いまさら聞けない基礎統計学4 ～変数を理解しよう

ダイバーシティ推進室

衛生学・予防医学講座

各務竹康

変数とは？

- 変数の対義語→定数
- 定数とは“定まった数”

  周囲の条件などに左右されない数字
- 変数とは“変わる数”

  様々な条件で変化する

その年に聞く限り、数字が変わることはない
→定数

年齢(年末時)　＝　今年の西暦　－　生まれた年

聞く相手によって数字が変わる
→変数

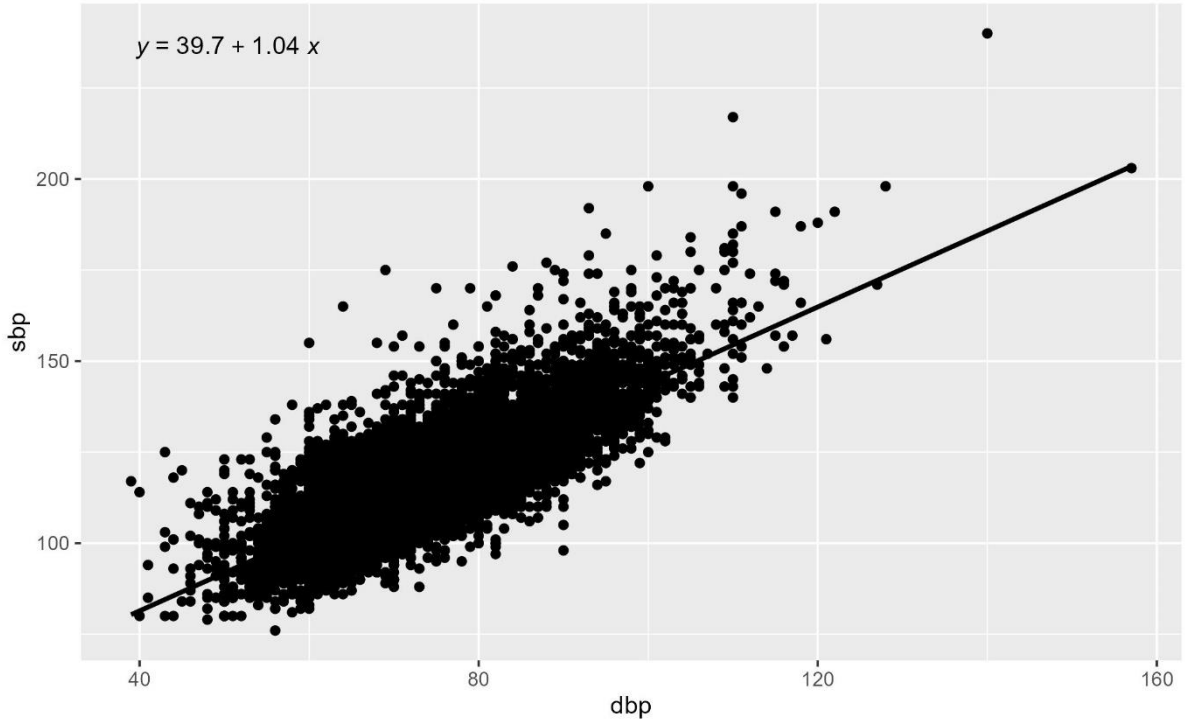
研究における“条件”

- サンプルの違い

- 収集する情報は、サンプルによって異なる
  (収集するまで分からない)
- 収集したデータは全て変数

統計解析は“定数探し”

- 収集データ→基本的に変数
- 変数同士の関連を見つけて、規則性を探る



例：収縮期血圧と拡張期血圧の関連

収縮期血圧＝39.7+1.04×拡張期血圧

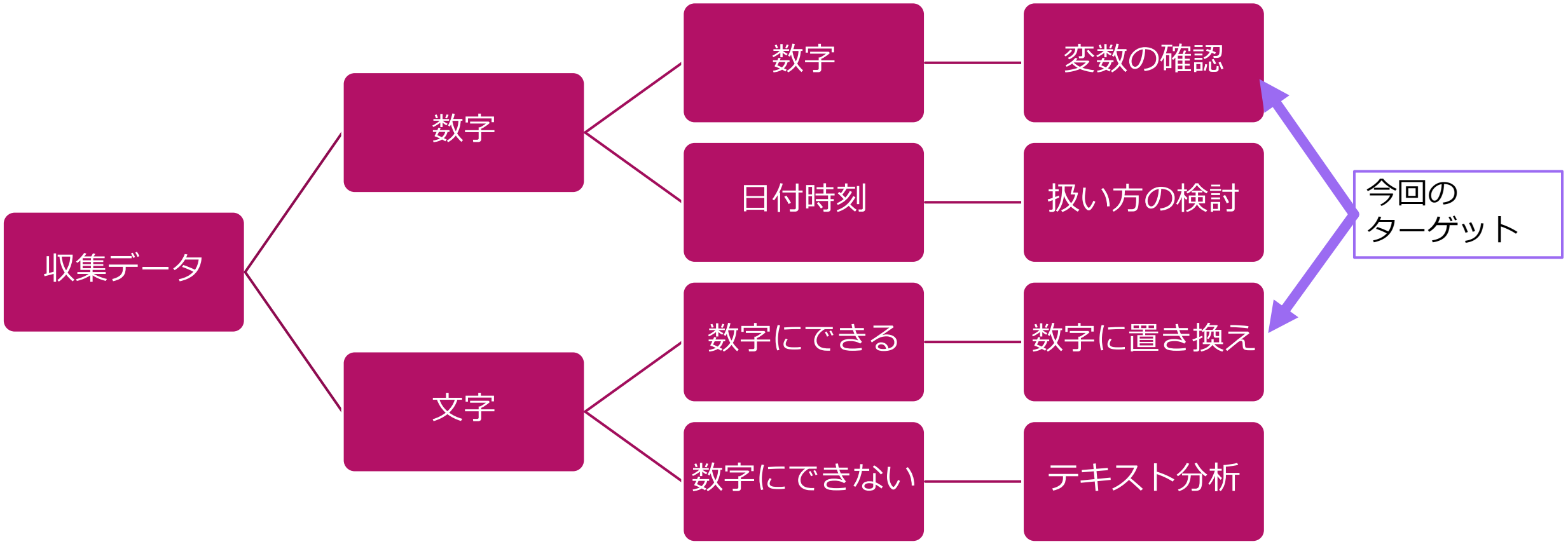
39.7および1.04が定数

目的(従属)変数と説明(独立)変数

- 統計解析の(よくある)命題
- ある変数は、他の変数から説明できるのか

  治療効果は、薬剤、患者背景から予測できるのか
- 予測したい変数：目的(従属)変数　objective (dependent) variable
- 予測のために用いる変数：説明(独立)変数 explanatory (independent) variable
- 目的⇔説明　　従属⇔独立　のどちらを用いても良いが、組み合わせは間違えないこと

データの確認


収集データ
数字
文字
数字
日付時刻
数字にできる
数字にできない
変数の確認
扱い方の検討
数字に置き換え
テキスト分析
今回の
ターゲット

「数字にできる」とは？

- 入力された文字に規則性がある

  季節、月、など

- 実は数字

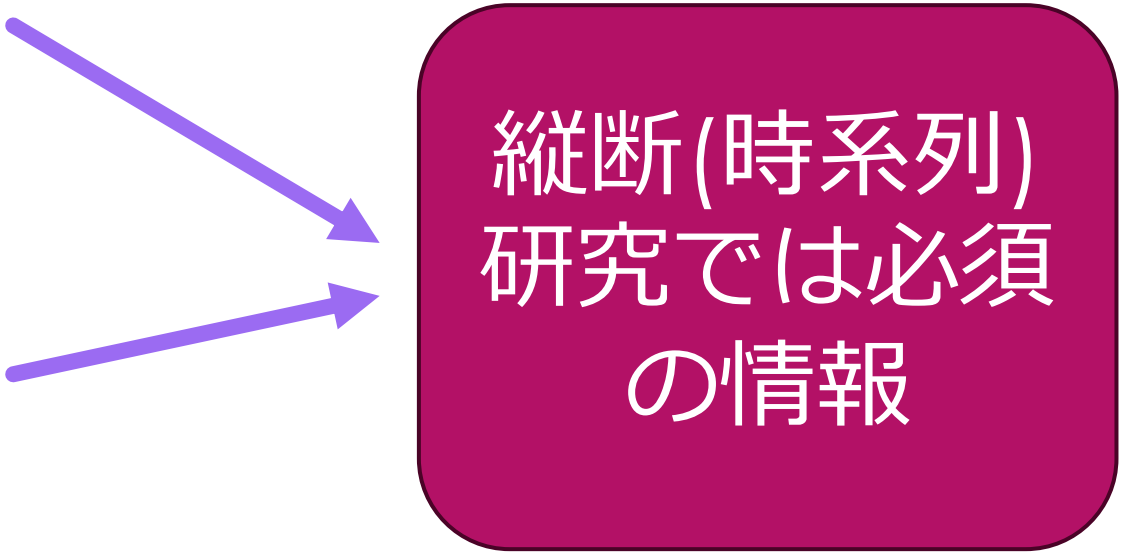
  2023年など単位がなければ単なる数字(入力時は単位を入れない！)

- 内容をカテゴリーに分類できる

  ポジティブな内容、ネガティブな内容、など

日付、時刻データ

- 日付データの扱い方
  - 日、月、曜日のカテゴリーとして扱う
  - 追跡開始日からの連続日として扱う
- 時刻データ
  - ○時台などのカテゴリーとして扱う
  - 追跡開始からの連続時間として扱う



変数の属性

- 連続量：数値データで示される
  - ✓ 比尺度
  - ✓ 間隔尺度

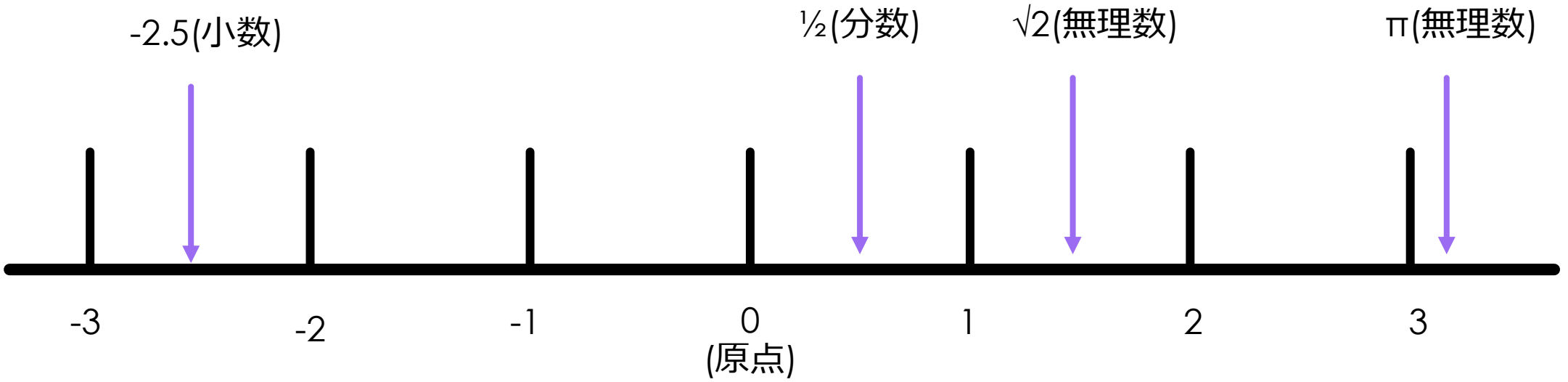
- 離散量：数値データや、数量的に測れないカテゴリデータがある
  - ✓ 順序尺度
  - ✓ 名義尺度

連続量：比尺度

- 2つの数字の比、差をとることができ、数字の0が原点を示す

6は4より2大きい

6は3の2倍である



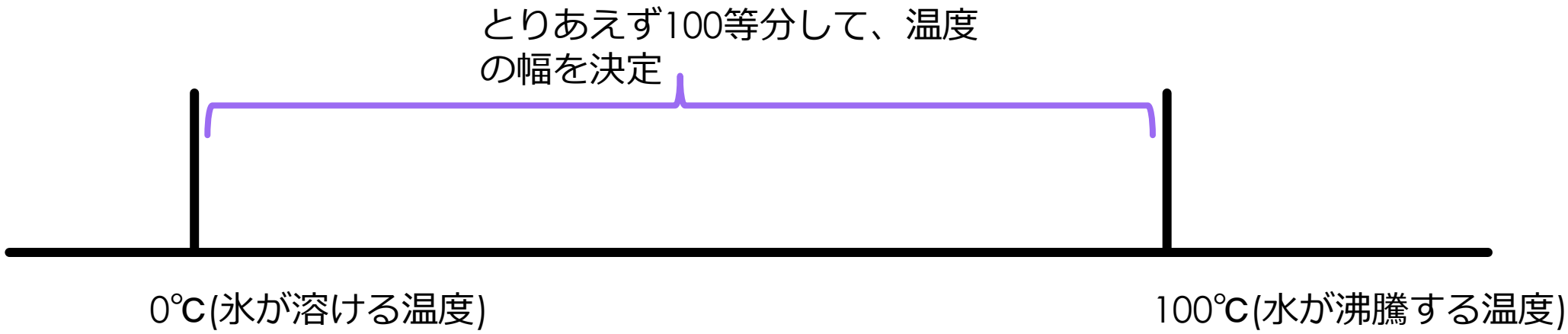
1本の数直線上に位置する

連続量：間隔尺度

- 2つの数値の差を取ることができるが、比は意味を持たず数字の0が原点を示さない

6は4より2大きい

6は3の2倍とは言えない



離散量：順序尺度

- データの並び方、順番には意味があるが、その差、比には意味を持たない

1は4より上位であることを示す

5は3より上位であることを示す

全日本大学駅伝の順位

1位：駒澤大学　　2位：青山学院大学　　3位：國學院大学

離散量：名義尺度

- 便宜的に数字を当てはめているだけで、その並び順には意味を持たない
- 2値データも名義尺度の1つである

1組　　2組　　3組　　4組

変数の扱い

- 連続変数(比尺度、間隔尺度)
- 順序変数
- カテゴリー変数(名義尺度)

- の3種類に分類する
- 連続変数をパラメトリックとノンパラメトリックに分けることも

統計ソフトでは

- データを読み込んだら、自動的に変数の種類を認識してくれる
  →アルゴリズムに沿うため、研究者の認識と異なる場合も
  →データを読み込んだらまずは変数の確認
- 多い分類
  連続量
  順序
  時系列
  カテゴリ
  文字列

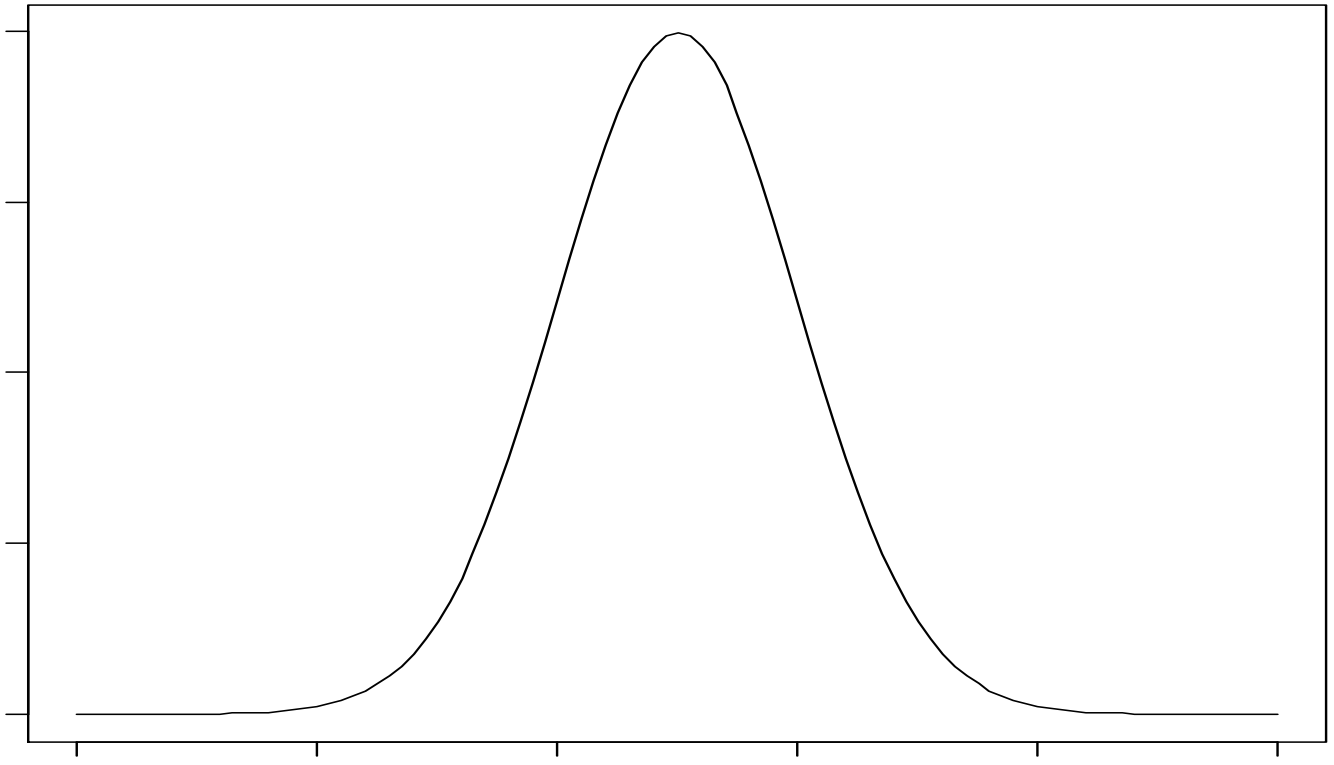
パラメトリック・ノンパラメトリック

- 分布や統計手法に対して用いる
- “Parametric”：特定のパラメーターに基づく統計的推論の方法を指す
- 母集団が何らかの規則性を持っている場合パラメトリックとみなし、パラメトリックな統計手法を用いる

何らかの規則性：正規分布

パラメトリックの対義語としてノンパラメトリック

正規分布とは

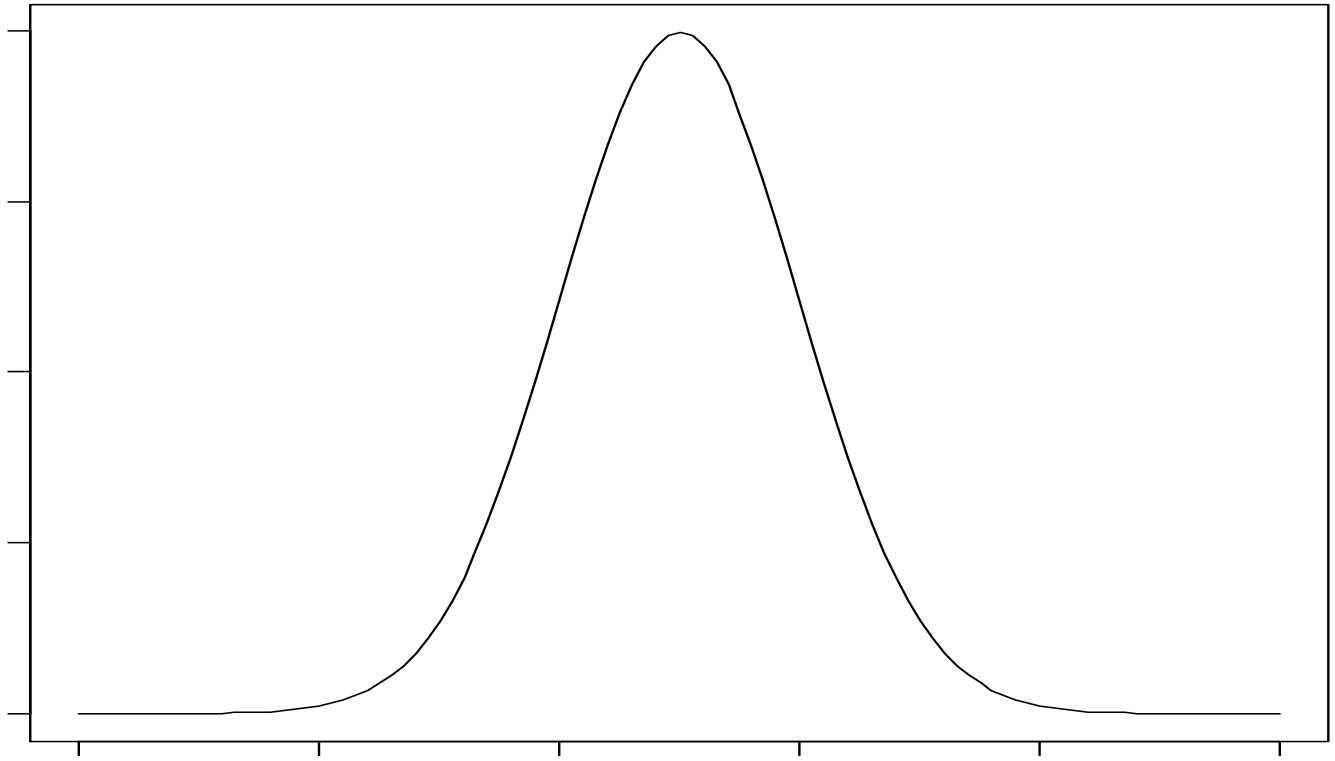


- 平均値、中央値、最頻値が一致する。
- 平均値を中心にして左右対称な分布をとる(グラフにすると左右対称な釣鐘状)
- 大規模データは正規分布を仮定して行う
- 様々な検定も、正規分布を仮定しているものが多い

正規分布を仮定すると便利

- 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



正規分布の確かめ方

- 統計的手法

  Kolmogorov-Smirnov検定

  Shapiro-Wilk検定

これらの検定は“サンプルから母集団の正規性を推定”
→サンプルの偏りが原因で本来母集団は正規分布なのに、正規性を棄却してしまう可能性

- 過去の論文などから母集団の推定

安易に統計で正規性を棄却するのも問題が？？

単変量解析

基本的な表示の仕方

	連続変数	順序変数	カテゴリ変数
度数分布表	×	△(項目が少ないとき)	○
グラフ	ヒストグラム 箱ひげ図 バープロット バイオリンプロット	箱ひげ図 バープロット(バルーンプロット) バイオリンプロット	×
統計量	最小値ー最大値 平均値±標準偏差 中央値(25－75パーセンタイル値)	最小値ー最大値 中央値(25－75パーセンタイル値) 最頻値	最頻値

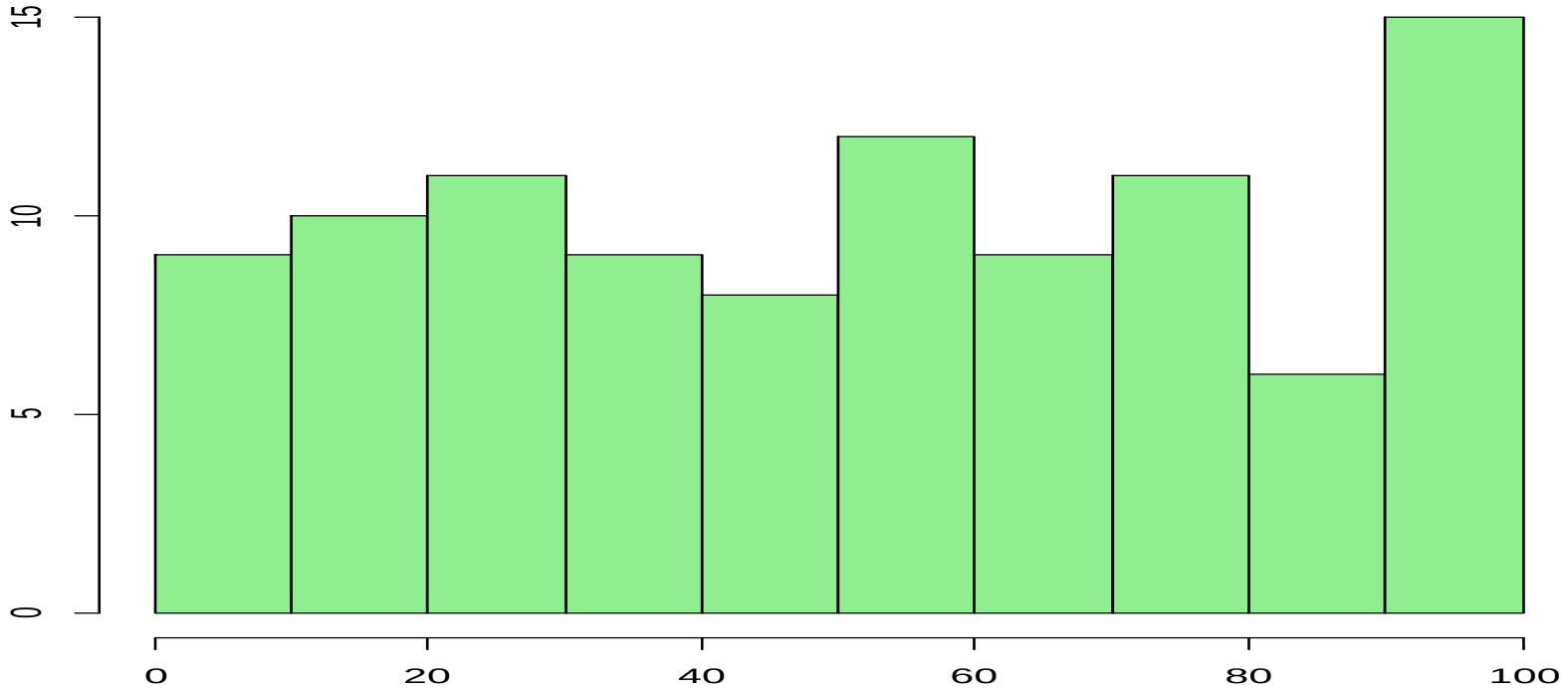
度数分布表

- 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

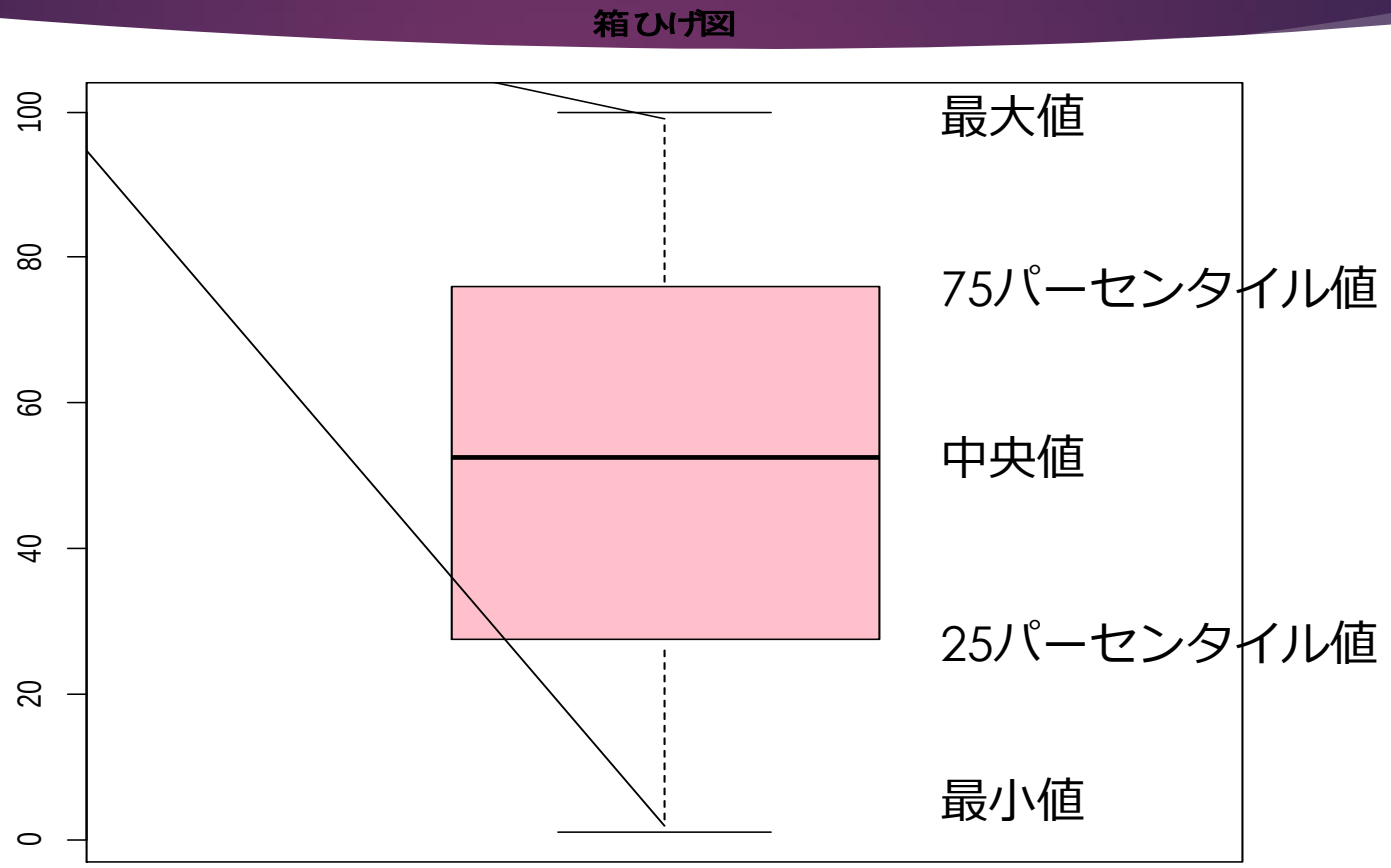
区分	出現回数	出現割合
A	○	○　%
B	△	△　%
C	■	■　%
D	X	X　%

ヒストグラム

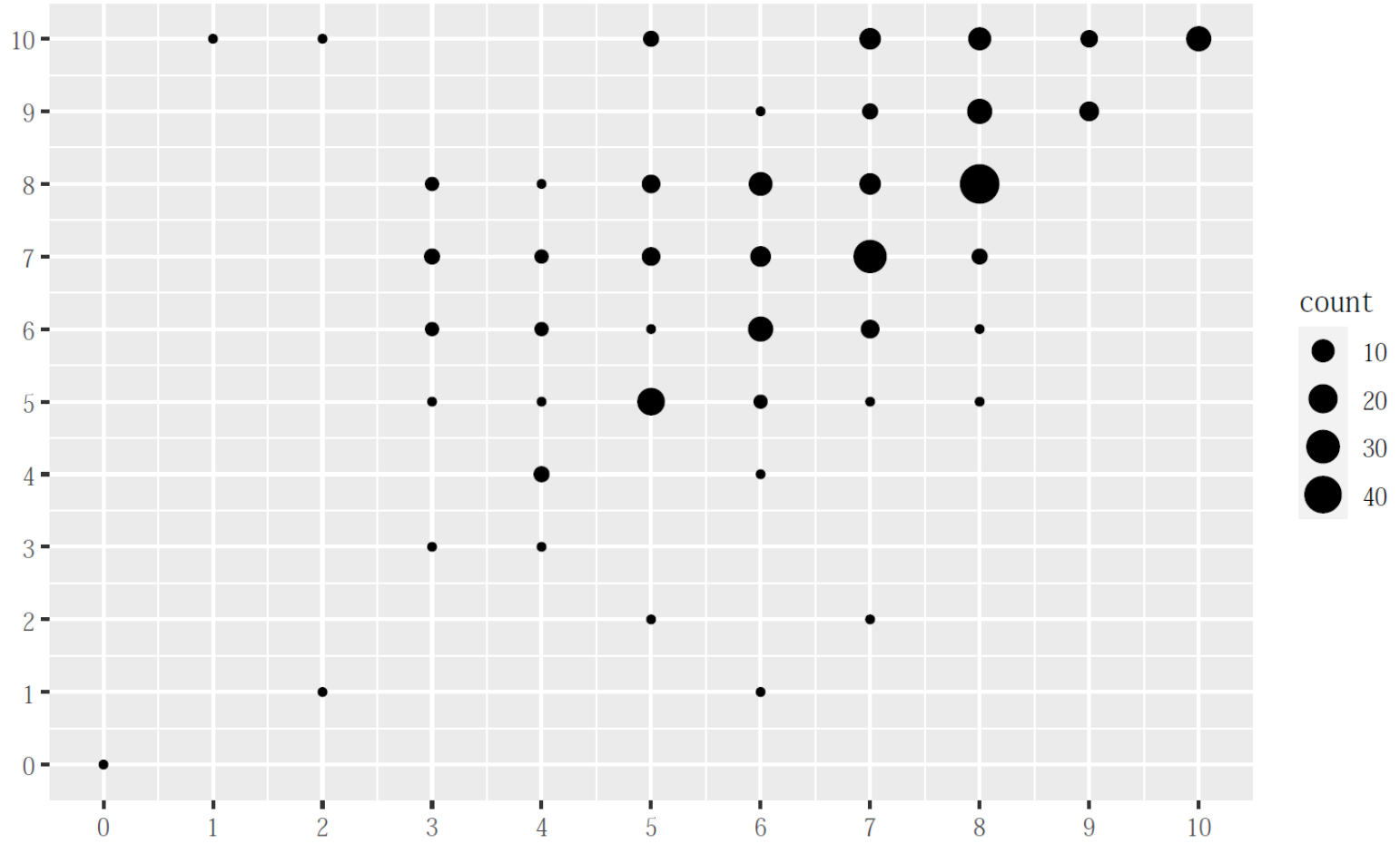
- 階級別の分布表を作成して、その度数をグラフ化する



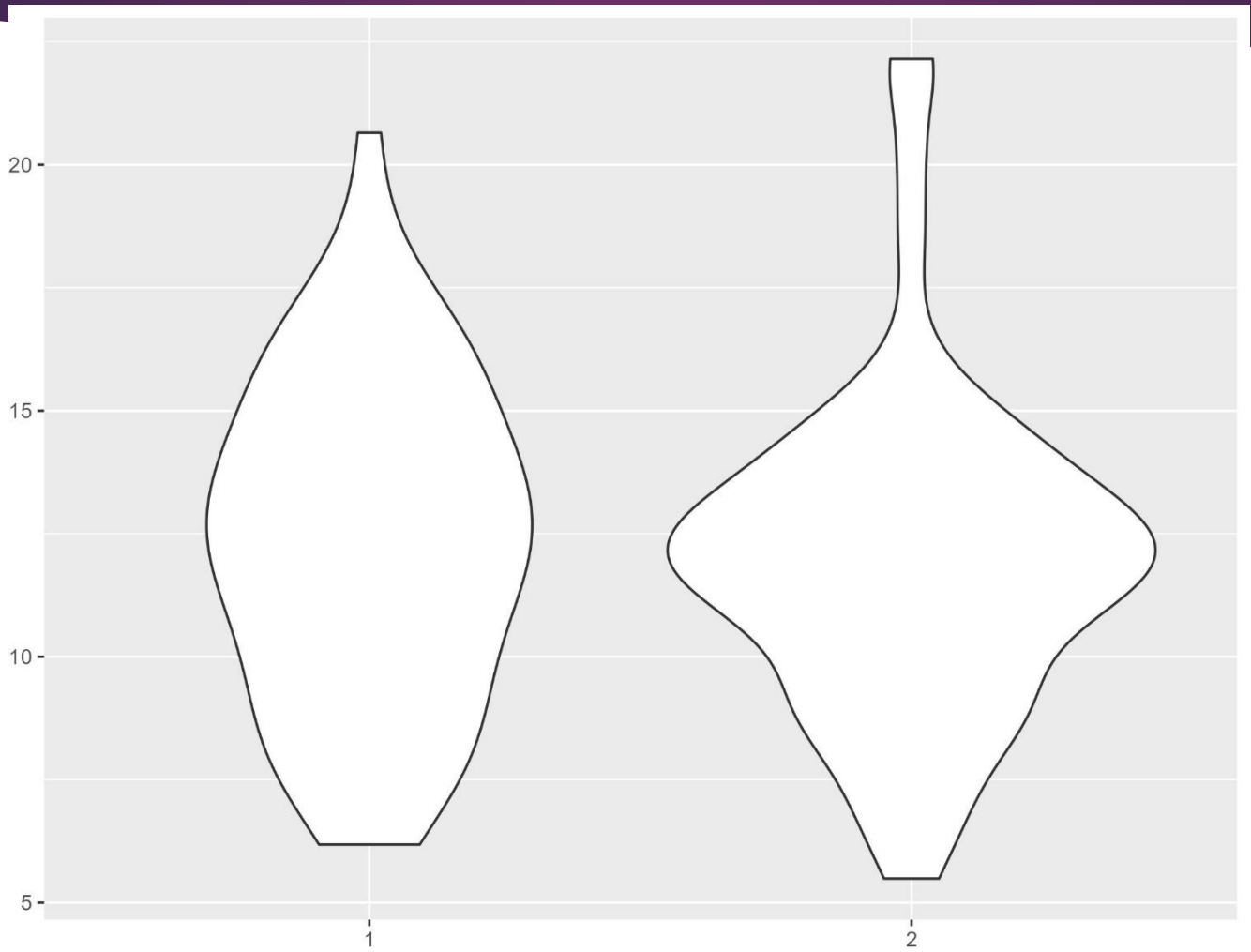
箱ひげ図



バルーンプロット



バイオリンプロット



平均値、中央値、最頻値

- 平均値：データの算術平均

$$\text{平均値} = \frac{\text{データの総和}}{\text{データ数}}$$

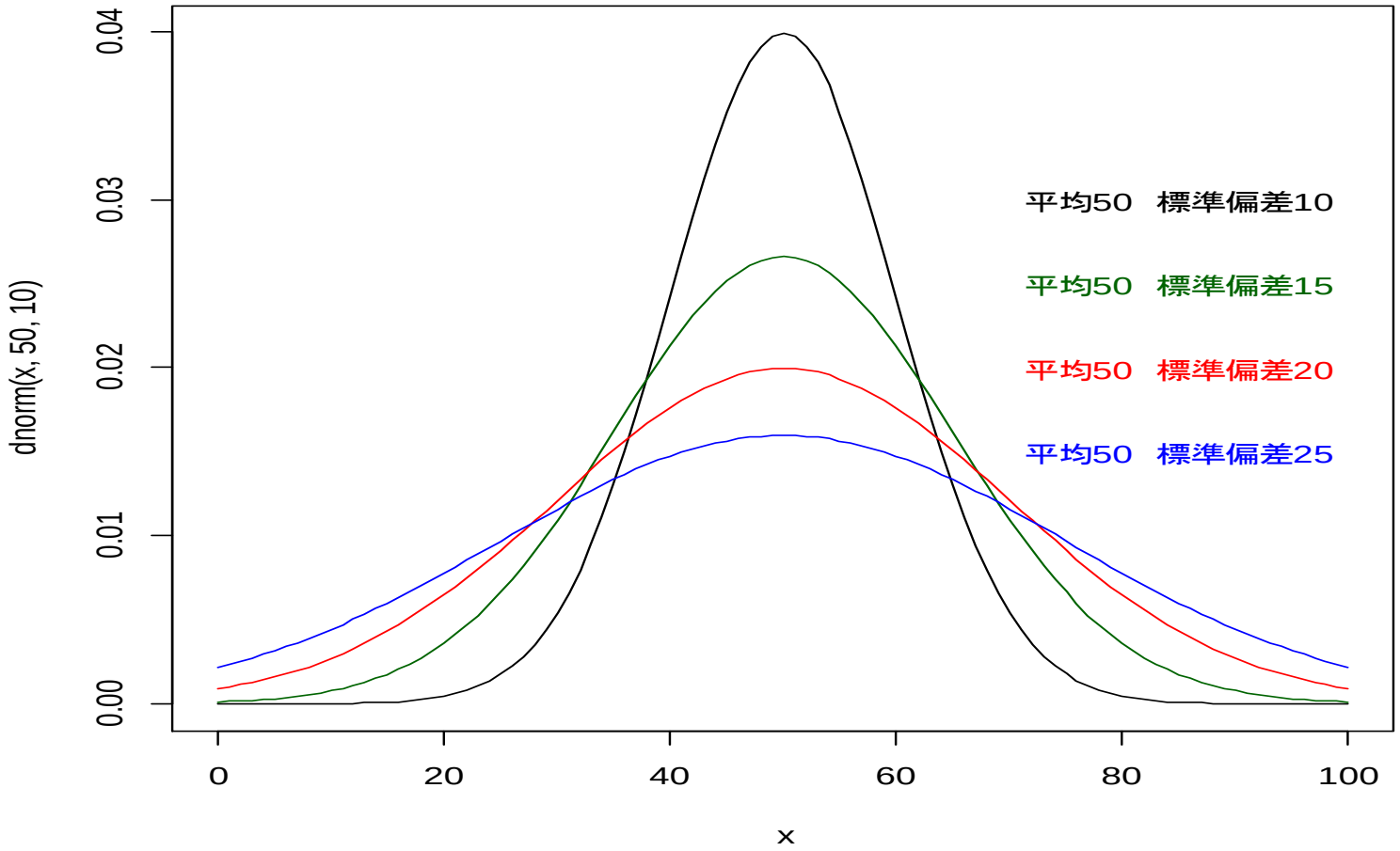
- 中央値：データを並べた時、真ん中の順位となる数値

- 最頻値：データのうち最も出現頻度の高い数値
  (カテゴリ変数、順序変数で使用される)

分散と標準偏差

- 標本のばらつきを示す指標
- 分散= $\frac{(\Sigma(\text{実測値}-\text{平均値}))^2}{n}$
- 標準偏差 = $\sqrt[2]{\text{分散}}$
- 分散は数値を2乗しているため、平均値などと単位が揃わない。標準偏差のほうが平均値などと単位が揃っているため、理解しやすい

分散(標準偏差)が異なると



標準誤差(Standard Error: SE)

- 平均値の信頼性
- 標本の平均値＝「今回」抽出した集団における平均値

- 条件を変えたら平均値も変動するかもしれない

- どの程度変動するのか？

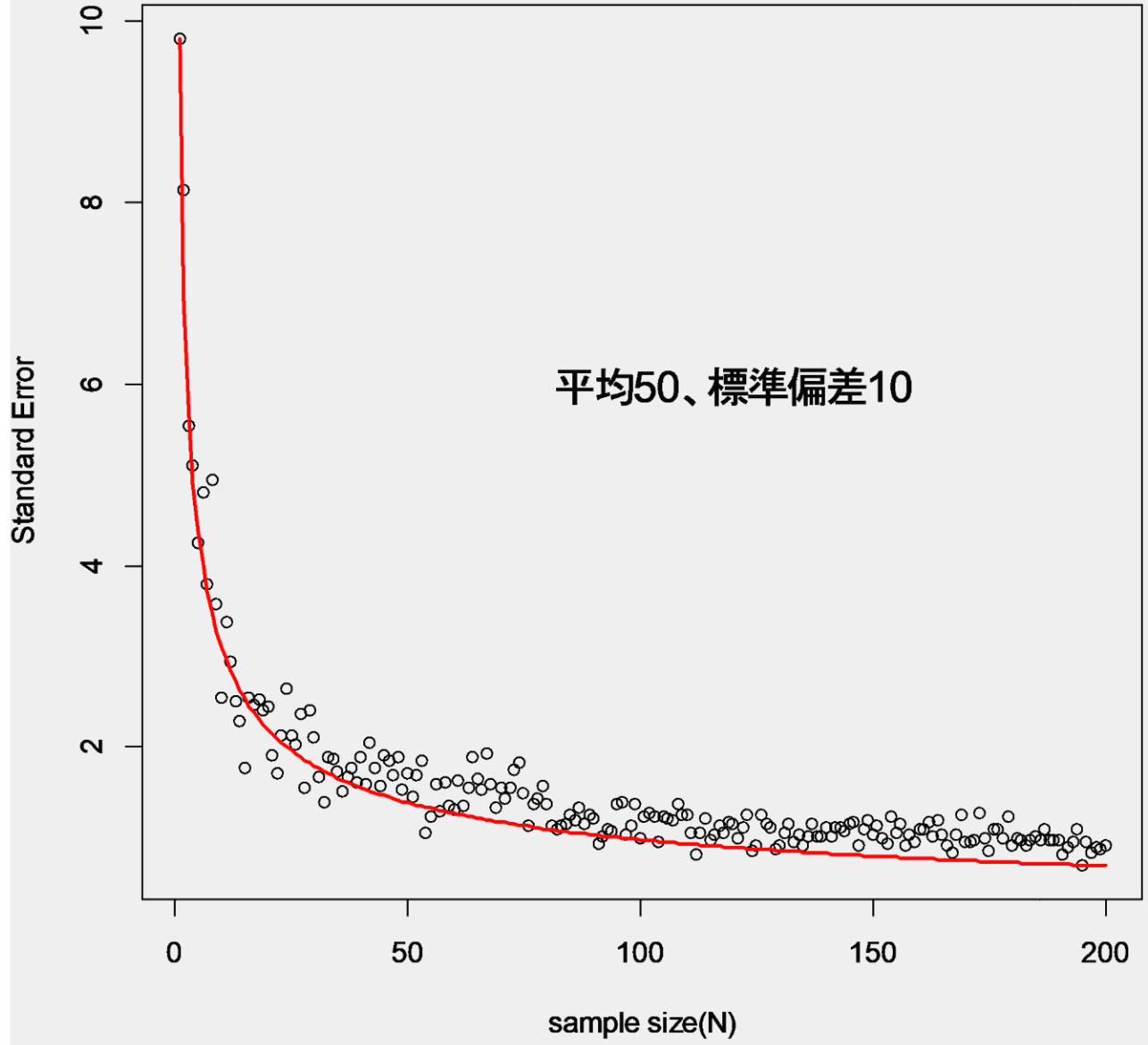
  → 標準誤差

標準誤差の計算式

- 標準誤差 $= \sqrt{\frac{分散}{標本数}} = \frac{標準偏差}{\sqrt{標本数}}$

- つまり・・・・・・
- 標準偏差が同じでも、標本数が多ければ標準誤差は小さくなる

サンプル数と標準誤差


Standard Error
10
8
6
4
2
0
50
100
150
200
sample size(N)
平均50、標準偏差10

最大値、最小値、パーセンタイル

- 最大値、最小値：言わずもがな

- パーセンタイル：全データの中で上位何%の位置にある数値
  主に、25パーセンタイル値ー75パーセンタイル値が用いられる
  (interquartile range: IQR)

有効数字とは

- 一般に測定においては、誤差を含む数字よりも上の桁のことを示す
- 体重を例として

  アナログの場合

  500g刻みで示されることが多い

  →100gの値までは有効？

  70.5ｋｇ(有効数字小数点1桁)

  デジタルの場合

  機種によって異なるが、10gまで正確に測定できるのであれば10gまで有効

  70.53kg(有効数字小数点2桁)

測定項目によって有効数字が異なる

- 肝機能(AST21)：有効数字1桁(1の位)
- 白血球数(WBC7500)：有効数字3桁(100の位)
- Hb(13.6)：有効数字小数点1桁(0.1の位)

- じゃあ、平均値はどこまで求めるべきか？

  有効数字より1桁小さい位

  例：体重はkg表示で通常小数点1桁→平均値は小数点2桁

2変量解析

数字と数字の関連を調べる

2変量解析

	連続変数	順序変数	カテゴリ変数
連続変数	相関 回帰	相関 傾向性の検定 群間の比較	群間の比較
順序変数		相関 群間の比較 クロス集計	群間の比較 クロス集計
カテゴリ変数			クロス集計

群間の比較はパラメトリック、ノンパラメトリック、対応のあり、なしで検定方法が異なる
クロス集計はその形状、対応のあり、なしで検定方法が異なる

相関と回帰

- 回帰：2変数の関係性に(数式的な)特徴があるか
- 相関：線形な回帰にどの程度2変数の関係性が集約しているか

相関はP値ではなく相関係数をみる

- 相関係数：どの程度強く関連しているのかを示す指標

  -1≦相関係数≦1

  ゼロが相関関係なし、-1が完全な負の相関、1が完全な正の相関

- 一般的に相関係数は、線形(直線)の相関があることが前提

2種類の相関係数

- Pearsonの相関係数

  正規分布×正規分布

  相関係数は“r”で表記

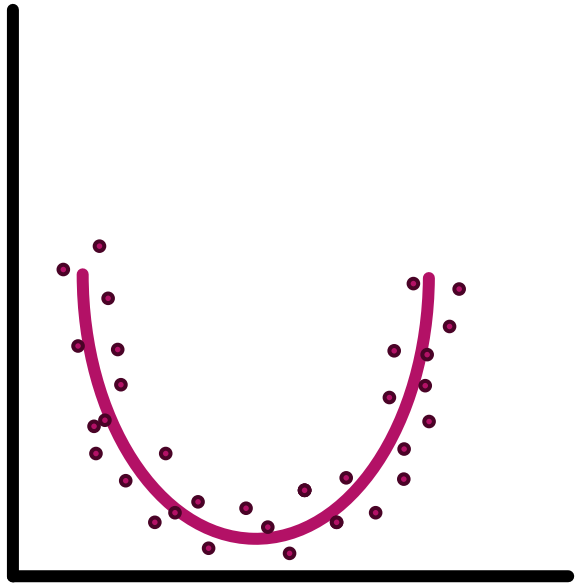
- Spearmanの相関係数

  一方でも非正規もしくは順序変数

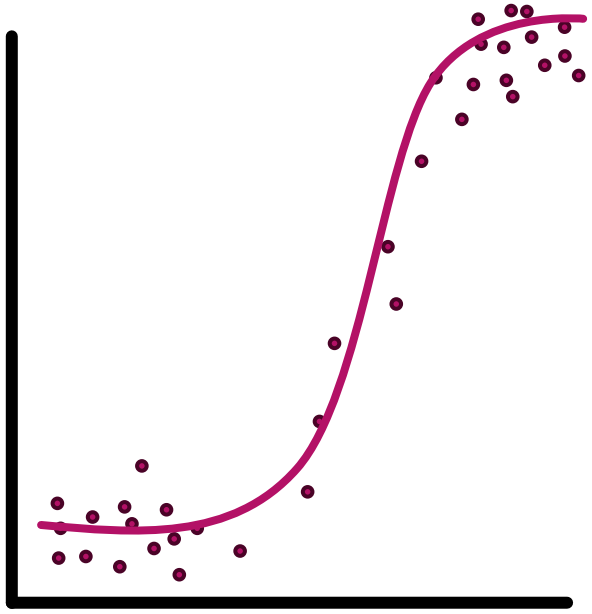
  相関係数は“ρ(ロー)”で表記

相関係数には数値的な区切りはないが、
目安として
｜相関係数｜＝0.7～1　強い相関あり
｜相関係数｜＝0.4～0.7　相関あり
｜相関係数｜＝0.2～0.4　弱い相関あり
｜相関係数｜＝0～0.2　ほぼ相関なし
が用いられることが多い

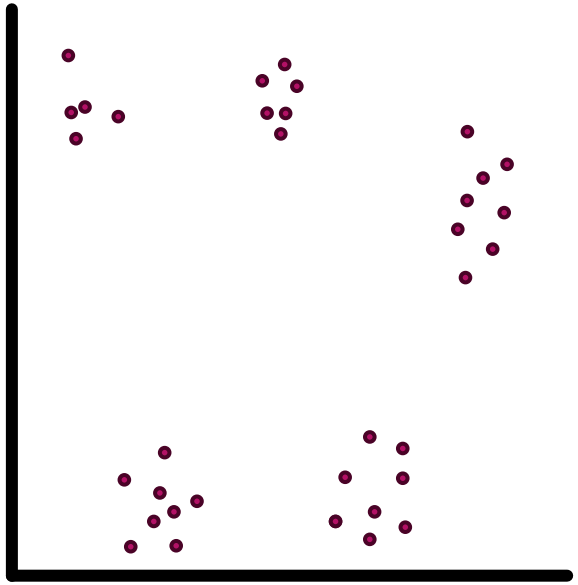
直線ではない相関とは？



二次関数の相関



ある箇所で急に分布が変化する



よくわからないけれど何らかの法則がありそう

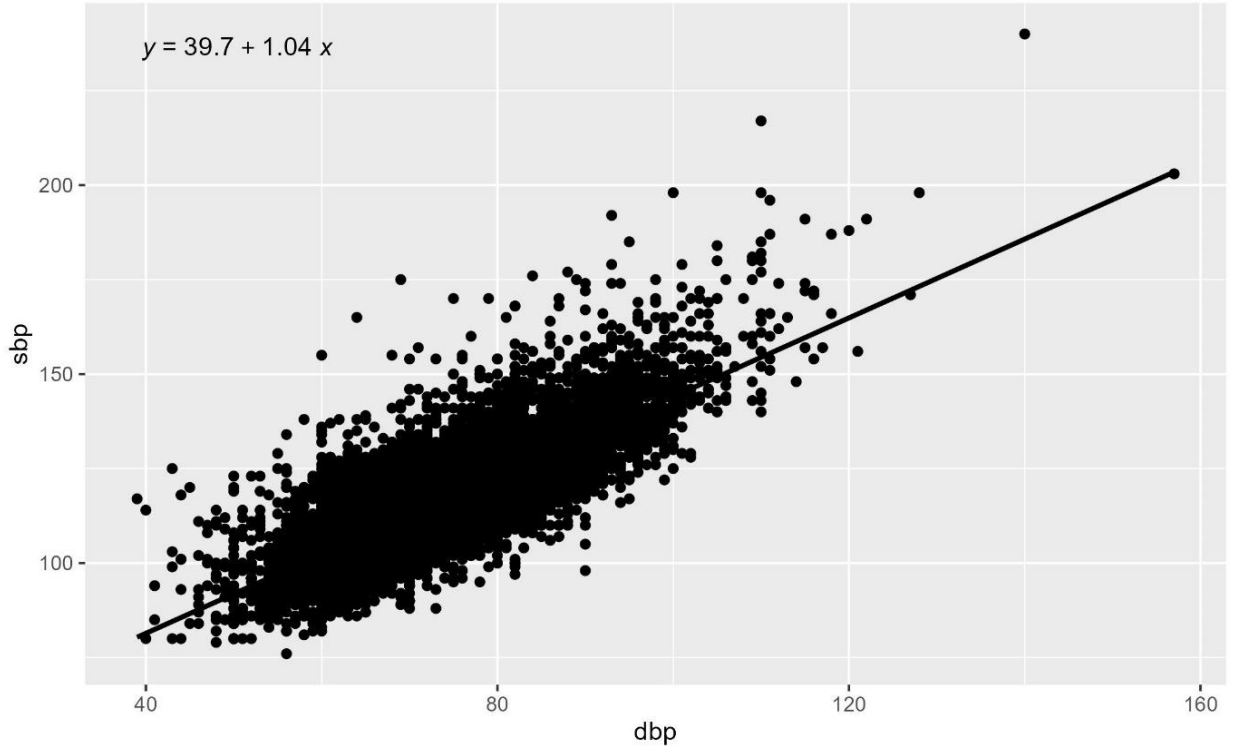
回帰

- 一般的には回帰は多変量解析で用いる

  重回帰分析、ロジスティック回帰分析、

  Cox回帰分析、など

- 2変量で回帰分析を行う場合は、回帰式を作成する



回帰式作成時の注意

- 回帰式は「片方の変数」から「もう片方の変数」を予測するための式
  ただし・・・かなり強引に回帰式は作成できる

- 回帰式作成の際には2点注意
  1.式の方向は正しいか？　因果関係が推察される場合、どちらを目的変数とするか
  2.モデルの適合度は？　R^2で表され、数字が大きいほど適合する

クロス集計

		変数A		
		1	2	3
変数B	1	○　(○%)	○　(○%)	○　(○%)
	2	○　(○%)	○　(○%)	○　(○%)
	3	○　(○%)	○　(○%)	○　(○%)

どのような分布になっているのかを見る
論文では縦(列)全体を100%となるように表記することが多い
　(集計の際にどちらを行、どちらを列にするのか要注意)
分布に偏りがあるかについて検定を行う

クロス集計に用いる検定

- カイ2乗検定、Fisherの正確検定
- McNemar検定　(マクニ(ネ)マー検定)
- カッパ統計量

クロス集計の検定は2×2分割表を前提としたものが多い
列、行のどちらかが2を超える場合は、カイ2乗およびその後の検定

カイ2乗検定

		現象2		合計
		あり	なし	
現象1	あり	a	b	a+b
	なし	c	d	c+d
合計		a+c	b+d	a+b+c+d

独立していた場合は
a:b = c:d　　a:c = b:d
の数式が成り立つはず

標本数が少ない時はP値が低めに出る傾向
→Yeatsの連続補正、Fisherの正確検定を用いる

カイ2乗検定におけるその後の検定
残差分析

- 行、列が3を超える場合、カイ2乗検定が有意であっても、どのセルで差が有るのかまでは分からない→その後の検定(残差分析)を行う

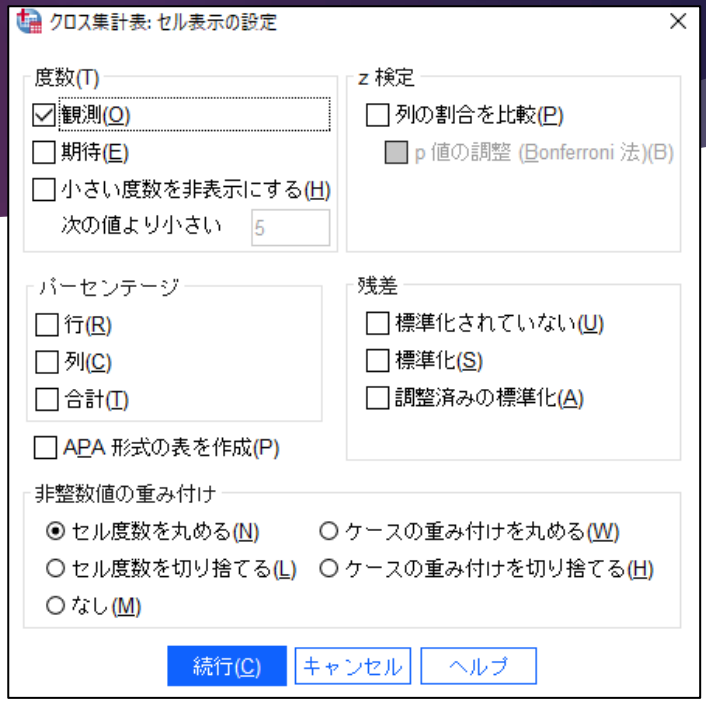
- カイ2乗検定で$p<0.05$の際に、どのセルの分布が有意に多い、もしくは少ないかを調べる検定
- 統計ソフト毎に、残差を調べる方法があるので、自身のソフトではどの用に分析するのか、調べてください。

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



- 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

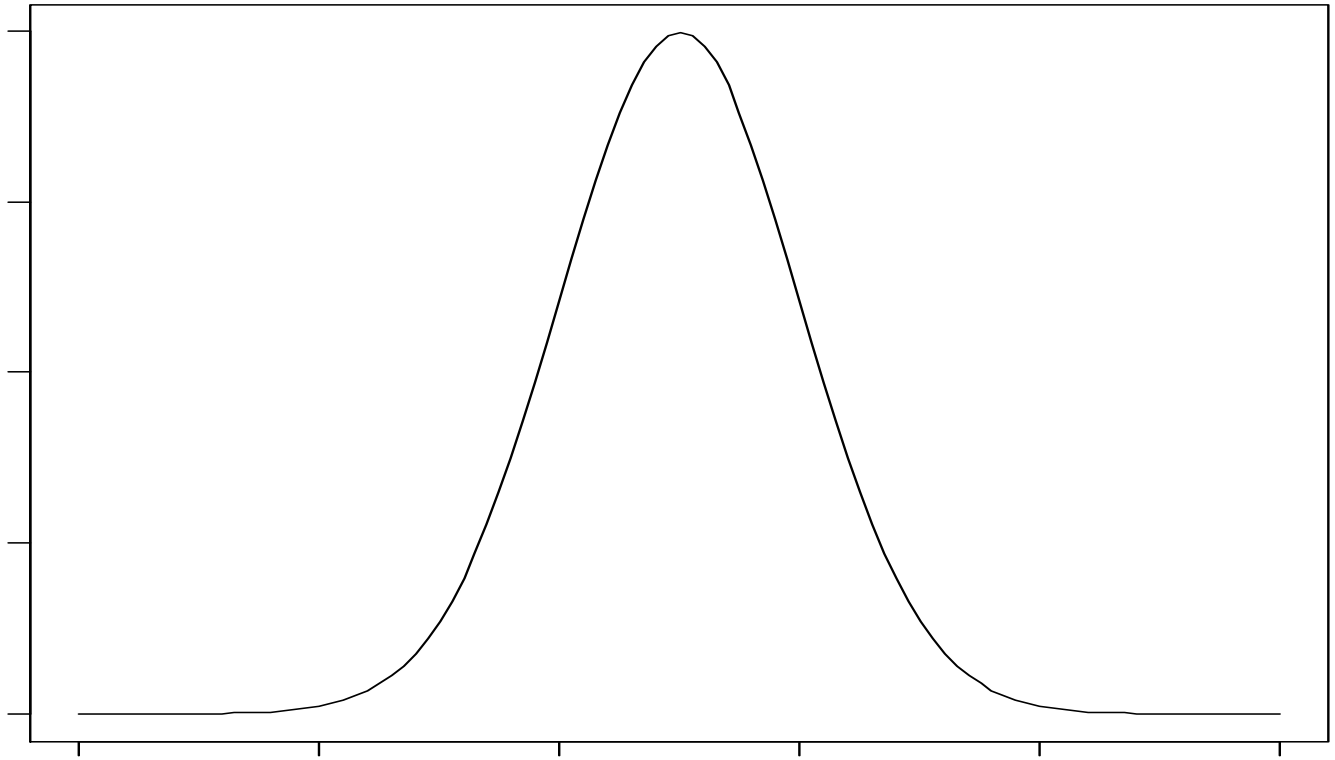
クロス表

			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



Mcnemar検定

		現象2		合計
		あり	なし	
現象1	あり	a	b	a+b
	なし	c	d	c+d
合計		a+c	b+d	a+b+c+d

対応していた場合は
bの数と c数が一致するはず

カッパ統計量

- 2種類の評価法(人)の一致率を調べる際に用いられる
- K値は0～1の間をとり、一般に0.6以上であれば一致度が高いとされる

		評価者2		合計
		悪性	良性	
評価者1	悪性	3	17	20
	良性	12	78	80
合計		15	85	100

K＝0.345

群間の比較

正規性	対応	群の数	用いる検定
あり(パラメトリック)	なし	2群	studentのt検定、Welchのt検定
		3群以上	一元配置分散分析(ANOVA)
	あり	2群	対応のあるt検定
		3群以上	反復測定分散分析(Repeated ANOVA)
なし(ノンパラメトリック)	なし	2群	Wilcoxonの順位和検定、Mann-WhitneyのU検定
		3群以上	Kruskal-Wallis検定
	あり	2群	Wilcoxonの符号順位和検定
		3群以上	Friedmanの検定

3群以上の比較を行う場合

- ANOVAやKruskal-Wallis検定で群間の差を示すだけではなく、どの群間に差があるのかの検定まで求められる

  →多重比較(その後の検定、post hoc test)がほぼ必須に

- ポイント：対照群があるのか

  対照群がある場合→対照群とその他の群それぞれの比較

  対照群がない場合→総当たり

統計量

古い教科書だと統計量とP値の対応表が付属でついていた
コンピューターでは統計量までしか計算できなかった時代の名残り

昔の論文では細かいP値まで計算できなかったため、P<0.05やNS(not significant:有意差なし)の表記があった
→今はP値を細かく計算できる時代のため、
　しっかりと表記しましょう

全体比較と多重比較

- 一元配置分散分析

  総当たり：Bonferroni, Tukey-Kramer, Holm, Bonferoni/Dunn, Williams

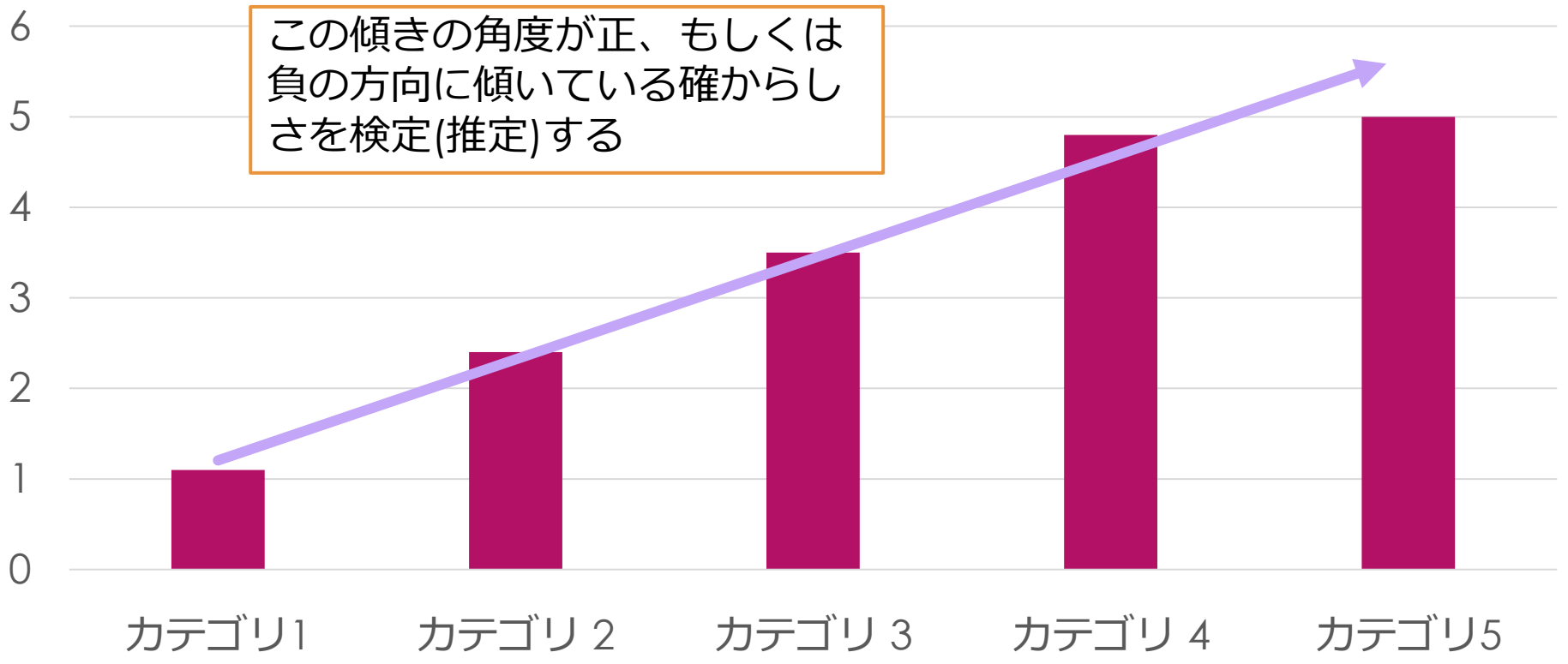
  対照群：Dunnet

- Kruskal-Wallis検定:

  総当たり：Steel-Dwass, Games-Howel, Scheffe

  対照群：Steel

傾向性の検定


この傾きの角度が正、もしくは負の方向に傾いている確からしさを検定(推定)する
6
5
4
3
2
1
0
カテゴリ1
カテゴリ 2
カテゴリ 3
カテゴリ 4
カテゴリ5

まとめ

- 変数の種類によって処理法は異なる
- 自身が取得した変数をどのように扱うのかで、結果に歪みが生じやすい
- 変数の性質に応じた適切な検定方法を選択

- 迷ったら悩まず相談を！