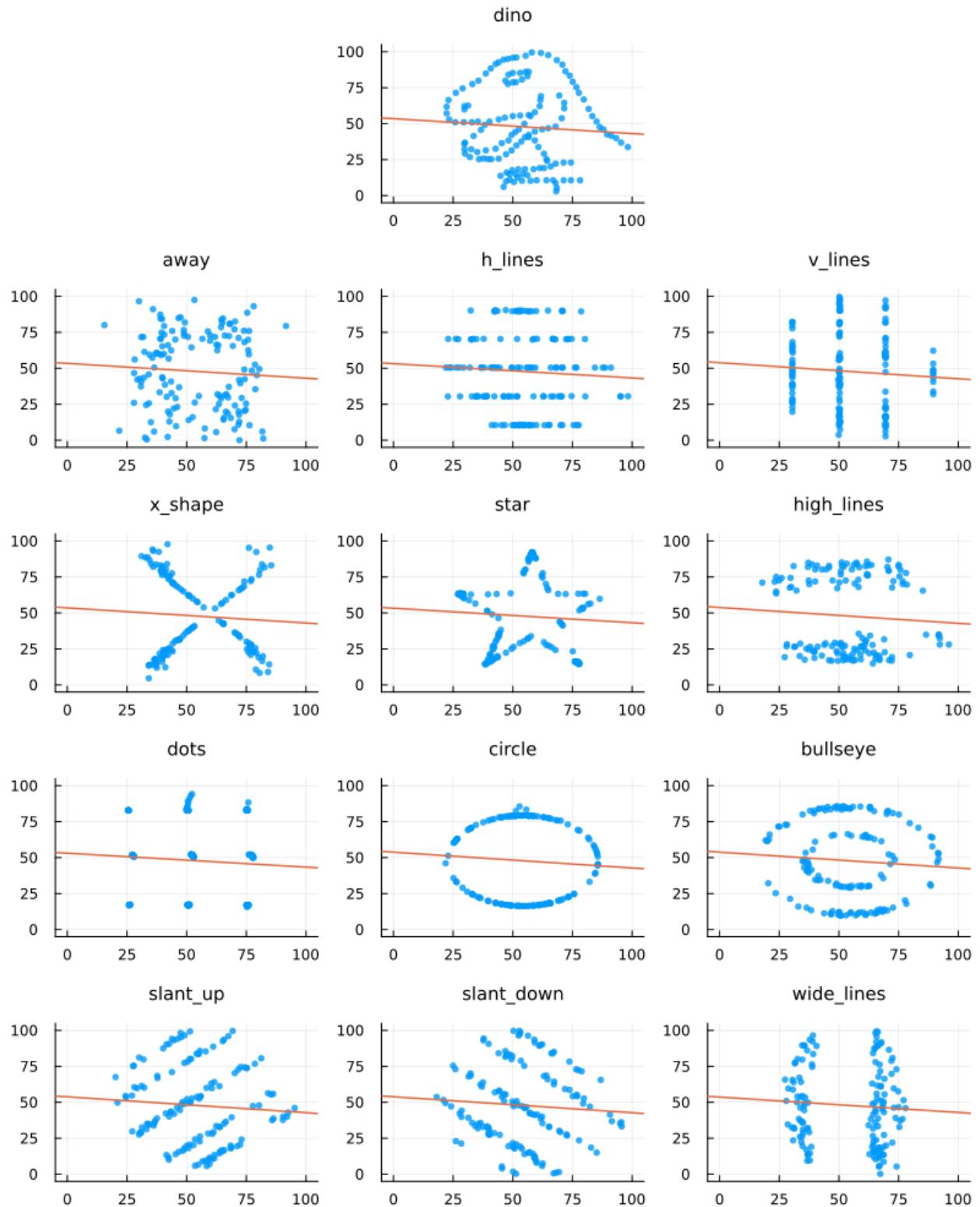
に英語による詳しい説明があり、

- <https://visualizing.jp/the-datasaurus-dozen/> (<https://visualizing.jp/the-datasaurus-dozen/>)

に日本語による簡単な解説がある。

In [14]: 1 Datasaurus_dozen

Out[14]:



以上の13個のデータの標本平均 \bar{x}, \bar{y} と不偏分散 s_x^2, s_y^2 と不偏共分散 s_{xy} はほぼ一致し、その結果、回帰直線 $y = \hat{\alpha} + \hat{\beta}x$ もほぼ一致する。

以上のデータの中に最小二乗法によって回帰直線を求めることが適切な場合はない。

左下の "slant_up" のデータでは、斜めに並んでいる5つのグループごとに傾きは正になっているが、回帰直線の傾きは負になっている。

このような場合が生じることを俗に **Simpson's paradox** と言うことがある。

私が読んだ範囲内での **Simpsonのパラドックス** の「正しい理解の仕方」に関する最良の解説は次のブログ記事である：

- KRSK, [データ分析の不思議、シンプソンのパラドックスを統計的因果推論から考える](https://www.krsk-phs.com/entry/simpsonparadox) (<https://www.krsk-phs.com/entry/simpsonparadox>), 2018-08-31

1.4 問題: データサウルス

「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%20estimation.ipynb>)

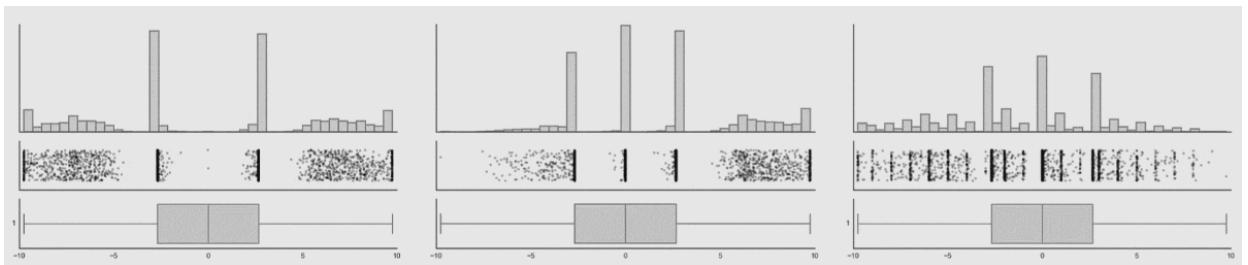
の最後の方にあるDataSaurusの例に関する問題をまだ解いていない人は解いてみよ。

1.5 箱ひげ図は十分に大きなサイズの単峰型のデータにのみ使用すること

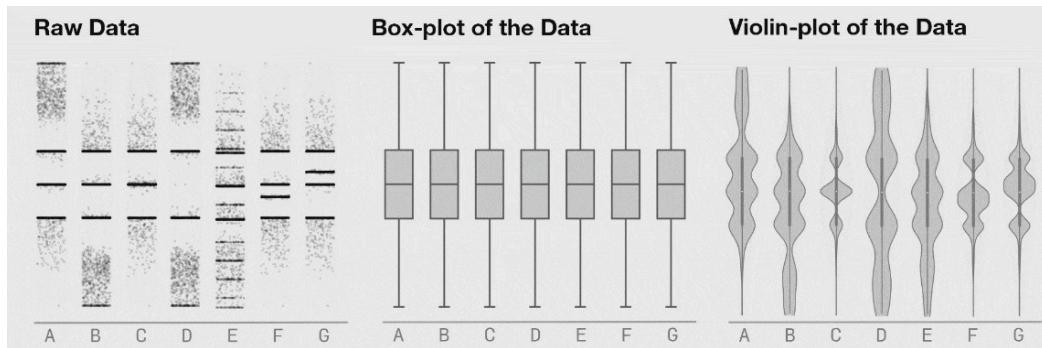
以下の図は既出のデータサウルスの解説ページ

- <https://www.research.autodesk.com/publications/same-stats-different-graphs/>
(<https://www.research.autodesk.com/publications/same-stats-different-graphs/>).

の最後の方にある動画のスクリーンショットとして取得したものである。



上段がヒストグラムで中断がなまデータの様子で下段が箱ひげ図(box plot)である。



左側がなまデータ(raw data)の様子であり、真ん中と右側はそれぞれそれに対応する箱ひげ図(box plot)とバイオリンプロット(violin plot)である。

以上の図から、なまデータの様子を見ておくことが必須であることがわかる。

どの場合にも、なまデータの様子が全然違うのに、箱ひげ図の様子はどれもほぼ同じになってしまっている。

箱ひげ図は十分な大きさのサイズを持つ单峰型のデータの場合にのみ適切なグラフの描き方である。

さらに、箱ひげ図に限らず、グラフのもとになったデータのサイズを必ず示しておくべきである。

「单峰型」とは「山が1つの」という意味である。

右側のバイオリン・プロットは箱ひげ図よりもずっとましたが、なまデータの様子を誤解なく伝えるだけの力はないように見える。

ヒストグラムはなまデータの様子を伝えるためには箱ひげ図やバイオリンプロットよりも適切な方法に成り得るが、横軸を区切る幅(bin)の設定に恣意性があり、 binの恣意性によって誤解を招くグラフを作ってしまう危険性がある。

データの様子をグラフで確認する場合にも、グラフの描き方によって誤解が生じる場合が出て来てしまうことを知っておく必要がある。

1.6 不適切な箱ひげ図の例

上で紹介した例はサイズは十分に大きいが单峰型でないデータで箱ひげ図を描いた場合(非適切な場合)の例になっている。サイズが小さ過ぎるデータの箱ひげ図を描いた場合(不適切な場合)の例が

- 藤井良宜、中学校数学におけるこれからの統計教育、特集統計教育はどう変わるかII、Rimse、理数教育研究所、2018年10月20日発行、pp.2-6. <https://rimse.or.jp/report/pdf/Rimse23.pdf> (<https://rimse.or.jp/report/pdf/Rimse23.pdf>)

のpp.2-6にある。以下の図はこの文献からの引用である。

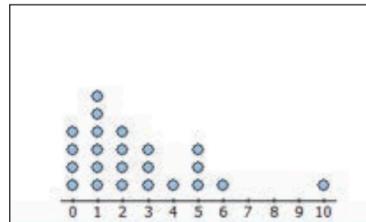


図1 あるクラスの1か月に読んだ本の冊数

これがなまデータである。データサイズが小さくて、しかも5に山がある可能性もあり、自信を持って单峰型のデータとは言い難い。しかし、上の文献ではこのデータに箱ひげ図を適用することを教えるべきだとしている。

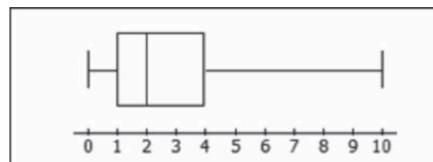


図2 図1のデータの箱ひげ図

この箱ひげ図による要約によってなまデータの様子を誤解なく伝えられるはずがない。

- この箱ひげ図には重要な情報である標本サイズが記されていない。
- この箱ひげ図を見ても、7,8,9冊の子が0人であることは分からぬ。

さらに次のような図も見つかる。

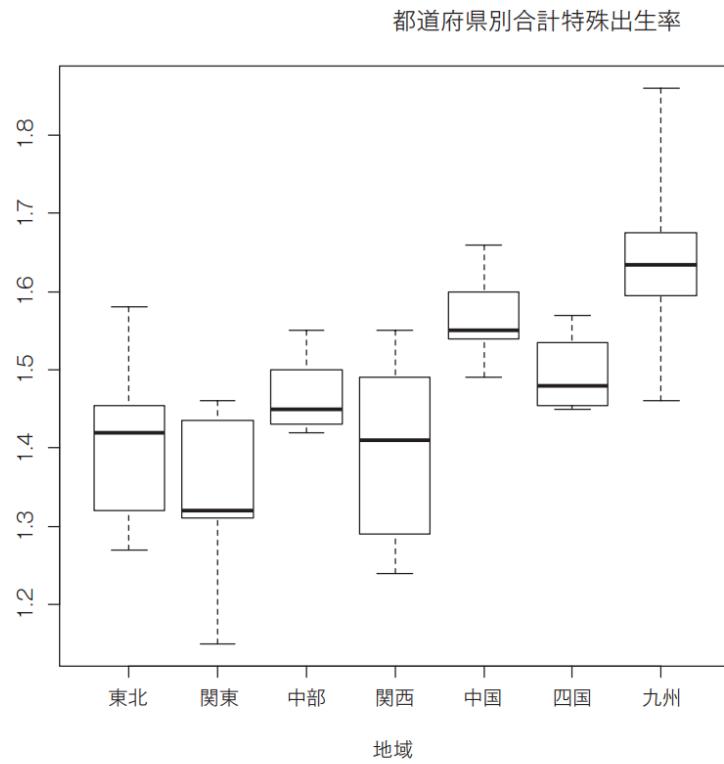


図3 地域別の合計特殊出生率

東北地方、関東地方、…の各々には数県しか含まれていない。このように、各々のグループがほんの数個の数値からなるデータをこのように箱ひげ図で「要約」することは不適切な行為である。

1.7 箱ひげ図などの正しい使い方

以下の表は

- https://twitter.com/T_Weissgerber/status/1087646465374281728
(https://twitter.com/T_Weissgerber/status/1087646465374281728)

からの引用である。

Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	<ul style="list-style-type: none"> Make all data points visible - use symmetric jittering Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points Only add error bars if the sample size is large enough to avoid creating a false sense of certainty Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	<ul style="list-style-type: none"> Make all data points visible (symmetric jittering) Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75 th , 50 th (median) and 25 th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	<ul style="list-style-type: none"> List sample size below group name on x-axis Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	<ul style="list-style-type: none"> List sample size below group name on x-axis The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	<ul style="list-style-type: none"> Do not use for continuous data

Source: @T_Weissgerber

例えば, box plot (箱ひげ図)については,

- サンプルサイズが大きい場合に使用
- 単峰型でないデータには使用しない

ということになっており, ベストプラクティス(最良実践)の項目には

1.8 ゴリラ

以下の例はこの一連のノート群の中では初めて紹介するものである。その例は論文

- Itai Yanai & Martin Lercher, [A hypothesis is a liability](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w) (<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w>), Genome Biology volume 21, Article number: 231 (2020)

で使われた擬似データセットのお話である。以下で使う図はこの論文からの引用になっている。

論文の著者たちは、学生達に以下のように伝えた:

データセットには1786人分のボディマス指数(BMI)とある特定の日にそれぞれが歩いた歩数(steps)が記録されており、男性と女性について2つのファイル分けてデータが格納されている。

学生に配布したデータセットの様子は以下の通り:

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
9		
12		
14		
15		
16		
21		
23		
26		
28		
31		
33		
34		
35		
36		
38		
39		
41		
44		
45		
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1
45	14259	20.0

(<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w/figures/1>)

そして、学生を次の2つのグループに分けた。

1つ目の「仮説あり」の(hypothesis-focused)グループの学生には、以下の3つの仮説を検討するように求めた。

- (i) 男性と女性の平均歩数に統計的に有意な差があること
- (ii) 女性では歩数とBMIの間には負の相関があること
- (iii) 男性では歩数とBMIの間には正の相関があること

さらに、データセットから結論できることが他にあるかどうかも質問した。

2つ目の「仮説なし」の(hypothesis-free)グループの学生には、単に「データセットから何を結論できますか?」と質問した。

その結果は以下の通り。

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

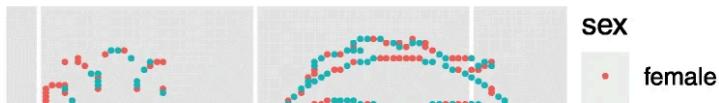
(<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w/figures/1>)

この表の日本語訳:

	ゴリラを発見できなかった	ゴリラを発見した
仮説ありグループ	14	5
仮説なしグループ	5	9

「ゴリラを発見できなかった」と「ゴリラを発見した」とは一体どういう意味なのだろうか?

その理由はデータの散布図を描くとわかる!



1.9 問題: ゴリラ

「検定と信頼区間: 比率の比較」のノート

(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/11%20Hypothesis%20testing%20and%20confidence%20interval%20Two%20proportions.ipynb>) の必修計算問題の節の解答例を参照して、 2×2 の分割表

	ゴリラを発見できず	ゴリラを発見した
仮説ありグループ	14	5
仮説なしグループ	5	9

のデータから「仮説あり」と「仮説なし」の2つのグループでゴリラを発見できない確率(比率)の違いについて以下を求めるよ:

- 「オッズ比パラメータは $OR = 1$ である」という仮説の Wald 版 P 値 P_{Wald} .
- オッズ比パラメータ OR に関する Wald 版 95% 信頼区間 CI_{Wald} .

上の問題の続き。可能ならば以下も求めよ.

- 「オッズ比パラメータは $OR = 1$ である」という仮説の Pearson の χ^2 検定版 P 値 P_{Chisq} .
- オッズ比パラメータ OR に関する Pearson の χ^2 検定版の 95% 信頼区間 CI_{Chisq} .
- 「オッズ比パラメータは $OR = 1$ である」という仮説の Sterne 版 Fisher 検定の P 値 P_{Sterne} .
- オッズ比パラメータ OR に関する Sterne 版 Fisher 検定の 95% 信頼区間 CI_{Sterne} .
- 「オッズ比パラメータは $OR = 1$ である」という仮説の Clopper-Pearson 版 Fisher 検定の P 値 P_{CP} .
- オッズ比パラメータ OR に関する Clopper-Pearson 版 Fisher 検定の 95% 信頼区間 CI_{CP} .

可能ならばそれぞれの場合の P 値函数のグラフをプロットしてみよ.

無理して 3 から 8 の小問を解いたり、P 値函数のグラフを描いたりする必要はないが、もしもそれらの答えを数値的に求めたり、グラフを描く方法をマスターできたりしたら、後で役に立つ可能性がある。

これらの問題の解答例はここには書かない。

2 印象操作のためにグラフを利用しない

統計学云々以前の問題として、印象操作のためにグラフを悪用することはやめた方がよい。

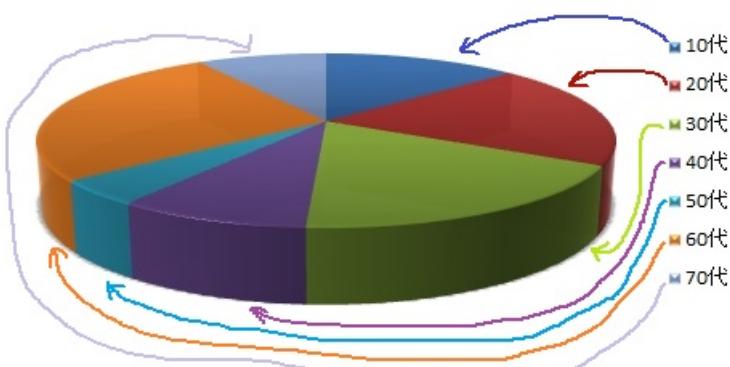
多くの場合に誤った印象を与えることになることで有名なグラフの書き方に 3 次元円グラフがある。(一般に円グラフはよくないグラフの書き方だとされている。)

2.1 3次元円グラフは絶対に避ける

3 次元グラフがダメな理由および円グラフ一般が良いグラフの書き方ではないことについては

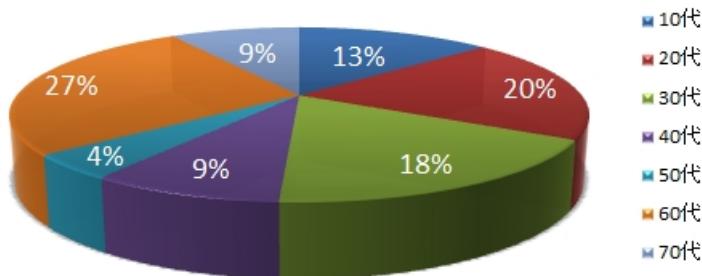
- [3D円グラフ撲滅委員会より、なぜ3D円グラフを使ってはいけないか](https://www.cresco.co.jp/blog/entry/950/) (<https://www.cresco.co.jp/blog/entry/950/>)

が見易く読み易い。以下の図はこのページからの引用である。ただし、矢線は筆者が追加した。



問題:隣り合っている20代と30代ではどちらの割合が大きいか?

見た目的には圧倒的に30代の割合の方が大きく見えるが、実際には20代が20%で30代は18%であり、30代の割合の方が少し小さい。手前の40代の方が奥側の10代や70代よりも割合が大きく見えるが、実際にはそうではない。



3 無作為抽出の失敗に注意を払う

「独立同分布」という条件を満たす確率分布を統計モデルとして採用した場合には、調べたい母集団からの適切な無作為抽出でデータを得ることが適切な統計分析のために必要になる。

しかし、実際にはそれはそう簡単ではない。

調べたい母集団よりも狭い集団からの偏ったサンプルの抽出になってしまっている危険がいつでも生じる。

そのような例として有名な事例を以下で紹介しよう。

3.1 1936年のアメリカ大統領選挙の結果の予測の失敗

1936年のアメリカ大統領選挙は民主党のルーズベルトと共和党のランドンの対決になった。

以下の得票率の予測の表は、鈴木督久『世論調査の真実』(日経BP 日本経済新聞出版本部, 2021年)の図表5-1の孫引きであり、そこには

(出所) 小山栄三(1946)『輿論調査概要』時事通信社(名称を一部省略)

と書いてある。

	ルーズベルト	ランドン
実際の得票率 (%)	60.2	39.8
リテラリー・ダイジェスト	42.6	57.4
ギャラップ	54	46
フォーチュン	74	26
クロスレー	52	48
バルチモア・サン	64	36
ファーム・ジャーナル	43	57
グラスルーツ	39.5	60.5

ここで、リテラリー・ダイジェスト、ギャラップ、…は予測を出した主体の名前(会社や雑誌の名前など)である。

例えば、リテラリー・ダイジェスト誌は、ルーズベルト候補は 42.6%，ランドン候補は 57.4% の票を得て、ランドン候補が勝利するだろうと予想した。しかし、実際にはルーズベルト候補が 60.2% の票を得て勝利することになった。

リテラリー・ダイジェスト誌は、1936年8月22日号で調査開始を宣言し、9回にわたる連載で最終的に237万6523票分のアンケート結果を集計して、上の予測を立てた。これは約1000万人を対象とする調査であった。

母集団中の6割がルーズベルト支持で残りの4割がランドン支持のとき、200万人を理想的に無作為抽出したならば、その中のルーズベルト支持者の割合が 43% 以下(86万人以下)になってしまふ確率はほぼ 0 である：

$$cdf(Binomial(2000000, 0.6), 860000) \approx 2.077 \times 10^{-50925}.$$

左辺は、ルーズベルト氏の支持者が6割いるときに、理想的に有権者全体から200万人を無作為抽出した中にルーズベルト支持者が86万人以下しか含まれていない確率を意味する。

のことから、リテラリー・ダイジェスト誌によるアンケート調査には大きな偏りがあったと考えられる。

リテラリー・ダイジェスト誌が調査対象としていた人達は富裕層が多く、偏っていたことが問題だった。

俗説によれば「リテラリー・ダイジェスト誌が調査対象としていた人達は富裕層が多く、偏っていたことが問題だった」と言われているが、これは誤りである。例えば次の論文を参照せよ：

- Dominic Lusinchi, "President" Landon and the 1936 Literary Digest Poll: Were automobile and telephone owners to blame?, Social Science History, 2012. [[Google Scholar](https://scholar.google.co.jp/scholar?cluster=5805358583625636408) (<https://scholar.google.co.jp/scholar?cluster=5805358583625636408>)]

リテラリー・ダイジェスト誌の予想が失敗した理由は無回答バイアス ([non-response bias](#) (https://en.wikipedia.org/wiki/Participation_bias)) であったと考えられている。

上の論文では1937年のギャラップ社の調査結果のデータから、リテラリー・ダイジェスト誌のアンケートが送付された人達の中で、アンケートを回答した人達と回答しなかった人達ではどちらをどれだけ支持しているかが違っていた(次の表は上の論文のTable 2より)：

	回答した (23.8%)	回答しなかった (76.2%)	全体 (100%)
ルーズベルト	42.9%	60.9%	56.6%
ランدون	57.1%	39.1%	43.4%

この表を誰を支持したかで分類し直すと次のようになる：

	ルーズベルト	ランدون
回答した	18.0%	31.3%
回答しなかった	82.0%	68.7%
全体	56.6%	43.4%

以上の表より以下が読み取れる。

- リテラリー・ダイジェスト誌のアンケートが送付された人達全体では、ルーズベルト候補の支持者の割合の方が高かった。
- しかし、ランدون候補支持者達のアンケートの回答率の方が高かったせいで、アンケートに回答した人達の中ではランدون候補が優勢になってしまった。

このようにアンケート調査結果の偏りは、アンケートを募った対象が偏っていることだけではなく、アンケートに答える傾向が強い人達が偏っていることによっても生じる。

この事例から、たとえサイズが200万を軽く超える **ビッグデータ** であっても、その内訳が偏っているならば、そのデータを使った予測は大外れしてしまうことがわかる。

```
In [15]: 1 @show 0.238 * 0.429 + 0.762 * 0.609
          2 @show 0.238 * 0.429 / (0.238 * 0.429 + 0.762 * 0.609)
          3 @show 0.762 * 0.609 / (0.238 * 0.429 + 0.762 * 0.609)
          4 @show 0.238 * 0.571 + 0.762 * 0.391
          5 @show 0.238 * 0.571 / (0.238 * 0.571 + 0.762 * 0.391)
          6 @show 0.762 * 0.391 / (0.238 * 0.571 + 0.762 * 0.391);
```

```
0.238 * 0.429 + 0.762 * 0.609 = 0.56616
(0.238 * 0.429) / (0.238 * 0.429 + 0.762 * 0.609) = 0.18034124629080117
(0.762 * 0.609) / (0.238 * 0.429 + 0.762 * 0.609) = 0.8196587537091987
0.238 * 0.571 + 0.762 * 0.391 = 0.43384
(0.238 * 0.571) / (0.238 * 0.571 + 0.762 * 0.391) = 0.31324451410658305
(0.762 * 0.391) / (0.238 * 0.571 + 0.762 * 0.391) = 0.6867554858934171
```

```
In [16]: 1 @time setprecision(50) do
          2     sum(pdf(Binomial(2000000, big"0.6")), k) for k in 0:860000
          3 end
```

```
44.532668 seconds (139.63 M allocations: 3.609 GiB, 4.94% gc time, 1.04% compilation time)
```

Out[16]: 2.0771983651745741e-50925

仮にルーズベルトへの支持率がちょうど5割で(実際には 60.2%), リテラリー・ダイジェスト誌のアンケート調査が母集団からの無作為抽出になっているならば、237万6523票分のアンケート結果中のルーズベルトの支持率が 42.6% 以下(1012399 票以下)になる確率(片側検定のP値)は

$$\text{cdf}(\text{Binomial}(2376523, 0.6), 1012399) \approx 1.057 \times 10^{-11348}.$$

で、これもほぼゼロになる。

```
In [17]: 1 round(Int, 2376523 * 0.426)
```

```
Out[17]: 1012399
```

```
In [18]: 1 @time setprecision(50) do  
2     sum(pdf(Binomial(2376523, big"0.5"), k) for k in 0:round(Int, 2376523 * 0.426))  
3 end
```

```
54.519709 seconds (164.12 M allocations: 4.237 GiB, 4.68% gc time, 0.18% compilation time)
```

```
Out[18]: 1.0568391602338336e-11348
```

3.2 1948年のアメリカ大統領選挙の結果の予測の失敗

以下の表は

- なるほど統計学園 アメリカ大統領選挙の番狂わせ（後編）標本調査における偏り2
(https://www.stat.go.jp/naruhoodo/15_episode/episode/senkyo2.html)

にあった表のフォーマットを変えたものである。そこには

候補者の当選予想と得票率のグラフ

（注）上記2人の他にも候補者がいたため、合計は100%にならない。

と書いてある。

	民主党 トルーマン	共和党 デューイ
実際の得票率 (%)	○ 49.5	45.1
ギャラップの予想	44.5	○ 49.5
ローバーの予想	37.1	○ 52.2
クロスレーの予想	44.8	○ 49.9

歴史的には1936年の大統領選挙でルーズベルトが当選することを正しく予測できたギャラップは1948年の大統領選挙では予測に失敗することになった。

ギャラップによる調査方法は無作為抽出法ではなく、割り当て法であった。詳しくは上のリンク先を参照して欲しい。

このように、アンケート調査の結果は偏っていることがあり、信頼できる予測を統計学的な方法で行うためにはデータの取得時に特別な注意が必要になる。

3.3 選挙の予測に関する参考文献

以上については以下の文献を参考にした。

- なるほど統計学園
 - アメリカ大統領選挙の番狂わせ（前編）標本調査における偏り1
(https://www.stat.go.jp/naruhoodo/15_episode/episode/senkyo1.html)
 - アメリカ大統領選挙の番狂わせ（後編）標本調査における偏り2
(https://www.stat.go.jp/naruhoodo/15_episode/episode/senkyo2.html)
- 鈴木督久、『世論調査の真実』、日経BP 日本経済新聞出版本部、2021年

歴史的には1936年の米大統領選挙の予測についてはギャラップ氏達の予測が有名である。その物語に興味がある人はこれらの文献を参照して欲しい。

しかし、上の文献には俗説が書かれているので取り扱いには注意が必要である。その点については次の論文を参照して欲しい：

- Dominic Lusinchi, "President" Landon and the 1936 Literary Digest Poll: Were automobile and telephone owners to blame?, Social Science History, 2012. [[Google Scholar](https://scholar.google.co.jp/scholar?cluster=5805358583625636408) (<https://scholar.google.co.jp/scholar?cluster=5805358583625636408>)]
- Peverill Squire, Why the 1936 Literary Digest poll failed, Public Opinion Quarterly, 1988. [[Google Scholar](https://scholar.google.co.jp/scholar?cluster=5805358583625636408) (<https://scholar.google.co.jp/scholar?cluster=5805358583625636408>)]

前者の論文には後者の分析をより精密にした結果が含まれている。

ウィキペディアでの説明も参考になる：

- [Wikipedia: The Literary Digest - Presidential poll](https://en.wikipedia.org/wiki/The_Literary_Digest#Presidential_poll) (https://en.wikipedia.org/wiki/The_Literary_Digest#Presidential_poll)

4 2つのグループを比較するときの注意

4.1 データは偏っていることが多い

我々は別のノートで「2つのグループの比率の違いは〇〇である」の型の仮説や「2つのグループの平均の違いは〇〇である」の型の仮説に関するP値や信頼区間の計算の仕方について学んだ.

- 検定と信頼区間: 比率の比較
(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/11%20Hypothesis%20testing%20and%20confidence%20interval%20Two%20proportions.ipynb>)
- 検定と信頼区間: 平均の比較
(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/12%20Hypothesis%20testing%20and%20confidence%20interval%20Two%20means.ipynb>)

例えば、風邪をひいた人に薬Xを与えることに効果があるかどうかを知るために、薬Xを与えたグループと薬を与えずに経過観察したグループに分けて取得した「5日以内に回復した人数」のデータを利用する場合には、薬Xを与えたかどうか以外の回復日数に影響を与える条件が2つのグループで揃っていることが必要である。

風邪症状が重い人達には薬Xを与える傾向が強く、症状が軽い人達は経過観察する傾向が強い場合には、薬Xに本当は効き目があったとしても、重い症状の人達に特に薬Xを与えてしまったせいで、経過観察グループとの比較で、回復にかかる日数が平均的に増えてしまうかもしれない。

統計学の現実での応用時にはこのような問題が発生することの方が普通なので注意が必要である。

例えば、〇〇ウイルス対策としてワクチンを接種したグループと接種しなかったグループを比較するために、〇〇ウイルスによって重症になった人数を調べたとしよう。

このときもしもワクチンを重症になりやすい人達に優先的に接種しているならば、ワクチンに効き目があったとしても、ワクチン接種群の方が非接種群よりも重傷者の割合が多くなってしまうこともあり得る。

このような問題は、〇〇ウイルスによって重症になりやすいか否かを完全に無視してワクチン接種するようにすれば解消されるのだが、人の命を守るために行われるべきワクチン接種ではそういうことは倫理的にするべきではない。

他にも様々な要因でデータが偏ってしまうことがある。

特別な場合を除けば、データは必ず偏っており、問題はその程度がどのくらい重大であるかということになる。

4.2 2つのグループを比較するときの注意に関する参考資料

次の動画が参考になるだろう。

- 佐藤俊哉, [ランダム化ができないとき](https://youtu.be/tUkyjZXU1vc) (<https://youtu.be/tUkyjZXU1vc>), 2018年10月25日

同じ著者による次の本も参考になる。

- 佐藤俊哉, 宇宙怪人しまりす 医療統計を学ぶ 検定の巻, 岩波科学ライブラリー 194, 2012年

データの偏りについて語るときの統計学的なキーワードの一つは **交絡** (confounding) である。「交絡」の正確な定義を知りたい人は次の講義動画の18:00あたりをみて欲しい。

- 佐藤俊哉, [交絡とその調整](https://youtu.be/ybdkQFEdCPM) (<https://youtu.be/ybdkQFEdCPM>), 2018年12月20日

5 P値や信頼区間の誤用

5.1 P値の定義

P値は以下の3つから計算される:

- データの数値 x
- 統計モデル $M(\theta)$
- パラメータ θ の値に関する仮説 $\theta = \theta_0$

P値は具体的には以下で説明されるように構成(計算)されるが、P値は

- 上の3つの間の相性の良さ(もしくは整合性もしくは適合度)

の指標の1つとして使われることになる。

データの数値 x とデータの数値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 $M(\theta)$ と具体的なパラメータ値 θ_0 が与えられたとき、仮説 $\theta = \theta_0$ のP値はその仮説下での統計モデル $M(\theta = \theta_0)$ 内でデータの数値 x 以上に極端な値が生成される確率またはその近似値として定義される。

「～以上に極端な」や「近似値」の意味を適切に与えるごとにP値の正確な定義が得られる。

例えば、データの数値が「 n 回中 k 回成功」の形をしていて、統計モデル $M(p)$ が試行回数 n 、成功確率パラメータ p の二項分布モデルのとき、 $p = p_0$ という仮説のClopper-Pearson型のP値は、モデル $M(p = p_0)$ 内でデータの数値 k 以下の値が生成される確率と k 以上の値が生成される確率の小さい方の2倍として定義される。実際には、P値の値が1を超えないように、その値と1の小さな方として、Clopper-Pearson型のP値が定義される。具体的に式で書くと以下の通り：

$$pvalue_{CP}(k|n, p = p_0) = \min \left(\frac{1}{2 \operatorname{cdf}(\operatorname{Binomial}(n, p_0), k)}, \frac{1}{2(1 - \operatorname{cdf}(\operatorname{Binomial}(n, p_0), k - 1))} \right).$$

「～以上に極端な」の大雑把な意味は「モデル内確率分布の平均値 np_0 からの距離がデータの数値 k 以上に大きい」という意味に設定しているが、モデルの分布が左右非対称になっていることを考慮してこのような定義になっている。この他にも

- ・「検定と信頼区間：比率の検定と信頼区間」のノート
(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/09%20Hypothesis%20testing%20and%20confidence%20inter%20Proportion.ipynb>)

では、Sterne型のP値と

- ・「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>)

で説明した二項分布の中心極限定理(正規分布近似)を使って定義されたP値を2種類定義した(Wilson型、Wald型)。

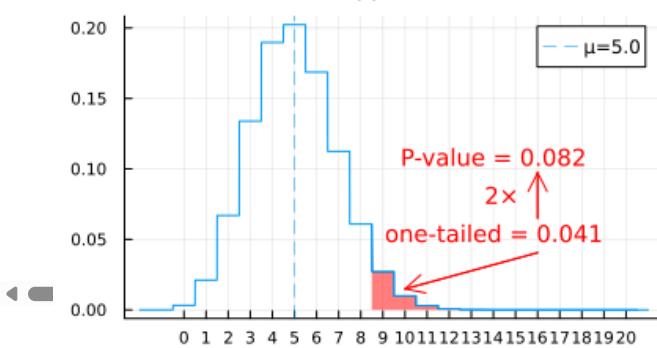
二項分布モデルの場合のそれら4種のP値の定義で使われた考え方は他の統計モデルでのP値を定義するときにも使われる。その意

In [19]: 1 illustrate_binomial_pvalues()

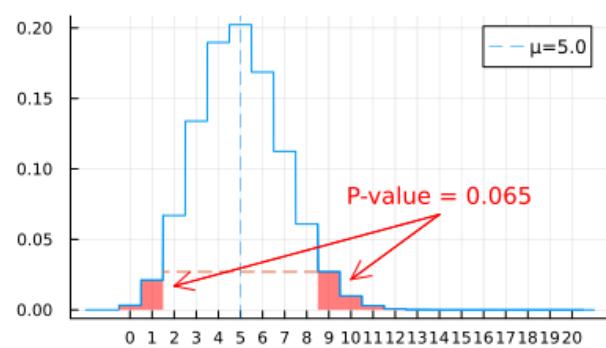
Out[19]:

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

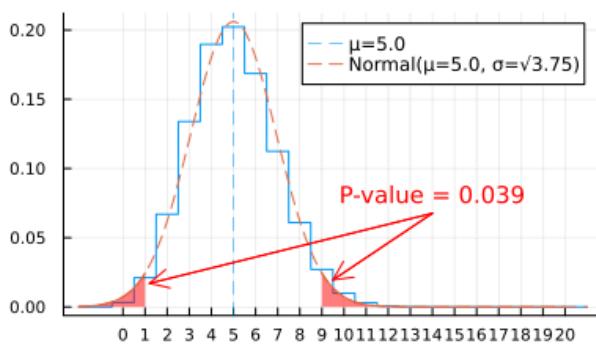
(1) Clopper-Pearson



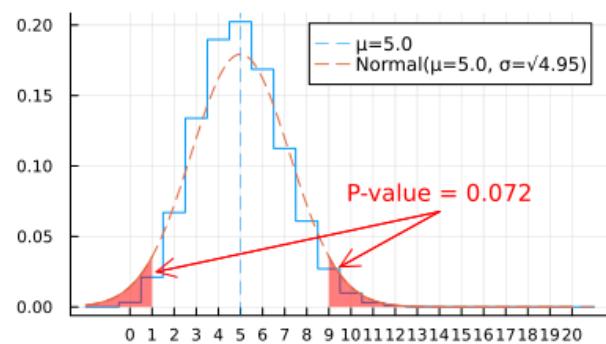
(2) Sterne



(3) Wilson



(4) Wald



5.2 P値の正しい使い方

仮説 $\theta = \theta_0$ のP値は、その仮説の下での統計モデル $M(\theta = \theta_0)$ とデータの数値 x の相性の良さ(整合性、適合度)の指標として使われる。

P値が非常に小さい場合には相性が非常に悪いと考える。

P値が0より十分に大きな場合には、モデル $M(\theta)$ の下での仮説 $\theta = \theta_0$ のデータの数値 x との相性が良いと考える。

しかし、相性の良さがある程度あることとそのパラメータ値が真実を表しているかもしれないことは別々の話になるので注意すること。

P値だけから得られる結論は弱い という点を強調したいがために、「P値は相性の良さの指標である」と述べていることにも注意せよ!

例えば、統計モデルとデータの数値のどちらかに問題があるせいで相性が良いかのように見えてしまったり、相性が悪いかのように見えてしまったりすることがあり得る。

P値を使った統計的推論は、使用した統計モデルが妥当でかつ、用いたデータの数値も信頼できる場合にのみ、真っ当な結論を出し得る。

実際には、データが偏っている可能性は常に心配する必要があるし、使用した統計モデルがそのようなデータの偏りを考慮していない不適切なモデルになっている可能性も常に心配する必要がある。モデルが妥当でなくなるパターンは多彩である。

注意: 使用している統計モデルが妥当であることの意味は、目的ごとにユーザー側が決める必要があり、さらに妥当であることの証拠の提出もユーザー側の責任になる。統計学入門の解説に書いてある統計分析の手続きに機械的に従えば適切な統計的推論が可能になるという考え方は完全に誤りである。使用した統計モデルの妥当性やデータの数値の信憑性のような重要な事柄については統計学のユーザー側が責任を持つ必要がある。

注意: P値 (P-value)はデータの数値とモデル+パラメータの値の **相性の良さの指標の1つに過ぎない**。別の考え方で相性の良さを測ることもあり得る。例えば、尤度 (likelihood)やその拡張版としてのベイズ統計における **事後分布** (posterior)の値もデータの数値とモデル(事前分布を含む)+パラメータの値の相性の良さの指標とみなされる。どの道具を使うにしても、安易に強い結論を出すことを控えて、謙虚になることが重要である。

5.3 復習: 累積分布函数 cdf と分位点函数 quantile の定義

X が確率分布 D に従う確率変数であるとき、 $X \leq x$ となる確率 $P(X \leq x)$ を x の函数とみなしたもの **累積分布函数** (cumulative distribution function, cdf) と呼び、

$$\text{cdf}(D, x) = P(X \leq x)$$

と書くことにしていたのであった。累積分布函数 $p = F(x) = \text{cdf}(D, x)$ の逆函数 $x = Q(p) = F^{-1}(p)$ (逆函数が存在しない場合には適当に代替物を定義する)を **分位点函数** (quantile function)と呼び、

$$x = \text{quantile}(D, p)$$

のように書くことにしていたのであった。よく使われる基本的な確率分布について、累積分布函数 cdf と分位点函数 quantile はコンピュータ上の各種の環境で利用可能な場合が多いので、便利に利用するとよい。

例えば、

$$\begin{aligned} \text{cdf}(\text{Binomial}(n, p), k) &= \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}, \\ \text{cdf}(\text{Beta}(a, b), k) &= \frac{\int_0^p t^{a-1} (1-t)^{b-1} dt}{B(a, b)}, \\ \text{cdf}(\text{Normal}(0, 1), z) &= \int_{-\infty}^z \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \end{aligned}$$

In [20]:

```
1 @show cdf(Binomial(100, 0.5), 40)
2 @show 1 - cdf(Beta(40+1, 100-40), 0.5)
3 @show cdf(Normal(0, 1), (40 - 100 * 0.5) / sqrt(100 * 0.5 * 0.5));
```

```
cdf(Binomial(100, 0.5), 40) = 0.02844396682049039
1 - cdf(Beta(40 + 1, 100 - 40), 0.5) = 0.02844396682049044
cdf(Normal(0, 1), (40 - 100 * 0.5) / sqrt(100 * 0.5 * 0.5)) = 0.022750131948179205
```

In [21]:

```
1 @show quantile(Beta(40, 100-40+1), 0.025)
2 @show quantile(Beta(40+1, 100-40), 0.975)
3 @show quantile(Normal(0, 1), 0.975);
```

```
quantile(Beta(40, (100 - 40) + 1), 0.025) = 0.3032947687028774
quantile(Beta(40 + 1, 100 - 40), 0.975) = 0.5027908495776652
quantile(Normal(0, 1), 0.975) = 1.9599639845400576
```

5.4 信頼区間の定義

パラメータ θ に関する信頼水準(信頼度、信頼係数) $1 - \alpha$ の信頼区間は

- データの数値 x が与えられたときの、P値が α 以上になる $\theta = \theta_0$ の値全体のなす集合

と定義される。値 θ_0 が信頼区間に含まれることは

- データの数値 x と統計モデル $M(\theta)$ とパラメータ値に関する仮説 $\theta = \theta_0$ の相性が閾値 α によって良いと判定されることを意味している。

データの数値 x とデータの数値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 $M(\theta)$ についてP値 $pvalue(x|\theta = \theta_0)$ が定義されているとき, 有意水準 $0 < \alpha < 1$ について, パラメータ θ に関する信頼度 $1 - \alpha$ の信頼区間は, $pvalue(x|\theta = \theta_0) \geq \alpha$ となるパラメータ値 θ_0 全体の集合として定義される:

$$\text{confint}(x|\alpha) = \{ \theta_0 \mid pvalue(x|\theta = \theta_0) \geq \alpha \}.$$

例えば, 二項分布モデルのClopper-Pearson型のP値の定義から得られる信頼区間はClopper-Pearsonの信頼区間と呼ばれ, 具体的には,

$$1 - \text{cdf}(\text{Binomial}(n, p_L), k - 1) = \alpha/2, \\ \text{cdf}(\text{Binomial}(n, p_U), k) = \alpha/2.$$

によって p_L, p_U を定義することによって得られる区間

$$[p_L, p_U] = \{ p_0 \mid p_L \leq p_0 \leq p_U \}$$

になる. どうやって p_L, p_U を求めるかが問題になるが, この場合には, 二項分布とベータ分布の累積分布函数のあいだの素晴らしい関係

$$1 - \text{cdf}(\text{Binomial}(n, p), k - 1) = \text{cdf}(\text{Beta}(k, n - k + 1), p), \\ \text{cdf}(\text{Binomial}(n, p), k) = 1 - \text{cdf}(\text{Beta}(k + 1, n - k), p)$$

を使えば, ベータ分布の分位点函数(quantile function, 累積分布函数 cdf の逆函数)を使って,

$$p_L = \text{quantile}(\text{Beta}(k, n - k + 1), \alpha/2), \\ p_U = \text{quantile}(\text{Beta}(k + 1, n - k), 1 - \alpha/2).$$

によって, p_L, p_U を効率的に求められるのであった.

Sterne型のP値に対応する信頼区間の計算はこのようにきれいにはまとまらない. 正規分布近似を使って定義されたP値から得られる信頼区間の計算は標準正規分布の分位点函数の $1 - \alpha/2$ での値

$$z_{\alpha/2} = \text{quantile}(\text{Normal}(0, 1), 1 - \alpha/2)$$

の計算に帰着することになる. 二項分布モデルの場合の信頼区間にに関する非常に詳しい説明が

- 「検定と信頼区間: 比率の検定と信頼区間」のノート
(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/09%20Hypothesis%20testing%20and%20confidence%20inter%20Proportion.ipynb>)

にある.

二項分布モデルの場合の信頼区間の構成法のパターンは他の統計モデルにおける信頼区間の計算でも同様に使われる. 標準正規分布が t 分布で置き換わる場合もあそ

5.5 信頼区間の正しい解釈の仕方

P値は 統計モデル内部 の確率分布における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値は現実における何らかの意味での確率ではなく, 数学的フィクションであるモデル内部における確率またはその近似値であるにすぎない.

ゆえに, 信頼度 $1 - \alpha$ の信頼区間すなわち $100(1 - \alpha)\%$ の信頼区間における $1 - \alpha = 100(1 - \alpha)\%$ は モデル内部 における確率またはその近似値になる.

すなわち, X をモデル $M(\theta = \theta_0)$ に従う確率変数とするとき, $pvalue(X|\theta = \theta_0) \geq \alpha$ となる確率は $1 - \alpha$ またはその近似値になる.

これは, 信頼度 $1 - \alpha$ の信頼区間 $\text{confint}(X|\alpha) = \{ \theta_0 \mid pvalue(X|\theta = \theta_0) \geq \alpha \}$ に θ_0 が含まれる確率が $1 - \alpha$ またはその近似値になることを意味している.

そのとき, 確率変数 X の値は確率的にランダムに決まると考えるので, その函数である信頼区間 $\text{confint}(X|\alpha)$ も確率的にランダムに決まる. すなわち区間の両端の点がランダムに決まる. そのような状況でその区間に定数 θ_0 が含まれる確率が $1 - \alpha$ またはその近似値になる. しかもその確率は数学的フィクションであるモデル内部における確率に過ぎない.

現実の未知の母集団から標本を何度も取り直すことによって 95% 信頼区間の 95% の意味を説明しようとしている解説をよくみかけるが, ミスリーディング(misleading, 誤誘導的)なので注意が必要である.

この節における以上の部分もミスリーディングであるという考え方もある.

この節での以上の解説では 95% 信頼区間の 95% をどのような確率(またはその近似値)とみなすことができるかを説明した.

確かにそういう知識を持っていた方が良いのだが、信頼区間については「P値はデータの数値とモデル+パラメータの値の相性の良さの指標である」というP値の解釈の基本に戻った方が分かり易い。

P値から真っ当な結論を出すためには、データの数値が信頼できてかつ、使用した統計モデルも何らかの意味で妥当でなければいけないのであった。以下ではこの条件を仮定する。

このとき、パラメータ θ に関する $100(1 - \alpha)\%$ 信頼区間は、閾値 α によってデータと相性が悪いと判断されずにすむパラメータ値 θ_0 全体の集合になる。

P値から導かれる結論は弱いので、信頼区間に含まれるパラメータ値についてはその現実における適切さについて判断保留することが正しい。

例えば、パラメータ θ がウェブページのデザインAのデザインBに対する優位性の指標であるときに、パラメータ θ に関する信頼区間については以下のように考える。

- (1) 信頼区間がゼロを含むならばデザインAとデザインBは同等の働きをしている可能性を排除できないと考える。
- (2) 信頼区間がたとえ正の値を含んでいても、単にデザインAの方がデザインBよりも優れている可能性を排除できないと考えるだけにとどめ、デザインAの方がデザインBよりも優れているとは安易に結論を出さない。
- (3) 信頼区間の左端の値が十分にゼロよりも大きいならば、デザインAの方が優れていると判断できる。ただし、上で「以下ではこの条件を仮定する」と述べていたことが重要であることに注意せよ! データが信用できなかったり、モデルが妥当でなかったりする場合にはこのような判断はひどい失敗を招く可能性がある。

5.6 P値の誤用の例

仮想的に、薬Aを与えた m 人と与えなかった n 人について〇〇日以内に回復した人の人数がそれぞれ a 人と c 人というデータが得られたとする。

その状況を2つの二項分布モデルを用いて分析するとする。薬Aを与えたときに〇〇日以内に回復する確率はパラメータ p でモデル化されており、与えなかった場合の回復確率はパラメータ q でモデル化されているとする。

そのモデルを使って仮説 $p = q$ のP値が計算されたとする。(P値の構成法は複数あるがどれか1つを選んでいると仮定する。)

5.6.1 P値は仮説が正しい確率ではない

そのP値は「薬Aを与えた場合と与えない場合で〇〇日以内の回復確率は等しい」という仮説が成立する確率ではない。

仮説 $p = q$ のP値の大雑把な定義は「データの数値 (m, a, n, c) 以上に $p = q$ に見える状況から離れた値が $p = q$ という仮説下での2つの二項分布モデル内で生じる確率」である。

P値は仮説が正しい確率では全然ないし、P値はモデル内確率に過ぎない。

注意: 「ベイズ統計では仮説が正しい確率を扱える」と言っている人達がいるが、ベイズ統計における「仮説が正しい確率」はモデル内確率分布である事後分布で測った確率であるに過ぎない。ベイズ統計においても使用した統計モデルが不適切ならば事後分布で測った確率も信用できないものになる。P値を使った場合に生じる困難がベイズ統計を使えば解消されるかのような主張は誤りである。

5.6.2 単独のP値のみを使って重要な決定をしてはいけない

事前に有意水準と呼ばれる相性の良さに関する閾値 α を決めてあたと仮定する(よく使われているのは $\alpha = 5\%$)。

さらに、データ中の回復者の割合を比較すると薬Aを与えた側の a/m の方が与えなかった側の c/n より大きい状況を仮定する。

(1) 帰無仮説 $p = q$ のP値が有意水準 α 未満になったとする。

そのときP値が α 未満になっただけで、「薬Aには効き目がある」と判断してはいけない。

P値以外の他の条件も確認する必要がある。

例えば、データから予想される薬Aの効き目は実生活においては無視できるほど些細な大きさなのに、P値が α 未満になる場合があり得る。

その他にも注意するべきことが沢山ある。

- データの数値が偏っている可能性も心配する必要がある。
- 使用した統計モデルが不適切な可能性にも配慮する必要がある。

統計学を使った分析は「すっきりしない」結果に終わることが多い。

分かり易さを求めすぎないように注意が必要である。

特に統計学を「科学的な御墨付きを得るための道具」扱いしてはいけない。

(2) 帰無仮説 $p = q$ のP値が有意水準 α 以上になったとする.

そのときに「統計的に有意ではない」という理由で、「ゆえに薬Aには効き目がない」と判断してはいけない.

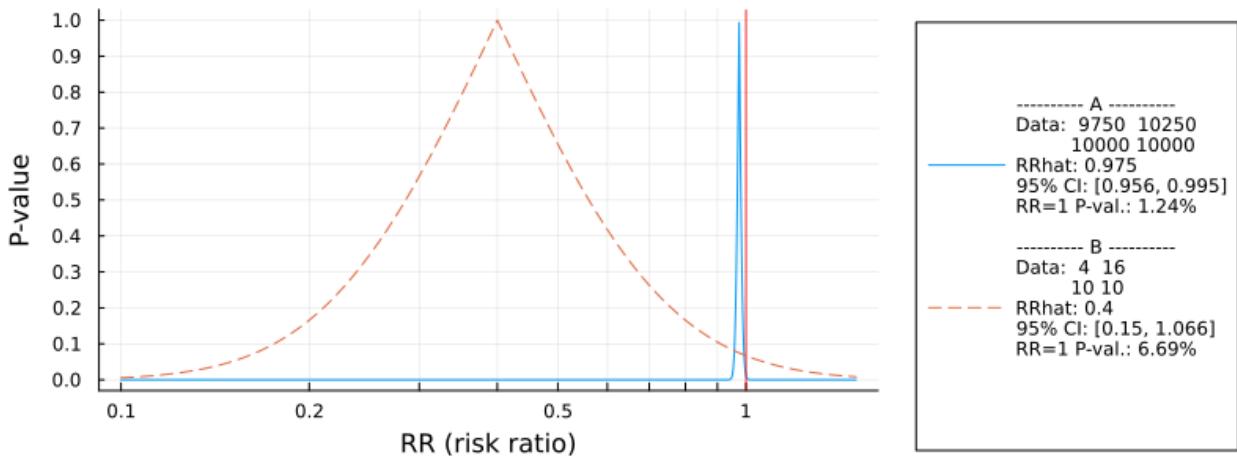
なぜならば、P値が有意水準 α 以上であることは、仮説 $p = q$ の下での統計モデルとデータの数値の相性の良さが閾値 α によって無さすぎると判定されないことを意味するに過ぎないからである。

「ゆえに薬Aには効き目がない」と結論を出すことは誤りで、「薬Aに効き目があるかどうかを閾値 α によって判定できなかった」と判断を保留しなければいけない。

ある研究においてP値が有意水準 α 以上になったとしても、その後の研究によって「薬Aには十分な大きさの効果がある」という結論が出る可能性があり得る。

In [22]: 1 risk_ratio_example1

Out[22]:



上のグラフは仮想的なデータに関するP値函数のプロットであり、薬Aと薬Bの効き目を推定しようとしている状況を想定している。

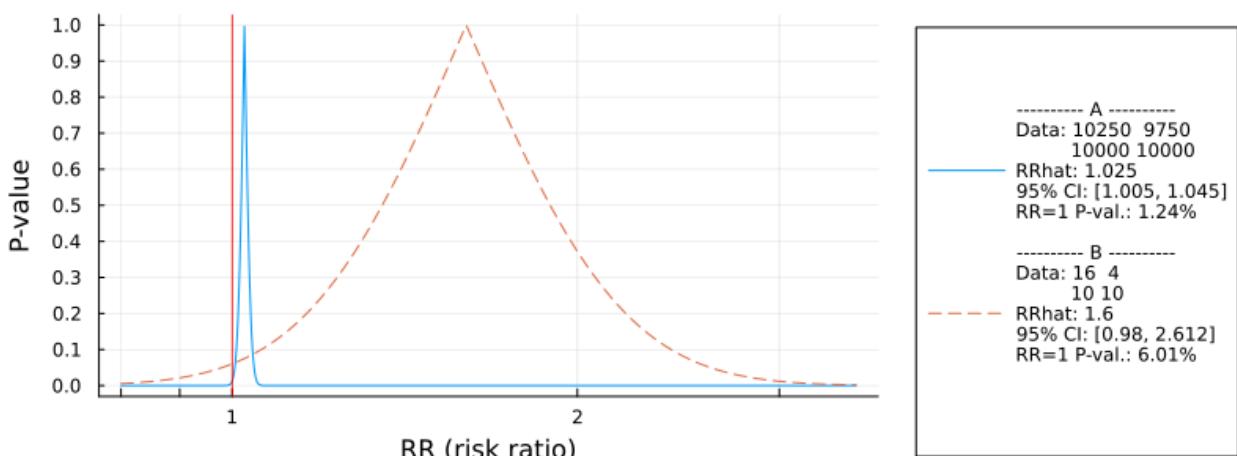
「薬Aを処方してもリスクは変わらない」という仮説のP値は1.24%で、有意水準5%で棄却される。しかし、リスクはほんの少ししか減らないように見える。

「薬Bを処方してもリスクは変わらない」という仮説のP値は6.69%で、有意水準5%で棄却されない。しかし、リスクは大きく減る可能性がある。薬Bについては、効果の大きさを期待しつつ判断を保留する必要がある。

このようなときに、「リスクは変わらない」という仮説のP値が5%を切ったかどうかだけで、薬Aには価値があり、薬Bには価値がないと判断することは誤りである。

In [23]: 1 risk_ratio_example2

Out[23]:



上のグラフも仮想的なデータに関するP値函数のプロットであり、物質Aと物質Bにさらされたときのリスクの大きさを推定している状況を想定している。

「物質Aにさらされてもリスクは変わらない」という仮説のP値は1.24%で、有意水準5%で棄却される。しかし、物質Aにさらされることによってリスクは少ししか増えないと考えられる。

「物質Bにさらされてもリスクは変わらない」という仮説のP値は6.01%で、有意水準5%で棄却されない。しかし、物質Bにさらされることによってリスクが大きく増える可能性があるかもしれない。物質Bについては、さらされることのリスクの大きさに注意しながら、判断を保留し、必要ならば研究を継続する必要がある。

このようなときに、「さらされてもリスクは変わらない」という仮説のP値だけを見て、物質Aにさらされることによって増えるリスクは気にする必要はあるが、物質Bについては気にする必要はないと考えることは明らかに間違っている。

5.6.3 扱う仮説達の中で正しいものの割合も重要

テストする薬達の中で実際に効くものの割合が w であると仮定する。

データ取得と仮説検定による判定によって(P値が有意水準未満なら効くと判定して良いと仮定している)、効かない薬が誤って効くと判定される確率は α (有意水準)で、効く薬が正しく効くと判定される確率は $1 - \beta$ (検出力)であるとする。

以上の設定は実践的に仮説検定が使われる状況としてはほぼ理想的な状態であると考えられる。

このとき、全確率の表は以下のようになる:

	効くと判定	効かないと判定
効く薬	$(1 - \beta)w$	βw
効かない薬	$\alpha(1 - w)$	$(1 - \alpha)(1 - w)$

この表より、

$$R_1 = (\text{効くと判定された薬の中での効かない薬の割合}) = \frac{\alpha(1 - w)}{(1 - \beta)w + \alpha(1 - w)}.$$

以下では $\alpha = 2.5\%$ (両側検定での有意水準 5% の片側検定換算)、 $1 - \beta = 80\%$ (よく使われている標準的な検出力)と仮定する:

$$R_1 = \frac{0.025(1 - w)}{0.8w + 0.025(1 - w)}.$$

例えば $w = 0.5$ ならば $R_1 \approx 3\%$ になる(以下のグラフも参照)。

そのことから、効くと判定された薬の中での実際には効かない薬の割合 R_1 を 3% 程度以下に抑えなければ、テストする薬の中での効く薬の割合を半分以上にすればよいことが分かる。

しかし $w = 0.05$ ならば $R_1 \approx 37\%$ になる。

このことから、もしもテストする薬の中での効く薬の割合が 5% しか無ければ(こういう場合は珍しくないだろう)、仮説検定によって効くと判定された薬の中での実際には効かない薬の割合が 37% 程度と高くなってしまう。

このようにテストする薬の中での効く薬の割合が小さい場合には、仮説検定を理想的に利用できたとしても、効くと判定した薬の中での効かないものの割合あ大きくなってしまう。

P値を使った判断では以上のようなことも気にする必要がある。

P値単体には科学的に信頼できる結論を出す力はない!

P値はそれ以外の情報と組み合わせて使う必要がある。

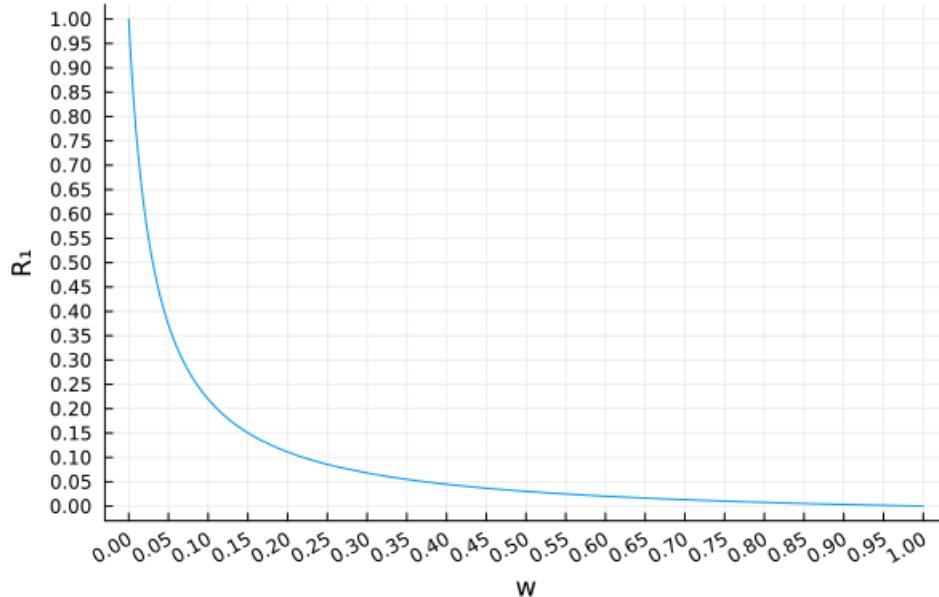
```
In [24]: 1 R_1(α, β, w) = α*(1-w) / ((1-β)*w + α*(1-w))
2 α = 0.025
3 β = 0.2
4
5 @show R_1(α, β, 0.5)
6 @show R_1(α, β, 0.05)
7
8 plot(w → R_1(α, β, w), 0, 1; label="")
9 plot!(xtick=0:0.05:1, ytick=0:0.05:1, xrotation=30)
10 plot!(xguide="w", yguide="R₁")
11 plot!(size=(600, 400))
12 title!("R₁ for α=$(round(α; sigdigits=3)), 1-β=$(round(1-β; sigdigits=3))")
```

$R_1(\alpha, \beta, 0.5) = 0.030303030303030304$

$R_1(\alpha, \beta, 0.05) = 0.37254901960784315$

Out[24]:

R_1 for $\alpha=0.025$, $1-\beta=0.8$



5.6.4 実際の検出力が低くなつたせいで再現率が低くなる場合がある

上の理想化された設定においてテストする薬が実際に効く場合であつても、検出力 $1 - \beta$ の確率でしか薬が効くと判定されない。

再現実験では、すでに得られている効果の大きさの点推定値を使って、検出力を例えれば 80% になるような標本サイズを計算してそのサイズのデータを取得することが多い。しかし、すでに得られている効果の大きさの推定値が過大評価になつていて、実際の効果の大きさが想定よりも小さい場合には、そのようにして求めた標本サイズでの検出力は 80% よりも低い値になつてしまふ。そのような場合には再現確率は想定よりも低くなつてしまふ。

標本サイズが小さくてかつP値も小さい場合には数学的必然として効果の大きさの点推定値は大きくなつてしまふ。だから、小さ過ぎる標本サイズと小さなP値の組み合わせの報告における効果の大きさの点推定値は過大評価になつてしまいがちである。そのような効果の大きさの推定値を用いて検出力 80% が保証される標本サイズを求めるとき、本当に必要な標本サイズよりも小さくなつてしまふ、再現実験にも失敗する確率が高まる。

こういう類のことが「再現性の危機」に関する議論で無視されていることが多いという問題については次の論文を参照せよ。

- Carol Ting and Sander Greenland, Forcing a deterministic frame on probabilistic phenomena: a communication blind spot in media coverage of the “replication crisis”, Science Communication, 2024. [Google Scholar \(<https://scholar.google.co.jp/scholar?cluster=6028850411184048306>\)](https://scholar.google.co.jp/scholar?cluster=6028850411184048306)

「再現性の危機」については杜撰な議論が横行しているので注意が必要である。特に「1つの研究だけで科学的に確立された結果が得られるのではない」という科学的な常識が無視されていることが問題である。この点については次の「再現性を期待しなければ再現性危機は存在しない」論文を参照せよ：

- Valentin Amrhein, David Trafimow, and Sander Greenland, Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication, The American Statistician, Volume 73, 2019. [Google Scholar \(<https://scholar.google.co.jp/scholar?cluster=3023429090384832581>\)](https://scholar.google.co.jp/scholar?cluster=3023429090384832581)

5.6.5 すべてを正直に報告しなければいけない

実際には薬Aだけではなく、薬B,C,D,...と多数の薬を試してみた場合には、その中でP値が $\alpha = 5\%$ 未満になった薬Aのみについて結果を報告するというようなことをしてはいけない。

葉A,B,C,...のすべてに効き目がゼロ(害もゼロ)であるとき、データから計算したP値が $\alpha = 5\%$ 未満になるものはその中に 5% 程度ある。有意水準 $\alpha = 5\%$ という閾値はかなり緩く、色々試してみれば容易にP値を $\alpha = 5\%$ 未満にできる。

都合の良い結果だけを選択的に報告してしまうと、その研究結果がどれだけどのように信頼できるかが分からなくなってしまう。

自分が所属する研究室のボスが部下に対して、「統計的に有意になった結果のみを報告すればよい」と指示している場合には、研究不正を意識的に指示している疑いがあるので、適切な窓口に相談した方が良いと思われる。

5.7 P値や信頼区間の誤用に関する参考資料

次の文献は必読である：

- 統計的有意性とP値に関するASA声明 (<https://www.biometrics.gr.jp/news/all/ASA.pdf>)、(翻訳: 佐藤俊哉, 2017年4月23日, 英語の原著は2016年)

ただし、この声明における

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。

という説明の中の「データ」は「統計モデル内部で生成された仮想的なデータ」であることに注意する必要がある。次のように言い直した方が誤解が少ないように思われる。

- おおざっぱにいうと、P値とは特定の統計モデル内で生成された仮想的データの統計的要約(たとえば、2グループでの比較での標本平均の差)が現実で観察されたデータの値以上に極端な値になる確率もしくはその近似値である。

P値に関するASA声明の翻訳者によるP値に関する解説動画が次の場所にある。

- 佐藤俊哉、[仮説検定とP値の誤解](https://youtu.be/vz9cZnB1d1c) (<https://youtu.be/vz9cZnB1d1c>)、2017年10月19日

分かり難いとよく言われている「信頼区間」については、検定(もしくはP値函数)と信頼区間の表裏一体性に基いて信頼区間について理解しようとすることが正しい理解の方針で思われる。検定と信頼区間の表裏一体性については、すぐ上の動画の48:37以降で解説されている。

同著者による次の本には非常に役に立つ話が書いてある：

- 佐藤俊哉、『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年03月01日 ([公式サイト](https://www.asakura.co.jp/detail.php?book_code=12297) (https://www.asakura.co.jp/detail.php?book_code=12297))

可能ならば以下の論文も参照せよ：

- Valentin Amrhein, Sander Greenland. Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. First Published June 3, 2022. [[doi](https://doi.org/10.1177%2F02683962221105904) (<https://doi.org/10.1177%2F02683962221105904>)]

この一連のノートの内容はこの論文が出る前に計画されていたのであるが、P値の使い方については内容的に非常に近いものになっている。

さらに次の文献も参考になる：

- アレックス・ラインハート著 (西原史暁訳)、『ダメな統計学 Statistics Done Wrong』、2014年12月28日 日本語版公開 [[PDF](https://repun-app.fish.hokudai.ac.jp/pluginfile.php/6536/mod_resource/content/1/%E3%82%BF%E3%82%99%E3%83%A1%E3%81%AA%E7) (https://repun-app.fish.hokudai.ac.jp/pluginfile.php/6536/mod_resource/content/1/%E3%82%BF%E3%82%99%E3%83%A1%E3%81%AA%E7])

特に第5章「P値と基準率の誤り」は必読である。

