

いまさら聞けない基礎統計学5 ～グラフによるデータの可視化

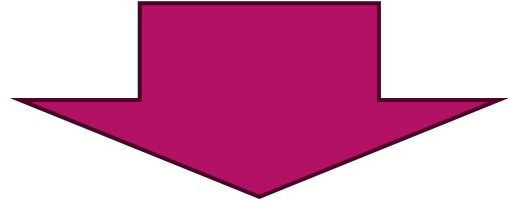
ダイバーシティ推進室

衛生学・予防医学講座

各務竹康

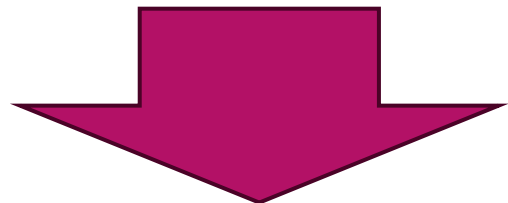
情報の取得

生体試料(血液、毛髪、汗など)、
紙媒体、web情報



情報の入力 (データベース化)

エクセルなど、情報を一元化し
て管理



分析

多くの場合、研究仮説に基づい
て適切な手法を用いる

結果の記載にグラフを用いる

得られた情報は仮説検証に適切
であるか？
→データの確認

グラフによって何を見るのか？

- ▶ “データの性質”
- ▶ 仮説に沿った検証が可能であるか

最近使用したグラフ

テンプレート

縦棒

折れ線

円

横棒

面

散布図

マップ

株価

等高線

レーダー

ツリーマップ

サンバースト

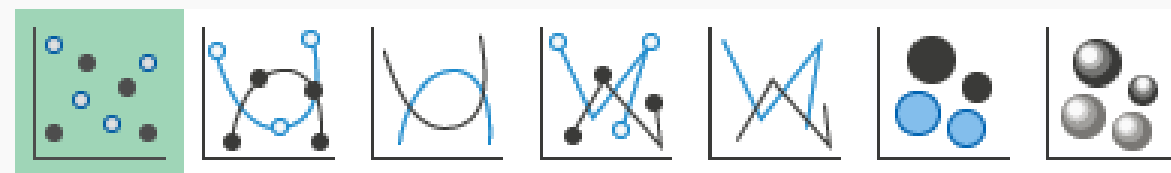
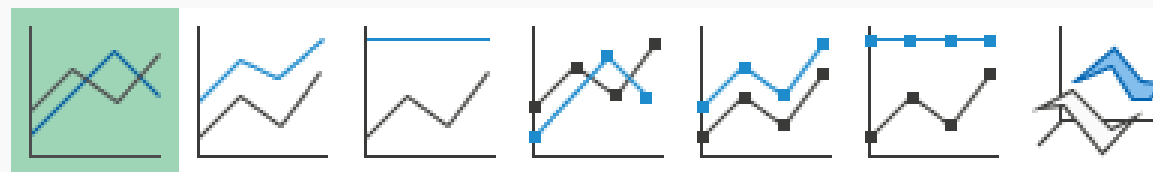
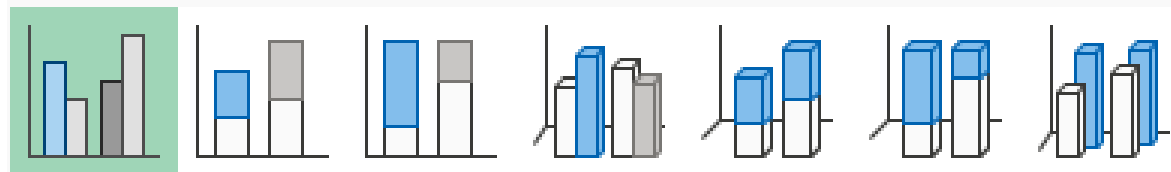
ヒストグラム

箱ひげ図

ウォーターフォール

じょうご

組み合わせ



結局どれを使うといいのか、
わからなくないませんか？

本日の目標

- ▶ 様々なグラフの長所、短所を学ぶ
- ▶ 集まったデータを確認する際に、適切なグラフの選択ができる
- ▶ 学会発表、論文作成において、効果的なグラフを使用できる

一変数のグラフ

得られたデータの性質を理解する

復習：3種類の変数

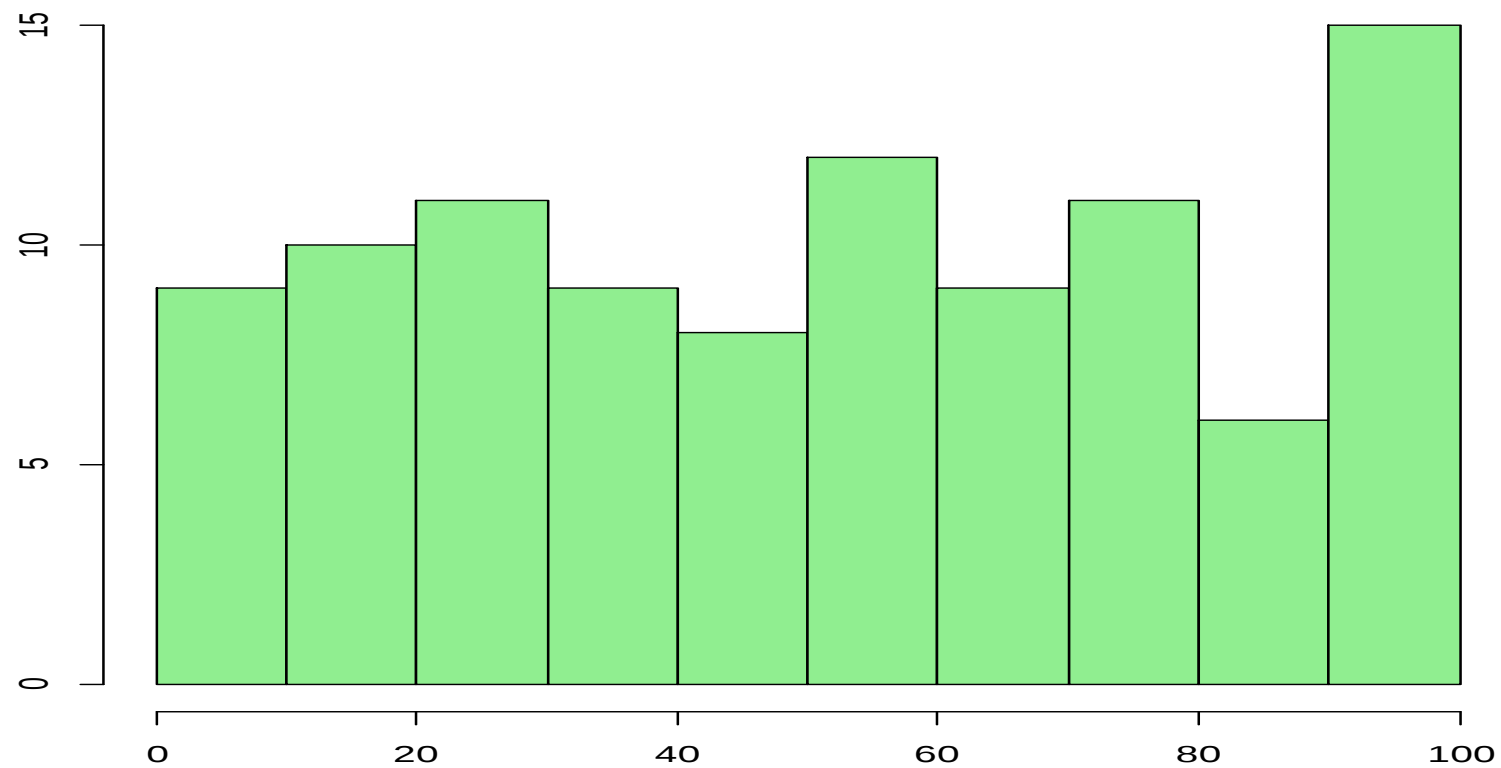
- ▶ **連続変数**：全ての数字が順番に並んでおり、その順番、間隔が意味を持つ。比尺度(原点“0”に意味があり、変数同士の足し算、掛け算が可能)と間隔尺度(“0”は連続する数字の一箇所で、変数同士の足し算はできるが掛け算は通常不可)に分類される
- ▶ **順序変数**：数字の並び順には意味があるが、数字と数字の間には特定の指標を持たず、数字同士の足し算、掛け算は通常行わない
- ▶ **カテゴリ変数**：便宜的に数字が割り振られているのみで、数字同士の関係は意味を持たない

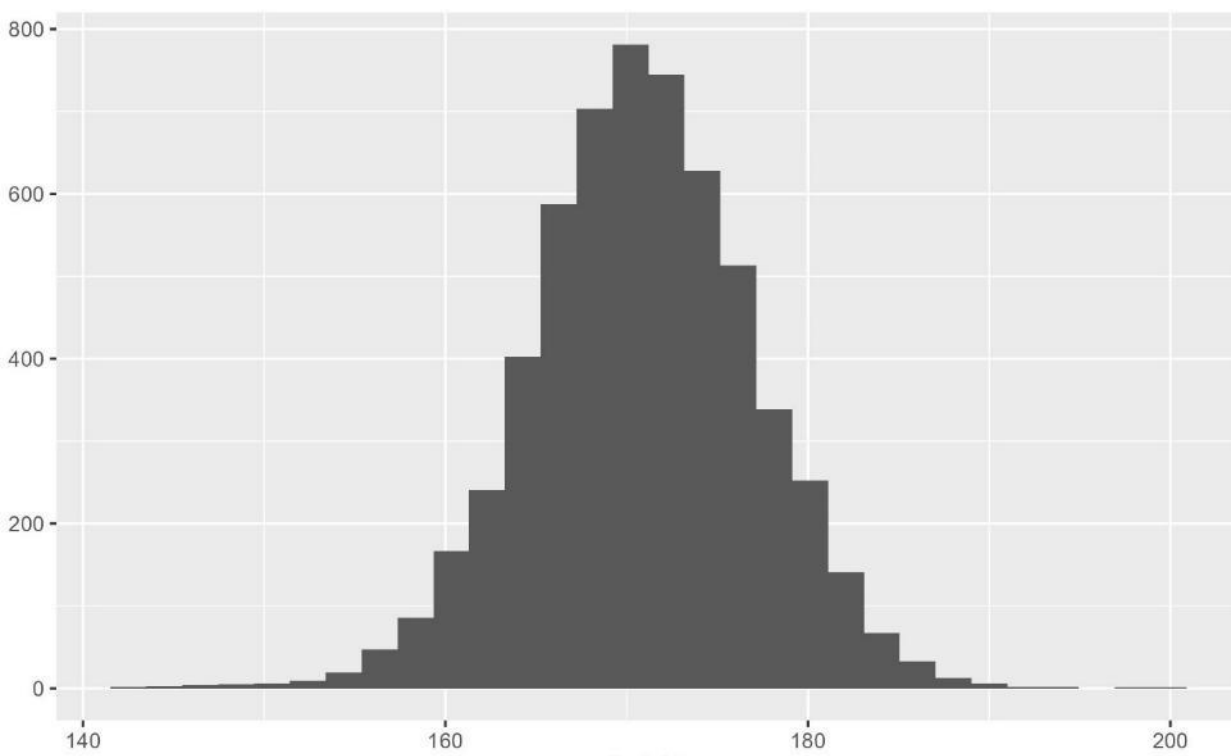
データの種類とグラフ

連続変数	順序変数	カテゴリ変数
ヒストグラム 箱ひげ図 ドットプロット 密度曲線 など	箱ひげ図 棒グラフ	棒グラフ

ヒストグラム

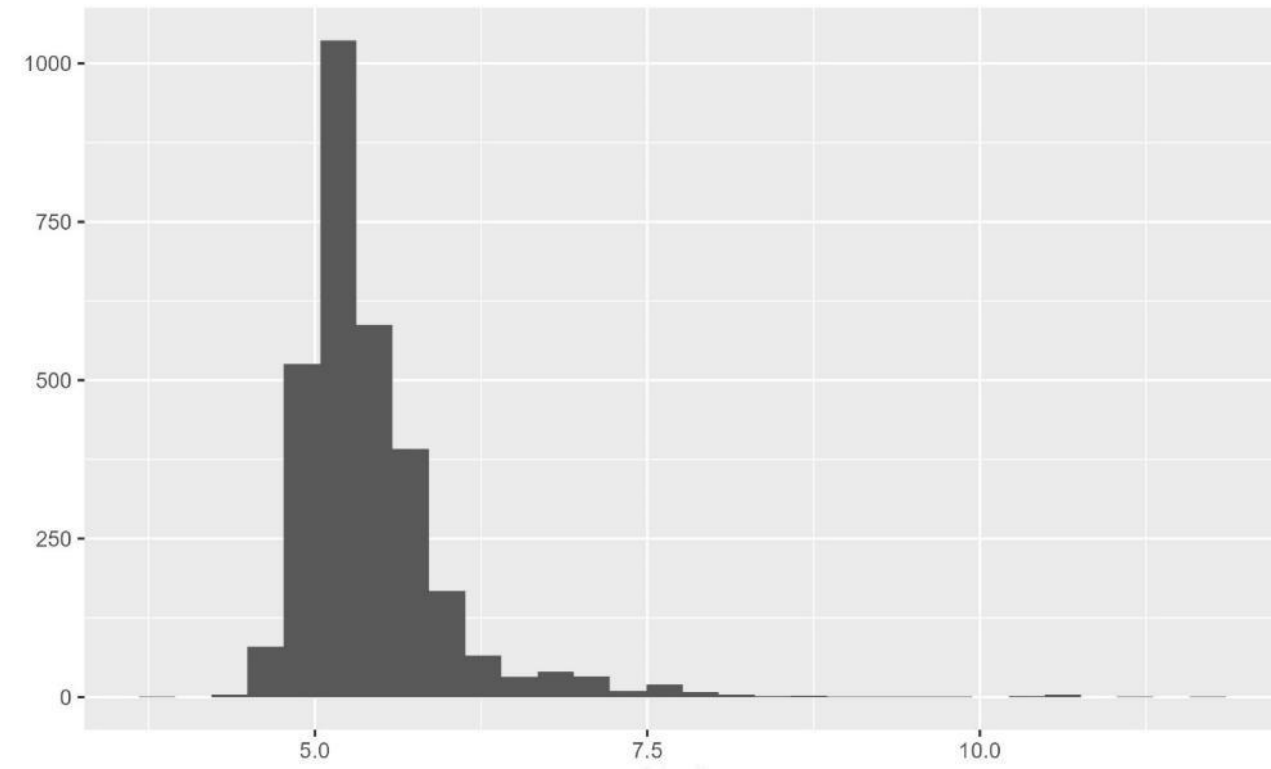
- ▶ 階級別の分布表を作成して、その度数をグラフ化する





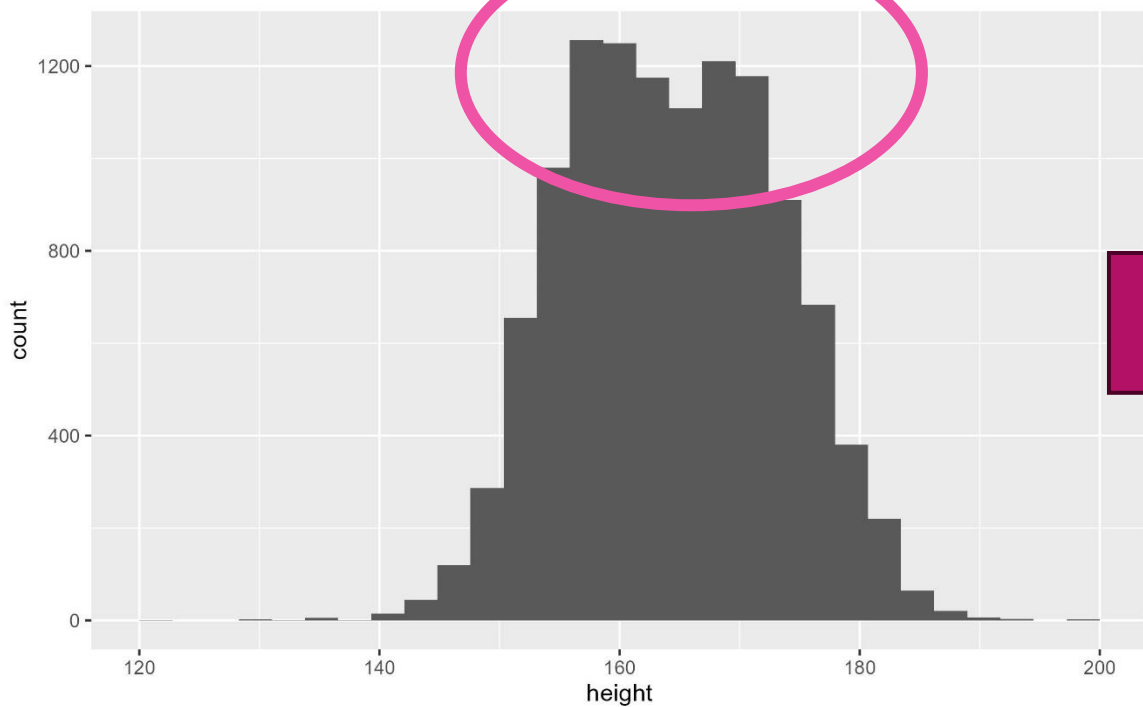
身長のヒストグラム
ほぼ正規分布をしている

HbA1cのヒストグラム
正常域に多くが集まり、異常値
が右にのびている

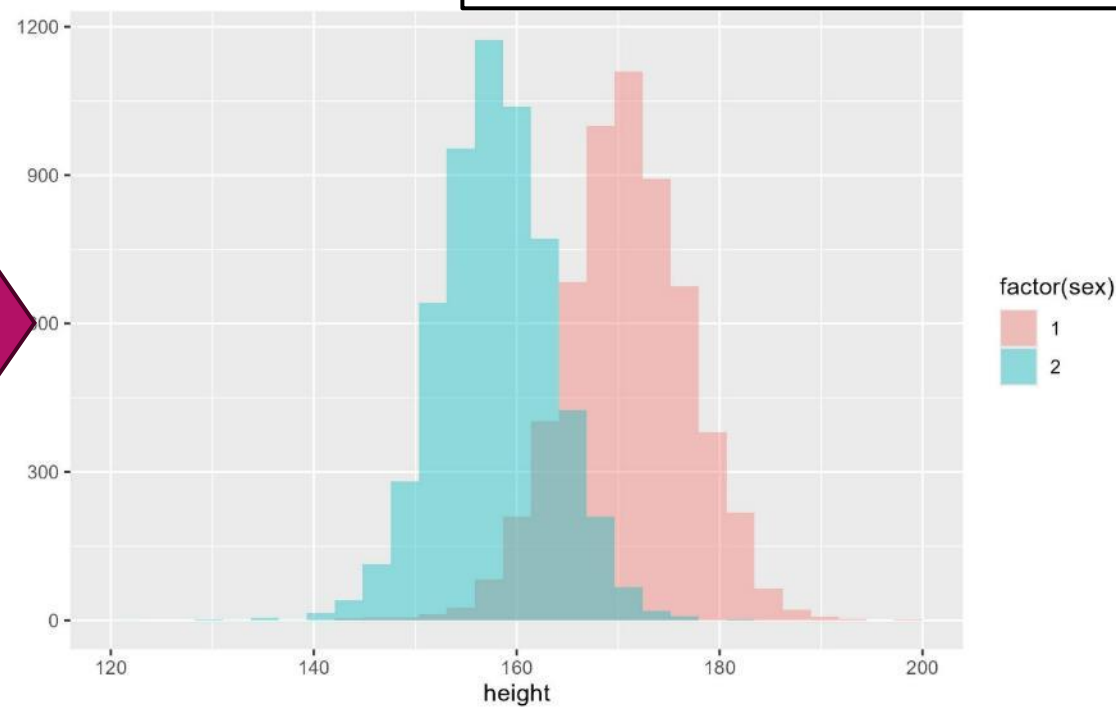


形の不自然さから性質がわかる

山の頂点が2つ？



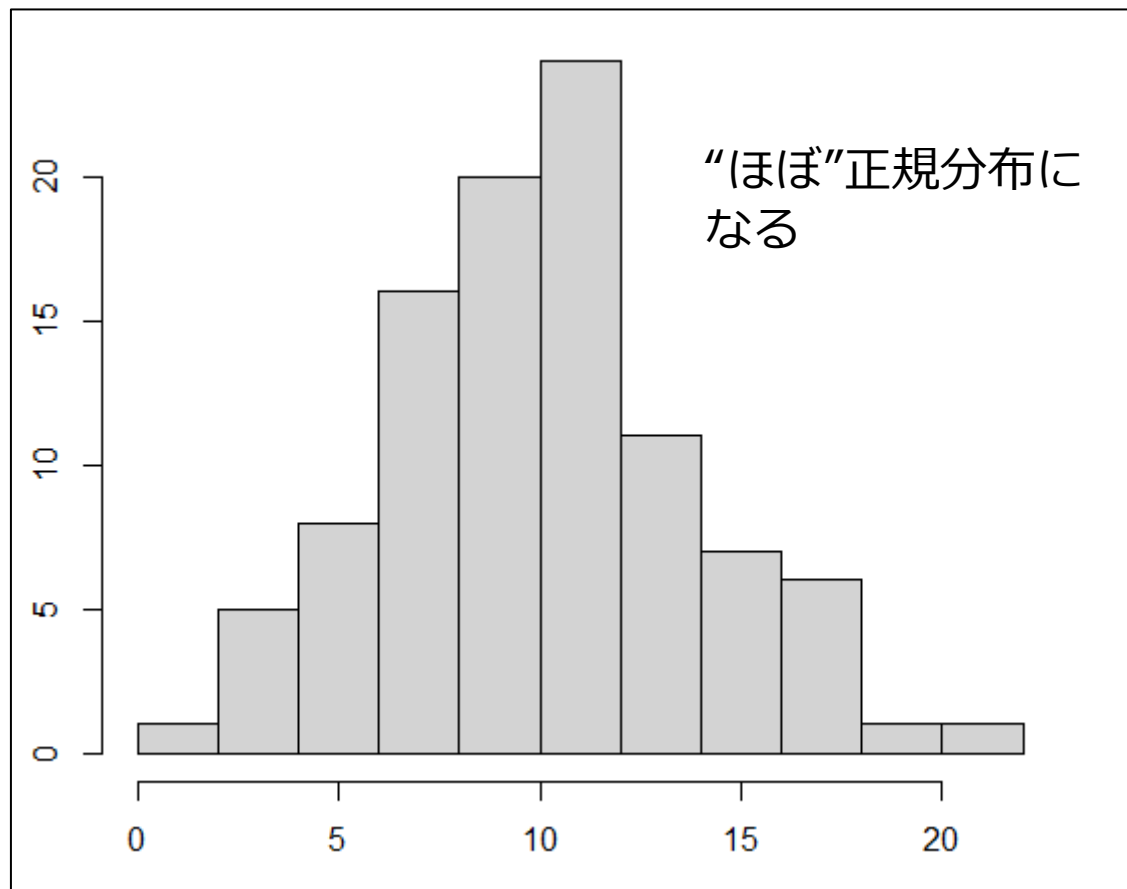
左の図は男女一緒に集計したもの。男女に色分けをすると2つの正規分布



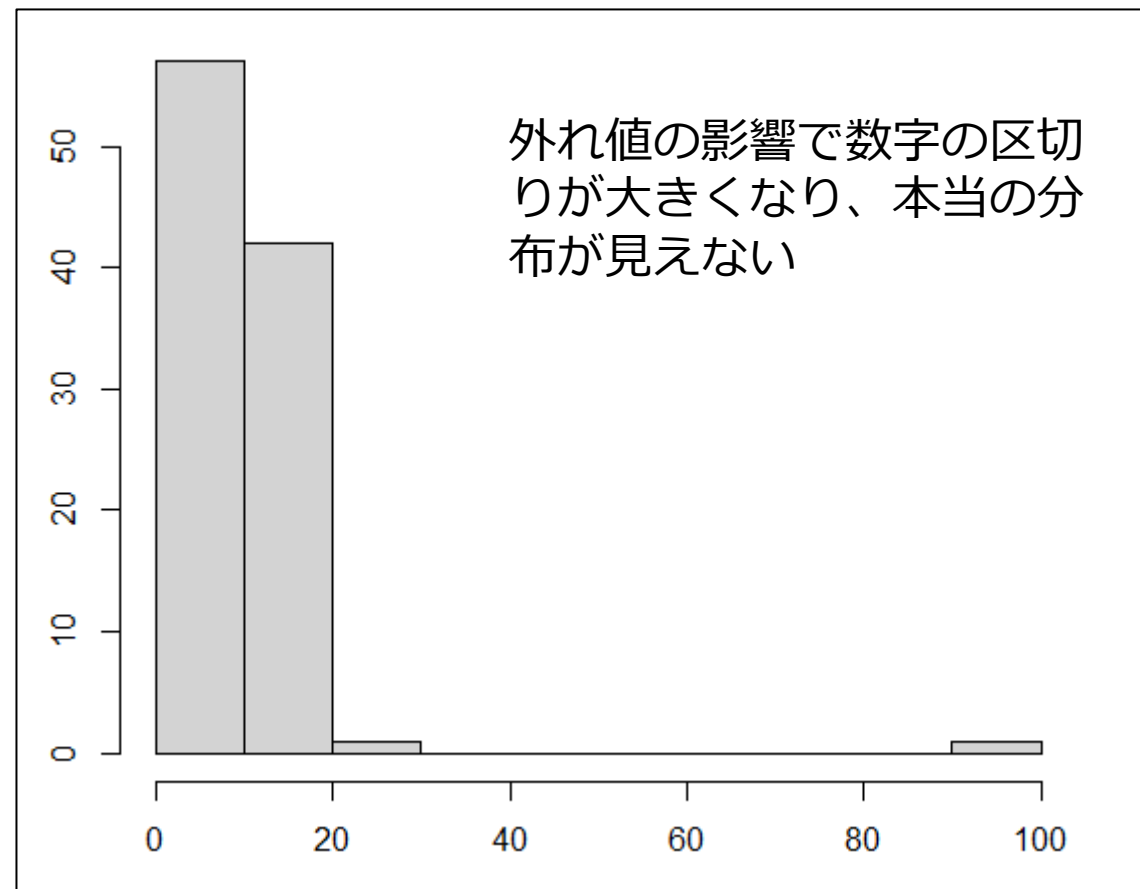
ヒストグラムの落とし穴

- ▶ 連続変数をどの“幅”で刻むか
- ▶ 区切り幅の選択次第で、ヒストグラムの形は大きく変わる
 - ヒストグラムの作成時は、いくつか区切り幅の異なるグラフを作成し、適切な区切り幅を選択
 - 恣意的にはならないように
- ▶ 統計ソフトの多くは、最大値と最小値から自動的に設定するアルゴリズム搭載
 - 自力で設定する必要性も高い(例：極端な外れ値がある場合)
- ▶ それなりに“センス”を要します(私にあるかは謎)

平均値10、標準偏差4の乱数を100個作成

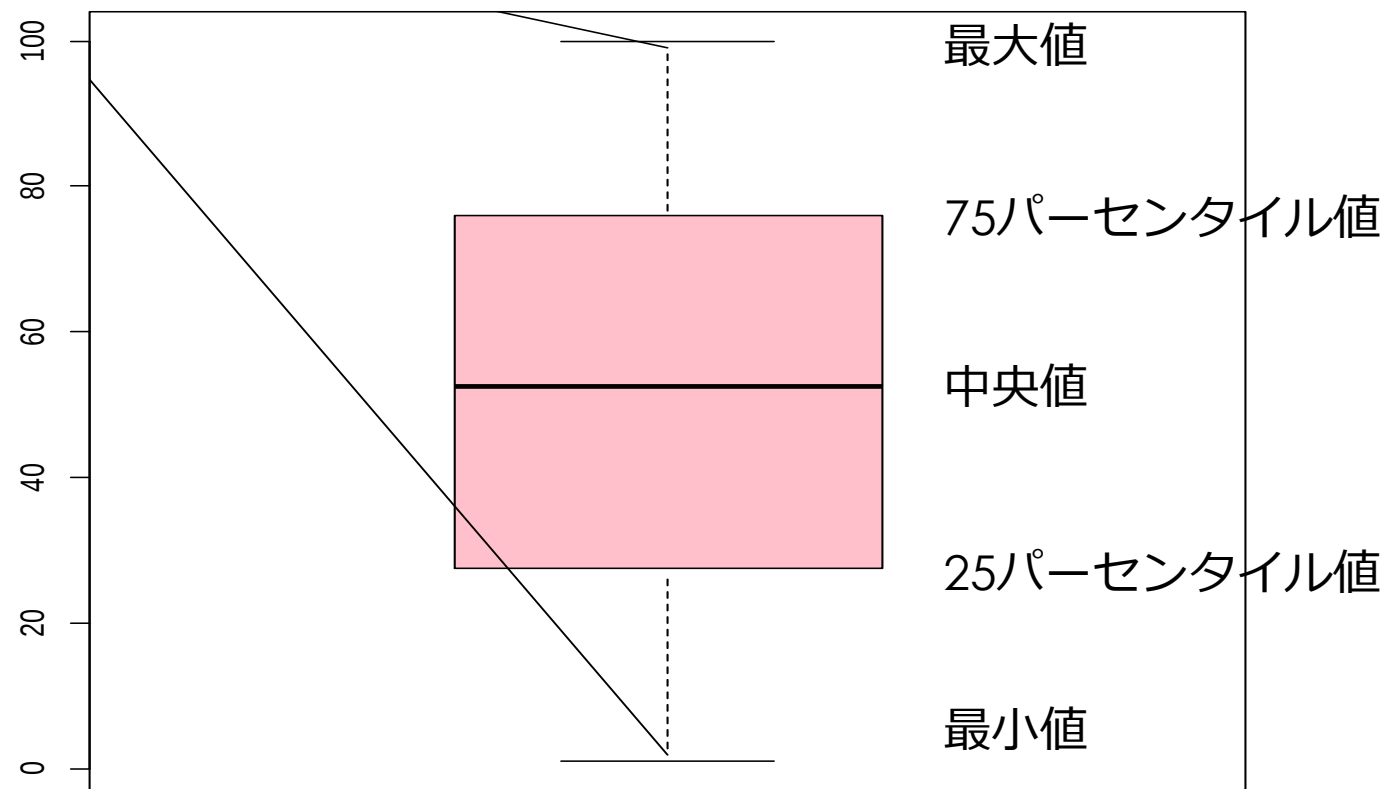


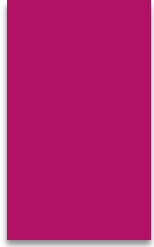
左の乱数に外れ値(100)を追加



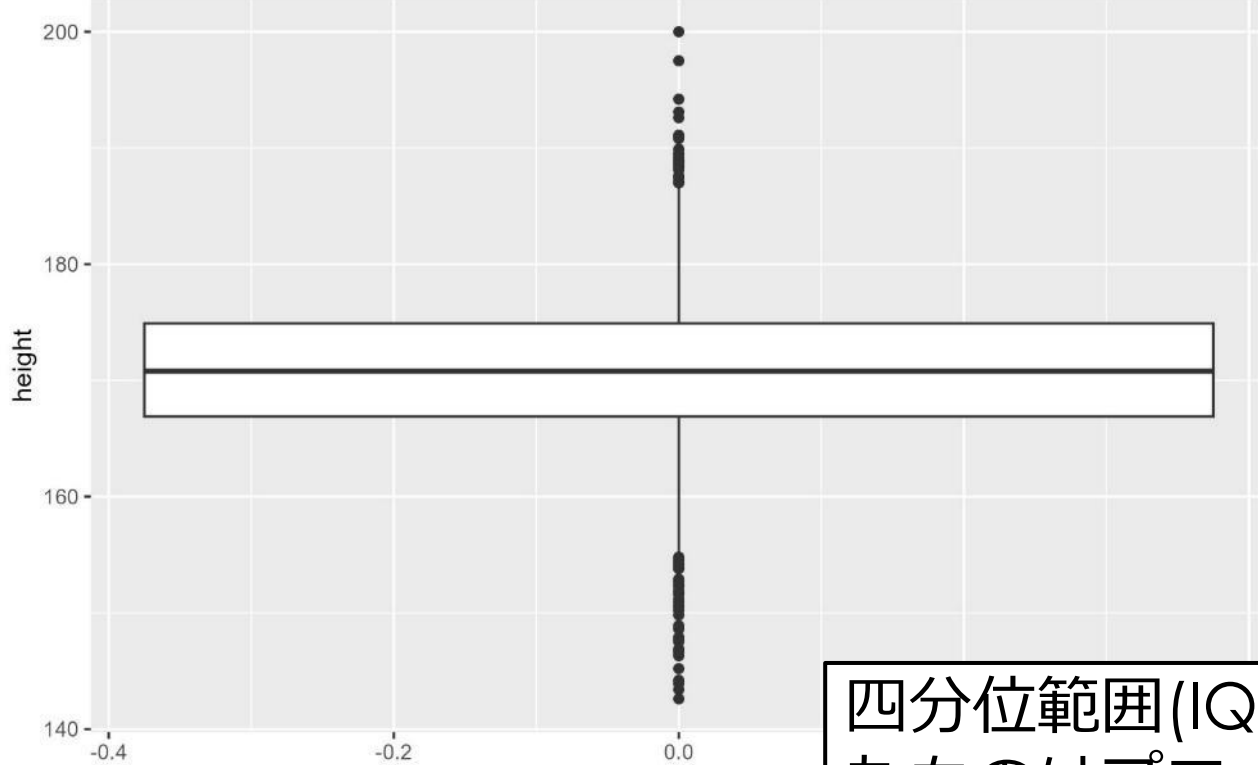
箱ひげ図

箱ひげ図



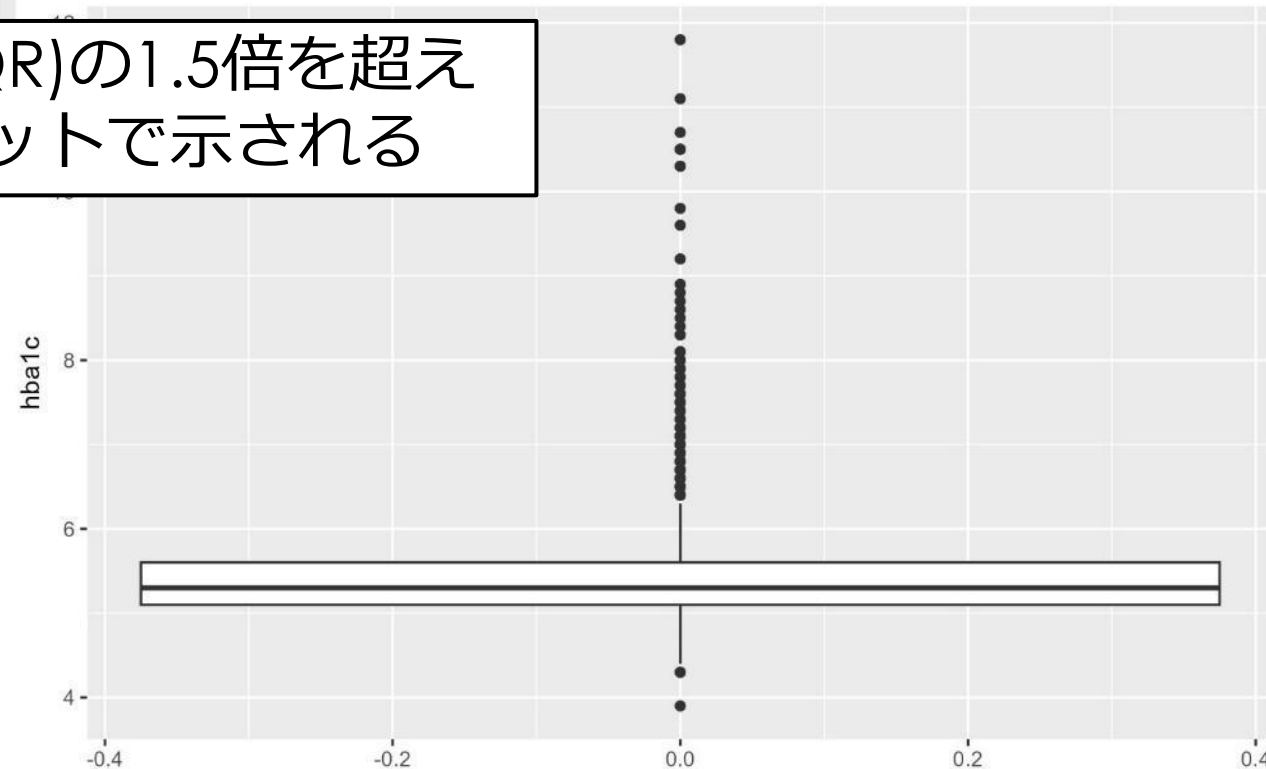


身長の箱ひげ図
ほぼ上下対象な形



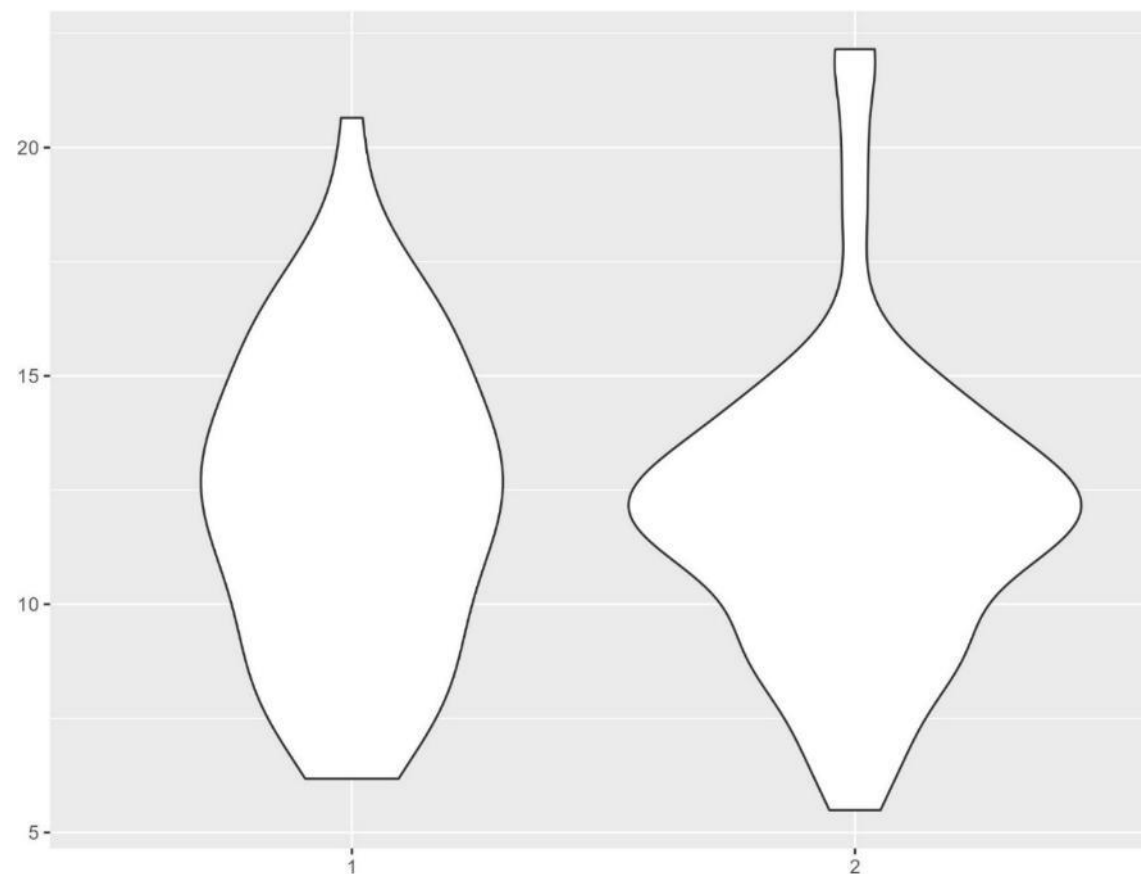
四分位範囲(IQR)の1.5倍を超えたものはプロットで示される

HbA1cの箱ひげ図
中央値は箱の中央より下に位置
しており、上のプロットが多い

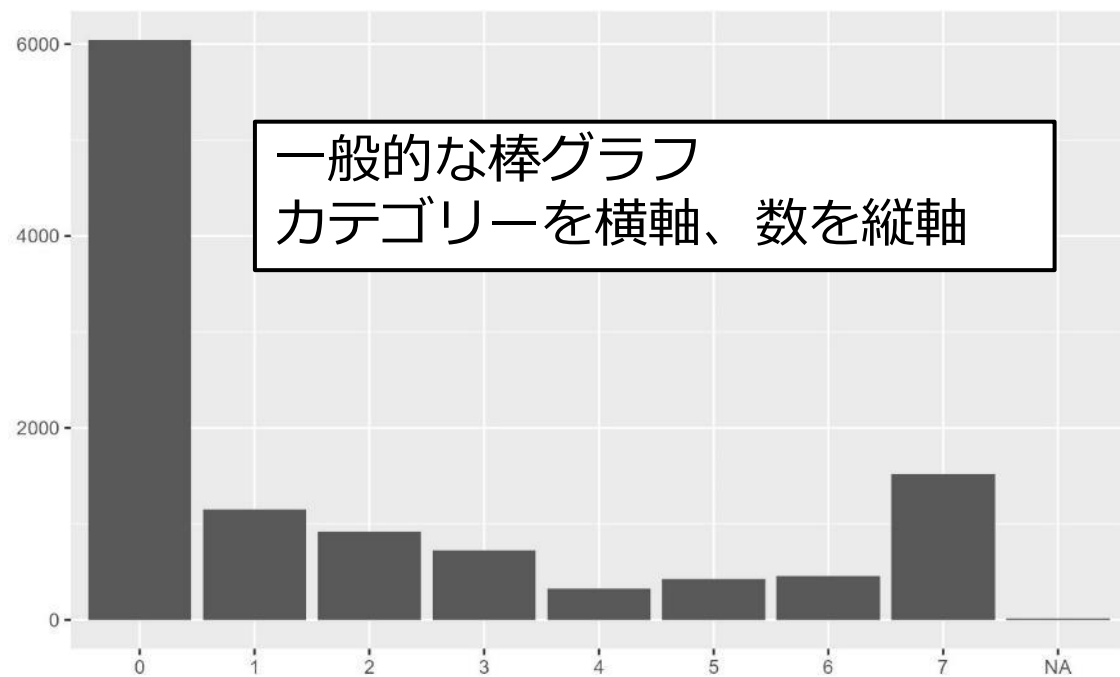


バイオリンプロット

- ▶ 箱ひげ図は、四分位範囲、外れ値の確認に向いているが、ボリュームゾーンがわかりにくい
- ▶ バイオリンプロットはボリュームゾーンの把握に適している(論文ではほぼ見られない)



棒グラフ



ヒストグラムと棒グラフ

- ▶ ヒストグラム：連続的に分布しているデータを、範囲ごとに区切って、それぞれの区分毎の集団の大きさを示す。
- ▶ 棒グラフ：あらかじめ決められているカテゴリに分類された集団の大きさを示す
- ▶ X軸の示すものがすでに決まっているか、決まっていないか

グラフを作成するには

- ▶ 当然ながら、記述統計も確認

平均値±標準偏差

中央値およびパーセンタイル値

最大値、最小値(特に外れ値の存在)

- ▶ グラフ“だけ”でも記述統計“だけ”でもデータの本当の姿はわからない

→併用することで十分な理解へ

二変数のグラフ

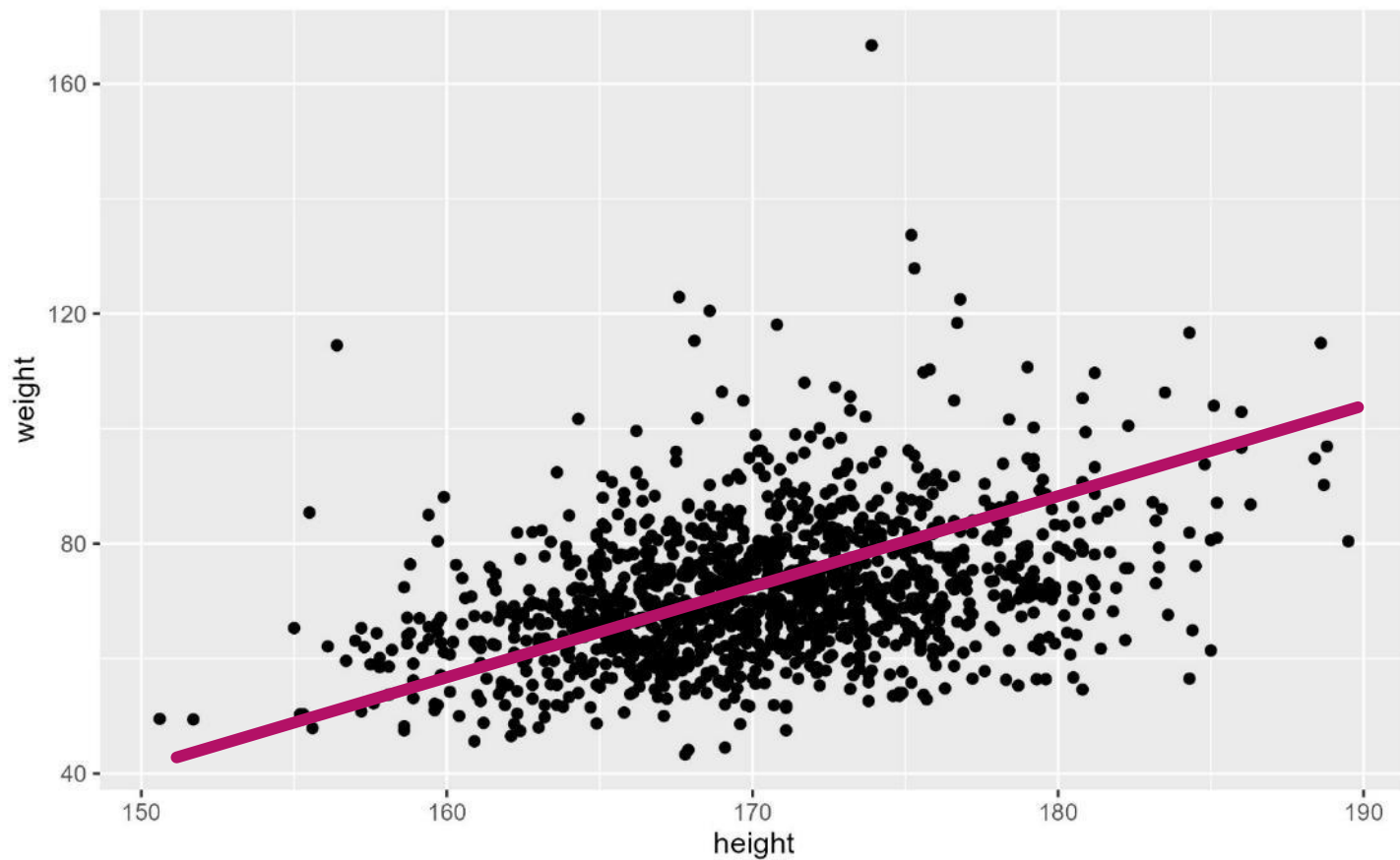
2 種類のデータ同士の関連を知る

データ形式同士

		X軸		
		連続変数	順序変数	カテゴリ変数
Y軸	連続変数	散布図	箱ひげ図 棒グラフ 折れ線グラフ	箱ひげ図 棒グラフ 折れ線グラフ
	順序変数		箱ひげ図 棒グラフ 散布図(バルーンプロット)	
	カテゴリ変数			

連続変数×連続変数の場合、片側を順序変数、
カテゴリ変数に変換する場合も多い

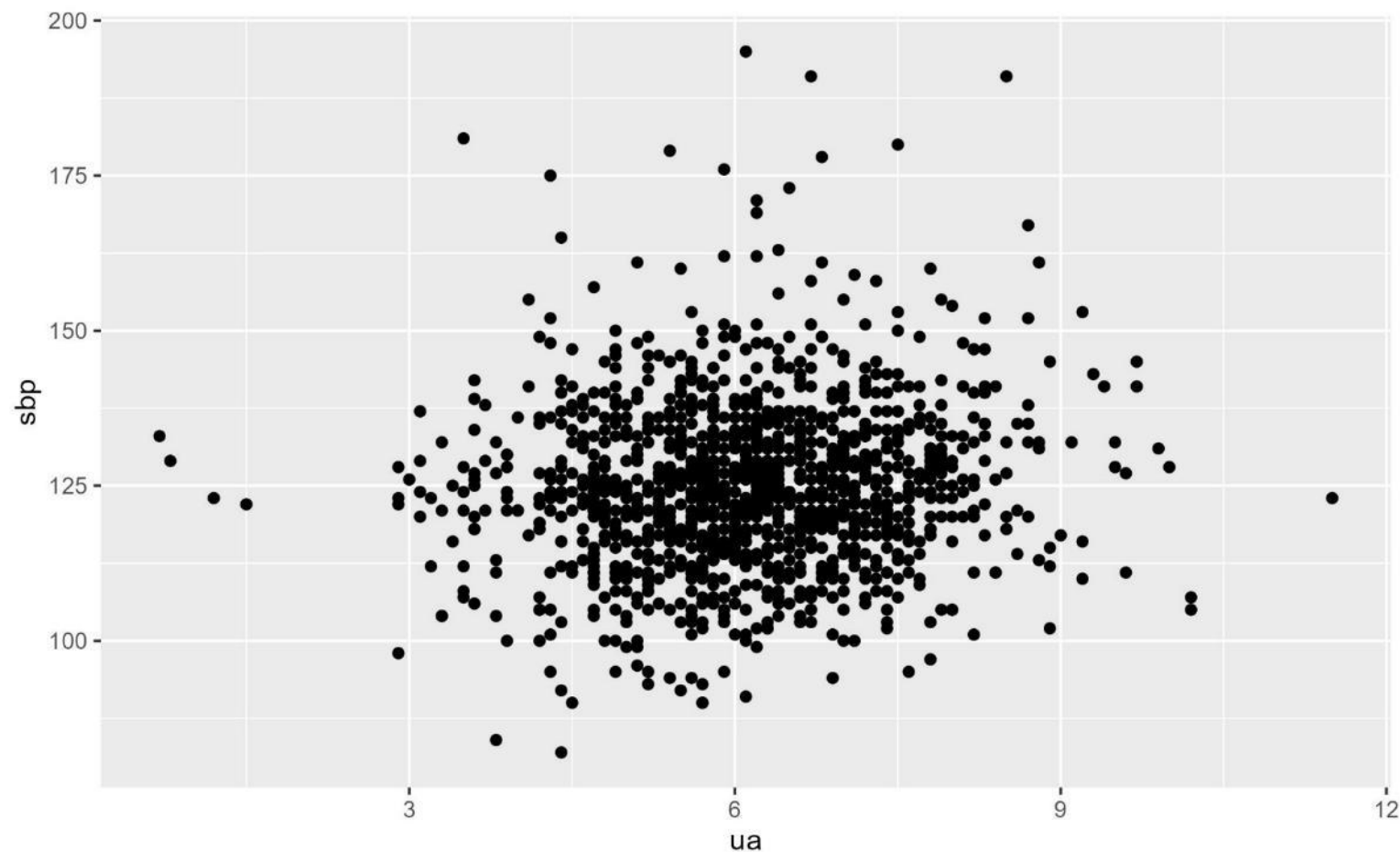
散布図



身長と体重の散布図
概ね相関する

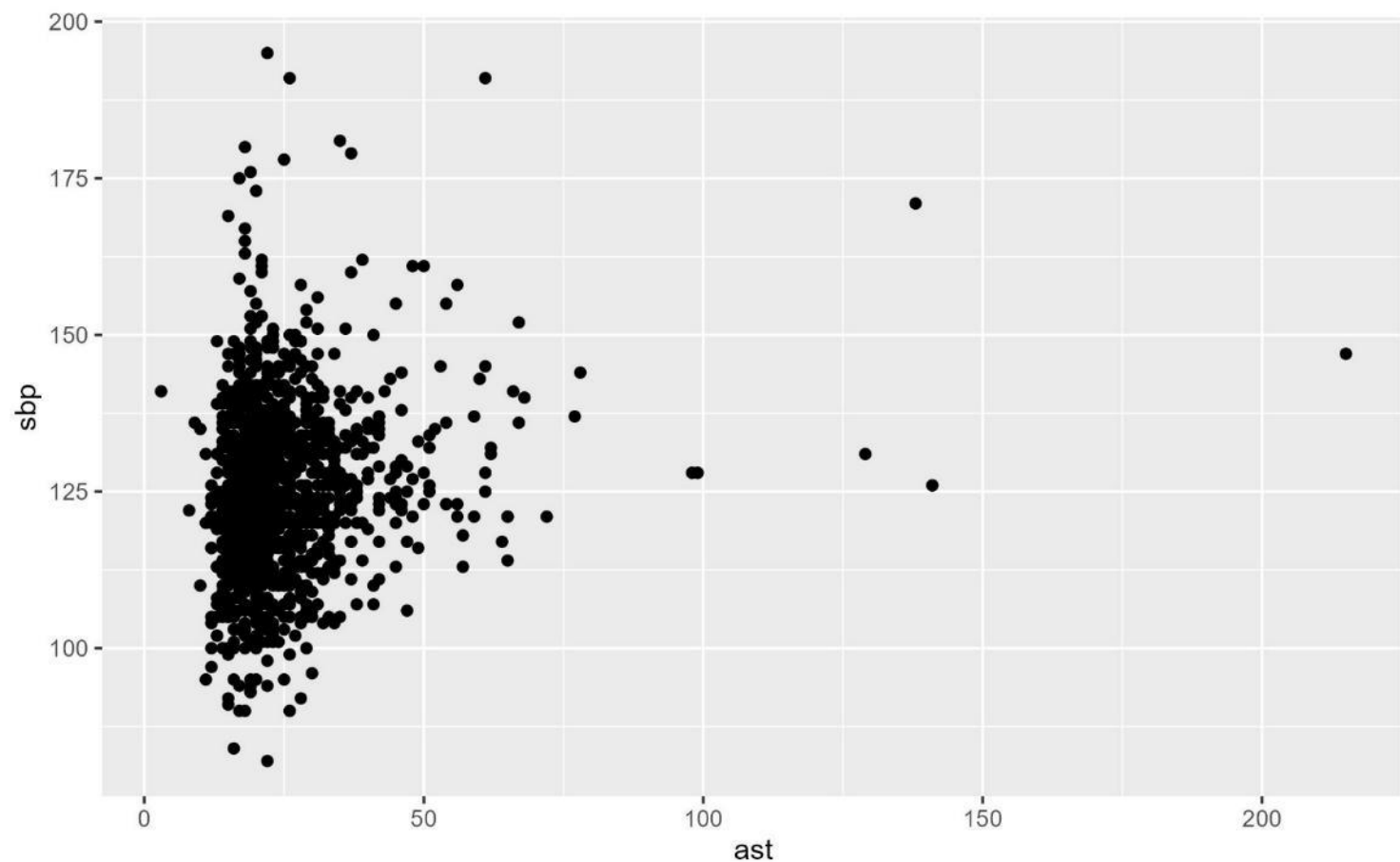
点が密集するところの
密度で相関しているか
どうか推定できる

相関がほとんどない散布図



尿酸と収縮期血圧の散布図
ほぼ同心円状に分布

こんな散布図も



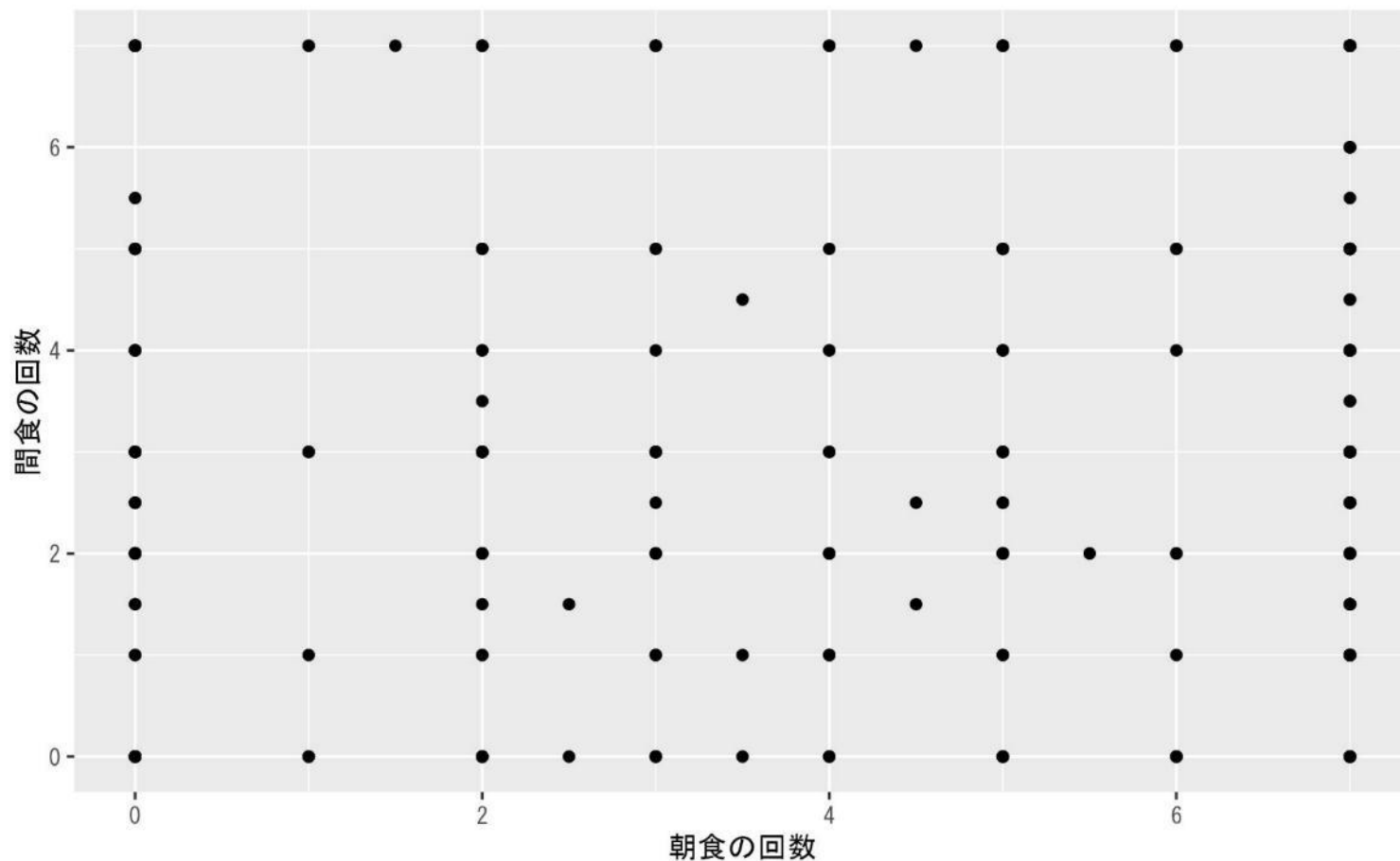
ASTと収縮期血圧の散布図

半円型の分布？

散布図を描くことで

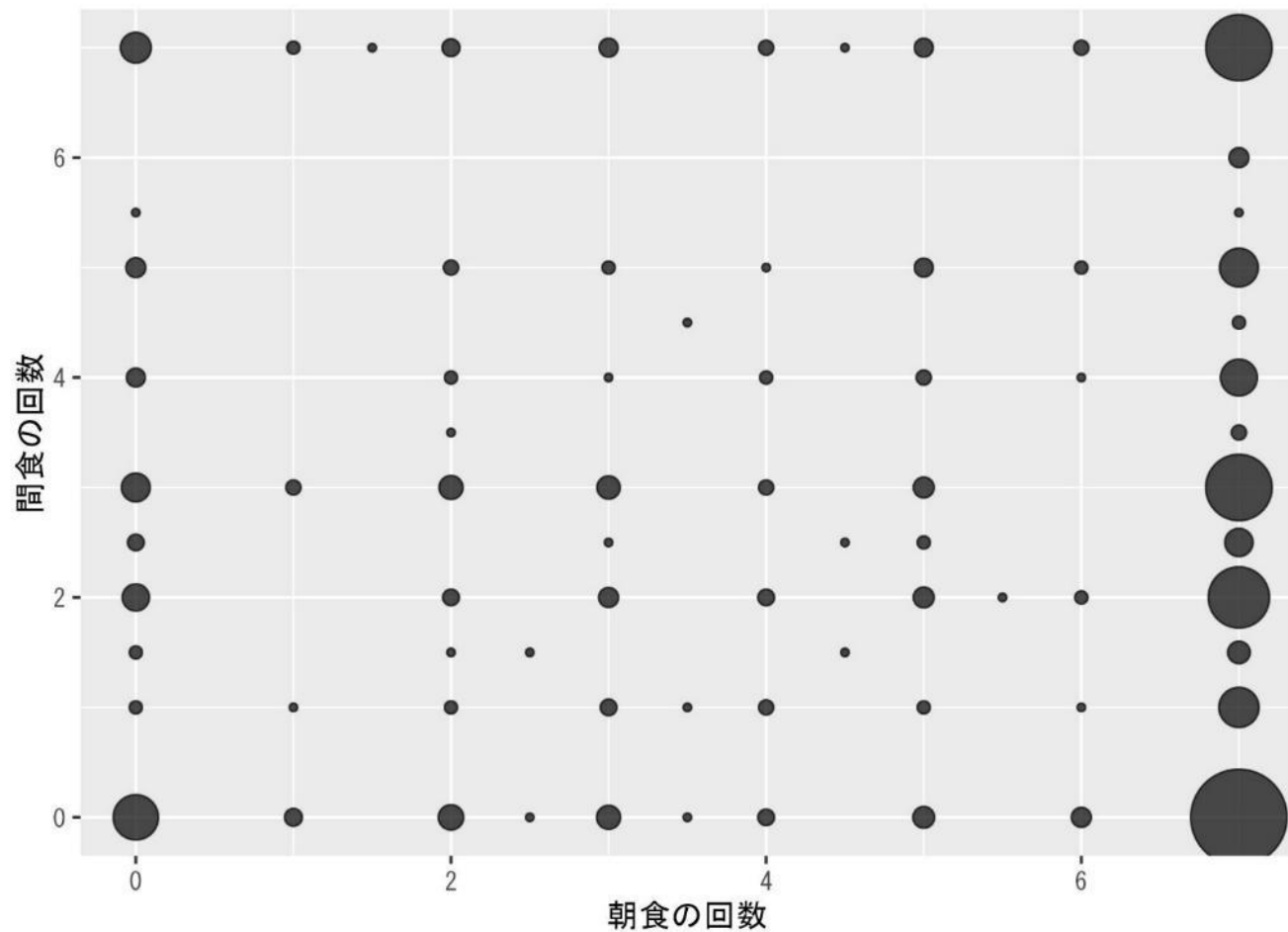
- ▶ 一方の変数の分布もおぼろげに見えてくる(濃い場所はn数多い、など)
- ▶ ただし、ボリュームゾーンの細かいことまではわかりにくい
- ▶ あくまでも“2変数”の関連を見ることが目的なので、各々の分布については、ヒストグラム等で確認しましょう。一気に変数同士の関連に行かないように

カテゴリ変数同士の散布図



カテゴリがそんなに多くない場合、格子状に配置され、どの数値がおおきいのか、ボリュームゾーンが分かりにくい

バルーンプロット



バルーンプロットで、個数を点の大きさで示すことにより、分布を可視化できる。

count

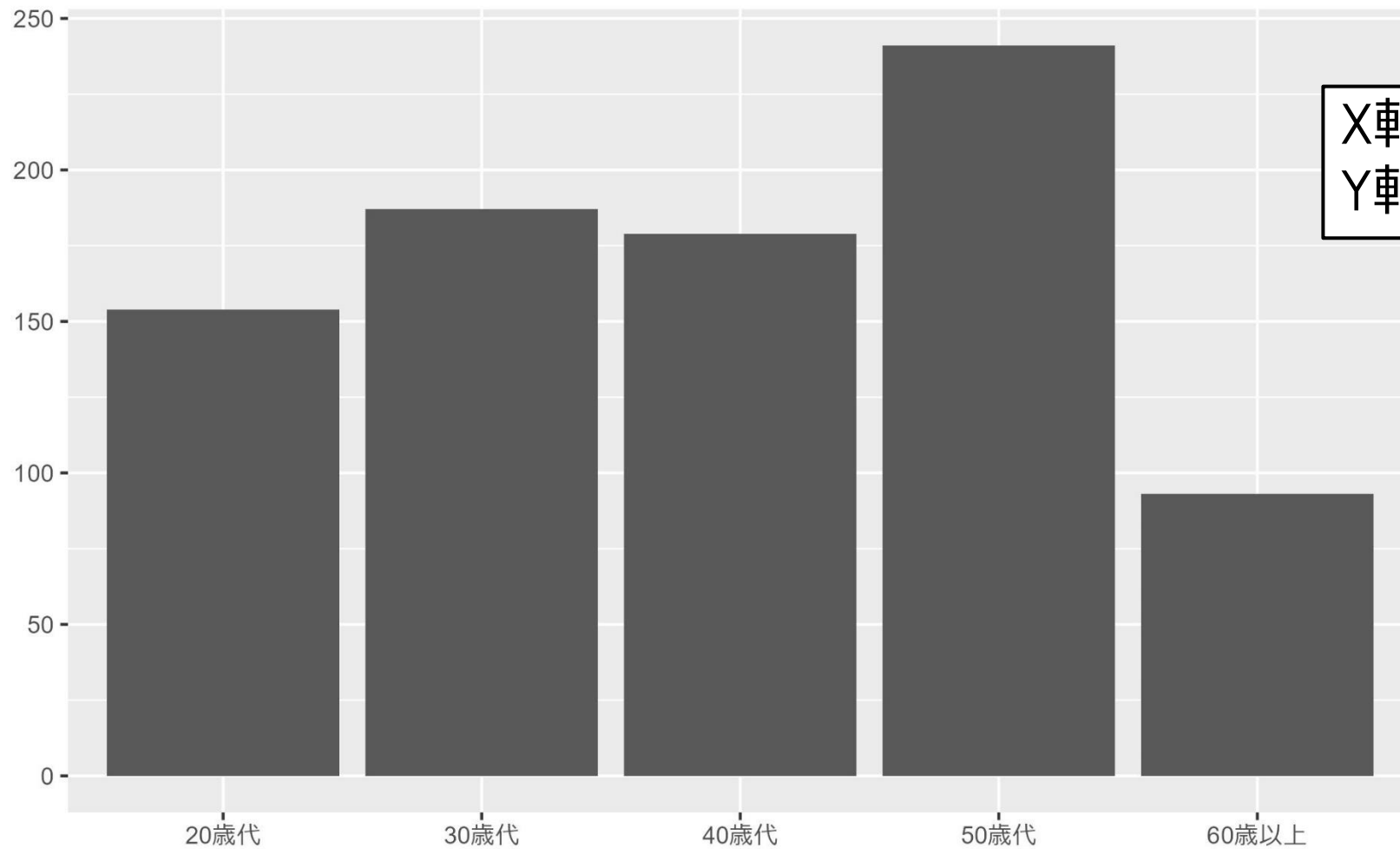


そんなことしなくても

- ▶ 順序変数×順序変数、カテゴリ変数×カテゴリ変数の場合、グラフで可視化するよりも度数分布表を作った方がはやい(かつ分かりやすい)
- ▶ 「まずグラフで分布をみなきゃ」でなく、統計量だけではわからない分布をみるためにグラフを使う

手段と目的を逆にしないように

棒グラフ



X軸に回答者の年齢(10歳刻み)、
Y軸に人数

積み上げ棒グラフ

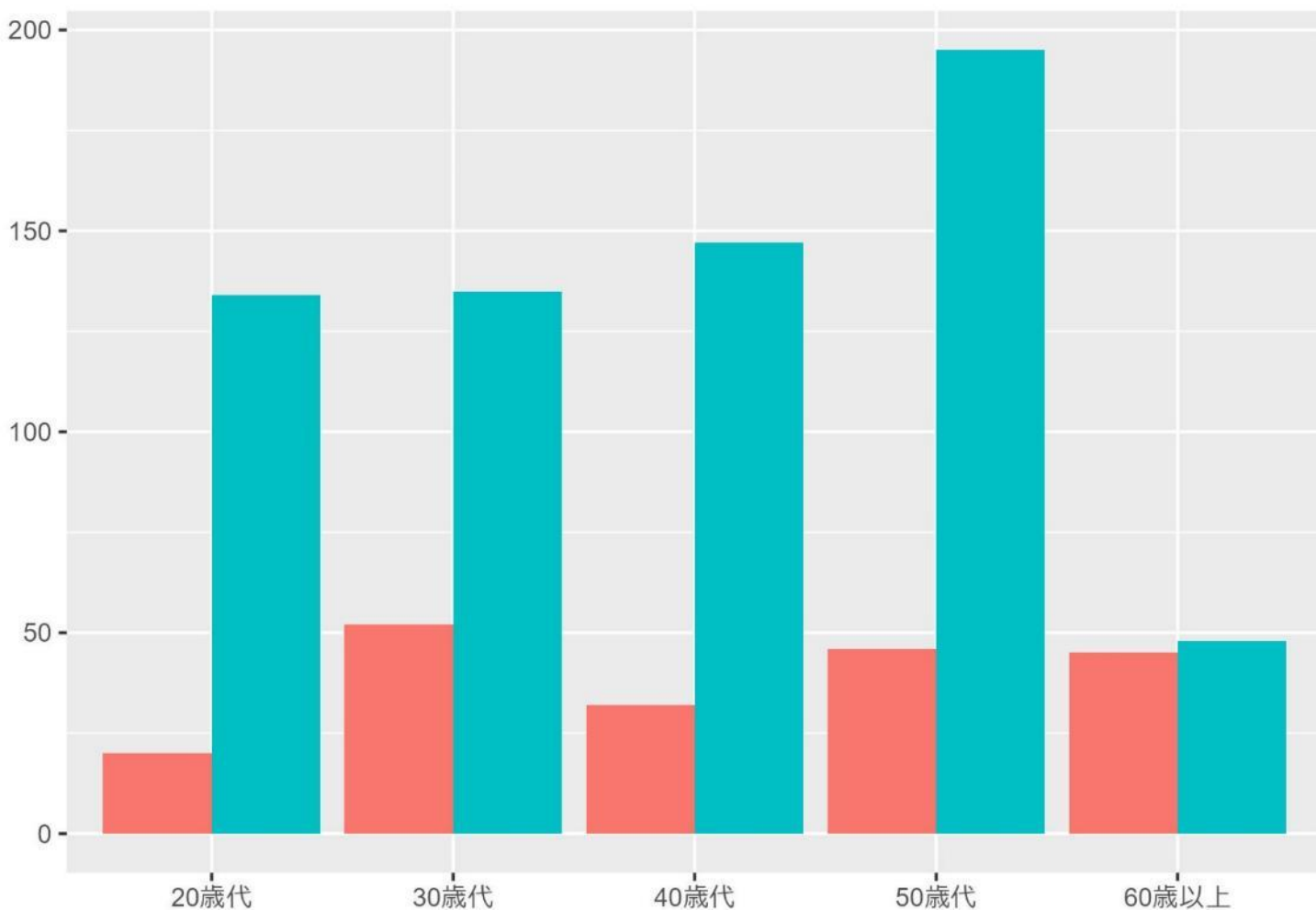


カテゴリごとに縦に積み上げることで、全体の中での割合を視覚化できる

ageg



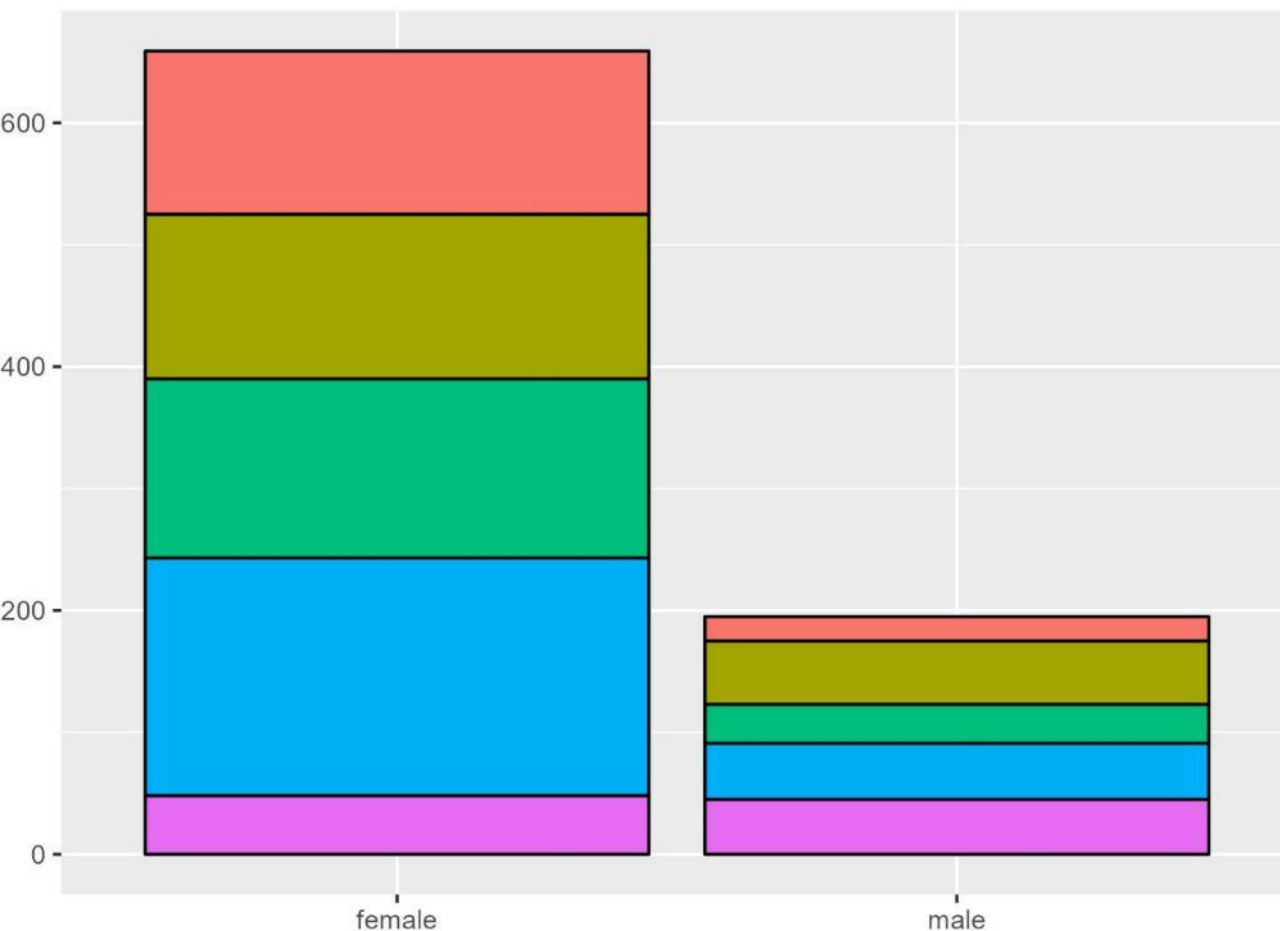
棒グラフ同士の比較



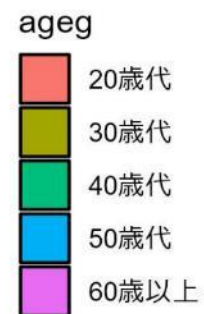
棒グラフはこのような異なる属性
同士の比較を行うことも可能

男性
女性

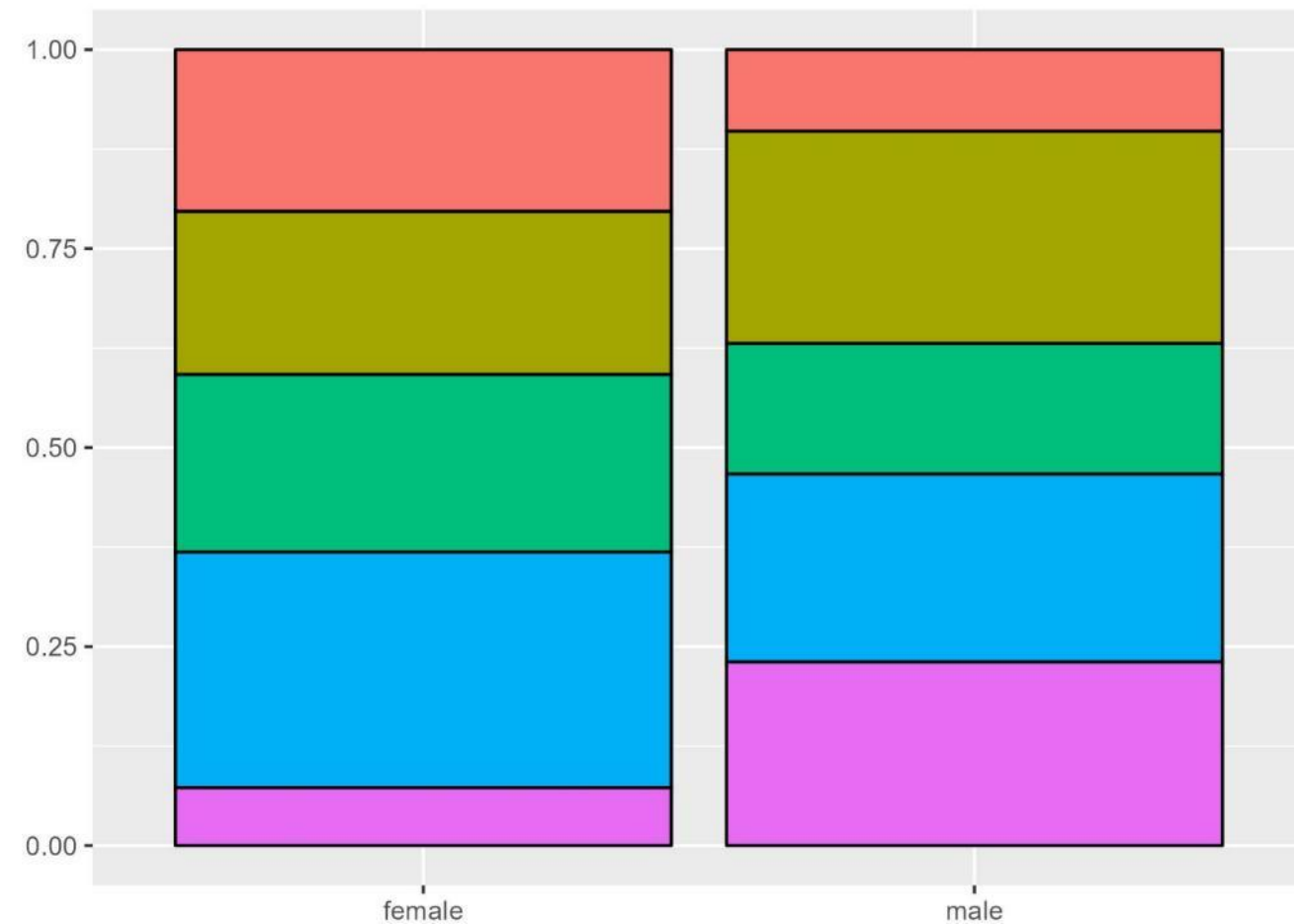
積み上げ棒グラフ同士の比較1



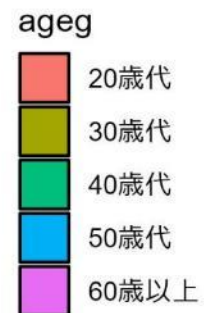
積み上げ棒グラフ同士を比較する場合、属性ごとの人数が大きく異なると割合の比較が困難になる



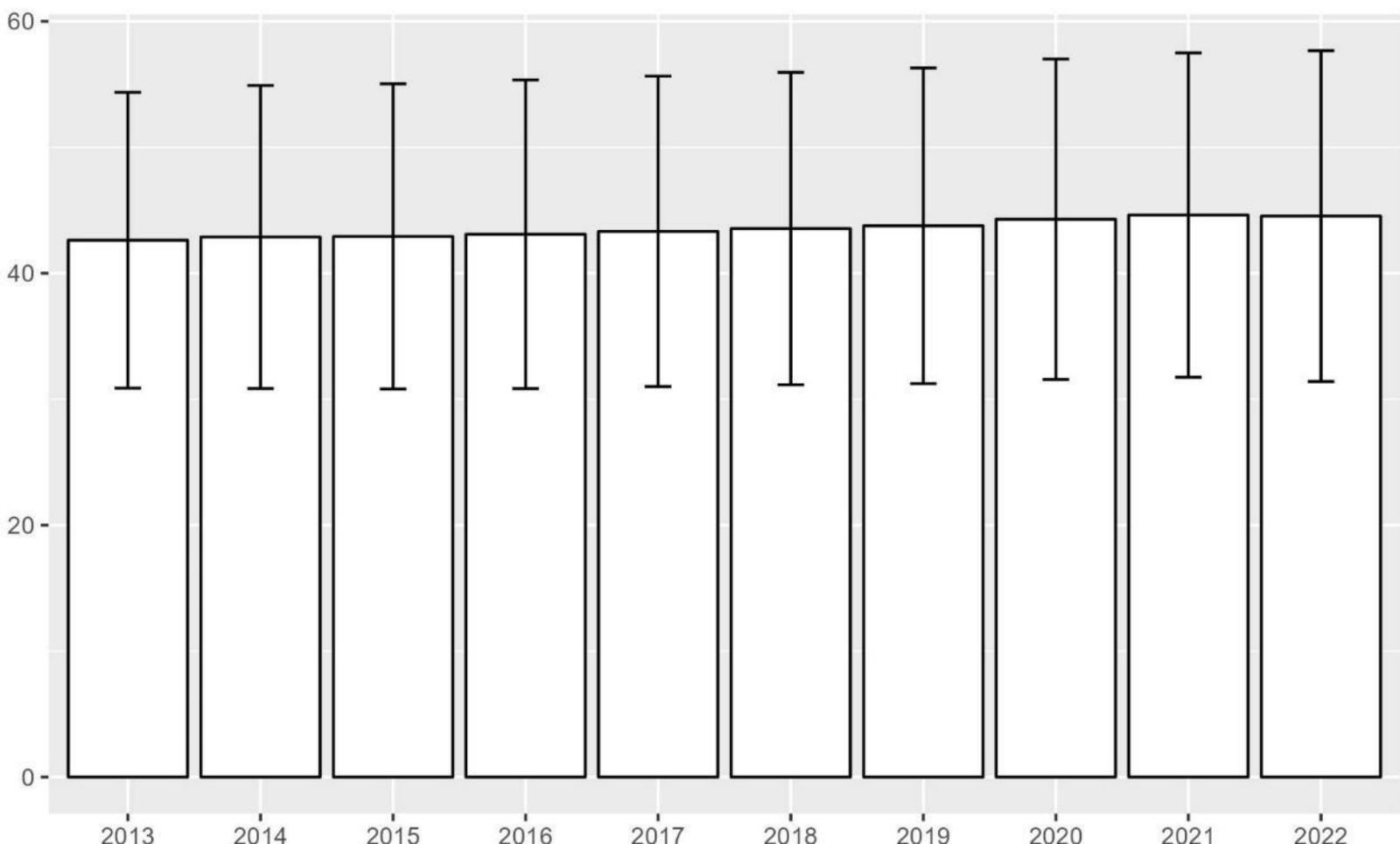
積み上げ棒グラフ同士の比較1



このように縦軸を100%にすることで、母数が異なる属性同士の比較が可能となる

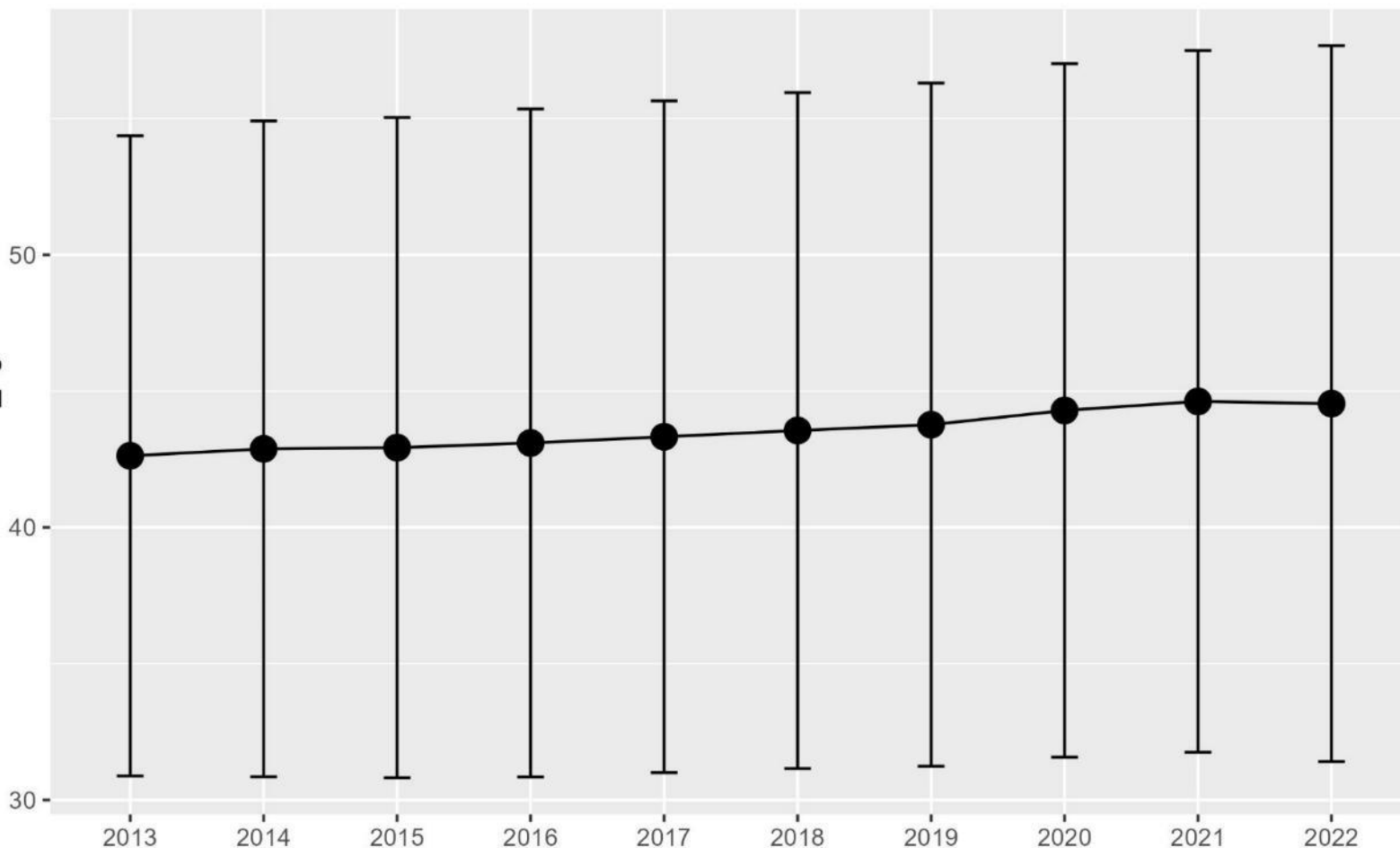


経年データの示し方



とある健診機関の年別受信者の平均年齢と標準偏差
エラーバーを上のみ記載する人も多いが、標準偏差は $\pm$ (プラスマイナス)で表記することを考えると、図に描くときも、上下にエラーバーを伸ばすのがいいか？

経年データの示し方



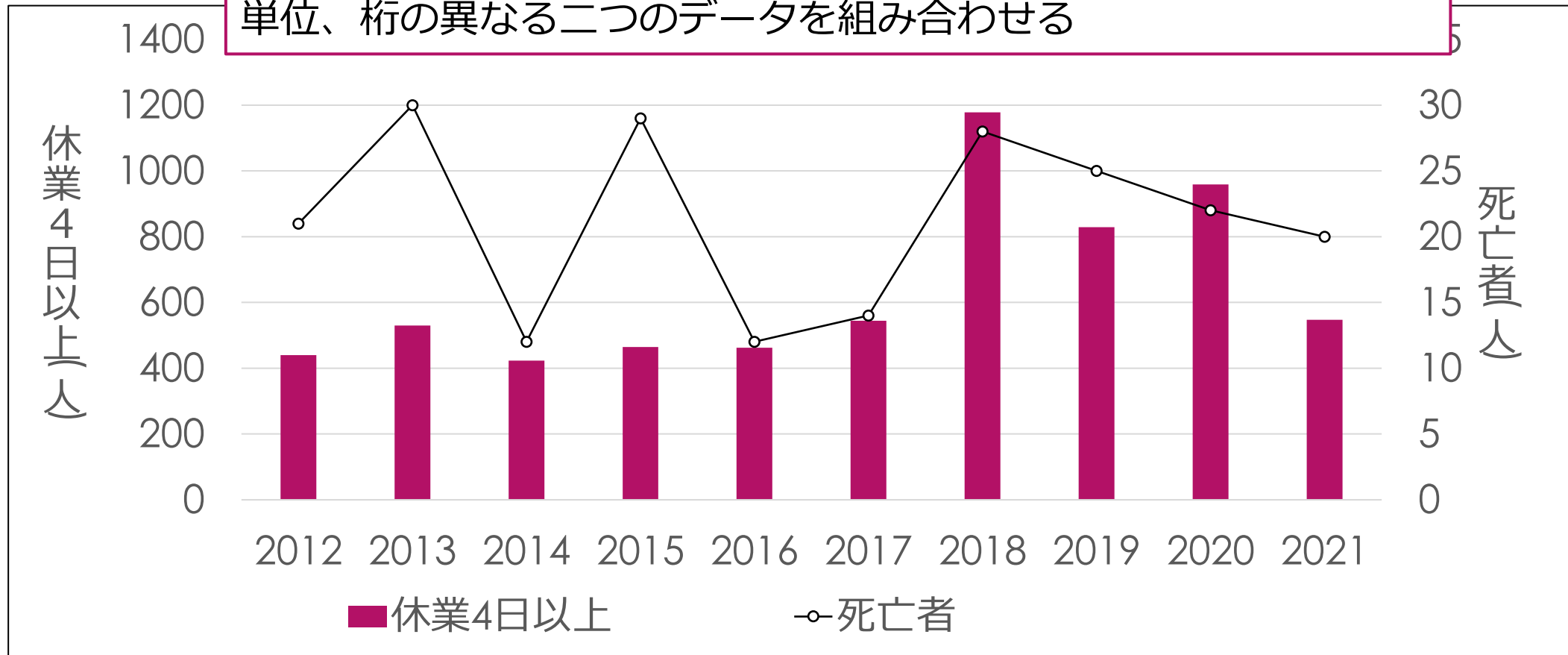
経年(経時)データは折れ線グラフで示すことも多い。
分布ではなく、平均値(標準偏差)の比較に意味があるようなときは、ヒストグラム、箱ひげ図よりもこちらを用いる。

棒グラフと折れ線グラフの組み合わせ

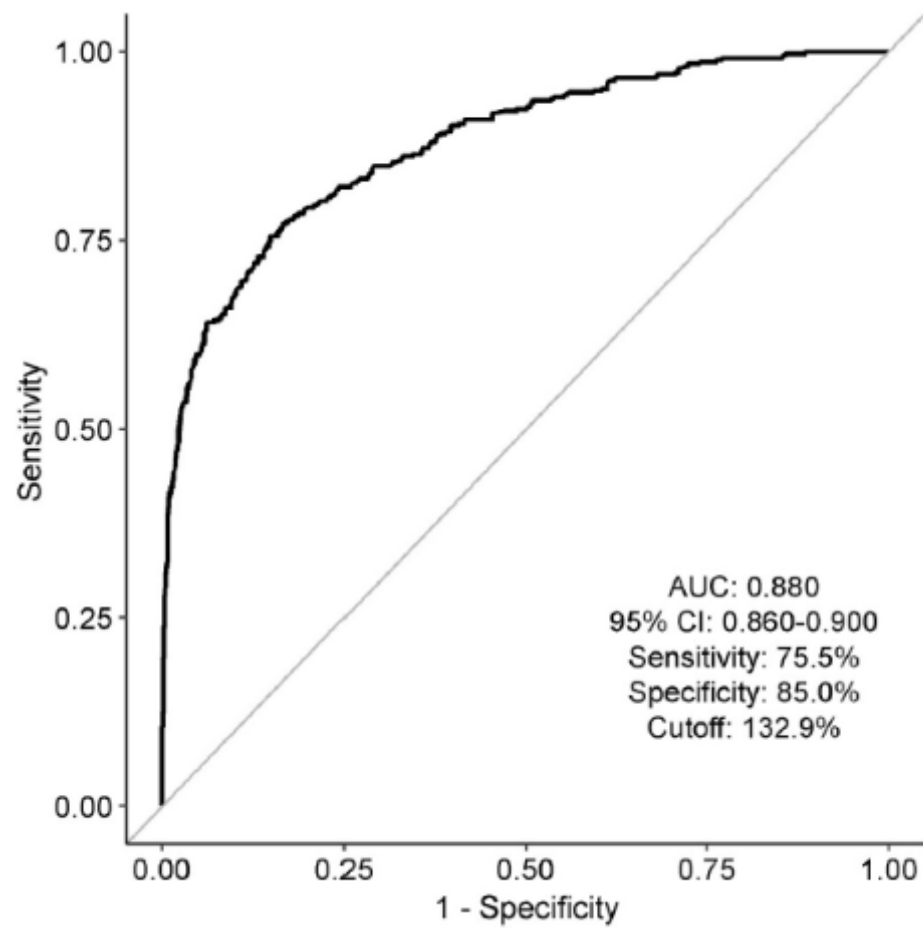
熱中症の労働災害発生

左が棒グラフで休業者、右が折れ線グラフで死亡者

単位、桁の異なる二つのデータを組み合わせる



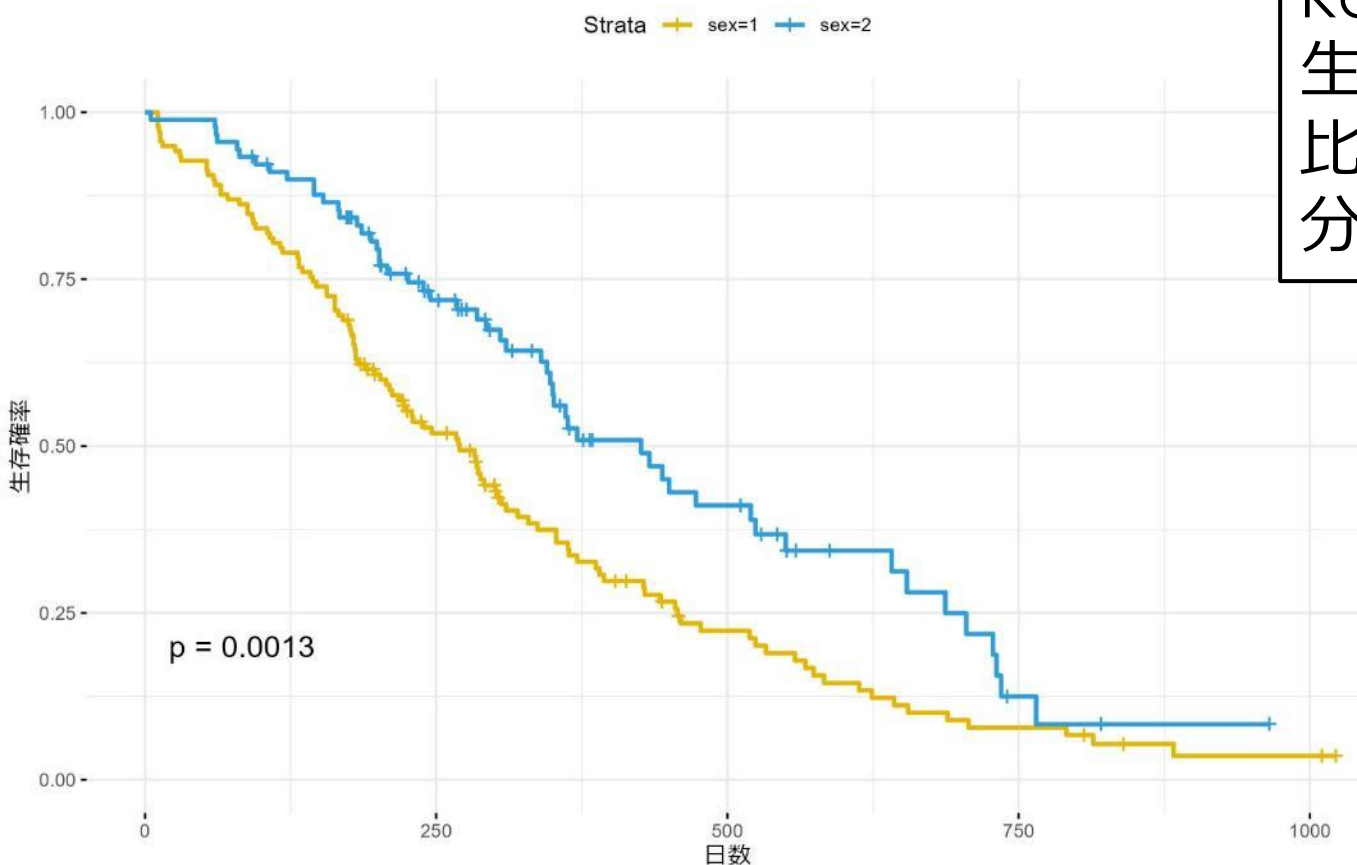
その他のグラフ



ROC(receiver operation curve)
曲線。
検査などの基準値(カットオフ値)を
決める際に用いる。

その他のグラフ

性別で層別したKaplan-Meier曲線



Kaplan-Meier曲線。
生存時間の追跡に用いる。群別の
比較にはLog-rank検定、Cox回帰
分析(多変量解析)を使用する。

グラフで説明する

“見せる”ではなく“魅せる”

グラフ作成について参考になる論文

- ▶ “Development of a framework for graph choice and construction”
- ▶ Advances in physiology Education 40: 123-28, 2016

手順1：計画

図表で何を表現するか、どのようなメッセージを込めるか決定する

- ▶ 研究課題、仮説を明確にして何を図表で表現したいのか明らかにする
- ▶ 独立変数、従属変数を同定する
- ▶ 変数を分類、階層化する
- ▶ どのデータをグラフに取り入れるか決定する
- ▶ データを表現するのにどの図表を用いることが適切か決定する

手順2：実行

実際に図表を作成する

- ▶ 軸ラベルを決定する
- ▶ 必要であれば軸の編成を行う
- ▶ データを挿入するために必要な軸の大きさを決める
- ▶ 適切な単位等を挿入する
- ▶ 図の説明と共に論文に図表を組み込む
- ▶ 図の内容を表すタイトルをつける

手順3：推敲

技術的に図表の内容を検討する

- ▶ 仮説を元に図表の体裁を整理する
- ▶ 図表の持ち味を考える
- ▶ 図表の弱点を考える
- ▶ 図表の主要メッセージを考える
- ▶ 現在採用した図表以外の表現方法を考えてみる

図表タイプ	用途	利点	欠点
棒グラフ	割合、記述集計(エラーバー含む)の比較	多集団データを見やすくする 積み上げ棒グラフは同一集団内複数レベルの比較に便利	データの細かい数字が曖昧になる 分布が分かりにくい
箱ひげ図	記述集計(パーセントイル表記)の比較	他集団データの比較と記述に便利 分布の偏り、外れ値が分かりやすい	小集団には使用できない
ヒストグラム	連続変数の分布様式の記述	連続変数の分布形態および頻度を見るために便利	分割サイズの決定(重要情報の記載)が研究者に委ねられる

図表タイプ	用途	利点	欠点
ヒストグラム	連続変数の分布様式の記述	連続変数の分布形態および頻度を見るために便利	分割サイズの決定(重要情報の記載)が研究者に委ねられる
折れ線グラフ	値の変化を記載、比較	時系列データの推移を視覚化しやすい	複数データを記載するときにその関連がわかりにくい
ドットプロット	小集団の分布様式を個人ごとにプロットしての記述	小集団の分布を比較するために便利	大集団になると点の重なりが多くなるので表記できない
散布図	2変数の関連を示すデータ分布	2変数の関連性、傾向を確認しやすい	同じ箇所に複数の方がいる場合に表記しにくい

学会と論文で使い分ける

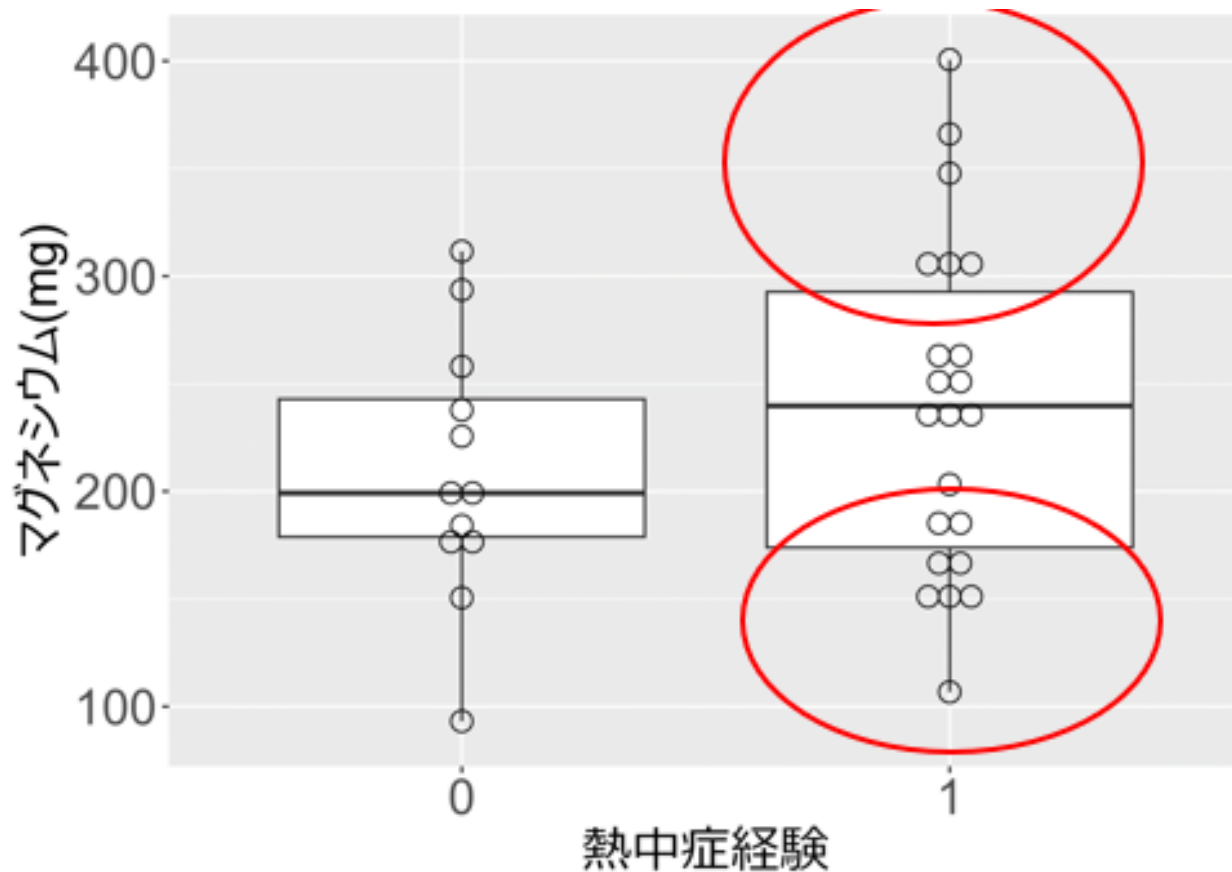
論文

- ▶ できるだけ装飾は少なくシンプルに
- ▶ 白黒になることも念頭に置きながら(グラデーションよりもパターンを使うなど)

学会

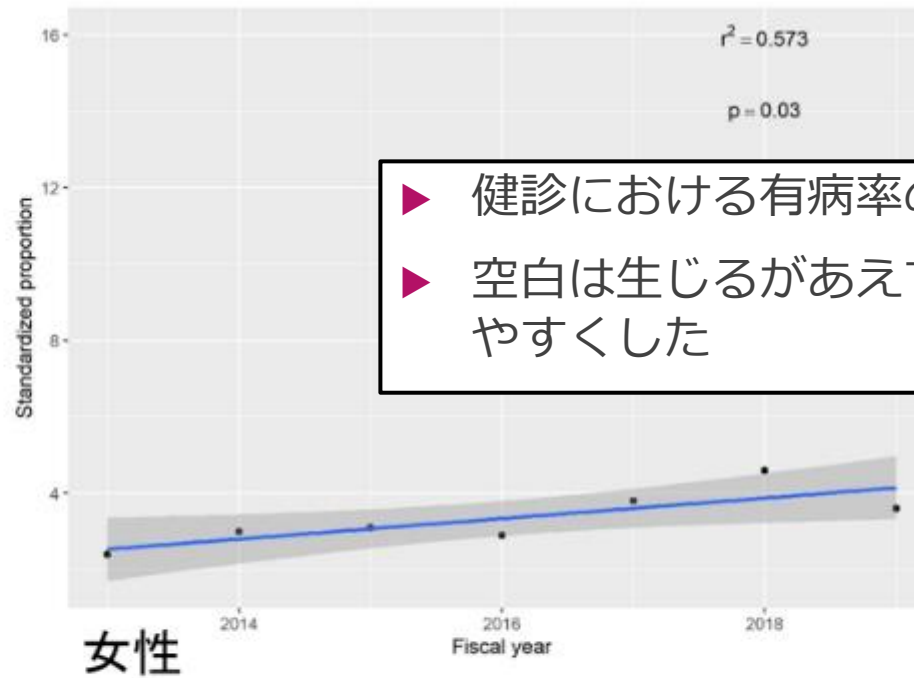
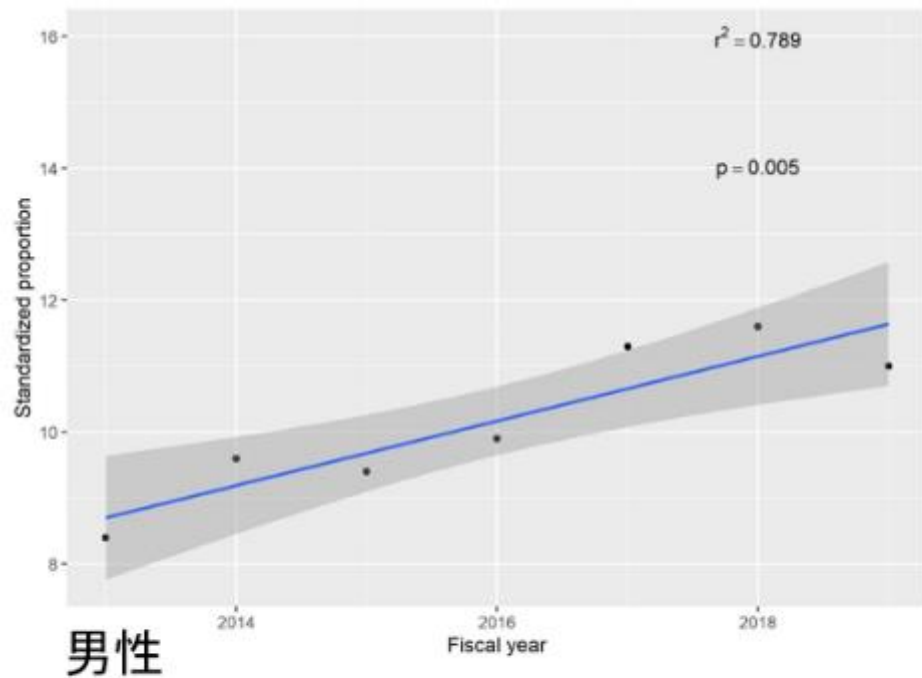
- ▶ 配色も効果的に用いる
- ▶ 人目を引くためにも、最も強調すべきところを分かりやすく

私の工夫事例1



- ▶ 労働中熱中症経験の有無と、摂取ミネラルの関連
- ▶ 箱ひげ図とドットプロットを組み合わせ、分布の幅をより視覚化

私の工夫事例2



- ▶ 健診における有病率の年次推移
- ▶ 空白は生じるがあえて軸を揃えて比較しやすくした

糖尿病有病率の年次推移グラフ
(青の直線は近似直線、グレーは95%信頼区間)

最後に

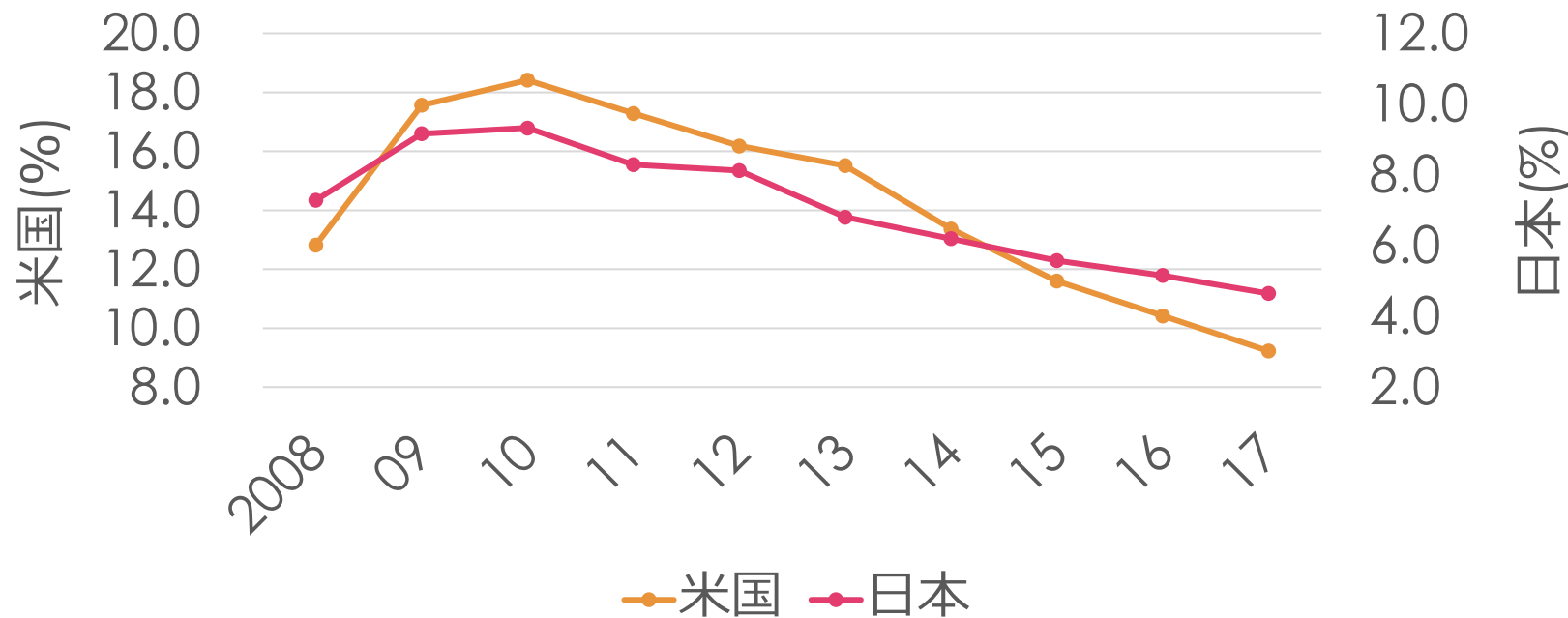
不誠実なデータの見せ方とは？

グラフは客観的に作成

- ▶ グラフは軸、単位などを変化させることで、いくらでも恣意的にデータを見せることができる。
- ▶ 本人にはそのつもりがなくても、ソフトで機械的に作成する際、アルゴリズムの影響を受けることも。
- ▶ グラフを作成して終わり、ではなく、グラフが客観的な基準で作成されているか、確認を忘れないように。

よく使われる“不誠実な”手法

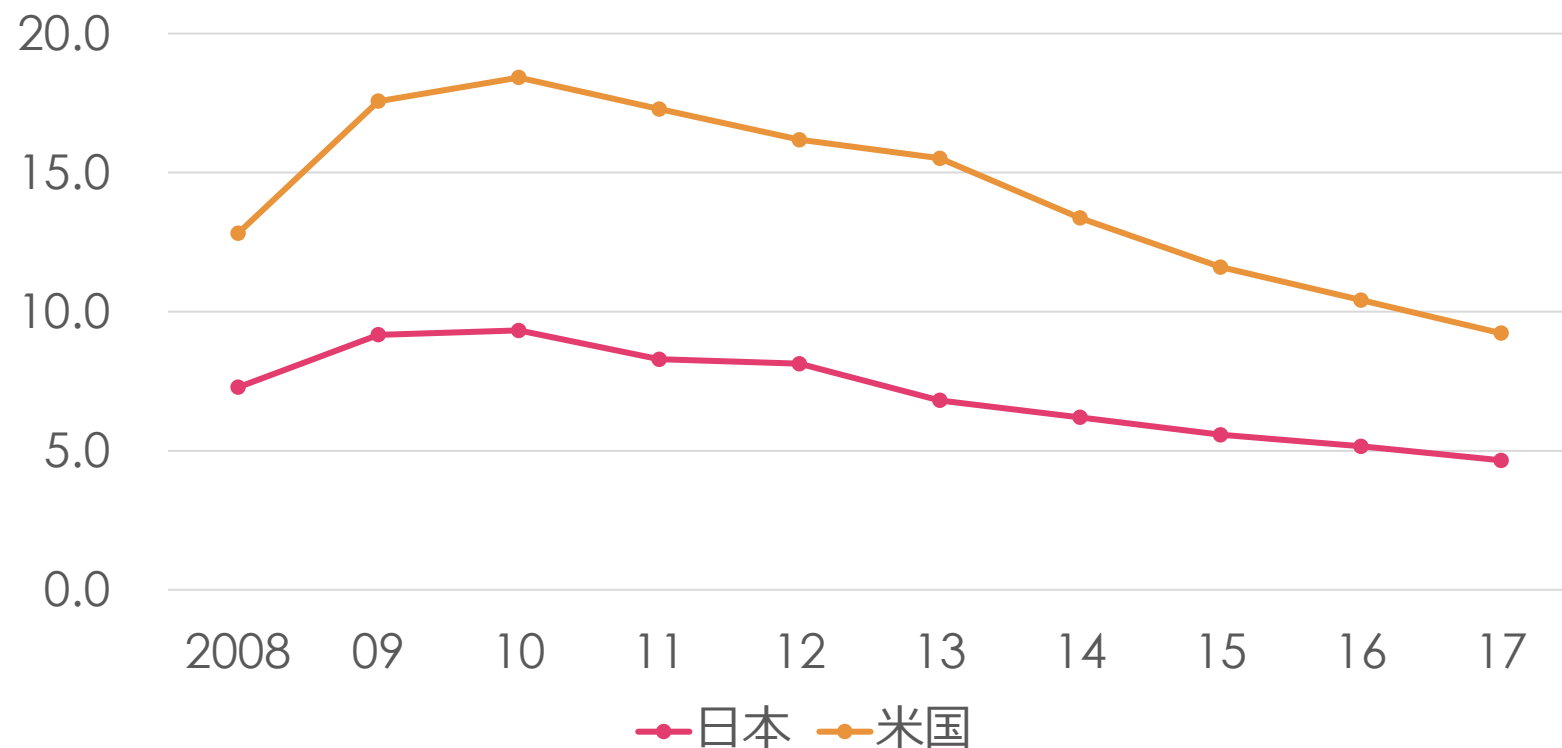
若者(15-24歳)の失業率推移



2009年のリーマン
ショックで一時的に失
業率が高くなった後、
アメリカの方が日本よ
り低下しているように
見えるが???

同条件の比較は単位をそろえる

若者(15-24歳)の失業率の推移



単位をそろえると、そもそものスタートラインが異なる。軸をいじくることでデータを不誠実に見せることも可能。

こんなことも

A党の支持率倍増！
一方B党は支持率
10ポイント減！

政党支持率

5%

50%

A党

B党

20△年



A党

B党

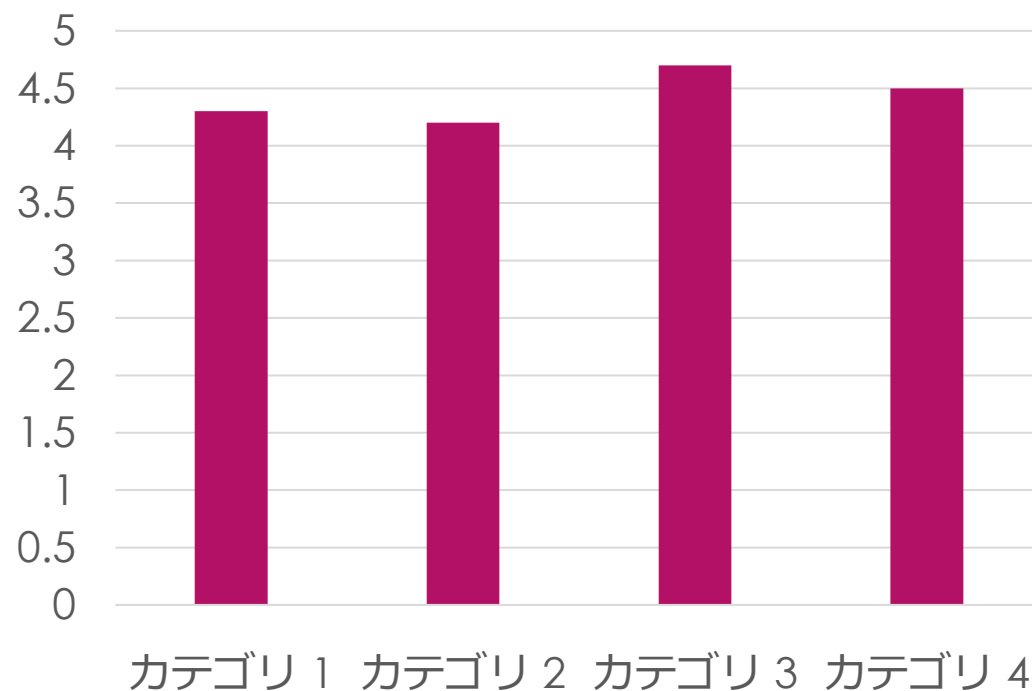
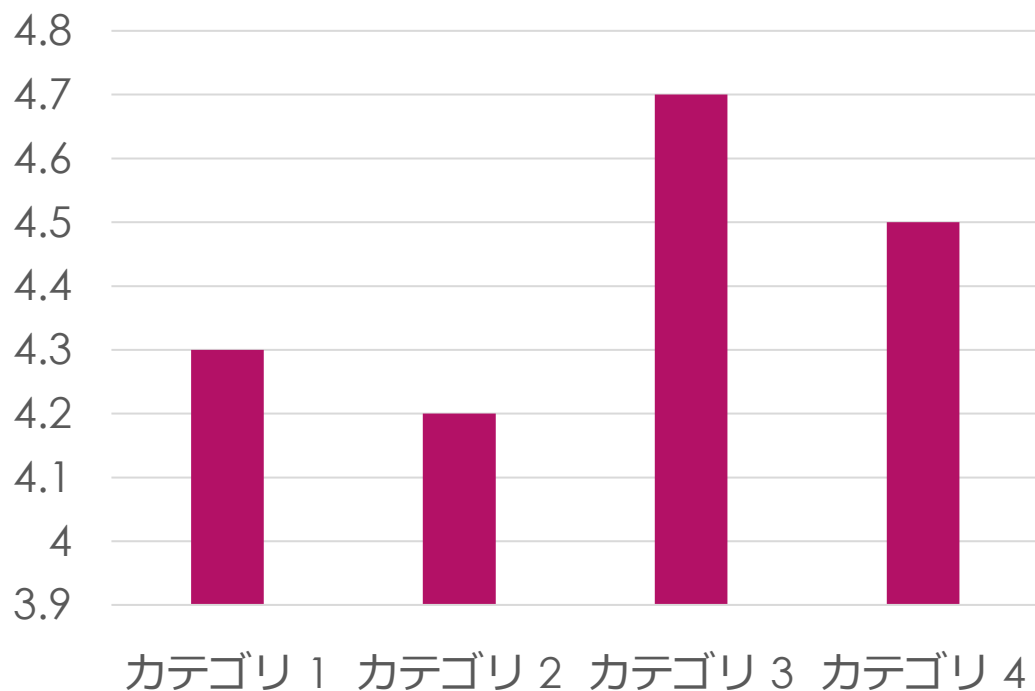
20□年

10%

40%

- ▶ そもそもグラフの幅と実際の数字が乖離している
- ▶ 2つを比較しているように見えるが変化を示すために用いる単位が異なっている

よくある(?)手口



カテゴリ間の差が大きいように見えるが、実際は微々たる差で、表示幅を狭くすることで際立たせている

まとめ

- ▶ 統計解析と同じくグラフも“簡単”にできるが、ソフト任せにすると、データが歪んで見えることもある
- ▶ 解析者の主観でデータを歪ませることも可能
- ▶ 解析前は自身が理解しやすく、発表時は他者が理解しやすく
- ▶ グラフ作成時は、データを俯瞰できているか、何度も確認を