いまさら聞けない基礎統計学
～データの入力と分析の初歩～

男女共同参画支援室
衛生学・予防医学講座
各務竹康

昨年のセミナー以降

- ありがたいことに、1年で10件近くの相談を頂きました

- うち5件、研究デザインの話(これから調査を始める)

- うち2件(!)、査読に対する対応

2件の内訳

- 査読者のコメントにパニック

- 査読者
    - ○○の結果について、解釈が超越している
    - ○○の手法は妥当なのか？
    - ○○の説明が足りない
- 相談者
    - 自分の知らない特別な何かが必要？？？

多くの場合・・・・

- 統計「用語」アレルギーが大きいのでは？
- 査読者が統計に不案内
- 追加の分析を求める場合は、具体的(この部分についてこのように、など)に指摘することが多い
- 分析の妥当性を鋭く問われる場合も

一方で

- 集計済みデータに難渋

- データの元をたどれないので、入力ミスか、外れ値か判定困難
- データの形が融通効かない
- 統計ソフトで操作しにくい形式

今日のメニュー

1.データはこうやって入力する

　　ここがうまくいかないと分析“前”の労力が数十倍に

1.2変量解析

　　分析方法をソフトで選ぶ時の考え方

データはこうやって入力する！

ここがうまくいかないと分析“前”の労力が数十倍に

データは

- 紙(アンケートなど)
- エクセル(他者からの提供)
- その他

- など様々な形式

- 情報を電子データで一元管理する

データ入力

- 主にエクセルを使用します

- データ入力の簡便さ
- データ操作
- 統計ソフトへの取り込み

統計ソフトに取り込む際に

- ソフト毎に“癖”があるので、その癖に対応できる形式を
- 全てのソフトに対応できる形がベスト！

- 最近は互換性も高くなっているが、癖を作らないに越したことはない

データ数が多くなるほど

- 目測でのデータクリーニングは困難
- ビッグデータはソフト上でクリーニングを行うことが多い

- ただし、データ入力に失敗するとそもそもクリーニングの前の段階でつまづき、膨大なデータを目視で修正する必要性も

ちなみに

- データ数(サンプルサイズ)はどのように見積もるか？
- 細かく説明したらそれだけで軽く1時間は使いますので、簡単に
- 何を調べたいかで必要数は異なる！

例えば

- 動物実験の場合(比較項目以外全てが統制されている)

  n=5×2群でも十分に比較できる

- 記述(そもそもどのような集団なのか知りたい場合)

  事前に調べたい人を何人集めたいのか、何%その集団にいそうなのか見積もる

- 比較を行う場合

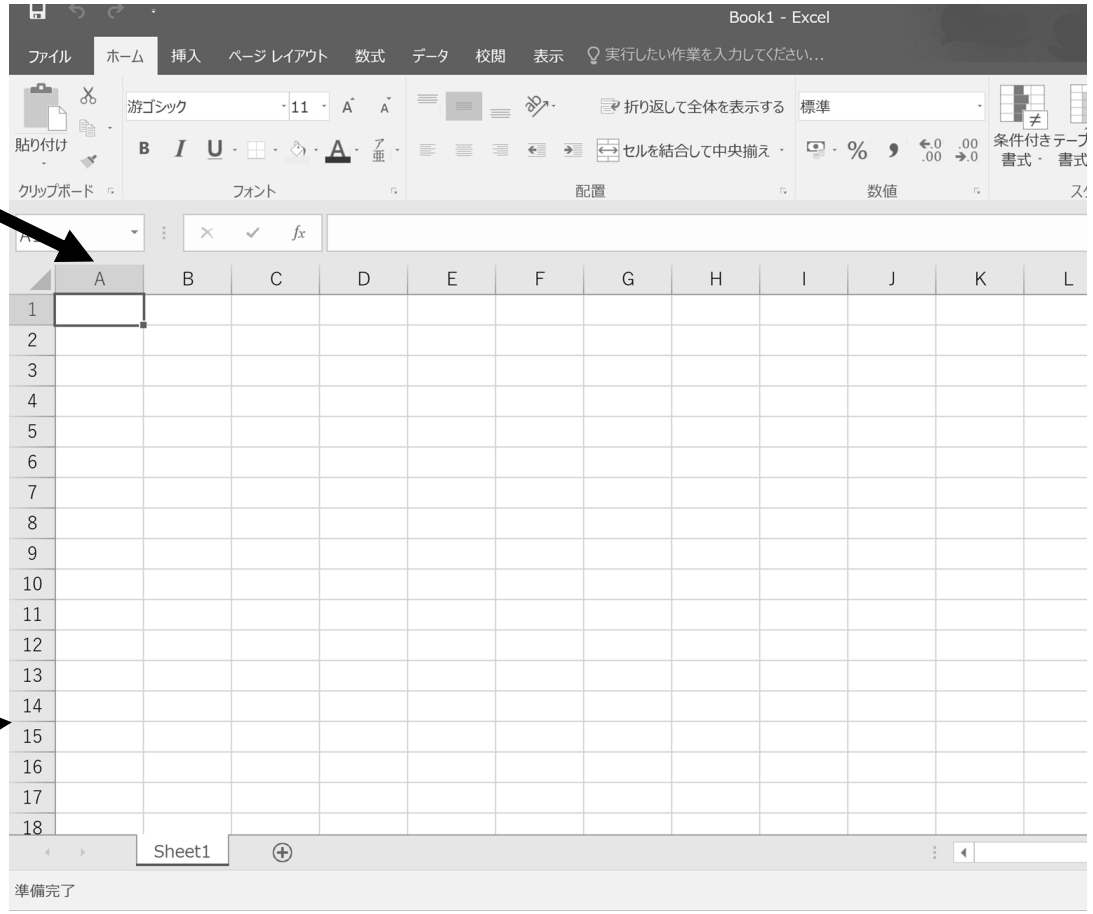
  事前のサンプルサイズ見積もり(最低症例数)の見積もりが重要

- 多変量解析を行う場合

  サンプル数が交絡因子数の必要条件を満たすように

前提

列(アルファベット)



行(数字)

原則

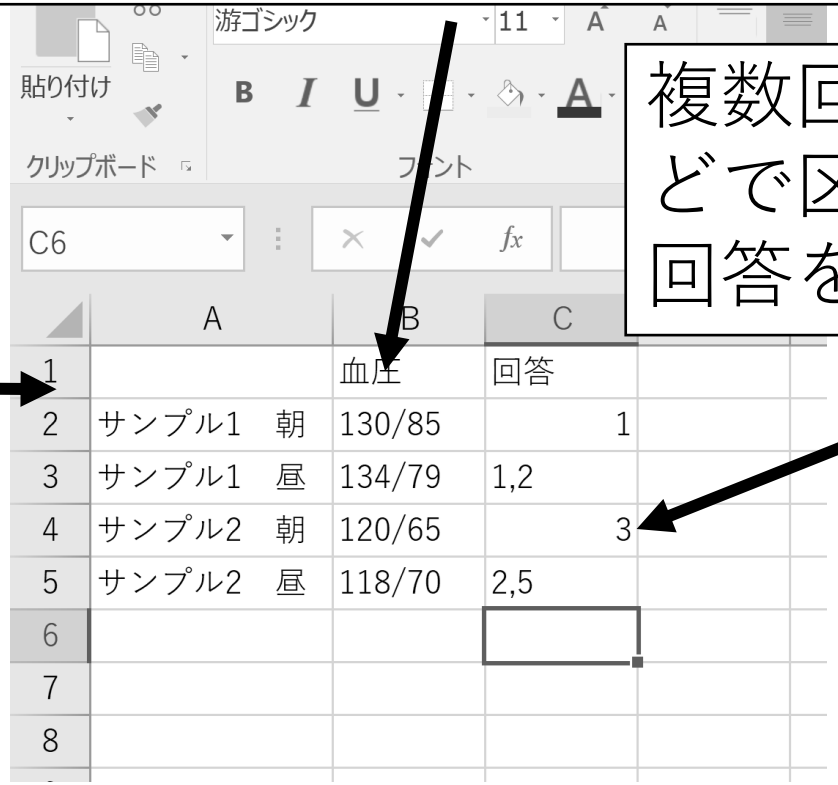
- 1人(1サンプル)のデータは全て1行に収める
- 1つのセル(マス)には1つのデータ
- 1行目に項目名を入力すること多い
- 項目名は数字から始めない
    ソフトに読み込むときに混乱する

ダメな入力例

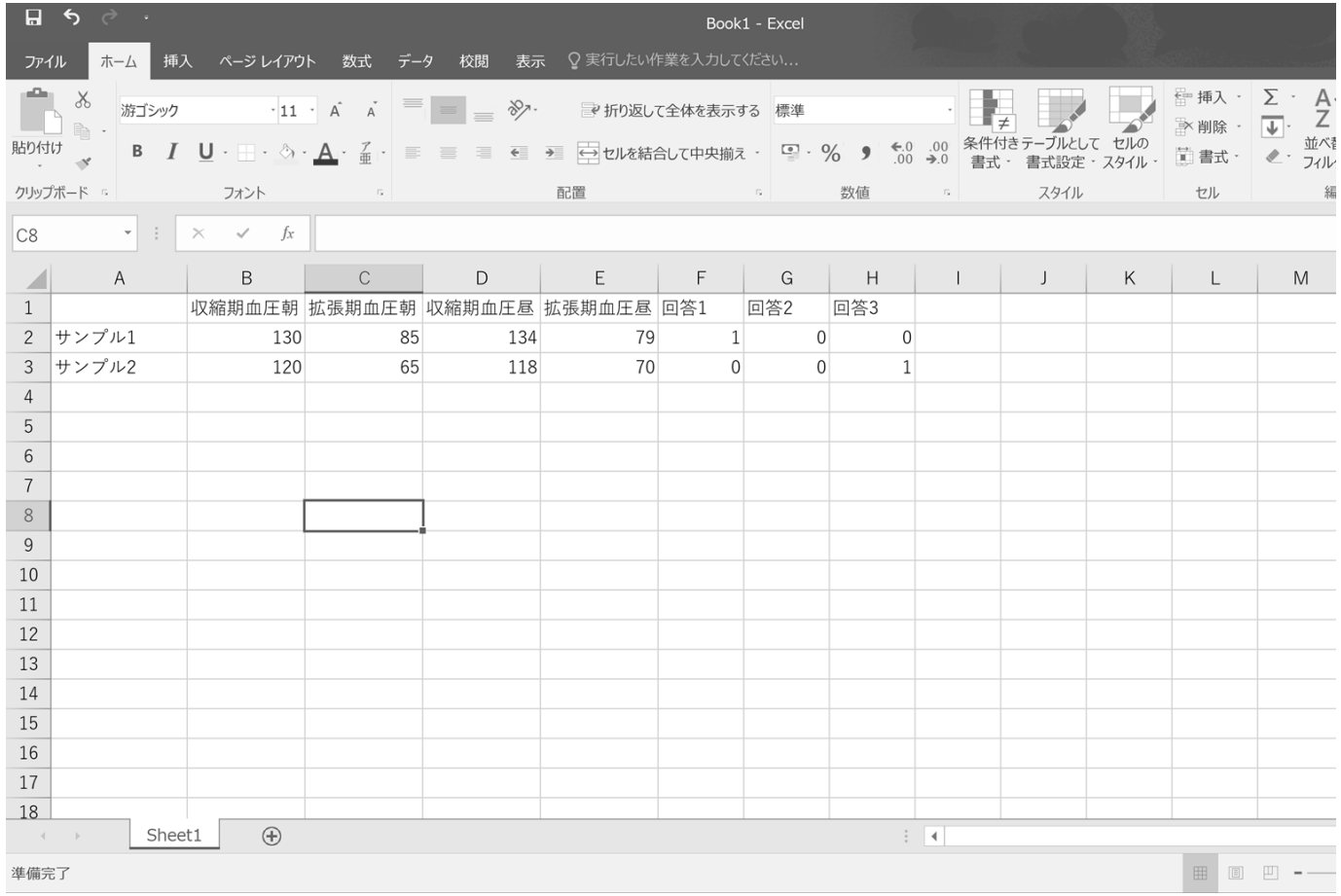
“/”を使うことで文字データになる(集計不可能)

複数回答の場合”,”などで区切って全ての回答を入力しない

1つのセルに複数の情報(サンプル名、取得時期)が含まれている



このように直す

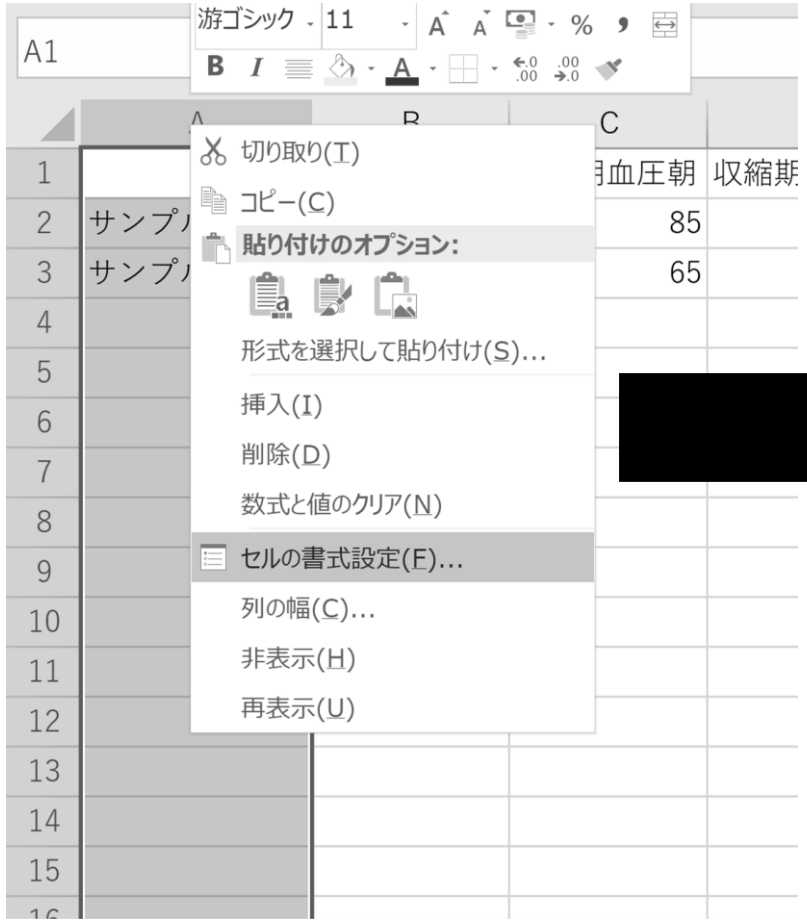


	A	B	C	D	E	F	G	H	I	J	K	L	M
1		収縮期血圧朝	拡張期血圧朝	収縮期血圧昼	拡張期血圧昼	回答1	回答2	回答3					
2	サンプル1	130	85	134	79	1	0	0					
3	サンプル2	120	65	118	70	0	0	1					
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													

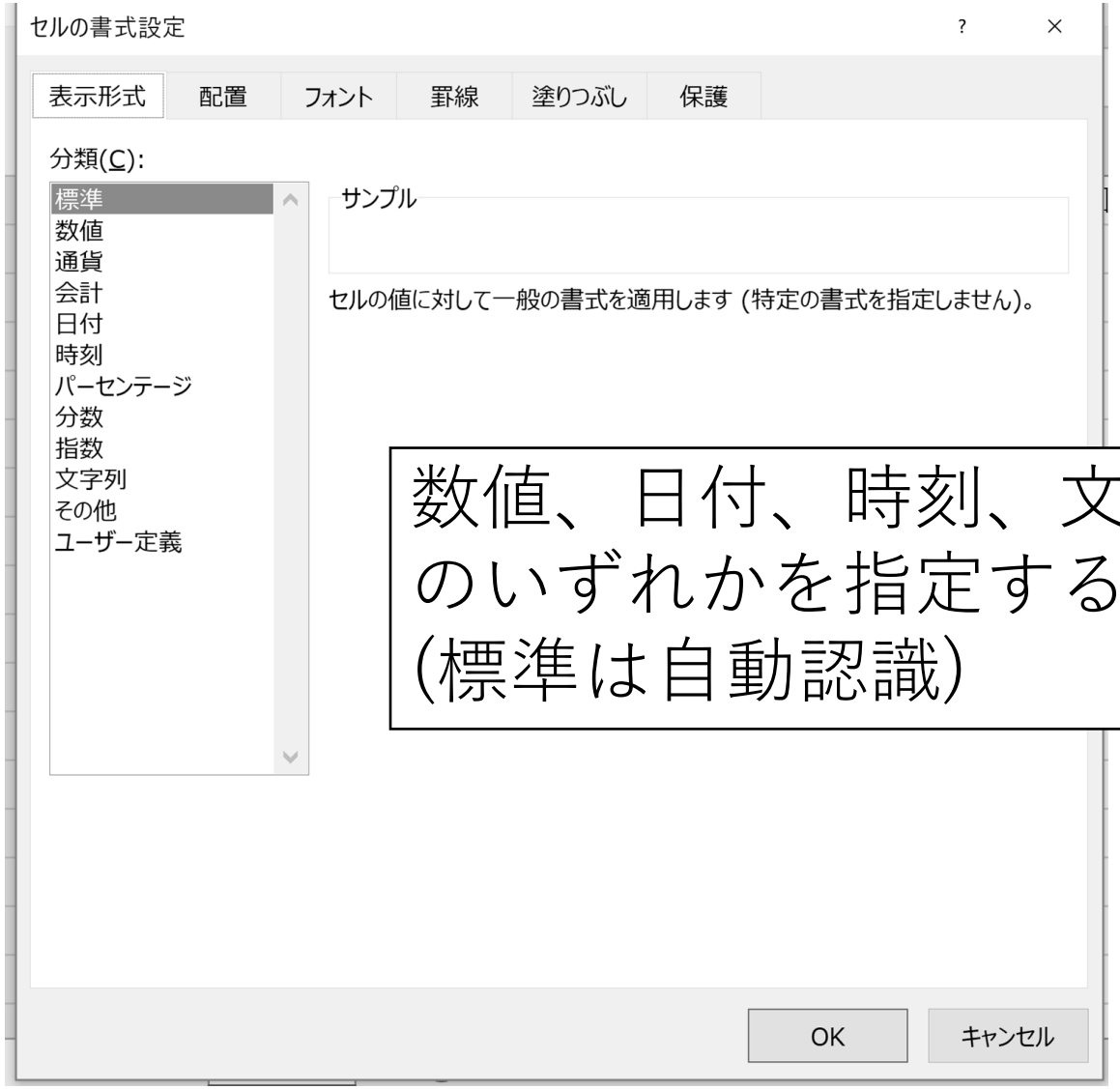
データの形1

- 数字 (1, 2, 3, …)
- 文字 (a, b, c, …)
- 時間 (YYYY/MM/DD, hh:mm)
- 空白

データ形式を確認



列の英字を右クリック



数値、日付、時刻、文字列のいずれかを指定する
(標準は自動認識)

データの形2

- WideデータとLongデータ(tidyデータとも言う)

- Wideデータ：とにかく1検体(1サンプル)の情報を横に記載していく
- Longデータ：情報の属性が似ている場合、属性の情報を与えながら全て行を変えて記載

- WideデータよりもLongデータの方が、統計ソフトと相性がよい(分析が複雑になるほど手順数が異なる)

Wideデータの例

id	sbp1	dbp1	Q1-1	Q2-1	sbp2	dbp2	Q1-2	Q2-2
A								
B								
C								

Longデータの例

id	time	sbp	dbp	Q1	Q2
A	1				
B	1				
C	1				
A	2				
B	2				
C	2				

入力の法則1

- 出来るだけ数字で入力
  カテゴリーも変数化　(男：1　女：2　等)
  (対応表を作成)
- 1行目の項目名は数字ではじめない＆出来るだけ英数で
  文字だとソフトによっては読み込みエラーとなる
  (不都合なら日本語でもしょうがない(裏技あり))
- 単位は全て省略！

GoogleフォームでのWeb調査

- CSVファイル形式(エクセルで開くことができる)で出力できる
- 選択式の場合、記載内容(1.そう思う、など)がそのまま出力されるため、加工が必要な場合も
- できるだけ加工不要な出力形式で作成するように

入力の法則2

- 単一回答：1セルに回答を入力
- 複数回答：選択肢毎に列を作成し(設問1-1, 1-2, 1-3など)、それぞれの選択肢の列に選択あり1、なし0のように2値で入力する

- あり、なしなど2値変数の場合は1,2よりも0,1の方がよい(解析の都合上)

欠損値

- 欠損値が生じないようなデータの取りかたをする
  と言ってもどうしても避けられないときも
  というわけで

- その後の集計にどのように活かすかを考えながら入力規則を作る

欠損値の入力

- 欠損値の存在が重要な場合(欠損値の数を集計する必要がある場合)

  ありえない数字を当てはめる

  1～10の選択肢に対して99を入力するなど

- 欠損値を省いて集計する場合

  そのまま空欄にする

例

選択肢	人数(割合)
1	○ (○%)
2	△ (△%)
3	□ (□%)
欠損値	◆ (◆%)

欠損値を含めた集計
(設問ごとの回答率を
比較したい場合など)

選択肢	人数(割合)
1	○ (○%)
2	△ (△%)
3	□ (□%)

欠損値を除いた集計
(純粋に回答者の中での
分布をみたい場合など)

縦断研究の場合

- 時間経過が大切

必要な情報
- 追跡開始時期
- 追跡終了時期
- イベント発生時期

- の3点を欠かさない

細胞実験の場合

- 計算の「単位」はどのようになっているのか？

- 例：10の培地で培養
  発現(生存)した培地の数を数えるのか
  培地ごとの発現(生存)数(割合)を数えるのか

- 分野による「お作法」は指導者に要確認！

入力ミス

- 厳禁！
- 入力時に全てのデータを細かく確認する！
- 解析ソフトで記述集計をする(最大、最小、度数分布)

  ありえない外れ値

  欠損値

  数字⇔文字
- 元データを確認

ありえない値

- 入力ミス　　←簡単に修正できる
- 測定ミス　　←再測定出来るなら修正可能
- 不適切サンプル　←除外を検討
- 本当の値　　←覚悟を決める

- 上記のどれであるか、適切に見極める
- 他から想像して数字を捏造しない

データクリーニング

- 解析途中でのデータ不備発見
  - 結果が大きく変更されるかも
  - 除外基準などに影響
  - さかのぼってデザインの見直しも起こるかも

- 記述集計の段階で何度もデータの不備をチェックする
- 適切なデータセットであるかを十分に吟味してから統計解析を行うこと

2変量解析

分析方法をソフトで選ぶ時の考え方

2変量解析

- 連続変数×連続変数

相関係数(Pearson, Spearman)

- 連続変数×カテゴリ変数

今日のメインターゲット

グループ分けの変数
2群、3群以上

- カテゴリ変数×カテゴリ変数

確認すること1

- パラメトリック？ノンパラメトリック？

確かめる方法
1.正規性の検定(Kolmogorov–Smirnov検定、Shapiro–Wilk検定)
2.ヒストグラムを作成して目視確認
3.過去の文献でどのように扱われているか
など様々です

確認すること2

- 群間に前後関係はあるのか？

ある場合の例
- 介入の前後
- 薬の投与時、投与後30分、60分、90分の血中濃度

対応あり

ない場合の例
- 男女での比較
- 3群に分けてそれぞれ投与する薬剤量が異なる

対応なし

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群以上の分析を行うとき

- 一元配置分散分析、Kruskal-Wallis検定
  複数の群を統合した全体としてばらつきの有無を検定している
  どの群とどの群に有意な差があるのかは、検定していない！

- **多重比較**で群間の検定を行う
  注意：いきなり多重比較を行わない
  まずはANOVA等で全体としてのばらつきに有意差があること
  とを確認してから

T検定(2群間)を繰り返してはいけない

- P<0.05を有意水準とすることが多い
- T検定を2回繰り返した場合、αエラー(間違って有意となる)可能性は

  $1-0.95^2=0.0975$　→　約10%

  繰り返し検定を行うと、全体としての有意水準が上がってしまう

- 2群間における検定での有意水準を下げることで、全体としての有意水準を5%とする作業を行う　→　**<u>多重比較</u>**

多重比較を行う時

- ANOVAを用いた際の多重比較、Kruskal-Wallisを用いた際の多重比較など、統計ソフトに実装されていることが多い

- まずは使用する統計ソフトがどのような多重比較を実装しているのか、確認してください

確認すべきシンプルな1点

- 対照群はあるのか？

- 多重比較のアルゴリズムは複数あるが、大きく対照群を有する場合(一つの群と他の群との比較)と、対照群を有しない場合(全ての群を総当たり)に分かれる

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

ANOVAは必須？

- 極端な場合、5群のうち1群だけ異なる分布の場合、ANOVAで有意差が出ない可能性もある。
- ANOVAは群が多くなるほど特異的な群を見つけにくくなる

- そのような群を見つけるためには、Dunnet, Tukey–Kramer , Bonferroniなどが適している(計算のアルゴリズムは省略)

多重比較の方法が問題になることは？

- 統計ソフトにより、実装されている多重比較は異なる
- 査読者が全ての統計ソフトを網羅しているとは限らない

- 投稿前に注意すること
    パラメトリック、ノンパラメトリックの区別はできているか
    対照群ありか、総当たりか
    適切に解釈ができているか

では、表記の方法は？

- こちらも分野のお作法次第

- グラフが一般的な場合
- 表が一般的な場合

- 総当りの多重比較は記載方法の工夫が必要

まとめ

- データ入力がスマートな形で出来たら、その後の分析も楽になる
- データ入力の段階から分析(統計ソフトでの操作)を考えて

- 変数の扱い方は、分野ごとの癖もある
- お作法については事前に学ぼう

- 解析の際には、抑えるべき場所を見極める
- 迷ったらすぐ相談！