2021년도 대한재활의학회 추계학술대회
편집위원회 워크숍

내게 맞는 통계 소프트웨어 찾기

2021년 10월 29일(금)

황 승 식
서울대학교 보건대학원
cyberdoc@snu.ac.kr

표 1 – 역대 뉴잉글랜드 의학저널에 가장 흔히 사용된 통계 기법 다섯 가지

과거(1978-1979)	현재(2005)	조만간
• *t*-검정 (44%) • 분할표 (27%) • 피어슨 상관 (12%) • 비모수 검정 (11%) • 생존 기법 (11%)	• 생존 기법 (61%) • 분할표 (53%) • 다중 회귀 (51%) • 검정력 분석 (39%) • 역학적 통계 (35%)	• 다중 비교 • 고효율(high-throughput) 데이터 분석 • 고급 실험 설계 • 베이지언 분석 • 순열/붓스트랩 기법

왼쪽에서 두 컬럼은 Switzer와 Horton이 *CHANCE* 2007년에 발표한 "당신의 의사가 알아야 하는(그러나 아마도 모를) 통계학"에서 인용. 괄호 안은 해당 기법을 적용한 논문의 백분율. 오른쪽 컬럼은 빠르게 대중화되고 있는 기법에 대한 예측.

출처: Brieger K and Hardin Johanna. Medicine, Statistics, and Education: The Inextricable Link. *Chance* 2012;25(3):31-34

기타 자료

웹세미나

인포그래픽

Q&A 포럼

연구 윤리의 이해 | 학술 출판

연구재현성 문제

Last updated Oct 25, 2019

연구 결과의 재현성은 과학에서 기초 중의 기초라고 할 수 있습니다. 자신 이외에 어느 누구도 같은 과정으로 같은 결과를 얻지 못한다면 연구 결과의 참신성과 신뢰도를 인정 받을 수 없습니다. 발표된 연구결과가 정말로 원했던 결과인지, 우연에 의한 것은 아닌지 혹은 오류가 아닌지 구분할 수 없게 되기 때문입니다.

작년 네이처(Nature)지에서 약 1,500명의 연구자들에게 재현성에 대한

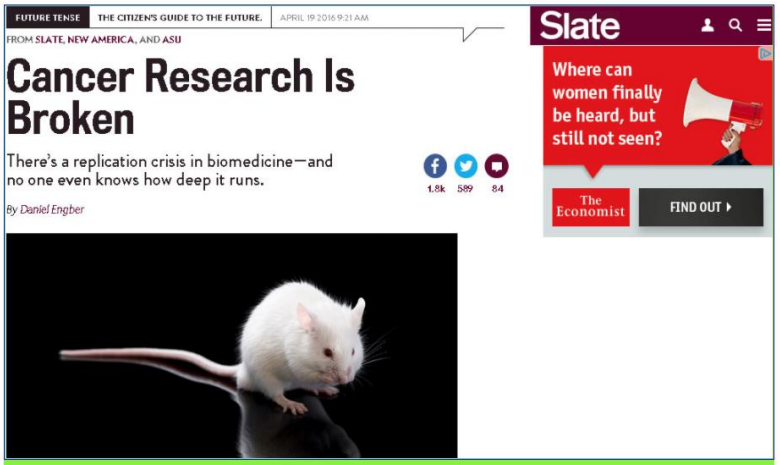
박준석의 "'실험 재현성의 위기' 바로보기"

발표된 실험 결과가 재현되지 않는 '재현성의 문제'가 요즘 실험 과학계에서 화두입니다. '재현성의 위기'가 무엇이며, 과학자들은 그 위기를 극복하고자 어떠한 노력을 기울이고 있는지를 심리학 박사과정 박준석 님이 정리합니다.

재현성 위기는 과학불신과 연구낭비를 초래한다

박준석 | 2016. 04. 27

[2] 재현성 위기의 실태와 그 결과



» 지난 4월 19일, 미국 언론매체인 <슬레이트(Slate)>는 '암 연구가 잘 작동하고 있지 않다'(Cancer research is broken)라는 제목의 글에서, 의학 및 관련 생명과학 분야의 연구 일반의 재현성에 대한 문제를 제기했다. 출처/ http://www.slate.com

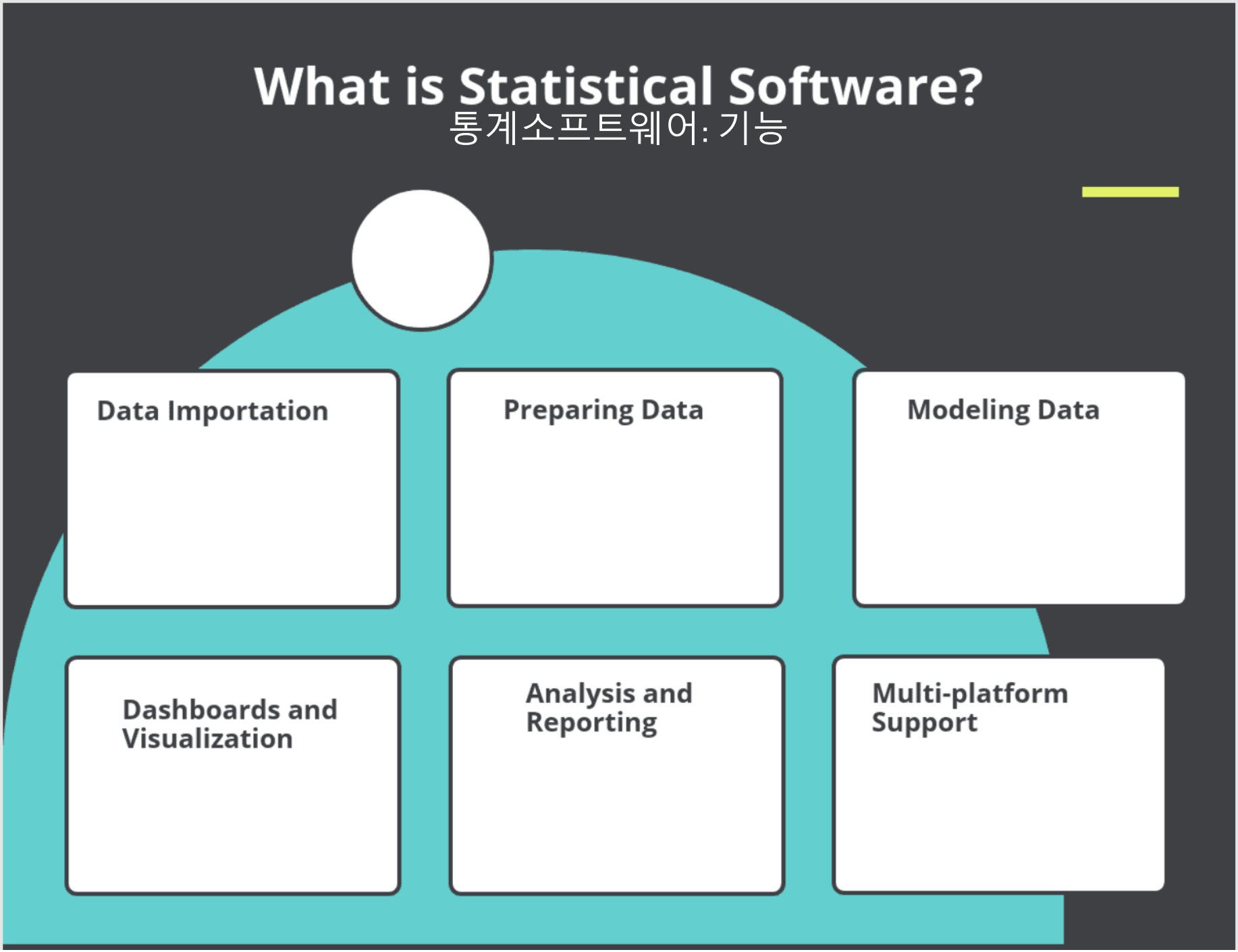
'재현성 위기'(reproducibility crisis)라는 용어가 본격적으로 사용되기 시작한 것은 2010년대 초였다. 이런 데에는 심리학 분야에서 있었던 몇몇 사건이 결정적 역할을 했다.

이를 테면 지난 2011년, 해당 분야 최고의 권위를 자랑하는 <성격 및 사회심리학회지(Journal of personality and social psychology)>에 인간에게 예지력이 있다는 주장을 담은 논문이 출간되었다. 이 연구에서 연구자들은 컴퓨터 화면에

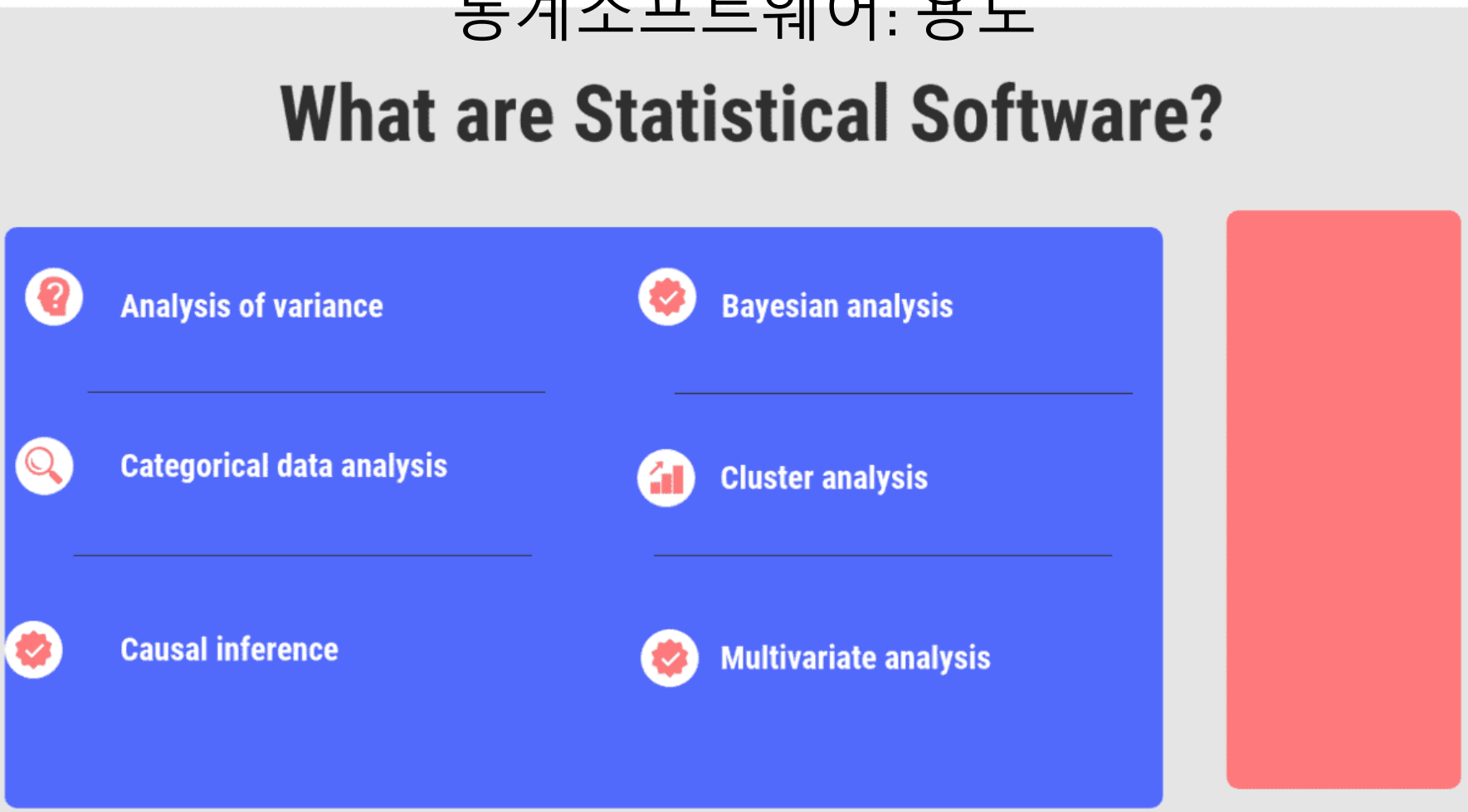
검색
자유게시판 "너른마당"
최근글
· [알림] 사이언스온이 미래&과학으로 ...
· 뇌과학, 인공지능과 우리
· '내 연구를 소개합니다', 연구자 위한...
· '잘해야 해!' 질식할듯한 긴장이 만든 ...
· '부산행'의 좀비와 감염병 인식: 의학...
인기글
최근 댓글
한겨레 미래&과학
5.2천 좋아요
좋아요
한겨레 미래&과학
23시간 전
2분46초 동안 600여미터를 날며 공중 탐사촬영을 했다. 그동안 퍼시비런스와 가까운 거리를 유지해 왔던 인지뉴이티가 처음으로 퍼시비런스의 시야를 벗어난 지점까지 자율적으로 임무를 수행하는 데 도전해 성공했다.

What is Statistical Software?

통계소프트웨어: 기능



통계소프트웨어: 용도


What are Statistical Software?
Analysis of variance
Bayesian analysis
Categorical data analysis
Cluster analysis
Causal inference
Multivariate analysis

Top Statistical Software

IBM SPSS Modeler, Minitab, Develve, XLSTAT, Forecast Pro, Analyse-it, SmartPLS, PolyAnalyst, Regression Analysis of Time Series, SAS Visual Statistics, Stata, AcaStat, MATLAB, EViews, JMP, Mathematica, Qlucore, MedCalc, NCSS, EasyFit, MaxStat, Data Desk, StatPlus, GAUSS, Statgraphics Centurion, TurboStats, Genedata Analyst, NLOGIT, Analytica, SigmaPlot, GeneXproTools, WinSPC, GraphPad InStat, UNISTAT, StatsDirect, Statwing, StatXact, statistiXL, Statistix, Number Analytics, LIMDEP, SUDAAN, PASS, NLREG, ESBStats, Origin, Maple, SuperCROSS are some of the best statistical software .

Top Free Statistical software

SAS University Edition, GNU PSPP, Statistical Lab, Develve, Shogun, DataMelt, GNU Octave, SOFA Statistics, Dataplot, SciPy, Zelig, Scilab, Gretl, OpenStat, Past, MacAnova, MaxStat Lite version, SageMath, Epi Info, NIMBLE, Arc, ADaMSoft, CumFreq, OpenMx, Salstat, Statcato, Stan, IDAMS, OpenEpi, BV4.1, pbdR, GNU Data Language, Dap, Simfit, First Bayes, MicrOsiris, Ploticus, NCAR Command Language, Perl Data Language, Yorick, EasyReg, IVEware, ViSta, StatCVS, WinBUGS, JAGS, WINPEPI, ADMB are some of the top free statistical analysis software.

Computer Methods and Programs in Biomedicine 27 (1988) 83–94
Elsevier

CPB 00911

Section III. Experiences with systems and programs

Evaluation of statistical packages for suitability for use by clinical investigators in medicine

Linda S. Chan and Bernard Portnoy

University of Southern California School of Medicine and Los Angeles County-University of Southern California Medical Center, Los Angeles, CA, U.S.A.

3. Evaluation criteria

For evaluating the suitability of a statistical software package, five general areas of concern are addressed:

1. availability of data-management features;
2. availability of statistical-analysis features;
3. ease of use or user-friendliness;
4. documentation; and
5. quality of the procedures.

TABLE 3

Comparison of overall suitability ranking and cost

Package name	Suitability ranking	Cost ($) *
PC STATISTICIAN	3	299
CRISP	1	495
MICROSTAT	2	99
SYSTAT	5	495
MSUSTAT	6	187
STATPAK	4	199

* Approximate cost as of early 1986.

2004년 3월 29일
덴마크 병원에서 회의 장면

(From Svend Juul, Stata and the Newcomer)

A: 우리 연구진의 연구 능력을 향상시키고 싶습니다. 기능이 좋고, 설치가 쉬우며, 가격도 적당한 통계 패키지가 필요합니다. 아, 그래프를 그리는 능력도 포함해서요.

B: Stata가 정답이겠네요.

A: 그렇진 않은 듯한데요. 사람들 말이 처음 배우기가 어렵다고 합니다. 전임 연구자라면 몰라도, 우리 병원의 젊은 의사 연구자들은 사용법을 익히기 위해 몇 주 씩 투자할 여력이 없어요.

B: 제가 듣기에 SPSS가 더 사용하기 편하다고 들었는데요.

A: 하지만 그건 너무 비싸요. Excel은 어떤가요?

Dembe et al. BMC Health Services Research 2011, 11:252
http://www.biomedcentral.com/1472-6963/11/252

BMC
Health Services Research

RESEARCH ARTICLE Open Access

Statistical software applications used in health services research: analysis of published studies in the U.S

Allard E Dembe[1*], Jamie S Partridge[2] and Laurel C Geist[3]

Table 1 Mention of Statistical Software in HSR Articles, 2007-2009, by Year

	2007	2008	2009	2007-2009
Total articles	393	374	372	1139
Excluded articles	111	66	85	262
Included articles	282	308	287	877
Included articles not mentioning software	110	120	112	342
Included articles mentioning software	172	188	175	535
% of all included articles that mentioned software	61.0	61.0	61.0	61.0
Number of software mentions	212	224	201	637
Average number of software mentions per included article	1.2	1.2	1.1	1.2
% of articles in which Stata was used*	48.3	42.6	47.4	46.0
% of articles in which SAS was used*	37.2	43.1	47.4	42.6
% of articles in which SUDAAN was used*	10.5	5.3	2.9	6.2
% of articles in which SPSS was used*	4.7	8.5	4.0	5.8
% of articles in which other statistical software was used*	22.7	19.7	13.7	18.5

* Note: percentages add up to more than 100% because some articles mentioned the use of more than one statistical software application.

Open Access Original Article

DOI: 10.7759/cureus.12639

Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis

Emad Masuadi [1], Mohamud Mohamud [2], Muhannad Almutairi [3], Abdulaziz Alsunaidi [3], Abdulmohsen K. Alswayed [3], Omar F. Aldhafeeri [3]

1. Research Unit/Biostatistics, King Saud bin Abdulaziz University for Health Sciences, College of Medicine/King Abdullah International Medical Research Centre, Riyadh, SAU 2. Research Unit/Epidemiology, King Saud bin Abdulaziz University for Health Sciences, College of Medicine, Riyadh, SAU 3. Medicine, King Saud bin Abdulaziz University for Health Sciences, College of Medicine, Riyadh, SAU

Corresponding author: Emad Masuadi, masuadie@ksau-hs.edu.sa

Review began 12/23/2020
Review ended 01/09/2021
Published 01/11/2021

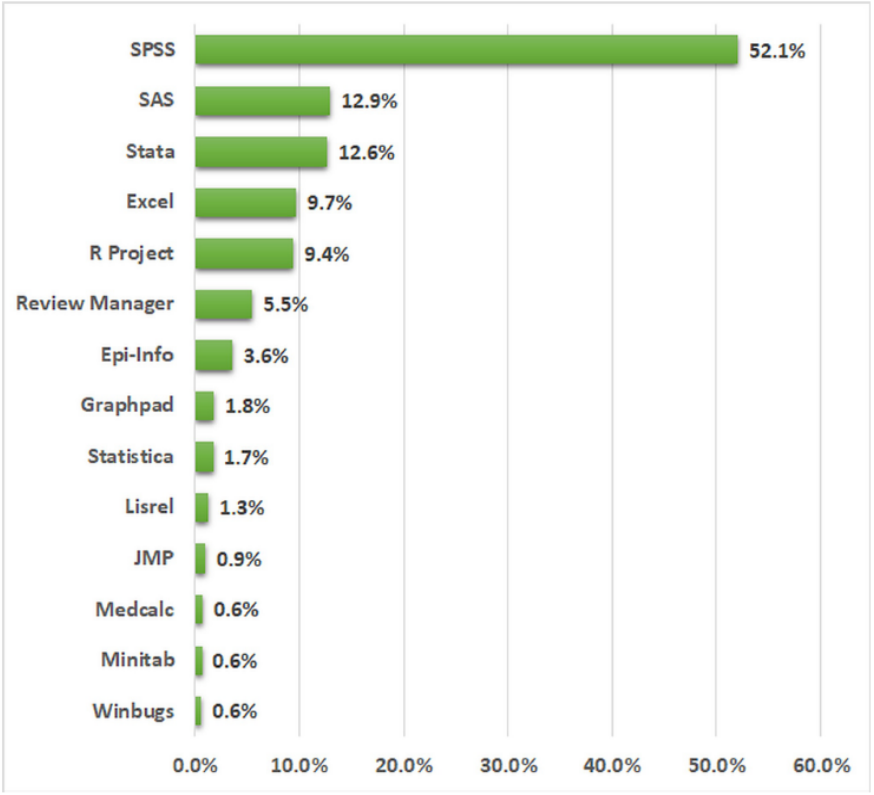


FIGURE 2: Percentages of the statistical software used in the reviewed articles.

The total percentage was 113.3% since some articles used more than one software.

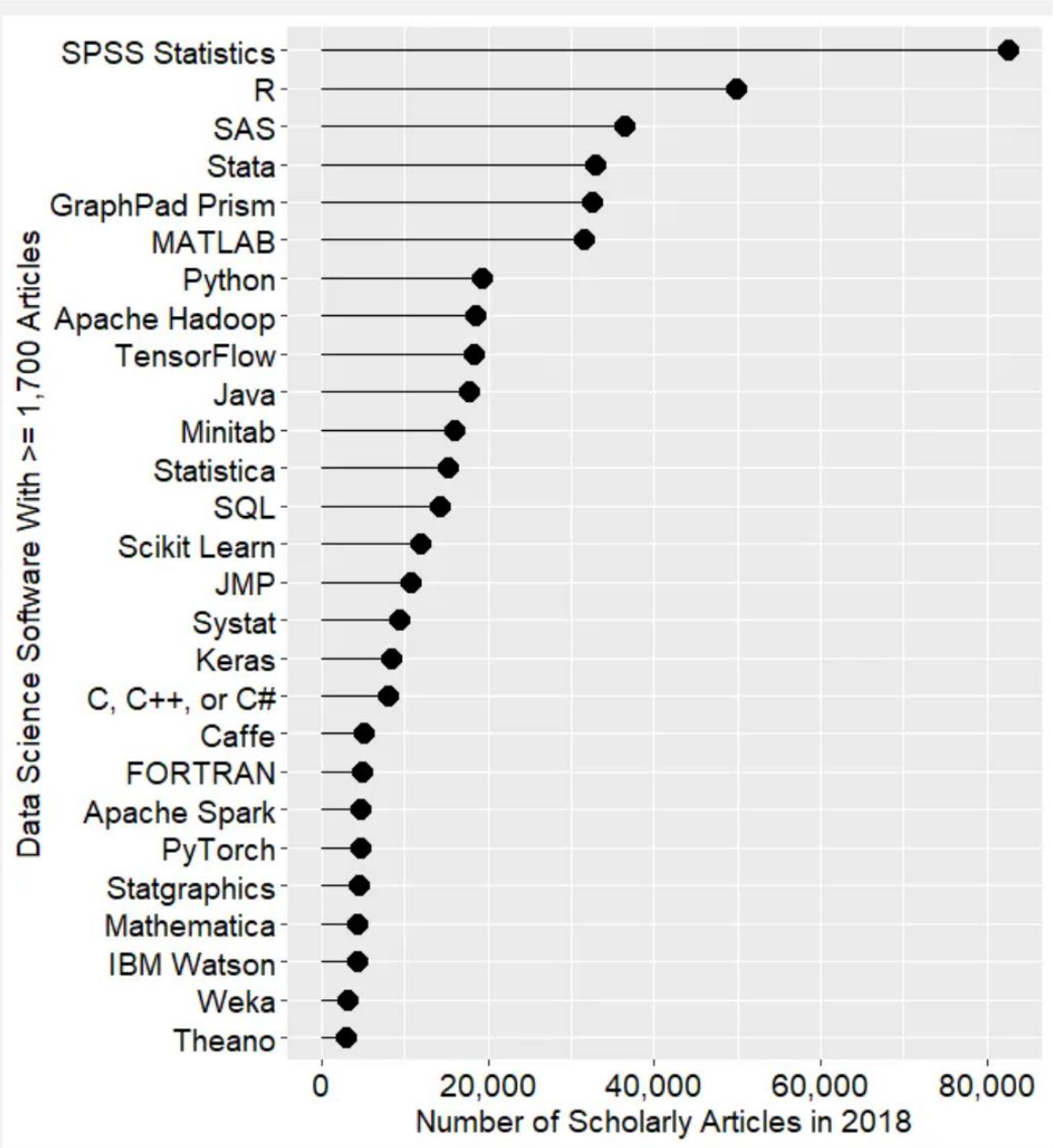


Figure 2a. The number of scholarly articles found on Google Scholar, for data science software. Only those with more than 1,700 citations are shown.

How does Stata compare with SAS and SPSS?

- Mitchell, MN (2007). *Strategically using General Purpose Statistics Packages: A Look at Stata, SAS and SPSS.* Statistical Consulting Group: UCLA Academic Technology Services.

Statistical Consulting Group
UCLA Academic Technology Services

Technical Report Series

Updated February 2, 2007 | Report Number 1, Version Number 1

Strategically using General Purpose Statistics Packages: A Look at Stata, SAS and SPSS

Michael N. Mitchell

라이센스 정책

- Stata >> SAS> SPSS
  - 비용: 영구 라이센스 Stata/SE 비용 약 150만원
  - 기간: 영구적!
  - 설치: 한 번 이상 설치 필요 없음
  - 추가 비용: 분리된 모듈에 대해 추가 비용 지불 불필요

설치와 업데이트

- Stata > SPSS >> SAS
  - DVD 한 장
  - 라이센스 코드만 정확하게 입력
  - 약 5분 소요
  - 200 MB 하드 디스크 공간 필요.

지원 구조

- Stata = SAS > SPSS
  - 기술 지원
  - 공식 웹사이트: www.stata.com
  - 설명 : **`webuse`** 명령어
  - 온라인 도움말: **`help`** 명령어 (예. **`help ttest`**)
  - 추가 서적: Stata Press
  - 부가 프로그램: **`findit`** 명령어
  - Stata 저널
  - 교육훈련: 온라인 코스

Features

- 분산 분석(ANOVA): SPSS = SAS > Stata
- 데이터 배포 용이성: Stata = SPSS > SAS
- 외부 데이터 불러오기: SPSS > Stata > SAS
  - Stat/Transfer 프로그램 사용 필요(별도 구입)
- 확장성: Stata > SAS > SPSS

Features

- 그래프
  - Stata > SAS >> SPSS
- 선형 혼합 모형(Linear Mixed Models)
  - SAS >> SPSS > Stata
- 로지스틱 회귀분석
  - Stata >> SAS = SPSS
    - **`logit, clogit, ologit, mlogit`** 명령어

예. 단순 로지스틱 회귀분석

- SAS
  - proc logistic data = "c:\mydata\hsb2" desc; model female = read / expb; run;
- Stata
  - logistic female read
  - logit female read
- SPSS
  - logistic regression female with read.

특징

- 결측 자료 처리: SAS = Stata >> SPSS
- 검정력 분석: SAS = SPSS > Stata
- 조사통계 자료 분석: Stata >> SAS = SPSS
- 생존 분석: Stata = SAS >> SPSS
- 가중치 분석: Stata > SAS >> SPSS

통계 패키지 조합을 고려한다면?

- Stata + SAS
  - 서로 단점을 훌륭하게 보완
    - 조사 통계 자료 분석: Stata > SAS
    - 방대한 규모 데이터셋 관리: SAS > Stata
    - 부트스트랩, 잭나이프, 몬테 카를로 기법: Stata > SAS
    - 복잡한 로데이터 불러오기: SAS > Stata
    - 가중치 처리: Stata > SAS
    - Stat/Transfer 이용 외부 파일 형식 불러오기
    - SPSS가 일부 종류의 분산분석에서 더 강력

마우스 지원

- 마우스 우선 사용자 환경은 결과 반복 및 재현 어려움.
  - SPSS > Stata >> SAS

명령어 구조 비교

- Stata > SAS >> SPSS
  - Stata, "적게 입력하고 적게 얻기"
  - SAS, "많이 입력하고 많이 얻기"
  - SPSS는 잘 정의된 명령어 구문 지원 부족
    - Stata

      ```
      mlogit y x1 x2 x3
      ```
    - SAS

      ```
      Proc logistic;
       model y=x1 x2 x3 / link=glogit;
      run;
      ```
    - SPSS

      ```
      NOMREG y WITH x1 x2 x3
       / PRINT = PARAMETER SUMMARY CPS MFI.
      ```

- 모형 추적 관찰
  - Stata > SAS >> SPSS
- 교육 용이성
  - Stata > SAS >> SPSS
- 표본수 산출
  - Stata=SAS >> SPSS
    . 기본 sampsi 및 생존분석용 stpower 명령어

결론

- 데이터 전처리에 집중하자.
  - SAS, Stata, R, …
- 범용 소프트웨어 하나를 익히자.
  - SAS, SPSS, Stata, R, …
- 연구재현성을 위해 명령문 방식을 익히자.
  - SAS, Stata, R, …

Kolabtree

Top 10 Statistical Tools Used in Medical Research

S/N	Product	Developer	Learning Curve	Cost (USD)	Open Source	Software license	Interface	Written in	Most Common Use Cases
1	STATA	StataCorp LLC	Steep	Academic starting at $595/ industry starting at $1,195	No	Proprietary	CLI/GUI	C	Clinical Data Analysis & Public Health
2	R	R Foundation	Steep	Free	Yes	GNU/PL	CLI/GUI	C with chunks in Fortran/C++	Meta-Analysis using special packages (Metafor & JASP)
3	GraphPad Prism	GraphPad Software, Inc.	Shallow	595	No	Proprietary	GUI	C/C++	Biological Labs, Research & Clinical Data Analysis
4	SAS	SAS Institute	Pretty steep	~$6000 per seat (PC version)/~$28K per processor (Windows server) first-year fees for BASE, STAT, GRAPH, and ACCESS modules. Modules are licensed individually. Subsequent year fees are roughly half.	No	Proprietary	CLI/GUI	C	Clinical Data Analysis, Health & Life Sciences
5	IBM SPSS	IBM	Shallow	$4,975	No	Proprietary	CLI/GUI	Java	Systematic Reviews, Surveys & Clinical Data Analysis
6	MATLAB	MathWorks	Pretty steep	$2150 (commercial), $99 (student), toolboxes additional	No	Proprietary	CLI	C++ & Java	Meta-Analysis & Clinical Data Analysis
7	JMP	SAS Institute	Shallow	$1995 (commercial) $29.95/$49.95 (student) $495 for H.S. site licence	No	Proprietary	CLI/GUI	C++	Clinical Data Visualisation and Analysis
8	Minitab	Minitab Inc.	Shallow	$895-$1395 perpetual, $542 or less concurrent annual, $29.99/$49.99/$99.99 academic	No	Proprietary	CLI/GUI	Fortran	Clinical Data Analysis & Healthcare Analytics
9	STATISTICA	StatSoft	Steep	>$695	No	Proprietary	GUI	C	Clinical Data Visualisation and Analysis
10	Excel	Microsoft Corporation	Shallow	$8.25 per month	Yes	Proprietary	GUI	C and C++ and C#	Clinical Data Analysis & Meta-Analysis (MetaXL add-in)

Note: ***Pretty Steep*** = Very difficult and gradual learning curve, ***Steep*** = Difficult and gradual learning curve, ***Shallow*** = Relatively easy and quick to learn

임상 연구자를 위한 Stata의 장점

1. '직관적인' 명령문 입력 방식으로 결과 재현 및 검증 가능
2. 표본수 산출 명령어 기본 내장
3. 역학통계 명령어 기본 내장(help epitab)
4. 회귀분석 모형 추적 관찰 및 가정 검증 편리
5. 범주형 자료분석(특히, 로지스틱 회귀분석) 명령어 강력
6. 진단검사의 타당도 및 신뢰도 평가 편리
7. 간단한 명령문으로 효과적인 프리젠테이션용 그래프 작성 가능

효과적인 통계 실무를 위한 10가지 규칙

PLoS Comput Biol 12(6): e1004961.

1. 통계 기법은 데이터가 과학적 질문에 답할 수 있게 해준다.
   - 마이크로어레이 데이터 분석 초보: "어떤 검정을 해야 하나요?", 고수: "어디에 분화된 유전자가 있나요?"
2. 신호는 항상 잡음을 동반한다.
   - 빅데이터 분석의 경우 바이어스 유발 잡음 걸러내는 일이 더욱 중요
3. 계획을 먼저, 진짜로 먼저 세워라.
4. 데이터 품질에 유의하라.
   - "쓰레기를 넣으면 쓰레기가 나온다."
5. 통계 분석은 연산 세트 이상이다.
   - 통계 소프트웨어는 분석 지원 도구이지, 정의 도구가 아님.

효과적인 통계 실무를 위한 10가지 규칙

PLoS Comput Biol 12(6): e1004961.

6. 단순하게 하라.
   - 무작위대조시험은 *t*-검정이나, χ^2-검정으로도 결론 가능.
7. 변이성 평가를 제시하라.
   - 표준오차나 신뢰구간을 반드시 제시.
8. 가정을 점검하라.
   - 통계 소프트웨어 제공하는 각종 모형 점검 시각화 도구 이용.
9. 가능하면, 반복해보라!
10. 자신의 분석을 재현가능하게 만들어라.
    - 데이터와 코드 공유