

いまさら聞けない基礎統計学2 ～データを俯瞰しよう～

男女共同参画支援室
衛生学・予防医学講座
各務竹康

データの表記

- “重要”：読者に伝わるデータの記載を行う
- “伝わる”とは、数字から全体像(母集団の特徴)が想像できること
- データが読みにくい→(ほぼ)問答無用で厳しめの査読対応
- 表記がまずい→著者は適切な解釈ができてない？
→リジェクトへの近道

研究は育児と同じ

- データの収集：妊娠

とにかく産まれてほしい

- 体調の管理
- 外的要因の除去
- あらゆる手段で守る
- でも回避不能な外的要因も

研究は育児と同じ

- データの分析：育児

元気に育ってほしい

- 能力の開花
- 長所を伸ばす
- 短所も上手に消す
- 大切に育てる

今日のメニュー

1. データを“俯瞰”する

集めたデータの特徴を知ること、分析方法を見極める

1. 統計結果の解釈

$P < 0.05$ が全てではない！

データを“俯瞰”する

適切にデータを提示しよう

データを視覚化する

- 細かい解析の前にまず実施すべき
→どの記述統計量を採用するか決めるため
- どのような分布をしているか
- 解析に耐えうるデータか
- 適切にデータが集まったか

変数の分類

- 連続変数

例

〇〇について、5点満点で・・・→連続変数

〇〇についての評価を次から・・・

非常に良い 良い 普通 悪い 非常に悪い

入力値は1,2,3,4,5 →順序変数

- 順序変数

- 名義変数

データの視覚化

- 度数分布表
- ヒストグラム
- 箱ひげ図

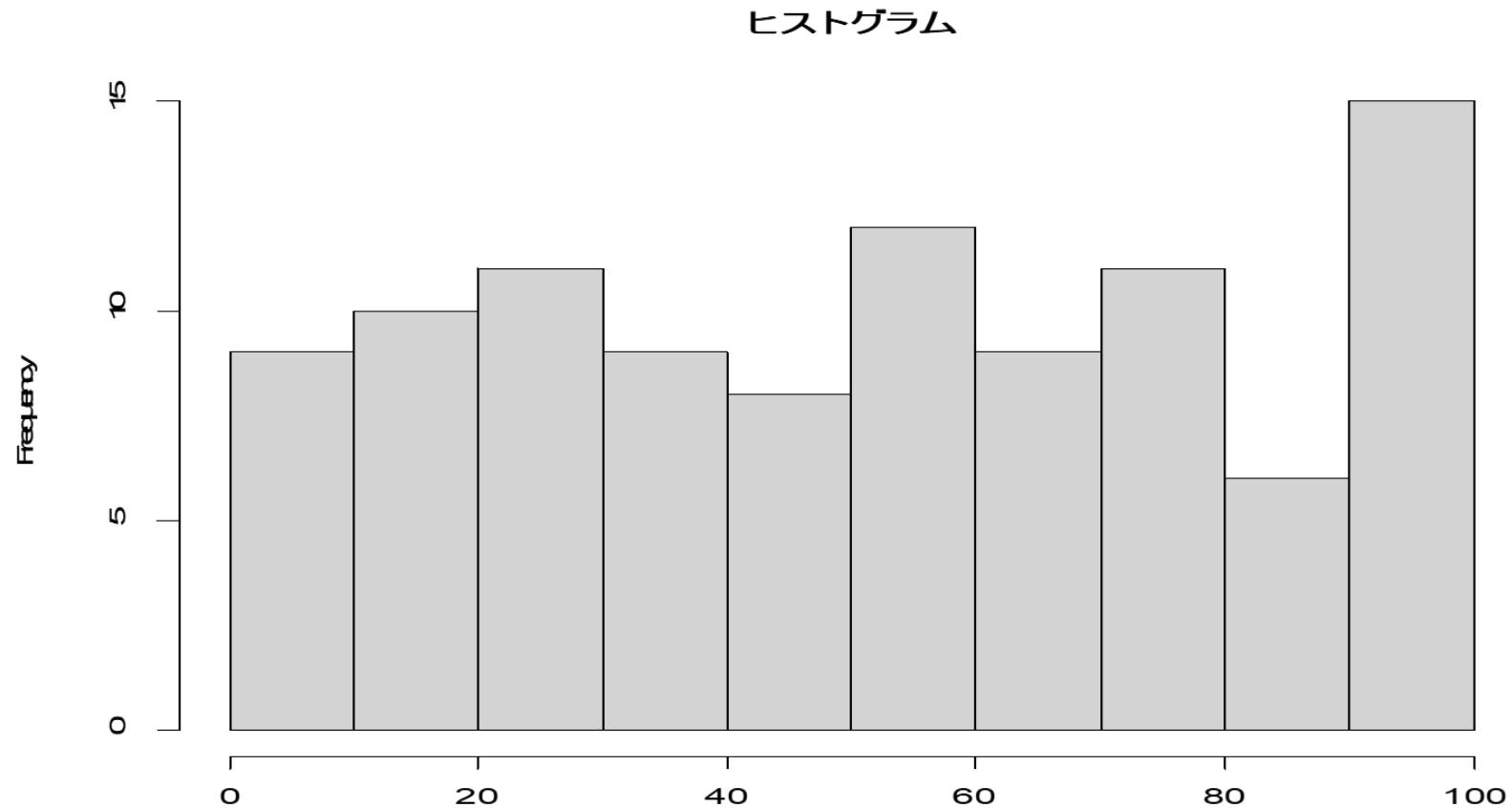
度数分布表

- データの範囲を複数に分割
- 区分別の出現回数と割合を示す

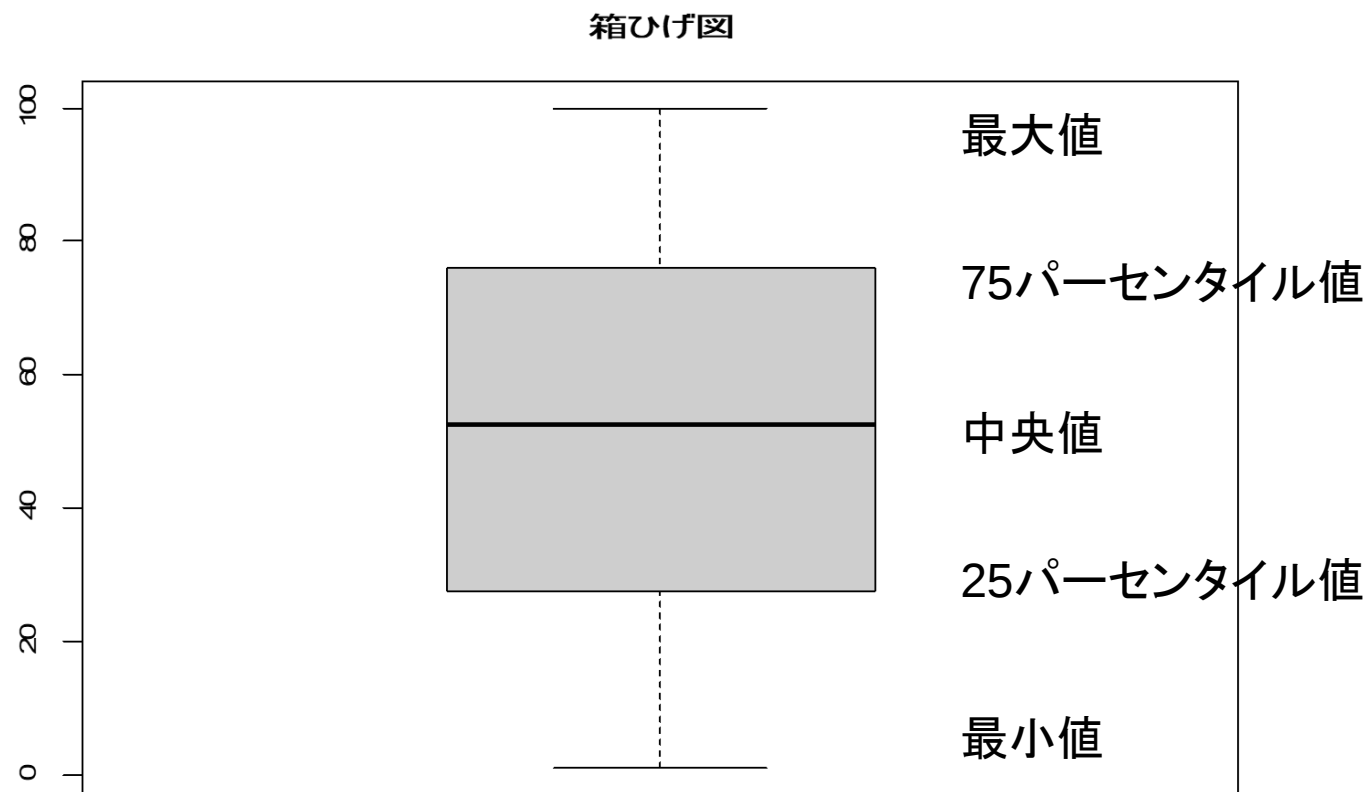
区分	出現回数	出現割合
A	○	○ %
B	△	△ %
C	■	■ %
D	×	× %

ヒストグラム

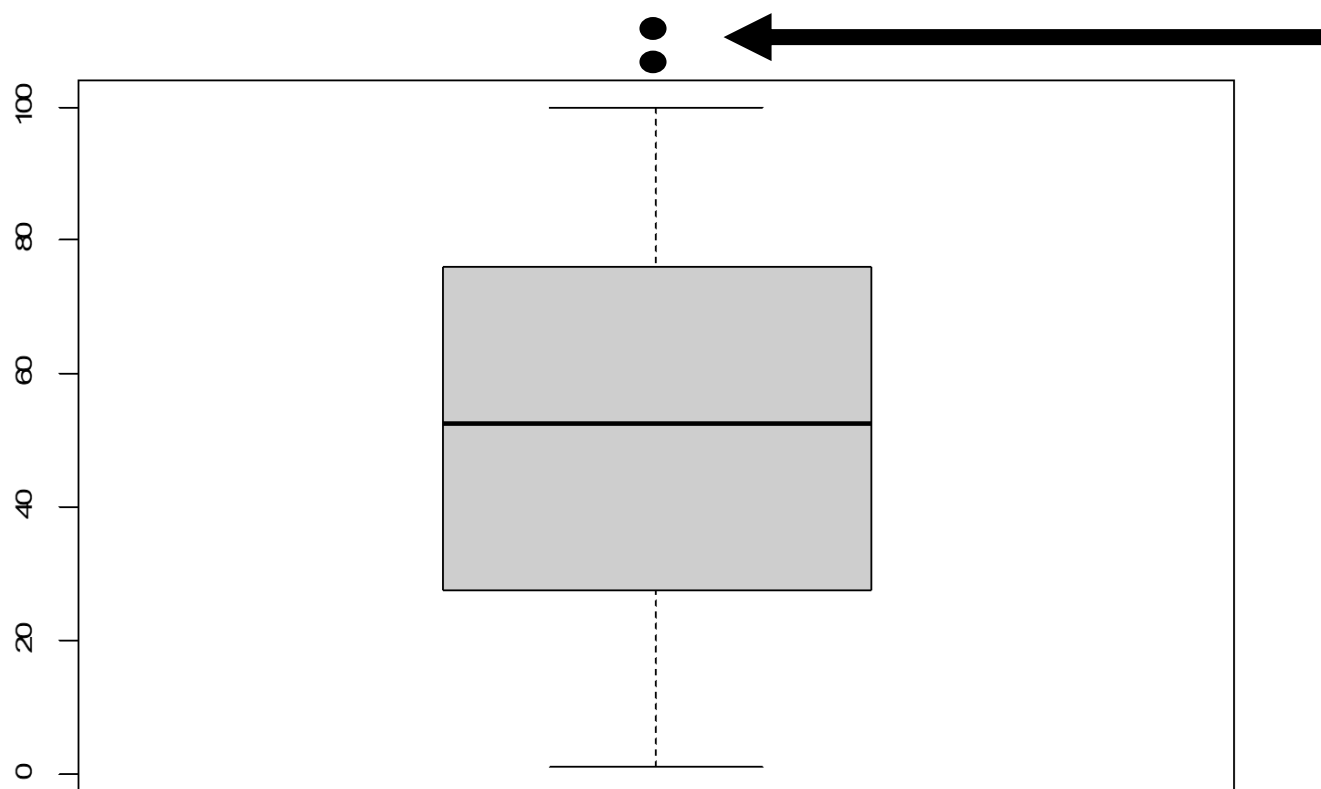
- 階級別の分布表を作成して、その度数をグラフ化する



箱ひげ図

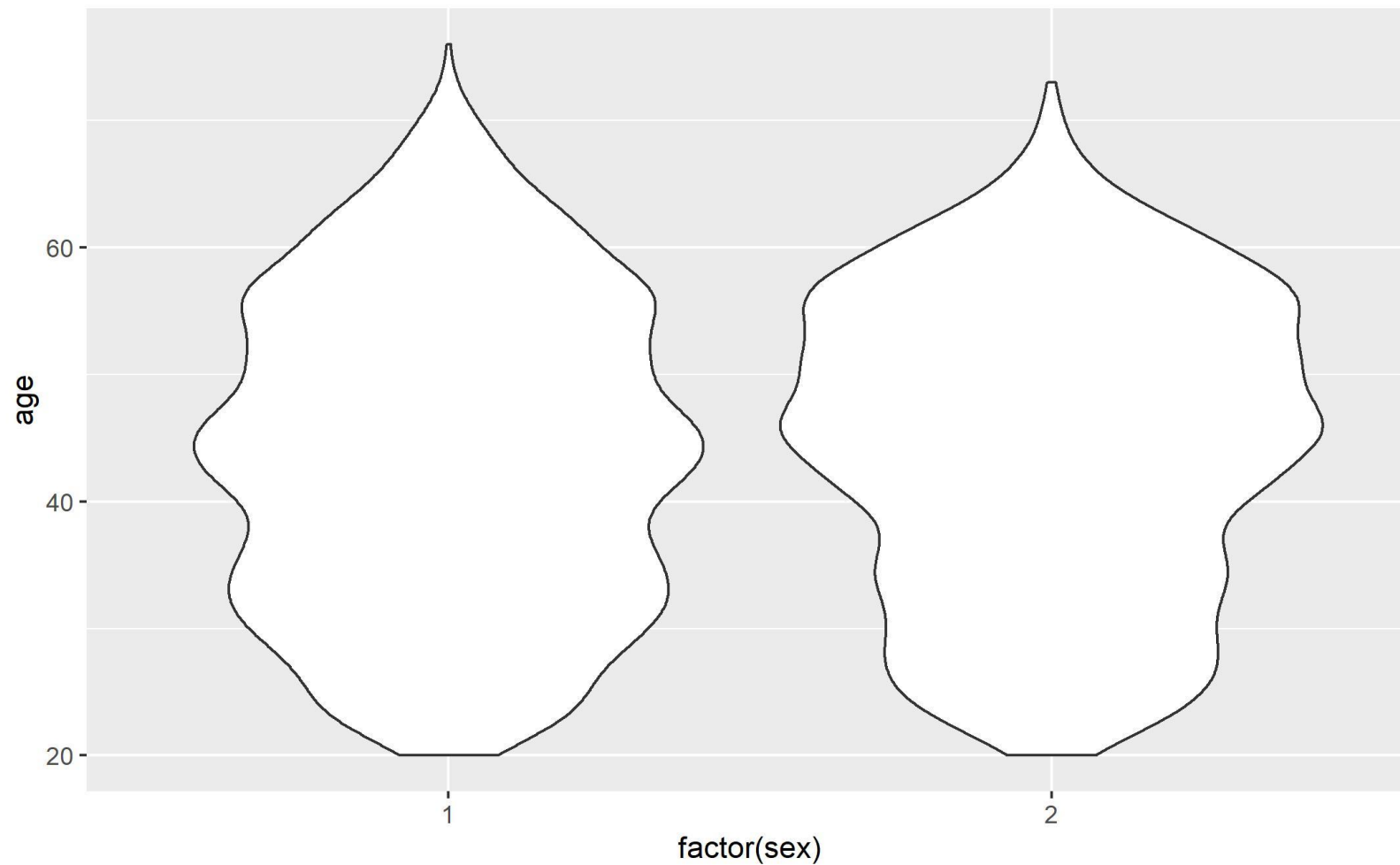


箱ひげ図



線が最大値、最小値を示さず、上に○などが表記されることも

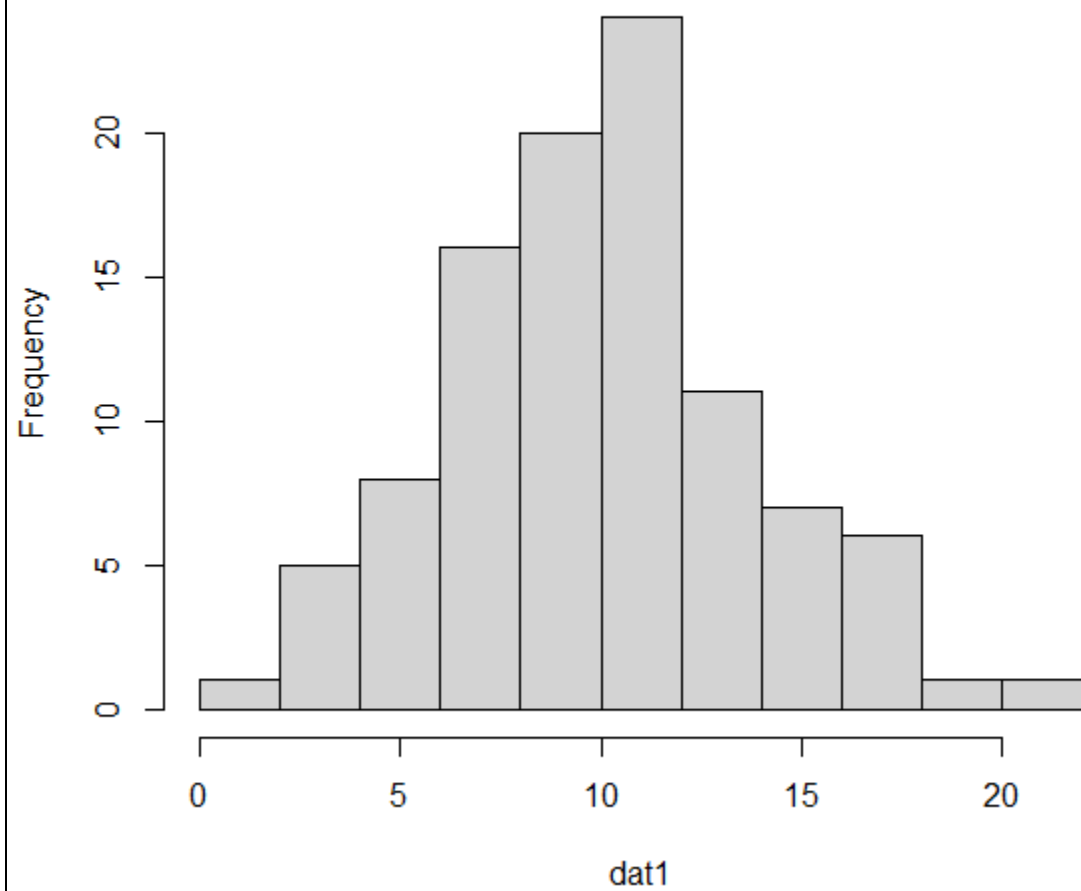
バイオリンプロット



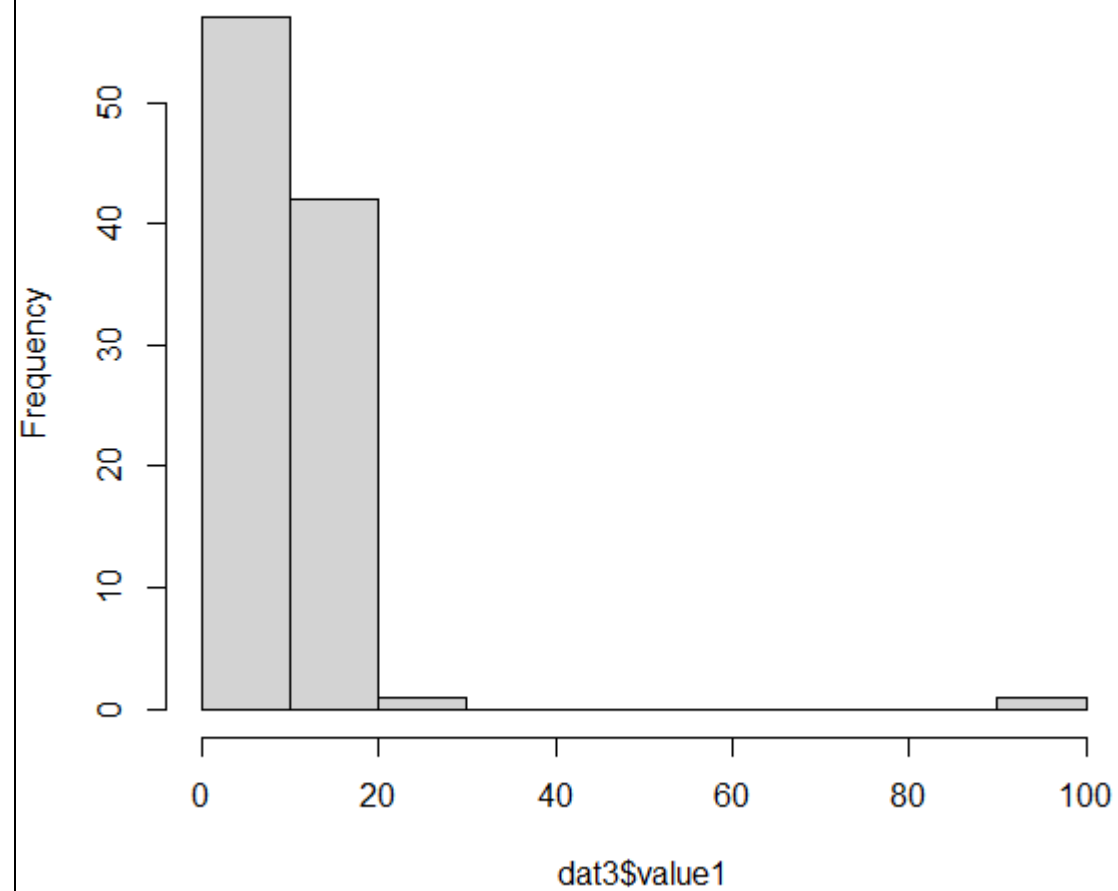
ヒストグラムの落とし穴

- 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

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



分布を見るときは

- どのような分布をしているのか確認する
正規分布
対数分布
その他の分布
- 極端な外れ値がないかも忘れずに

外れ値

- 極端な数値が出現すること
- 考えられる原因
 1. 入力ミス
 2. 測定方法の誤り
 3. 不適切なサンプル選択
 4. 真の値

外れ値を除外→除外する正当な理由があればOK

外れ値を(できるだけ)なくすため

- デザインの段階で制御、予備実験を丁寧に
- サンプリング(対象の選出)プロトコルを練る
- 例数を増やす

Table1の記載例

	Group A (n = 24)	Group B (n=30)	p value
age	48 ± 3.5	40 ± 4.0	0.004
	正規分布を仮定: 平均±SD		
sex (male)	12 (50.0)	10 (33.3)	0.337
	群のうち該当者の割合		
working duration (year)	10 (2 – 20)	10 (5-16)	0.582
	正規分布を仮定できない: 中央値(パーセンタイル)		
smoking			
current smoker	8 (33.3)	2 (6.7)	0.033
ex-smoker	5 (20.8)	12 (40.0)	
non-smoker	11 (45.8)	16 (53.3)	列(縦)に対しての割合

範囲の記載

最大値－最小値

パーセンタイル値(inter quantile range: IQR)

信頼区間

- (〇〇－▲▲)→マイナス値の場合、(-〇〇－-▲▲)
- (〇〇, ▲▲) や (〇〇 to ▲▲)の記載を採用することも

連続データのカテゴリ化

- 連続データ: 連続変数もしくは順位変数
 - カテゴリ化したデータ: 順位変数もしくは名義変数
-
- まずは2群に分けるのか、3群以上に分けるのか、決定
 - その結果で何を伝えたいのかを明確に

これはNG

- 分類の根拠が不明瞭
- カテゴリ化する際は、その根拠が「他者が納得」できることが必要
- 確かな根拠を持って分類する

カテゴリ化の基準

- 規定のカットオフ値がある場合

単純明快

カットオフ値を採用しない場合、相応の根拠が求められる

有意差を出すためにわざわざ境界を動かした？

片方の群が少なすぎる？

- 規定のカットオフ値がない場合

分け方が“恣意的”とみられない理由付け

カットオフ値の決め方

- 中央値(平均値)で分ける？
- まずは分布を確認
- 二峰性の分布だとわかりやすい
- それ以外の場合は？

決め方の例

- 過去の論文で用いられている値を参考にする
- 上位(下位)〇%を基準にする(パーセンタイル値)

$P < 0.05$

これで終わりではない

統計結果の解釈

P<0.05が全てではない！

例文

- 高血圧患者1万人に対して従来のA薬と新しいB薬の比較を無作為割付試験にて行いました。A薬を使った患者の平均血圧(SD)は125.3(2.1)mmHg、B薬を使った患者の平均血圧(SD)は120.5(4.2)mmHg、t検定の結果、両群の血圧は $p<0.05$ でB薬が有意に降圧作用が高く、優れた薬であることが明らかになりました。
- 上の文章、“統計的に”不自然なところを探してください

例(カイ2乗検定を例として)

		要因B	
		○	×
要因A	○	30%	20%
	×	20%	30%

Nを色々に変化させると

N=50		要因B	
		○	×
要因A	○	15	10
	×	10	15

N=100		要因B	
		○	×
要因A	○	30	20
	×	20	30

N=200		要因B	
		○	×
要因A	○	60	40
	×	40	60

N=300		要因B	
		○	×
要因A	○	90	60
	×	60	90

余談: カイ2乗検定の解釈

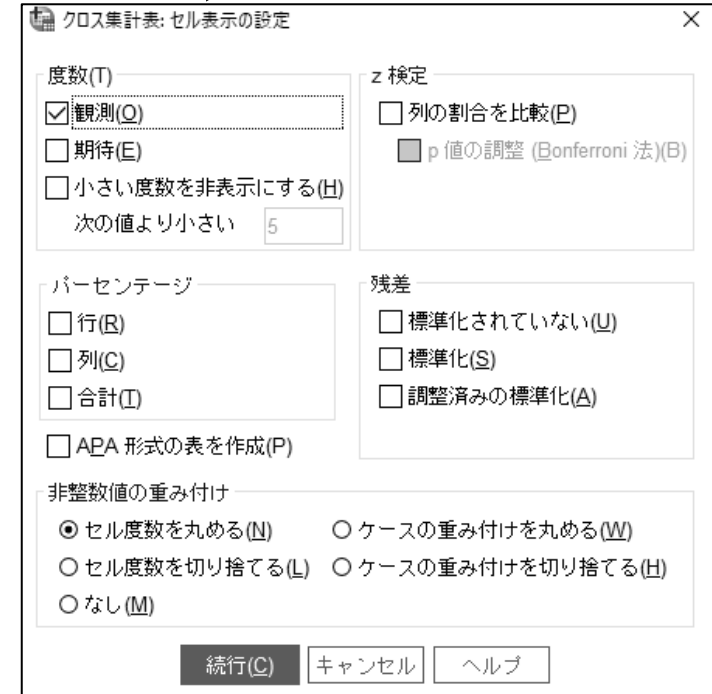
- 2×2 分割表: 解釈は容易
- $\bigcirc \times \bigcirc$ (丸の少なくとも1方が3以上): 分布の偏りが“ある”ことのみを示す
どのセルで多く、どのセルで少ないのかは示していない
- 残差分析

N=50		要因B	
		○	×
要因A	○		
	×		

N=50		要因B		
		○	×	△
要因A	○			
	×			

残差分析(カイ2乗検定の応用)

- SPSSでは
分析→記述統計→クロス集計表
- “セル”の項目より、残差(調整済みの標準化)にチェック



クロス集計表: セル表示の設定

度数(T)
☒ 観測(Q)
☐ 期待(E)
☐ 小さい度数を非表示にする(H)
次の値より小さい 5

z 検定
☐ 列の割合を比較(P)
☒ p 値の調整 (Bonferroni 法)(B)

パーセンテージ
☐ 行(R)
☐ 列(C)
☐ 合計(I)
☐ APA 形式の表を作成(P)

残差
☐ 標準化されていない(U)
☐ 標準化(S)
☐ 調整済みの標準化(A)

非整数値の重み付け
☒ セル度数を丸める(N) ☐ ケースの重み付けを丸める(W)
☐ セル度数を切り捨てる(L) ☐ ケースの重み付けを切り捨てる(H)
☐ なし(M)

続行(C) キャンセル ヘルプ

- 1.96より大きい場合は期待値より有意に大きい
- 1.96より大きい場合は期待値より有意に小さい

クロス表

			6の2		合計
			0	1	
ガイドライン	0	度数	272	70	342
		調整済み残差	-2.4	2.4	
	1	度数	24	9	33
		調整済み残差	-1.5	1.5	
	2	度数	141	14	155
		調整済み残差	3.3	-3.3	
合計		度数	437	93	530

1.96はどこから？

- 正規分布の場合
 - ±1SD: およそ70%(68.3%)
 - ±2SD: およそ95%(95.4%)
 - ±3SD: 99%以上(99.7%)が含まれる
- 正確に95%が含まれる範囲は±1.96
- 残差: 平均0、標準偏差1の場合の値

単変量解析の持つ意味

- $P < 0.05$ の持つ意味
異なる属性を持った2集団を比較した
「異なる属性」≠「検定した項目」
- 検定を行う項目“だけ”で比較した場合の統計学的な結果
- その他の項目については比較していない

人の集団を対象とした研究

- 様々な要因を考慮に入れる必要
結果に影響を与える要因「バイアス」
- バイアスを排除するためには？
デザイン段階で考慮
分析段階で考慮

分析段階で考慮

- 多変量解析
- ロジスティック回帰
- 重回帰
- Cox回帰
- Mantel-Haenszel検定
- 他

まとめ

- データの“ありかた”を正しく解釈→適切な集計
- 外れ値は本当に外れ？全てのデータに一度疑いの目を持つとう
- 分布の解釈を誤るとその後全てに尾を引くので、まずはデータ全体を俯瞰しよう