日時：2022年2月1日（火）15:30～18:00
会場：オンライン開催（Zoom形式）
主催：統計数理研究所 大学統計教員育成センター

第1回
「統計エキスパート育成システムの構築」
に向けたワークショップ
－大学統計教員育成センター設立を記念して－

副プログラム「(DS)データ科学」の紹介とQ&A への対応

狩野　裕 大阪大学

令和3年度
「統計エキスパート人材育成エコシステム」

- 「統計エキスパート」の最上位に位置する「大学統計教員」を育成する
  - ソースは必ずしも統計学・データ科学を専門としない大学院生または若手教員
  - 育成された大学統計教員のミッション
    - ビッグデータやAI等のイノベーションに寄与する統計学を用いた融合領域の研究を振興できること
    - 大学等で核となり修士水準の統計学の授業や研究指導を行い，統計エキスパートを再生産できること

- 「統計エキスパート」とは
  - 統計を駆使して学術研究や産業界等に貢献することができる人材
- https://www.ism.ac.jp/kouhou/news/20210712.html

大阪大学大学院の
データ科学教育

- 大学間連携共同教育推進事業
  - 2014/4　副プログラム「データ科学」を開始
- 概算要求
  - 2015/10　数理・データ科学教育研究センター（MMDS）新設
  - 3本の副プログラム
    - 「データ科学」「金融保険」「数理モデル」
- 独り立ちデータサイエンティスト育成プログラム
  - 2019/4　副プログラム「DSデータ科学」構築
  - 既卒社会人も対象に

大学院等高度副プログラム「データ科学」

http://www.sigmath.es.osaka-u.ac.jp/~Estat/subprogram.html

- 開設年度
  - 2014年度
- 7つのコース
  - 統計数理コース
  - 機械学習コース
  - 医学統計学コース
  - 保健医療統計学コース
  - 人文社会統計学コース
  - 経済経営統計学コース
  - ビッグデータ&データサイエンティストコース

- 提案(幹事)部局
  - 数理・データ科学教育研究センター(MMDS)
- 協力部局
  - 基礎工学研究科
  - 経済学研究科
  - 人間科学研究科
  - 医学系研究科
  - 工学研究科
  - 理学研究科
  - 情報科学研究科

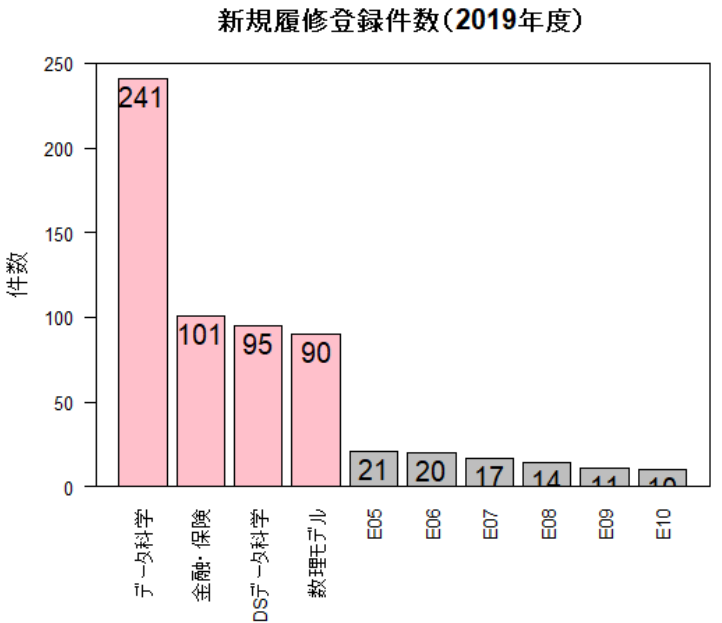
大学院等高度副プログラム「DSデータ科学」

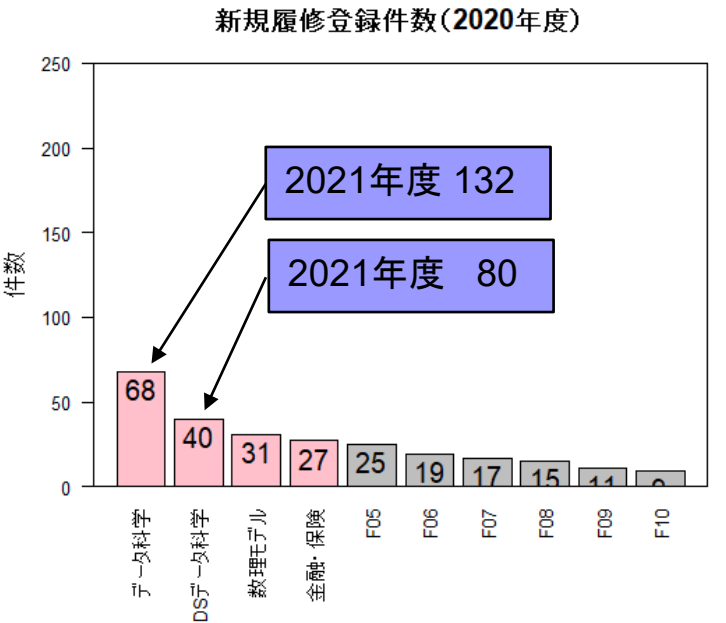
http://ds4.sigmath.es.osaka-u.ac.jp/

- 開設年度
  - 2019年度
- 6つのコース
  - DS統計数理コース
  - DS機械学習コース
  - DS医学統計学コース
  - DS保健医療統計学コース
  - DS人文社会統計学コース
  - DS経済経営統計学コース

- 提案(幹事)部局
  - 大学院基礎工学研究科
- 協力部局
  - 経済学研究科
  - 人間科学研究科
  - 医学系研究科
  - 工学研究科
  - 理学研究科
  - 情報科学研究科

MMDS+基礎工が提供する副プログラム





専攻別履修者数

副プログラム名称	コース	基礎工学研究科		他研究科			社会人	計
		数理	数理外	人文・社会科学系	理工情報系	生命系		
データ科学	統計数理コース	11	3	2	10	4	－	30
	機械学習コース	11	9	3	21	4	－	48
	人文社会統計学コース	1	0	3	2	1	－	7
	保健医療統計学コース	1	0	0	1	1	－	3
	経済経営統計学コース	0	2	9	5	1	－	17
	ビッグデータ＆データサイエンティストコース	7	1	3	12	4	－	27
	医学統計学コース							
DSデータ科学	DS統計数理コース	10	3	2	3	4	－	22
	DS機械学習コース	12	3	4	9	4	－	32
	DS医学統計学コース	5	0	0	2	2	－	9
	DS保健医療統計学コース	1	0	0	1	2	－	4
	DS経済経営統計学コース	2	1	2	2	1	－	8
	DS人文社会統計学コース	0	0	3	1	1	－	5
DSデータ科学（社会人対象）		－	－	－	－	－	12	12
計		61	22	31	69	29	0	212

カリキュラム例
「DS⁴」DS統計数理コース

概要（統計数理コース）

- 修了要件単位数
  - 12単位(6コマ)
  - 修了証を交付
- 選択必修科目A群(6単位以上)
  - DSインターンシップ
  - 実証型研究法
  - データ科学PBL
  - データ科学各論
  - 数理特論III(意思決定とデータ科学)
  - 実践データ科学演習@神戸大
- 選択必修科目B群(4単位以上)
  - 統計的推測
  - 多変量解析
  - 統計的多重比較特論1,2 @神戸大
  - モデリング基礎理論 @滋賀大
  - 数理統計学特論, 多変量解析特論,
  - ベイズ統計学特論 @同志社大
- 選択科目
  - 6科目

選択必修科目A群の詳細

- DSインターンシップ
  - 2週間 (1単位だが, 2単位と見なす)
  - マッチングはDS⁴ で実施
- 実証型研究法
  - 週末集中(3コマ/土曜×5回)
  - 研究法, コンサルテーション
- データ科学PBL
  - 夏季合宿ゼミ(3泊4日)
  - 課題解決・演習型授業
- データ科学各論
  - 週末集中(2コマ×8週)
  - DSが活躍する分野における実務家教員によるオムニバス講義
- 数理特論III(意思決定とデータ科学)
  - 前期ほぼ隔週開講
  - ビジネスアナリシスセンターによる講師陣の講義・実習
- 実践データ科学演習@神戸大
  - 地方自治体の課題とデータの提供を受け, PBL実習を実施

統計エキスパート人材の質問対応能力の醸成

- 授業を担当する場合，受講生からの質問に適切に解答できる能力が要求される
- 講述より質疑の方が辛いときがある
  - 国際会議の発表と似ている
- なぜか典型的な教科書に載っていない質問が多い
  - 初級者の質問は，素人だけに，予想外のものも多い
  - 経験のあるユーザの質問は細かく深い
  - 数式に加えて直観的説明を求められることがある
  - 質問者を満足させる回答が出せない質問もある
- Q&Aスキルは共同研究やコンサルテーションでも有用
- 質問に的確に回答すると同時に，関連した興味あるトピックを示して，さらに興味を惹き付けるように指導したい

大阪大学における
統計学・データ科学の教育FD

- 教員向けのFD
  - 授業の工夫や質問への答え方，学生とのやり取りなどについての，教員間の意見交換・情報交換の場
  - カリキュラムやテキスト選択なども議論
  - 年に数回
- 次スライド以降に，受講者から受ける典型的と思われる質問（一部）を一覧にまとめる
- 発表当日はこのリストからいくつか回答例を示したい．

リテラシー編

- 順序尺度のデータはヒストグラムか棒グラフか？
- いろいろな平均を習うけどテキストの後半で出てこない．たとえば調和平均とかはどこで役に立つのでしょうか？
- 収入の分布は歪んでいるのでその代表値は中央値だと習った．演習問題で，球団の選手年俸の平均値を使った分析があった．なぜ中央値ではないのか？
- ばらつきの尺度では標準偏差と四分位範囲をよく見る．
  これらの使い分けを教えてください．平均偏差はなぜあまり使われないのでしょうか？
- 事象の独立と排反がよくわかりません．視覚的に説明できないのでしょうか？
- 共分散の計算はできます．その意味が分からない．
- 相関係数や決定係数の大きさはいかほど？
- 外れ値と思われるデータでも自動的に削除するのはよくないと言われた．なぜか．具体例を挙げて説明して欲しい

初級編（統計検定2級レベル）

- 両側検定と片側検定の使い分け．不適切な例題が多いように思も思うが，よく分からない．
- カイ二乗検定やF－検定は片側検定でしょうか？
- H0を受容するときの表現は難しいが，不適切な説明のテキストがある．
- 独立二標本のt－検定で，合併した分散の意味が分かりません．二つの分散を単純に平均するのではだめなんですか？
- 確率変数の意味が分からない．確率変数が理解できないとデータ解析のどこで困るのでしょうか？
- 不偏分散の存在意義が分からない．計算はできるけど．標本分散では何故いけないのか？（言葉使いも変．不偏標本分散？）
- 不偏分散の平方根は標準偏差の不偏推定量にならない．
- ランダムと無作為の違いは？
- エラーバーとして示される統計量には標準偏差，標準誤差，信頼区間などがあるようだが，その使い分け方がわからない．

初級編(続)

- 母平均を二群比較するとき，標本サイズはなるべく揃えておくべきだと教えられた．何故でしょうか？確率比例抽出法の役割は何でしょうか？
- 層別抽出法は単純無作為抽出法より良いと習った．何故でしょうか．直観的にわかるように教えてください．無作為抽出法は標本抽出の王様ではないのでしょうか？
- 単回帰直線は，データの散布図の真ん中を通っていず，少し寝ている．直線を最小二乗法で決めるのは分かるが，データのバルクを捉えられていないのはまずいのではないか．
- Xが二値変数，Y|Xが正規分布に従う量的変数の場合に，ピアソンの相関係数を計算し，通常の母相関＝0の検定を行ったが問題ないか？poly(bi)serial correlationとは何でしょうか．
- 統計学で予測式を作り当たらなかったら誰が責任をとるのでしょうか
- 統計学とデータサイエンスの違いは？
- 結局，仮説検定で何が分かるのでしょうか？
- 正解が複数あり得る方法はサイエンスと言えるのか

中級編

- 分散分析･多重比較法
  - タイプⅠ, Ⅱ, Ⅲの平方和のわかりやすい説明，使い分け，を教えてほしい．理由も．
  - 交互作用の意味が分からない．交互作用という具体的な原因があるのでしょうか．
  - 主効果･交互作用にはいくつか異なった制約式がある．母数を一意に決めるためのものであり，数学的に同値であることは分かる．では，どのように使い分けるのか．異なった制約式を採用すると解釈が変わるのでしょうか．
  - 乱塊法計画ではブロック因子を入れると言われた．ブロック因子とは何か，分析にどのような影響があるのか．手続きだけでなく意味を教えてほしい．
  - 多重比較法は重要なんでしょうか．分散分析では，要因の水準間の比較には多重比較法が用いられるが，要因効果の検定には多重比較法はあまり使われない．何故なんでしょうか．
  - 一要因（3つ以上の群）で平均値を比較するとき、なぜ分散分析するのですか？直接多重比較しないのはなぜですか？
  - 分散分析と水準間多重比較の結果が一致しないとき，どのような処方箋があるか？
  - 分散分析における効果量はどのように評価すべきか．特にη^2と偏η^2の使い分けと目安．

中級編(続)

- 回帰分析
  - 回帰分析の説明変数が固定変数と扱われるのはなぜでしょうか．そもそも固定変数(確定変数)とは何でしょうか，名称も矛盾しています．
  - 重回帰分析を行った．本来正の偏回帰係数と思われるが負の値として推定された，どうしたらよいか？なお，多重共線性は起こっていないようである．
  - 重回帰分析を行った．標準偏回帰係数が1を超えている．どうしたらよいか？なお，多重共線性は起こっていないようである．
  - 抑制変数とは何でしょうか．何を抑制するのでしょうか．
  - 固定効果と変量効果の違いと使い分けは？
- 標本調査
  - 層別には事前と事後の2種類があると習った．統計的な性質として何が違うのか？
  - 標本抽出法をいくつも習った．これらの方法はすべて，各構成要素の抽出確率が等しいことを保証しているのでしょうか．だとすると，標本抽出法の理論的な違いはどこにあるのでしょうか？
  - 標本調査における回収率の低さはどの程度まで許されるのでしょうか．

中級編(続)

- 効果量の大きさについて具体的なイメージがわかない．Cohenが効果量を大中小と区別しているのを知っているが有効なのでしょうか？
- 帰無仮説からのずれを表す効果量を習った．一方で対立仮設の下での状況を表す量として非心母数がありますね．両者に関係があるのでしょうか．
- 偏相関係数は公式に基づいて計算はできる．この公式の直観的な意味を説明してほしい．
- 5-point Likert scale(5件法)で取ったデータを平均したり相関を求めたりしてよいのでしょうか．
- AICの数理的な導出は分かる．倹約の精神を活かしたモデル選択であることも分かる．しかし，複雑さがペナルティになることがどうしても腑に落ちない．
- 傾向スコアのモデル選択ではAIC，BICなどの情報量は使わないと言われた．本当か？
- サンプルサイズ n は普通与えられた定数とするが，確率変数である場合も多いように思う．
- データ科学・統計学と機械学習との違いは何でしょうか？
- Kernel法はなぜ非線形関係に対応できるのでしょうか？
- 二次元散布図に10次の多項式回帰を使ったら割とよく当てはまったが，機械学習的方法の方が良いのではと言われた．何故なんでしょうか？

回答例(2件)

- 結局，仮説検定で(or 統計で)何が分かるのでしょうか？
- 合併した分散の意味が分かりません．二つの分散を単純に平均するのではだめなんですか？

結局，仮説検定で(or 統計で)何が分かるのでしょうか？

- 質問の意図
  - 以下のようなことを言っているのではない
    - 分布やモデルの仮定の適切性，標本の適切性，サンプルサイズが巨大だが有意性を確認，推定効果量が小さい，解釈が不可能， p値が0.05ギリギリ，多重比較すると非有意に，帰無仮説が採択されたときは. . .
  - 統計法に関する根本的な問い

データ科学・統計学の活躍の場

- （科学的）理論研究の初期段階（探索的研究）において重要
  - 第一原理による説明に接続
- すぐには解明できそうにはないが，早期の対応が要求される現象の理解・予測・制御に有効
  - ランダム現象の予測モデル，因果モデル
    - Keep stayは有効だった？
    - 薬効・毒性
  - 注： AIとインターネット，情報端末等を活用した，ざっくりとした予測はより俊敏
- 原理原則の展開で説明困難な現象の説明・予測・制御に有効

合併した分散の意味が分かりません．二つの分散を単純に平均するのではだめなんですか？

- 二標本，共通分散，母平均の差の検定

$$U^2 = \frac{(n_1 - 1)U_1^2 + (n_2 - 1)U_2^2}{n_1 + n_2 - 2}$$

U^2って何？

- 初級者は，以下のような説明ではおそらく納得しない
  - 共通分散σ^2の不偏推定量
  - U^2がカイ二乗分布(÷自由度)に従う
  - 尤度比検定統計量
  - U^2 は共通分散σ^2のBLUE
    - 最良線形不偏推定量
  - こちらの方がベターだが:

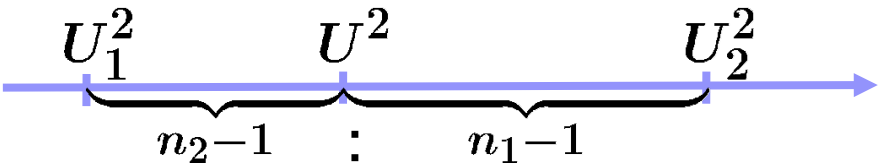
$$U^2 = \frac{(n_1-1)U_1^2 + (n_2-1)U_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2 + \sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}{n_1+n_2-2}$$

重み付き平均という説明

合併(Pool)した不偏分散は，二つの推定量の推定精度を考慮した重み付き平均，という説明は如何でしょうか．

$$U^2 = \frac{(n_1-1)U_1^2 + (n_2-1)U_2^2}{n_1+n_2-2}$$
$$= \frac{n_1-1}{n_1+n_2-2}U_1^2 + \frac{n_2-1}{n_1+n_2-2}U_2^2$$

- U^2は，U_1^2とU_2^2の重み付き平均．
- U^2は，U_1^2とU_2^2を$(n_2-1):(n_1-1)$に内分する点．
- 標本サイズが大きいとばらつきが小さくなり精度が高くなる．



ご清聴に感謝を申し上げます