

誤用を避けるための注意

- 黒木玄
- 2022-07-01~2022-07-01

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

- ▼ [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 信頼区間の定義](#)
 - [5.4 信頼区間の正しい解釈の仕方](#)
 - [5.5 P値や信頼区間の誤用に関する参考資料](#)

```
In [1]: 1 ENV["LINES"], ENV["COLUMNS"] = 100, 100
2 using Base.Threads
3 using BenchmarkTools
4 using DataFrames
5 using Distributions
6 using LinearAlgebra
7 using Memoization
8 using Printf
9 using QuadGK
10 using RCall
11 using Random
12 Random.seed!(4649373)
13 using Roots
14 using SpecialFunctions
15 using StaticArrays
16 using StatsBase
17 using StatsFuns
18 using StatsPlots
19 default(fmt = :png, size = (400, 250),
20         titlefontsize = 10, plot_titlefontsize = 12)
21 using SymPy
```

```
In [2]: 1 # Override the Base.show definition of SymPy.jl:
2 # https://github.com/JuliaPy/SymPy.jl/blob/29c5bfd1d10ac53014fa7fef468bc8deccadc2fc/src/types.jl
3
4 @eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::SymbolicObject)
5     print(io, as_markdown("\displaystyle " *
6         sympy.latex(x, mode="plain", fold_short_frac=false)))
7 end
8 @eval SymPy function Base.show(io::IO, ::MIME"text/latex", x::AbstractArray{Sym})
9     function toeqnarray(x::Vector{Sym})
10         a = join(["\displaystyle " *
11             sympy.latex(x[i]) for i in 1:length(x)], "\\\")
12         """\left[ \begin{array}{r}$a\end{array} \right]""
13     end
14     function toeqnarray(x::AbstractArray{Sym}, 2)
15         sz = size(x)
16         a = join([join("\displaystyle " .* map(sympy.latex, x[i,:]), "&")
17             for i in 1:sz[1]], "\\\")
18         """\left[ \begin{array}{r} * repeat("r", sz[2]) * "}" * a * "\end{array}\right]""
19     end
20     print(io, as_markdown(toeqnarray(x)))
21 end
```

```
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      $\hat{\alpha}$ ,  $\hat{\beta}$  = 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 = (a - 0.1(d-c), b + 0.1(d-c))
14    end
15
16    scatter(x, y; label="", msw=0, ms, alpha)
17    plot!(x →  $\hat{\alpha}$  +  $\hat{\beta}$ *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datasauRusパッケージをインストールしておくこと
2 datasaurus = rcopy(R"datasauRus::datasaurus_dozen")
3 datanames = unique(datasaurus.dataset)
4 xlim = ylim = (-5, 105)
5 xguide = yguide = ""
6 PP = []
7 for name in datanames
8     x = datasaurus.x[datasaurus.dataset .== name]
9     y = datasaurus.y[datasaurus.dataset .== name]
10    P = plot_ols(x, y; xlim, ylim, title=name, ms=3)
11    push!(PP, P)
12 end
13 layout = @layout[_ a _; b c d; e f g; h i j; k l m]
14 Datasaurus_dozen = plot(PP...; layout, size=(800, 1000));
```

In [14]:

```
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)
```

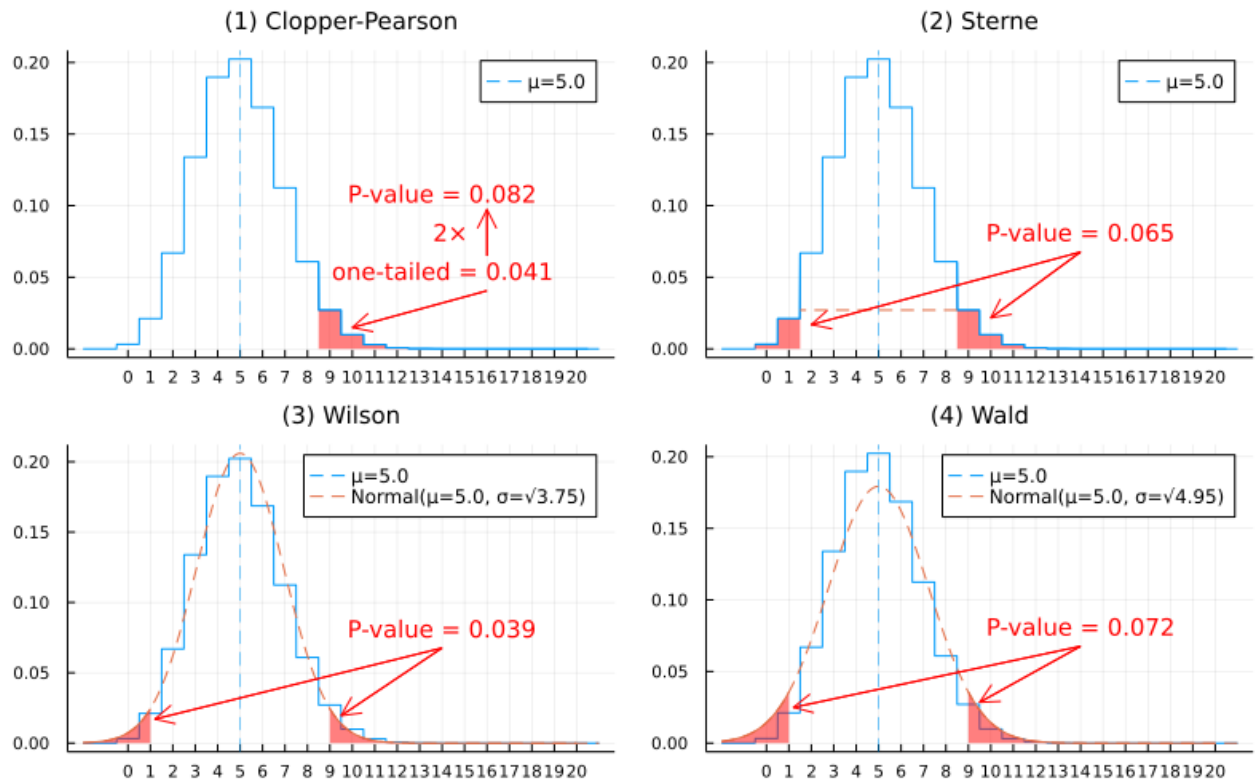
```

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

```

Out[14]:

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



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

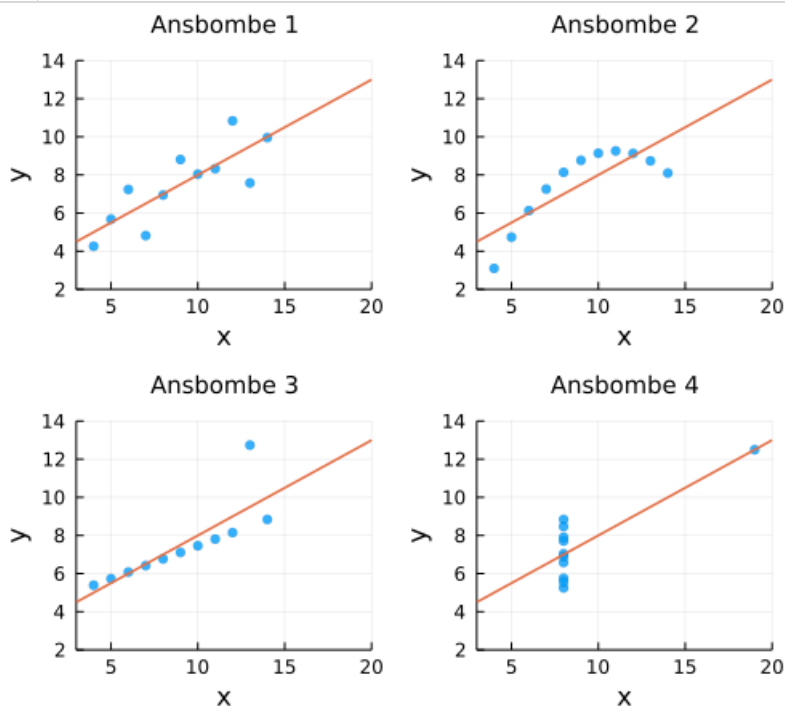
Anscombeの例 (https://en.wikipedia.org/wiki/Anscombe%27s_quartet)については

- 「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>)

の必修問題を参照せよ。

In [7]: 1 Anscombe_quartet

Out[7]:



4つ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。

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%2C%20likelihood%2C%20>

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

に英語による詳しい説明があり、

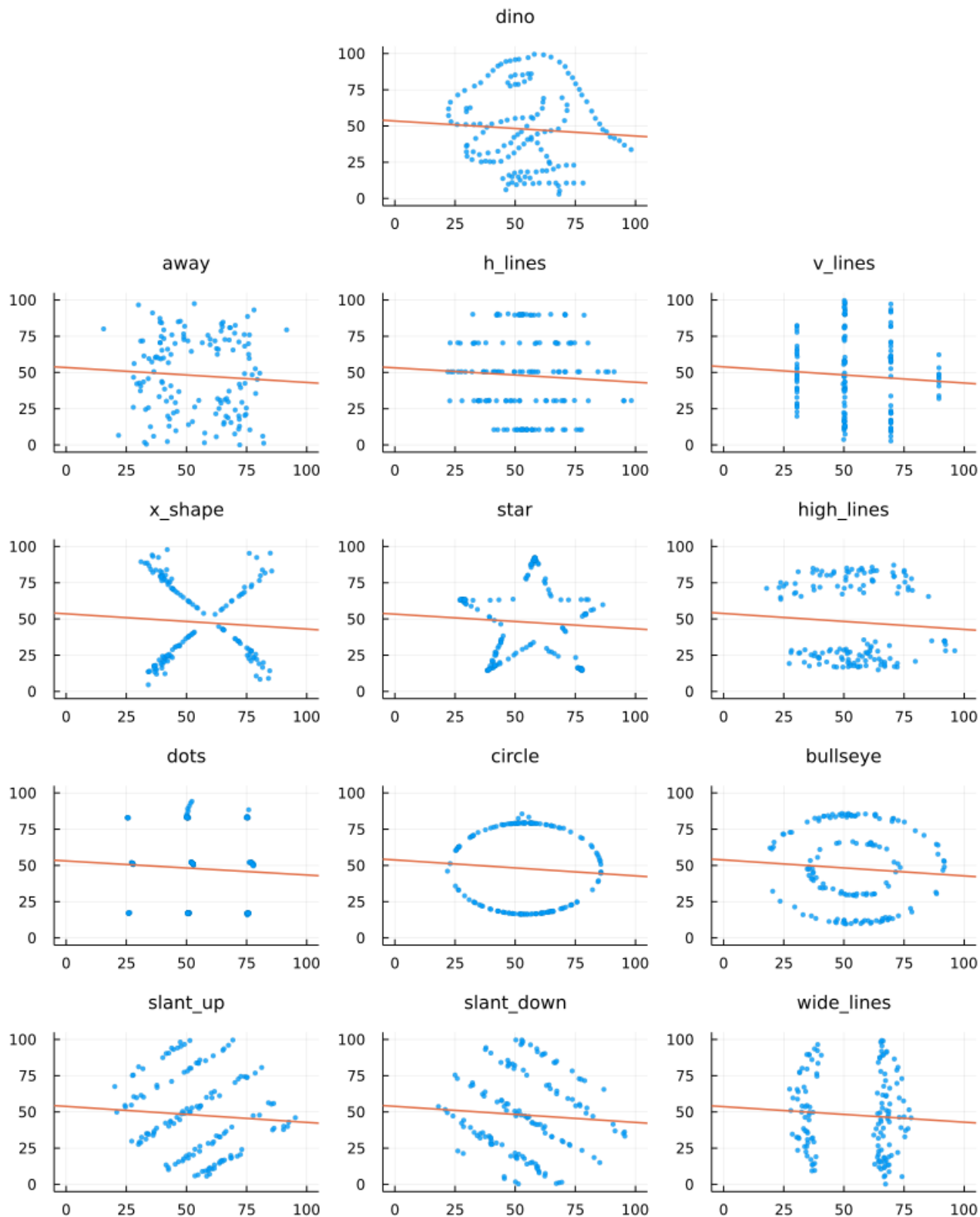
- <https://visualizing.jp/the-datasaurus-dozen/> (<https://visualizing.jp/the-datasaurus-dozen/>)

に日本語による簡単な解説がある。



In [8]: 1 Datasaurus_dozen

Out[8]:



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** とすることがある.

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)

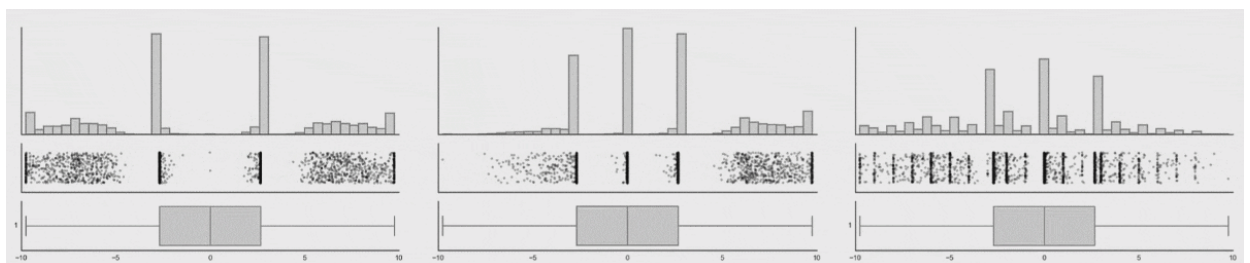
<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/06%20Conditional%20distribution%2C%20likelihood%2C%20estimation> の最後の方にあるDataSaurusの例に関する問題をまだ解いていない人は解いてみよう.

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

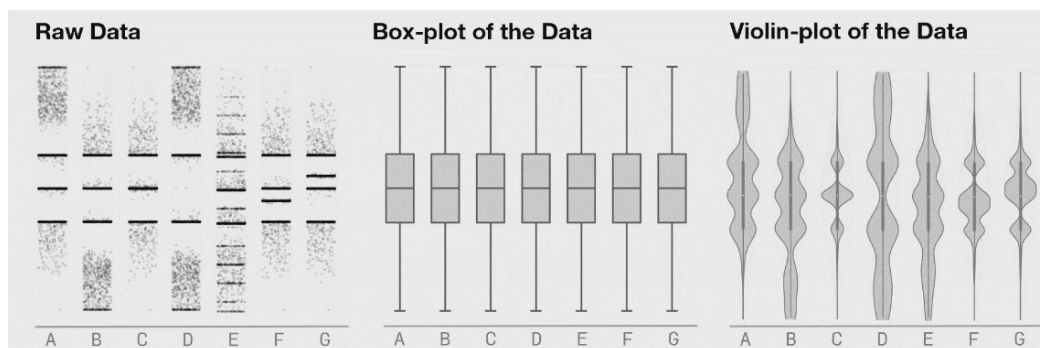
以下の図は既出のデータサウルスの解説ページ

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

「単峰型」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(ビン, bin)の設定に恣意性があり、ビン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単峰型でないデータで箱ひげ図を描いた場合(不適切な場合)の例になっている。サイズが小さす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。

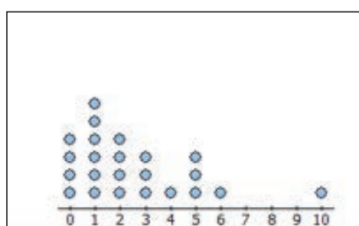


図1 あるクラスの1か月に読んだ本の冊数

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に山がもう1つある可能性もあり、自信を持って単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。

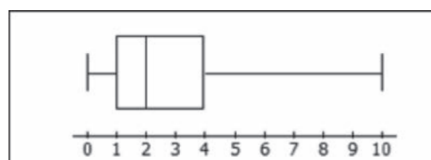


図2 図1のデータの箱ひげ図

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人であることは分らない。

さらに次のような図も見つかる。

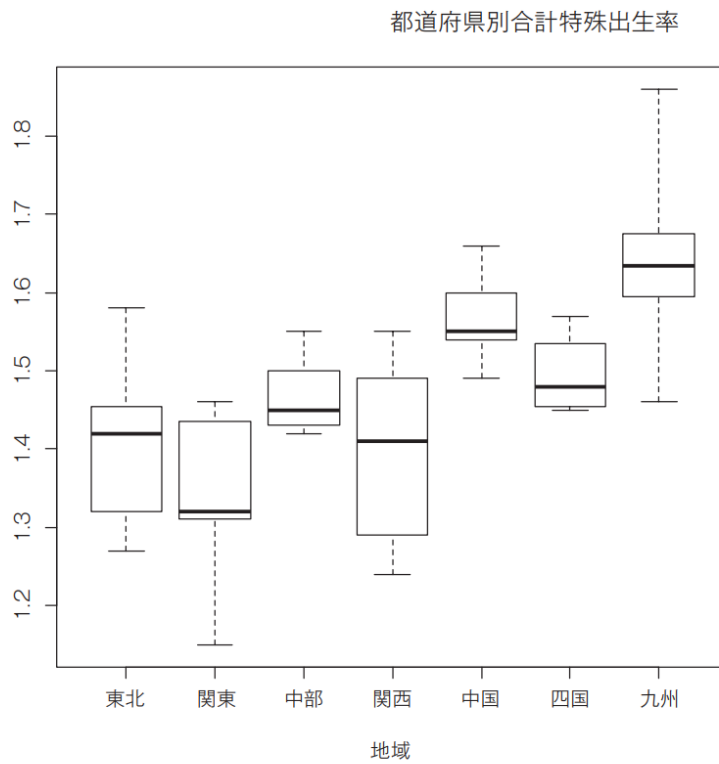


図3 地域別の合計特殊出生率

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.


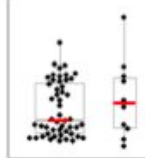

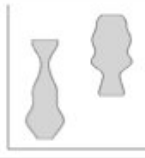

この不適切なグラフの描き方を見れば, データサイズの情報が非常に重要なことも分かる.

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

- x軸上のグループ名の下にサンプルのサイズも示しておくこと
- ひげが何を表しているかを凡例で示しておくこと

と書いてある.

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, 男性用と女性用の2つのファイルに含まれている.

学生に私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
12		
14		
15	1	15000 16.9
16	2	15000 16.9
21	6	14861 16.8
23	7	14861 16.8
26	8	14699 17.3
28	10	14560 20.5
31	11	14560 20.6
33	13	14560 20.5
34	17	14560 20.4
35	18	14560 20.4
36	19	14560 19.8
38	20	14560 19.7
39	22	14560 19.7
41	24	14560 19.6
44	25	14560 19.6
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
	44	14259 19.8

<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の)グループの学生は、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質問した。

2つ目の「仮説なし」の(hypothesis-freeの)グループの学生には単に「データセットから何を結論付けますか?」と聞いた。

その結果は以下の通り。

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

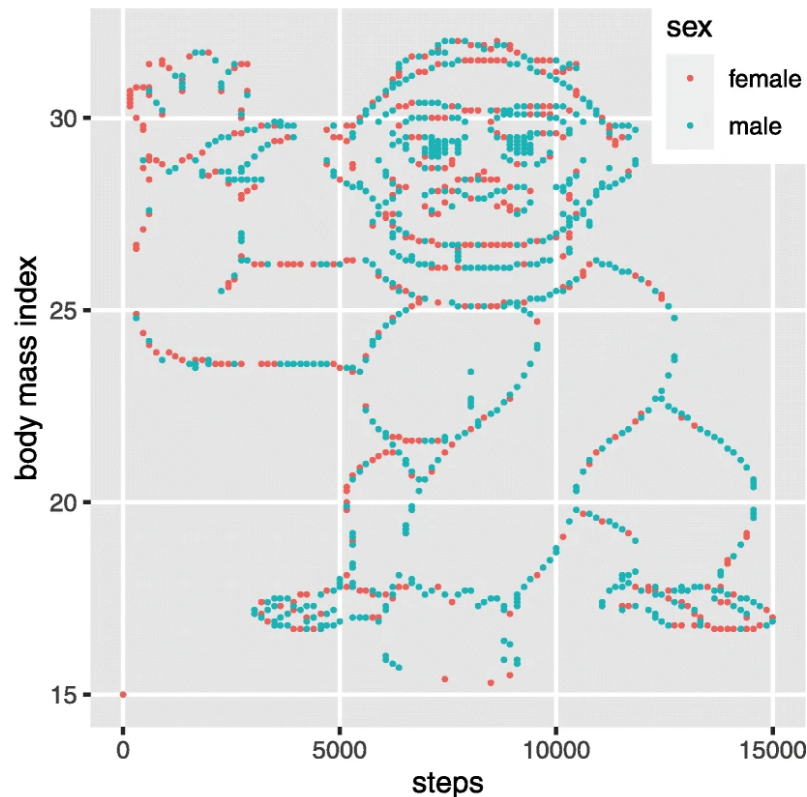
<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w/figures/1>

この表を翻訳すると以下の通り:

	ゴリラは発見されなかった	ゴリラが発見された
仮説ありグループ	14	5
仮説なしグループ	5	9

「ゴリラは発見されなかった」と「ゴリラが発見された」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!



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

学生に課された統計学のレポート課題のデータセットは実は偽物のデータセットであり、散布図を描くとゴリラの絵が現れるように作られていたのである!

レポートを提出した学生は2つのグループを合わせて33人で、そのうちの過半数の19人がゴリラを発見できなかった。

さらに、(i),(ii),(iii)の3つの仮説を検討するように指示を出された「仮説あり」のグループでは19人中14人もゴリラを発見できなかった! 3つの仮説の検討の指示によって誤誘導されてしまった格好になってしまっている。

このようなレポート課題を出すことは統計学の講義として教育的であり、学生の側も良い勉強になったと思われる。

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よ:

1. 仮説「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.
2. オッズ比パラメータ OR に関するWald版 95% 信頼区間.

可能ならば以下も求めよ.

3. 仮説「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.
4. 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.
5. 仮説「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.
6. 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.
7. 仮説「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.
8. 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.

可能ならばそれぞれの場合のP値関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値関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。

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。

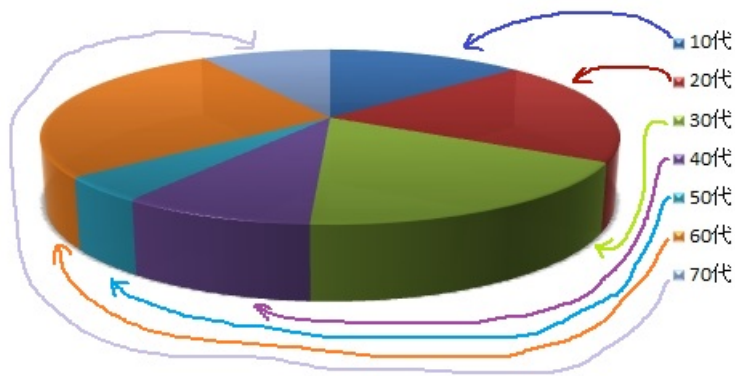
多くの場合に誤った印象を与えることになることで有名なグラフの描き方に3次元円グラフがある。(一般に円グラフはよくないグラフの描き方だとされている.)

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描き方ではないことについては

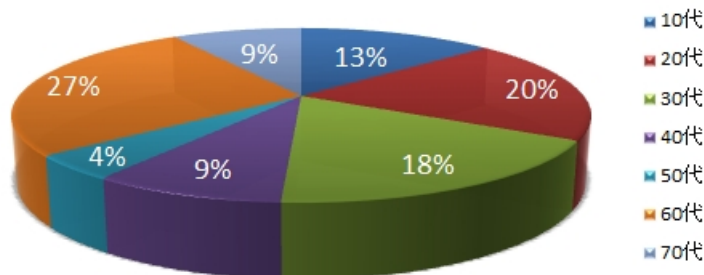
- [3D円グラフ撲滅委員会より、なぜ3D円グラフを使ってはいけないうか](https://www.cresco.co.jp/blog/entry/950/) (<https://www.cresco.co.jp/blog/entry/950/>).

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。



このように3次元円グラフは割合について間違った印象を与えるために利用できる!

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
パルチモア・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。

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)になってしまう確率はほぼ 0 である：

$$\text{cdf}(\text{Binomial}(2000000, 0.6), 860000) \approx 2 \times 10^{-50925}.$$

このことから、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。

この事例から、たとえサイズ200万の **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 [9]: 1 setprecision(50) do
        2   @time sum(pdf(Binomial(2000000, big"0.6"), k) for k in 0:860000)
        3 end
```

36.197282 seconds (106.64 M allocations: 3.172 GiB, 5.54% gc time, 0.01% compilation time)

Out[9]: 2.0771983651745741e-50925

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/naruhodo/15_episode/episode/senkyo2.html)
(https://www.stat.go.jp/naruhodo/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 デュエイ
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/naruhodo/15_episode/episode/senkyo1.html)
(https://www.stat.go.jp/naruhodo/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/naruhodo/15_episode/episode/senkyo2.html)
(https://www.stat.go.jp/naruhodo/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。

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)
(<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)
(<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いうこともあり得る。

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りえる。

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/ybdkQFEcCPM) (<https://youtu.be/ybdkQFEcCPM>), 2018年12月20日

5 P値や信頼区間の誤用

5.1 P値の定義

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：

$$\text{pvalue}_{\text{CP}}(k|n, p = p_0) = \min \left(\begin{array}{c} 1 \\ 2 \text{cdf}(\text{Binomial}(n, p_0), k) \\ 2(1 - \text{cdf}(\text{Binomial}(n, p_0), k - 1)) \end{array} \right).$$

「～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。その他にも

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

では、Sterne型のP値と

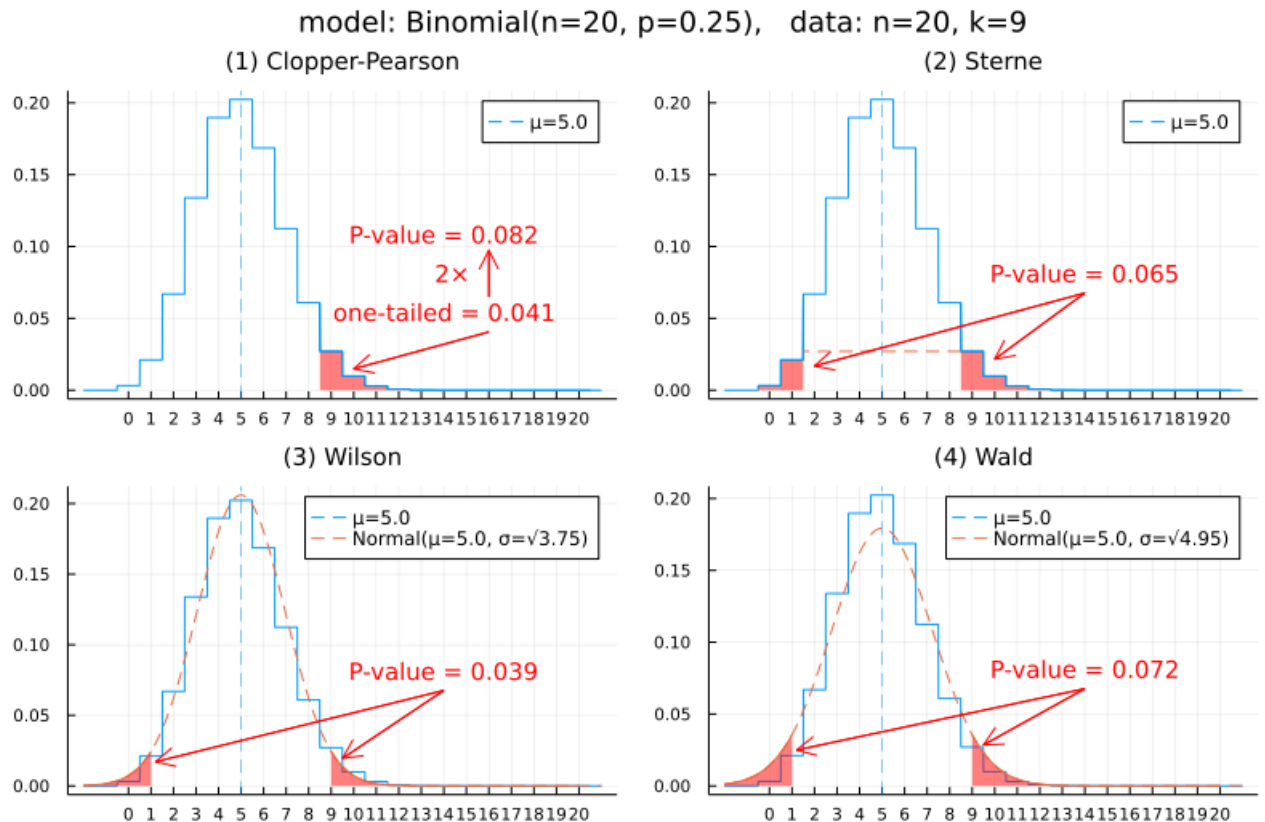
- ・ 「[大数の法則と中心極限定理](https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/05%20Central%20limit%20theorem.ipynb)」のノート (<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/05%20Central%20limit%20theorem.ipynb>)

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。その意味で非常に基本的である。

```
In [15]: 1 illustrate_binomial_pvalues()
```

```
Out[15]:
```



5.2 P値の正しい使い方

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

P値が非常に小さい場合には整合性がほとんどないと考える。

P値が 0 より十分に大きな場合には、モデル $M(p)$ の下での仮説 $p = p_0$ のデータの数値 x との整合性がある程度あると考える。しかし、整合性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の話になるので注意すること。

P値だけから出せる結論は常に弱い という点を強調したいがために、「P値は整合性の指標である」と述べていることにも注意せよ!

妥当ではないもののどうしのあいだで整合性がある場合もあり得ることに注意せよ。

例えば、仮説 $\theta = \theta_0$ 下で妥当でない統計モデル $M(\theta = \theta_0)$ とデータの数値 x のあいだに整合性があることもあるし、モデルと信頼できないデータの数値 x とのあいだに整合性があることもありえる。

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、真つ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他にもありえる。)

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何らかの意味のある統計分析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頼性のような最も重要な事柄についてはユーザー側が責任を負う必要がある。

5.3 信頼区間の定義

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 $\text{pvalue}(x|\theta = \theta_0)$ 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、 $\text{pvalue}(x|\theta = \theta_0) \geq \alpha$ となるパラメータ値 θ_0 全体の集合として定義される:

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

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関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関数(quantile function, 累積分布関数 cdf の逆関数)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関数の $1 - \alpha/2$ 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%20interval%20Proportion.ipynb)
(<https://nbviewer.org/github/genkuroki/Statistics/blob/master/2022/09%20Hypothesis%20testing%20and%20confidence%20interval%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 分布で置き換えられる場合もある。

5.4 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、 $\text{pvalue}(X|\theta = \theta_0) \geq \alpha$ 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 $\text{confint}(X|\alpha) = \{ \theta_0 \mid \text{pvalue}(X|\theta = \theta_0) \geq \alpha \}$ 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、その関数である信頼区間 $\text{confint}(X|\alpha)$ 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。

実はこの節における以上の解説全体がミスリーディングであるという考え方もできる。

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であるが、**信頼区間を解釈する場合には「P値はデータの数値とモデル+パラメータ値の整合性の指標である」というP値の解釈の基本に戻った方が分かり易い。**

P値から真つ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ある必要があるのであった。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 でデータと整合性がな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 が商用ウェブサイトのデザインAのデザインBに対する優位性の指標であるときには、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結論しない。
- (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。

Type Markdown and LaTeX: α^2

Type Markdown and LaTeX: α^2

Type Markdown and LaTeX: α^2

5.5 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) (<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、そこにおける

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。

における「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

分かり難いと言われている「信頼区間」については、検定(もしくはP値関数)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の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からの解説にある。

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。

In []:

1	
---	--