KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(온라인)

Introduction to Biostatistics

원성호 _ 서울대학교

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의개요

Introduction to Biostatistics

최근 생물정보 자료의 폭발적 증가와 함께 대규모 자료를 효율적으로 다루는 것은 생물정보 분야 연구자로서 반드시 지녀야 할 중요한 소양이 되었습니다. 데이터과학자가 생물정보 자료를 활용하여 의미있는 연구결과를 도출하기 위해서는 데이터의 편집 및 분석을 수행할 수 있는 능력이 필수적이며, 특히 통계학적 지식은 분석 결과를 해석하고 데이터를 이해하는데 필수입니다. 이번 BIML워크샵에서는 R 및 R 기반 엑셀 addin 프로그램, Rex, 를 활용하여 보다 쉽고 빨리 분석을 수행하고, 분석 결과를 해석하는데 필요한 기초적인 통계 지식을 강의할 계획입니다. 이 워크샵이 우리나라 생물정보 분야 연구자의 데이터 분석 수준을 높이는 계기가 될 수 있기를 희망하며 많은 분들의 관심과 참여를 부탁 드립니다.

강의는 다음의 내용을 포함한다:

- 가설검정
- 평균비교
- 회귀분석
- 군집분석

* 교육생준비물:

  노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

  R 프로그램 설치

  Rex 프로그램 무료버전 설치 (https://rexsoft.org/)

* 강의 난이도: 초급

* 강의: 원성호교수 (서울대학교 보건대학원)

Curriculum Vitae

Speaker Name: Sungho Won, Ph.D.

▶ **Personal Info**

Name	Sungho Won
Title	Professor
Affiliation	Seoul National University

▶ **Contact Information**

Address	1, Gwanak-Ro, Gawanak-Gu, Seoul, 151-742
Email	won1@snu.ac.kr
Phone Number	010-3442-1040

Research Interest

Biostatistics, Statistical Genetics, Genetic Epidemiology

Educational Experience

1999	B.S. in Biology Education, Seoul National University, Korea
2004	M.S. in Statistics, Seoul National University, Korea
2007	Ph.D. in Biostatistics and Epidemiology, Case Western Reserve University, USA

Professional Experience

2008-2009	Research Assistant, Dept. of Biostatistics, Harvard, USA
2009-2014	Assistant Professor, Dept. of Applied Statistics, Chung-Ang University, Korea
2014-2021	Associate Professor, Dept. of Public Health Sciences, Seoul National University, Korea
2021-	Professor, Dept. of Public Health Sciences, Seoul National University, Korea

Selected Publications (5 maximum)

1. Park J, Lutz SM, Choi S, Lee S, Park SC, Kim K, Choi H, Park H, Lee SY, Weiss ST, Hong SJ, Kim BS†, **Won S**†. Multi-omics analyses implicate EARS2 in the pathogenesis of atopic dermatitis. *Allergy* (SCI) 2021 Mar 31.
2. Cho J*, Park K*, Choi SM, Lee J, Lee CH, Lee JK, Heo EY, Kim DK, Lee YJ, Park JS, Cho YJ, Yoon HI, Lee JH, Lee CT, Kim N, Choi KY, Lee KH, Sung J, **Won S**†, Yim JJ†. Genome-wide association study of non-tuberculous mycobacterial pulmonary disease. *Thorax* (SCI) 2021 Feb;76(2):169-177. doi: 10.1136/thoraxjnl-2019-214430.
3. Kim KJ*, Park J, Park SC, **Won S**†. Phylogenetic tree-based microbiome association test. *Bioinformatics* (SCI). 2020 Feb 15;36(4):1000-1006. doi: 10.1093/bioinformatics/btz686.
4. Kim KW*, Park SC*, Cho HJ*, Jang H, Park J, Shim HS, Kim EG, Kim MN, Hong JY, Kim YH, Lee S, Weiss ST, Kim CH†, **Won S**†, Sohn MH†. Integrated genetic and epigenetic analyses uncover MSI2 association with allergic inflammation. *Journal of Allergy and Clinical Immunology.* 2020 Aug 11:S0091-6749(20)31106-4. doi: 10.1016/j.jaci.2020.06.040. Online ahead of print.
5. Song YE*, Lee S*, Park K, Elston RC, Yang HJ† and **Won S**†. ONETOOL for the analysis of family-based big data. *Bioinformatics* (SCI) 2018 Aug 15;34(16):2851-2853

생물통계학 기초

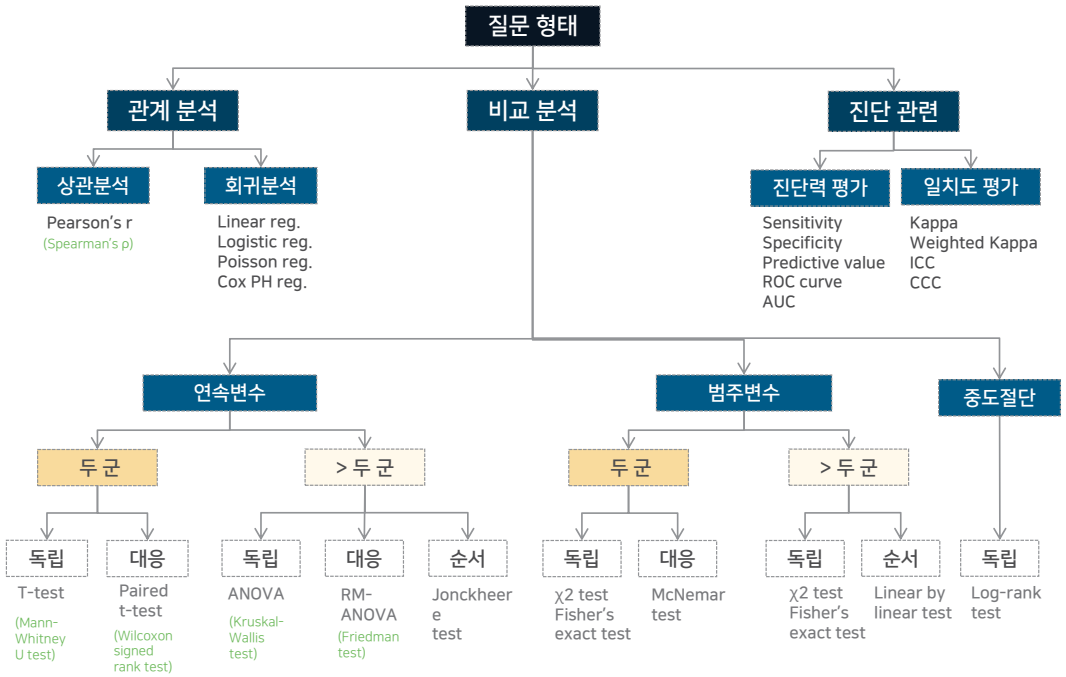
서울대학교 원성호

목차

통계분석 개요

통계분석 Flow chart

RexSoft /rexsoft.org



3

통계 분석 개요

평균 비교 개요

RexSoft /rexsoft.org

- **두 그룹에서의 평균 비교**
  - 독립적으로 추출된 경우: 독립표본 T검정 (모수), 순위합 검정 (비모수)
  - 짝을 지어 추출된 경우: 짝표본 T검정 (모수), 부호 검정 (비모수), 부호순위 검정 (비모수)

- **셋 이상 그룹에서의 평균 비교**

분산 분석 (ANOVA)

- 정규성 가정이 만족되는 경우
- 두 개 이상의 모집단을 비교하기 위한 방법으로 F-test를 이용한다.
- F-test 로 모집단간 평균에 차이가 있다는 결론을 내리면 추가적으로 어떤 모집단이 가장 큰 차이를 보이는가를 다중비교을 통하여 추가적으로 검정

K-독립표본 비모수검정 (크루스칼-왈리스 검정)

- 정규성 가정이 만족되지 않는 경우
- 관측값의 순위를 이용한 비모수적 방법

4

RexSoft
/rexsoft.org

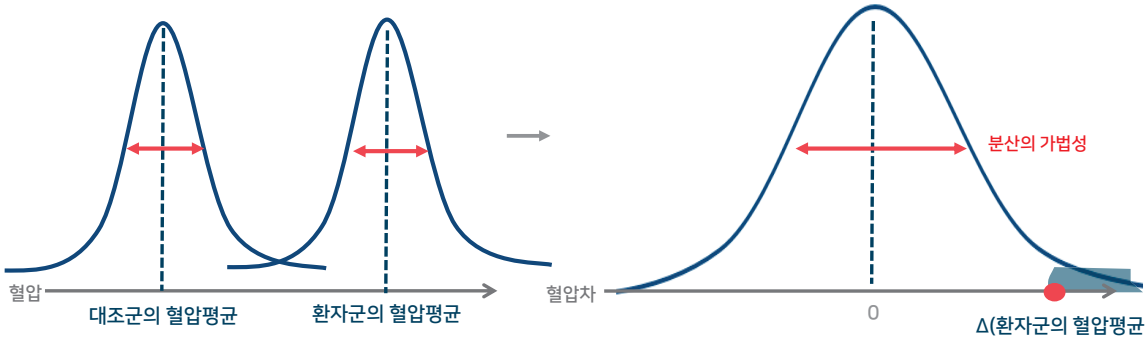
01 두 그룹에서의 평균 비교

5

독립표본 T검정

RexSoft
/rexsoft.org

- **독립표본 T검정 (등분산)**
  - H_0 : 환자군과 대조군의 혈압은 차이가 없다.
  - H_1 : 환자군과 대조군의 혈압은 차이가 있다.



$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

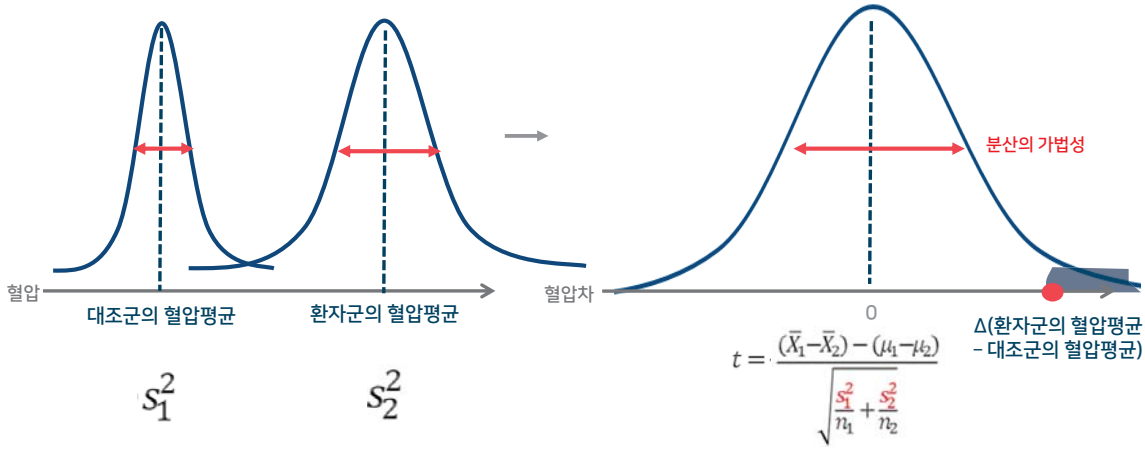
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

6

독립표본 T검정

▪ 독립표본 T검정 (이분산)

- H_0 : 환자군과 대조군의 혈압은 차이가 없다.
- H_1 : 환자군과 대조군의 혈압은 차이가 있다.

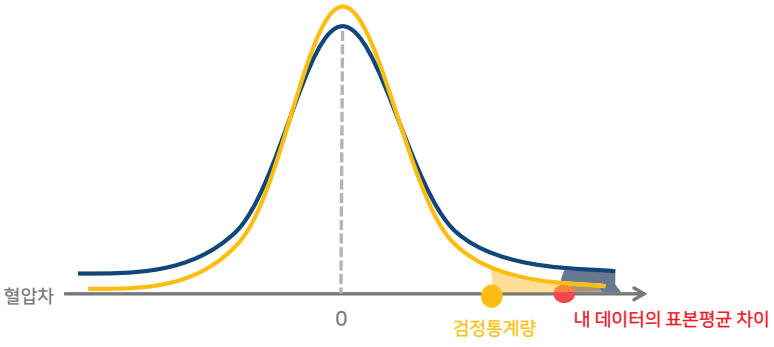


독립표본 T검정

▪ 독립표본 T검정의 예시

- H_0 : 환자군과 대조군의 혈압은 차이가 없다.
- H_1 : 환자군과 대조군의 혈압은 차이가 있다.

- 두 군의 평균 차이에 대한 분포로 변경하면, 일표본 T검정과 유사한 과정으로 검정

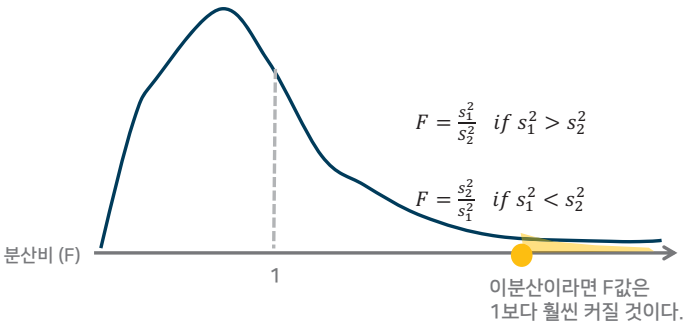


독립표본 T검정

▪ 등분산검정의 예시

- H_0 : 두 집단의 모분산에 차이가 없다. (등분산)
- H_1 : 두 집단의 모분산에 차이가 있다. (이분산)

- 두 군의 분산에 유의한 차이가 없을 경우, 등분산을 가정한 T통계량을 통해 가설 검정
- 두 군의 분산에 유의한 차이가 있을 경우, 이분산을 가정한 T통계량을 통해 가설 검정



2-독립표본 비모수검정

▪ 윌콕슨 순위합 검정의 예시

- 윌콕슨 순위합 검정 : 순위데이터로 변환하여 분포가 겹치는 정도를 나타내는 통계량을 계산
- H_0 : 환자군과 대조군의 혈압은 차이가 없다.
- H_1 : 환자군과 대조군의 혈압은 차이가 있다.

환자군	순위	대조군	순위
82		80	
100		87	
150		99	

2-독립표본 비모수검정

RexSoft /rexsoft.org

▪ 윌콕슨 순위합 검정의 예시

- 윌콕슨 순위합 검정 : 순위데이터로 변환하여 분포가 겹치는 정도를 나타내는 통계량을 계산
- H_0 : 환자군과 대조군의 혈압은 차이가 없다.
- H_1 : 환자군과 대조군의 혈압은 차이가 있다.

환자군	순위	대조군	순위
87	2.5	80	1
100	5	87	2.5
150	6	99	4
순위합 평균	13.5/3 = 4.5	순위합 평균	7.5/3 = 2.5

순위합 평균 비교

2-독립표본 비모수검정

실습

RexSoft /rexsoft.org

▪ 환자 12명과 정상인 11명이 있을 때 분변마이크로비옴 자료를 측정한 결과는 다음과 같다고 한다. 이때 alpha-diversity가 환자군이 대조군에 비하여 낮다고 할 수 있는지 유의수준 0.05에서 비모수 검정하라.

- 데이터: stat_example.xls
- 그룹 : 2=환자, 1=정상인

대조군	4.62	3.71	3.76	2.84	4.93	4.65	4.69	3.81	4.58	3.27	3.79	
환자군	4.88	3.87	4.50	4.05	5.09	3.94	4.62	4.10	4.40	4.98	5.01	5.12

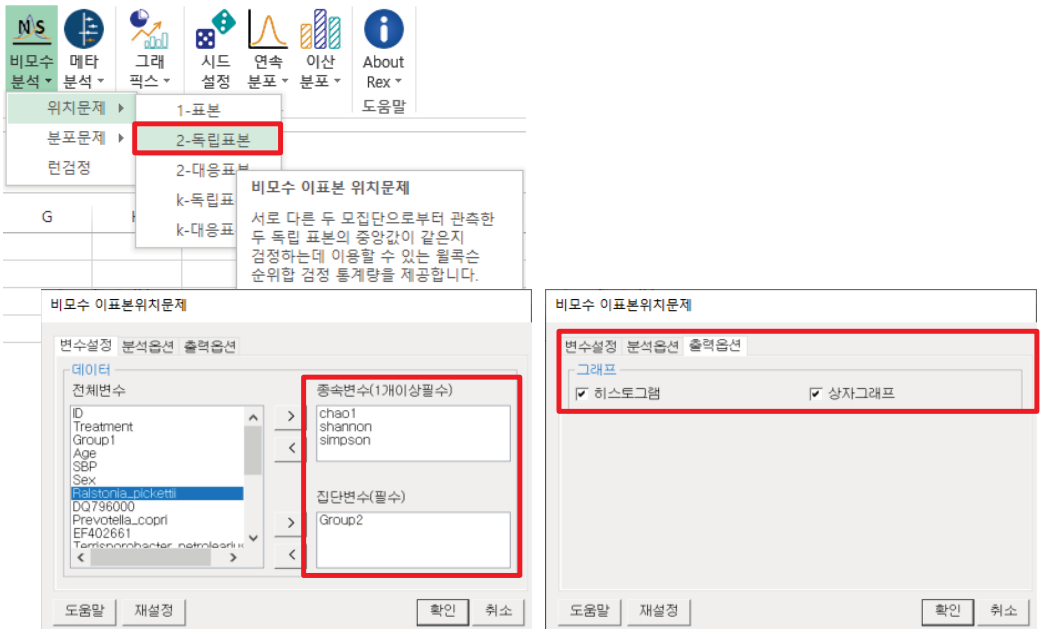
01 두 그룹에서의 평균 비교

2-독립표본 비모수검정

실습

RexSoft /rexsoft.org

입 력 Rex ► 비모수분석 ► 위치문제 ► 2-독립표본



01 두 그룹에서의 평균 비교

2-독립표본 비모수검정

실습

RexSoft /rexsoft.org

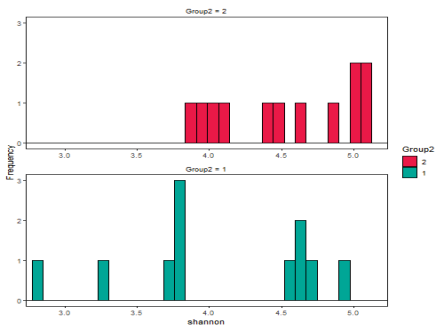
출 력

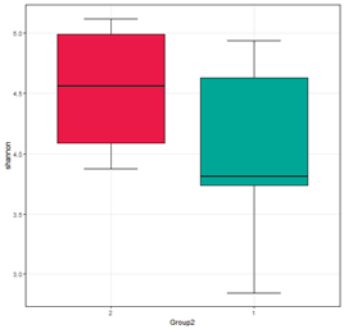
데이터 요약

	# of obs.	Mean	SD	Median	Q1	Q3
Group 2	12	4.54875	0.471234	4.55924	4.08765	4.98825
Group 1	11	4.05905	0.674026	3.80985	3.73707	4.63157

윌콕슨 순위합 검정 (Wilcoxon rank-sum test)

	WS	U	E(WS)	Var(WS)	Z(WS)	P-value
shannon	176	98	144	264	1.9695	0.0489





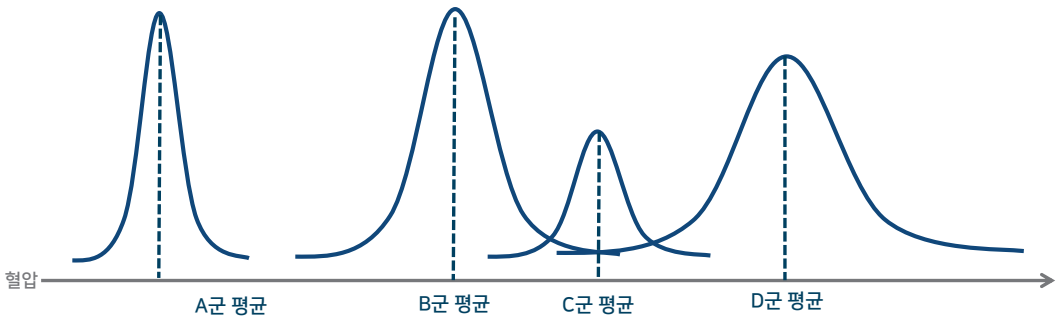
P<0.05, 두 군에서
유의한 차이가 난다고 할 수 있다

02 셋 이상 그룹에서의 평균 비교

분산분석 (ANOVA)

RexSoft
/rexsoft.org

- **분산분석 개요**



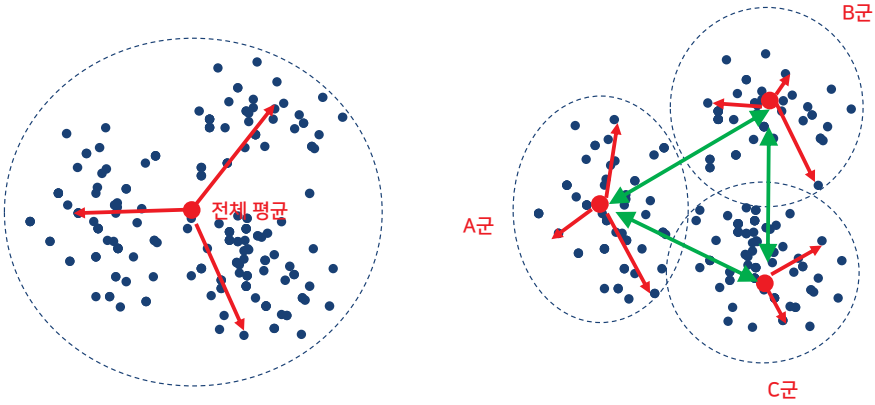
검정통계량 구성이 어려워지고,
두 쌍씩 짝지어 평균 비교를 할 경우 다중비교로 인해 제1종오류가 늘어난다

평균 말고 다른 걸로 비교할 순 없을까?

→ 분산 이용!!! (ANalysis Of VAriance)

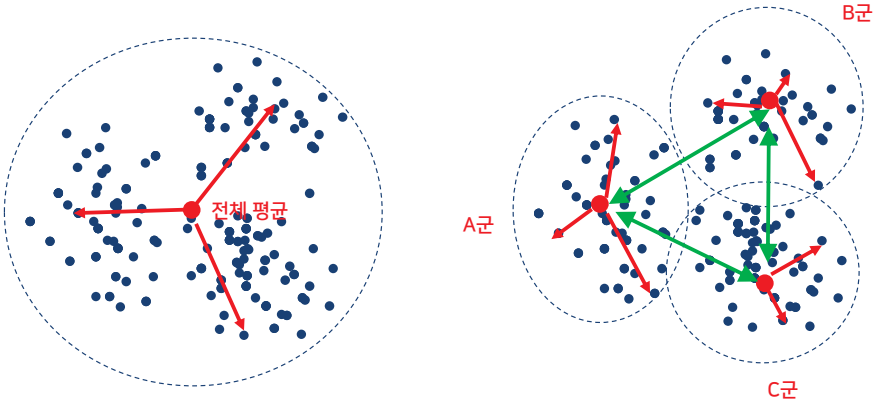
분산분석 (ANOVA)

- 분산분석 개요



분산분석 (ANOVA)

- 분산분석 개요

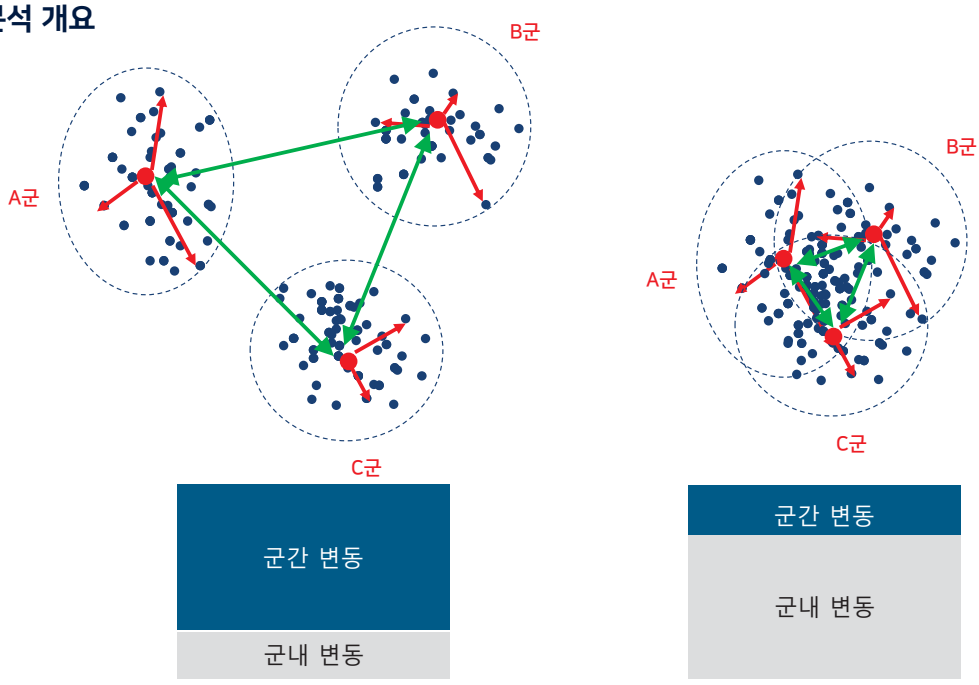


전체 분산 = 군내 변동 + 군간 변동

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{..}\right)^2 \qquad SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{.j}\right)^2 \qquad SSA = \sum_{j=1}^{k} n_j\left(\bar{x}_{.j} - \bar{x}_{..}\right)^2$$

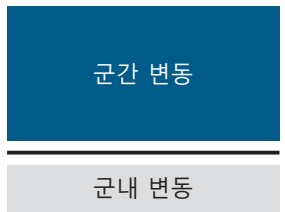
분산분석 (ANOVA)

- **분산분석 개요**



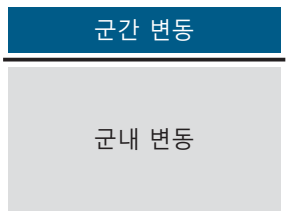
분산분석 (ANOVA)

- **분산분석 개요**



분산비로 검정통계량 F 를 만들자!

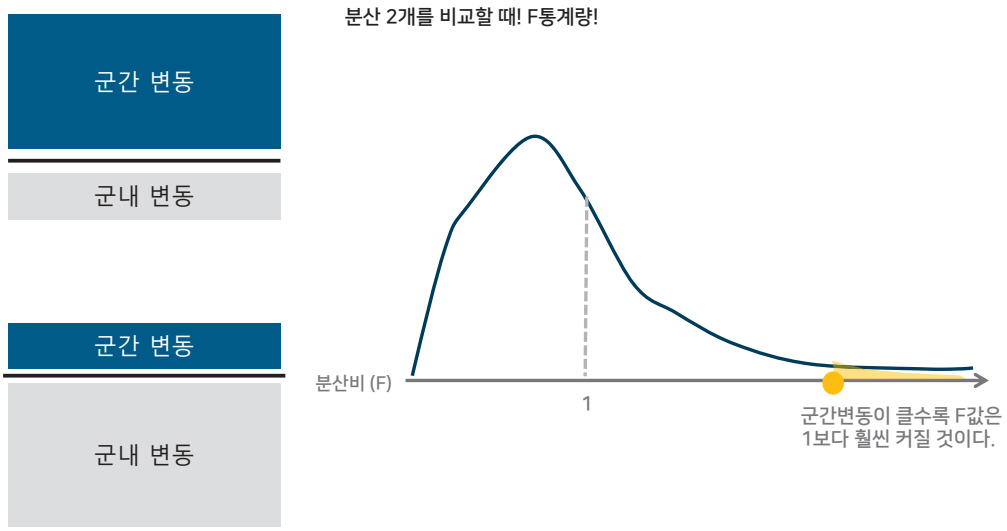
F가 1보다 크면 군간변동 > 군내변동
군에 의한 차이가 많으니, 이 종속변수는 군별로 차이가 나겠구나!



F가 1보다 작으면 군간변동 < 군내변동
군간변동이 우연에 의한 변동보다 작으니, 이 종속변수는 군별로 차이가 난다고 하기 어렵겠네!

분산분석 (ANOVA)

▪ 분산분석 개요



분산분석 (ANOVA)

▪ 분산분석표와 검정통계량

요인	제곱합	자유도	평균제곱합	F
집단 간 제곱합	$SSA = \sum_{j=1}^{k} n_j (\bar{x}_{.j} - \bar{x}_{..})^2$	$k-1$	$MSA = SSA/(k-1)$	$\frac{MSA}{MSW}$
집단 내 제곱합	$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})^2$	$N-k$	$MSW = SSW/(N-k)$	
총 제곱합	$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{..})^2$	$N-1$		

- 검정통계량 F가 자유도(k-1, N-k)인 F분포를 따르기 때문에 F검정이라고 부르기도 함
- 검정방법: F-검정통계량의 p-값이 유의수준 0.05 보다 작으면 H0을 기각

  ▶ '각 처리 그룹간에 차이가 있다'라고 결론

분산분석 (ANOVA)

▪ 분산분석표와 검정통계량

요인	제곱합	자유도	평균제곱합	F
집단 간 제곱합	$SSA = \sum_{j=1}^{k} n_j(\bar{x}_{.j} - \bar{x}_{..})^2$	$k-1$	$MSA = SSA/(k-1)$	$\frac{MSA}{MSW}$
집단 내 제곱합	$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{.j})^2$	$N-k$	$MSW = SSW/(N-k)$	
총 제곱합	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{..})^2$	$N-1$		

- 검정통계량 F가 자유도(k-1, N-k)인 F분포를 따르기 때문에 F검정이라고 부르기도 함
- 검정방법: F-검정통계량의 p-값이 유의수준 0.05 보다 작으면 H0을 기각

  ▶ '각 처리 그룹간에 차이가 있다'라고 결론

분산분석 (ANOVA)

▪ 예시 – Leukemia 연구

요인	제곱합	자유도	평균제곱합	F
집단 간 제곱합	55.684	2	27.842	125.43
집단 내 제곱합	15.316	69	0.222	
총 제곱합	71.000	71		

- 검정방법: F-검정 통계량에 대응하는 p-value = 1.04514E-23 << 0.05

▶ 'GeneBank accession number X03934 (T-cell antigen receptor gene T3-delta)의 발현 정도가 Leukemia의 종류에 따라 차이가 있다'고 결론

분산분석 (ANOVA)

▪ 사후분석 (Post-hoc Analysis)

- 일단 '그룹 간의 차이'가 있다는 결론을 내리게 되면 추가적으로 `어떤 처리가 가장 효과가 있는가?'등과 같은 추가적인 의문 발생

- 예를 들면 3가지 종류의 Leukemia 간에 차이가 있다는 결론을 내렸을 때
  - ✓ 세 종류의 Leukemia 모두가 아주 다른지
  - ✓ 또는 AML과 T- cell ALL이,
  - ✓ 또는 AML과 B-cell ALL이,
  - ✓ 또는 T-cell ALL과 B-cell ALL이 서로 다른가?

분산분석 (ANOVA)

▪ 다중비교 (Multiple Comparison) 보정방법

- 검정을 반복해도 검정통계량이 기각역에 쉽게 들어가지 않도록 엄격하게 패널티를 주는 방법
- 검정 다중성의 보정 대상에 따라 20 종류 이상의 다양한 방법이 있음
- 크게 3가지 종류로 나눌 수 있음

보정대상	내용	대표적 방법
유의수준	반복하는 횟수에 따라 유의수준을 작게 함으로써 기각역이 넓어지지 않게 하는 법	Bonferroni
검정통계량	집단 수에 따라 검정통계량을 작게 해서 기각역에 들어가기 어렵게 만드는 법	Scheffe Fisher's LSD
분포	반복하는 횟수가 늘어도 유의수준이 커지지 않는 독자적인 분포를 만들어 한계값을 통해 판정	Tukey, Dunnet

실습

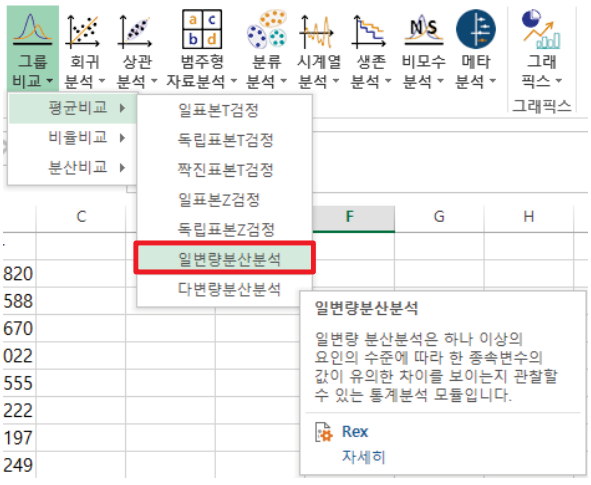
RexSoft /rexsoft.org

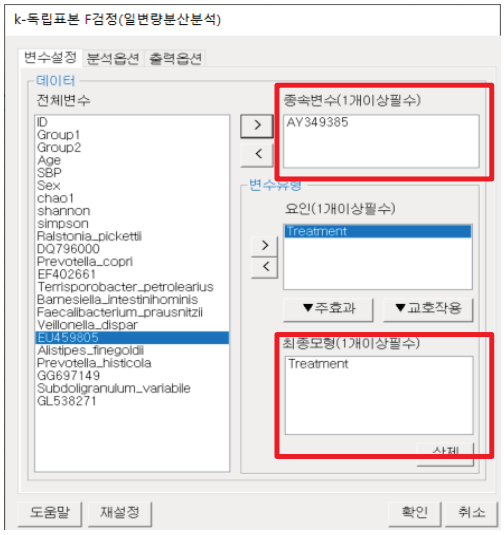
- **AY349385의 relative abundance를 비교하고자 한다. 3가지 그룹별로 abundance농도에 차이가 있는지 가설검정을 수행하시오.**
  - Treatment : 1, 2, 3
  - AY349385 : relative abundance

실습

RexSoft /rexsoft.org

입 력 Rex ► 그룹비교 ► 평균비교 ► 일변량분산분석



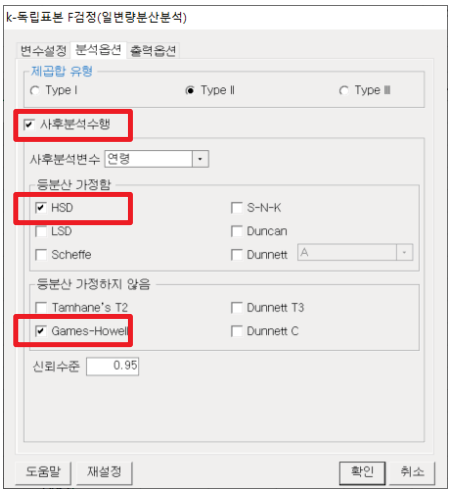


실습

RexSoft /rexsoft.org

입 력

Rex ► 그룹비교 ► 평균비교 ► 일변량분산분석



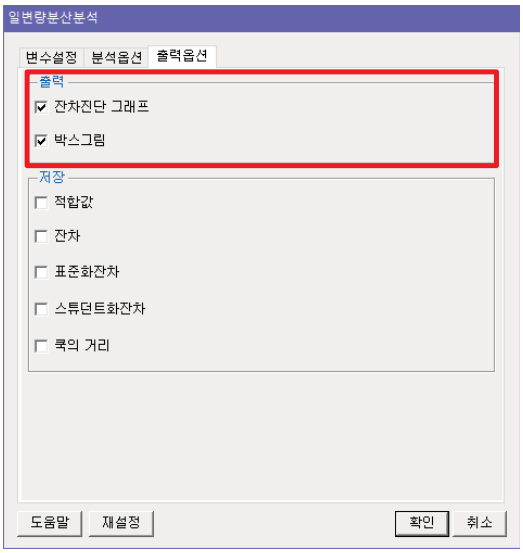
제곱합 유형	Type I	모형에 나열된 요인 순서대로 검정됨
	Type II	Type III와 유사하나, 주효과들이 교호작용은 고려하지 않고 서로 다른 주효과만 고려한 상태에서 검정됨
	Type III	모형의 각 요인들이 모형 내 다른 요인들을 고려한 상태에서 검정됨 주효과들은 교호작용들과 다른 주효과를 고려한 상태에서 검정됨
사후분석 방법	Tukey's HSD	등분산 가정 필요 / 군별 표본수 무관 / 검정력 중간
	Fisher's LSD	등분산 가정 필요 / 군별 표본수 동일 / 검정력 최대
	Scheffe	등분산 가정 필요 / 군별 표본수 무관 / 검정력 최소

실습

RexSoft /rexsoft.org

입 력

Rex ► 그룹비교 ► 평균비교 ► 일변량분산분석



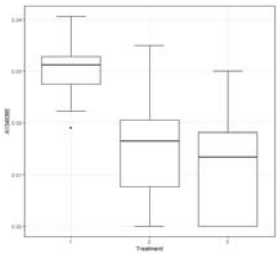
출력	잔차진단그래프	잔차, 쿡의 거리, leverage value에 대한 진단 그래프
	박스그림	요인 수준별 종속변수에 대한 박스그림
저장	적합값	모형에 의해 해당 요인에 대해 예측된 종속변수값
	잔차	적합값과 실제값의 차이
	표준화잔차	잔차를 표준화한 값/ 절대값이 2보다 큰 경우 이상값 의심
	스튜던트화잔차	모형에서 해당 값을 제외하고 계산한 표준화잔차 / 절대값이 2보다 큰 경우 이상값 의심
	쿡의 거리	모형에 의한 각 관측치의 전반적인 영향력을 측정하기 위해 잔차와 leverage value를 동시에 고려한 척도 / 1보다 크면 영향값 의심

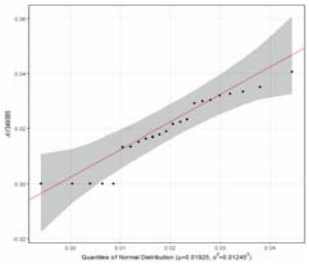
분산분석

실습

RexSoft /rexsoft.org

출 력





Leven's 검정

	F value	D.F1	D.F2	P-value
Treatment	0.9171	2	20	0.4158

P>0.05, 세 군에서 등분산 가정을 만족한다고 판단. 만약 만족하지 않는 Welch's ANOVA를 활용해야함.

ANOVA

	DF	Sum of squares	Mean squares	F	P-value
Treatment	2	0.0014	0.0007	6.7923	0.0056
Residuals	20	0.0021	0.0001		

P<0.05, 세 군 중에 적어도 한 그룹은 relative abundance에 차이가 있다고 판단

Welch's ANOVA

F	D.F1	D.F2	P-value
8.881	2.000	11.852	0.0044

P>0.05, 세 군에 relative abundance 에 차이가 없다고 판단

분산분석

실습

RexSoft /rexsoft.org

출 력

사후검정 (Post-hoc)

Method	Comparison	Difference	95% LCI	95% UCI	P-value
Tukey	1-2	0.0145	0.0010	0.0280	0.0338
	1-3	0.0180	0.0050	0.0310	0.0062
	2-1	-0.0145	-0.0280	-0.0010	0.0338
	2-3	0.0035	-0.0100	0.0170	0.7926
	3-1	-0.0180	-0.0310	-0.0050	0.0062
	3-2	-0.0035	-0.0170	0.0100	0.7926
GH	1-2	0.0145	-0.0002	0.0292	0.0533
	1-3	0.0180	0.0056	0.0304	0.0060
	2-1	-0.0145	-0.0292	0.0002	0.0533
	2-3	0.0035	-0.0129	0.0198	0.8398
	3-1	-0.0180	-0.0304	-0.0056	0.0060
	3-2	-0.0035	-0.0198	0.0129	0.8398

1-2 : p<0.05, 1군과 2군의 농도에서 유의한 차이가 난다고 판단
1-3 : p<0.05, 1군과 3군의 농도에서 유의한 차이가 난다고 판단

K-독립표본 비모수검정

- **크루스칼 왈리스 검정 (Kruskal-Wallis test)**
  - 자료 : 정규성 가정이 만족되지 않는 경우
  - 방법 : yij대신에 전체 관측값들 중에서 계산한 yij 의 순위(Rij) 를 사용하여 검정

AML	순위	B	순위	T	순위
2.57	22	2.01	6	2.01	6
2.05	10	2.37	20	2.37	20
2.06	11	1.97	5	1.97	5
1.64	1	1.85	2	1.85	2
1.90	4			2.01	6
2.14	14			2.37	20
2.28	18			1.97	5
2.03	8				

- **가정**
  - 영가설 (예: 세 그룹의 평균이 같다)
  - 검정 방법 :
    p-값이 유의수준 0.05보다 작으면 H0 을 기각
    ▶ '각 처리 그룹간에 차이가 있다'고 결론

K-독립표본 비모수검정

실습

- **약물처방에 따른 AY349385의 relative abundance를 비교하고자 한다. 3가지 그룹별로 abundance 농도에 차이가 있는지 가설검정을 수행하시오.**
  - Treatment : 1, 2, 3
  - AY349385 : relative abundance

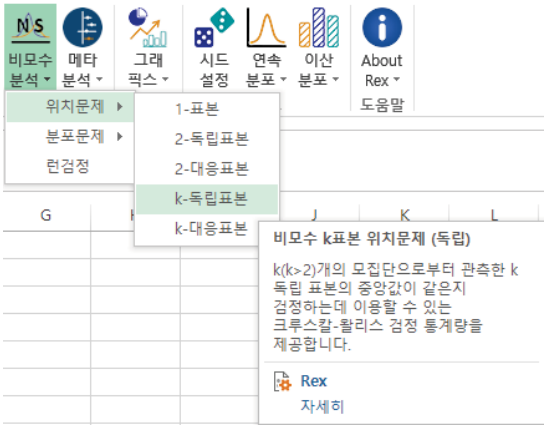
02 셋 이상 그룹에서의 평균 비교

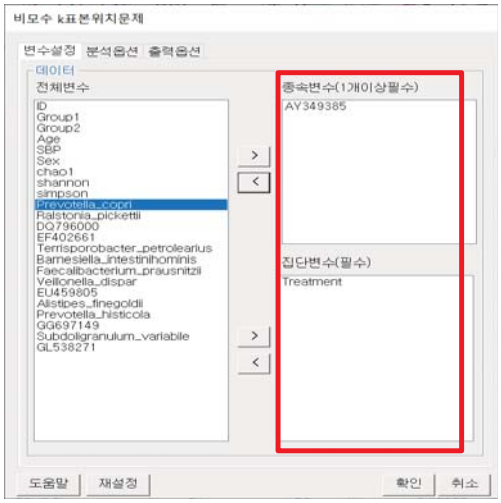
K-독립표본 비모수검정

실습

RexSoft /rexsoft.org

입 력 Rex ► 비모수분석 ► 위치문제 ► K-독립표본





35

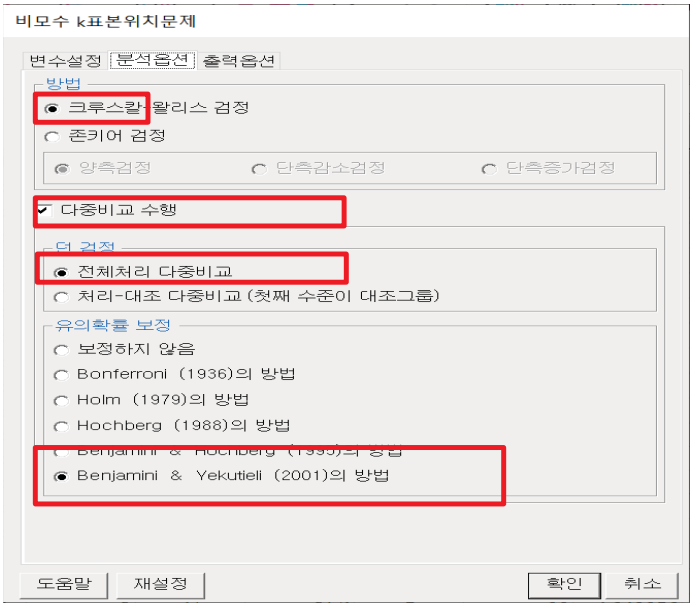
02 셋 이상 그룹에서의 평균 비교

K-독립표본 비모수검정

실습

RexSoft /rexsoft.org

입 력 Rex ► 비모수분석 ► 위치문제 ► K-독립표본



36

K-독립표본 비모수검정

실습

RexSoft /rexsoft.org

출 력

크루스칼-왈리스 검정 결과

	H	df	P-value
AY349385 by Treatment	9.7654	2	0.0076

p=0.0076, 세 군 중 적어도 한 쌍의 집단에서 유의한 차이가 난다고 판단

사후검정(Post-hoc)

	|Diff.Rank|	Z	P-value (adjusted)
1 - 2	7.7321	2.2137	0.0738
1 - 3	10.1250	3.0006	0.0148
2 - 3	2.3929	0.6851	0.9044

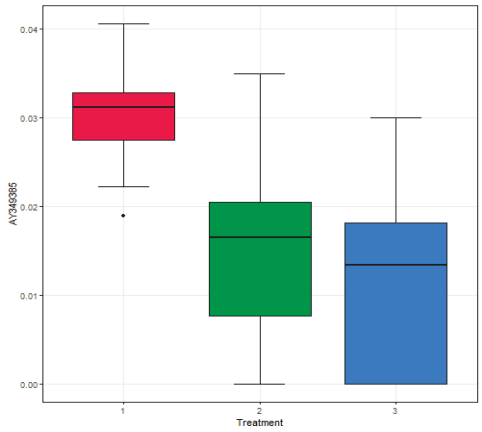
1-2 : p=0.0738, 1군과 2군의 균주 농도에서 유의한 차이가 없다고 판단
1-3 : p=0.0148, 1군과 3군의 균주 농도에서 유의한 차이가 있다고 판단
2-3 : p=0.9044, 2군과 3군의 균주 농도에서 유의한 차이가 없다고 판단

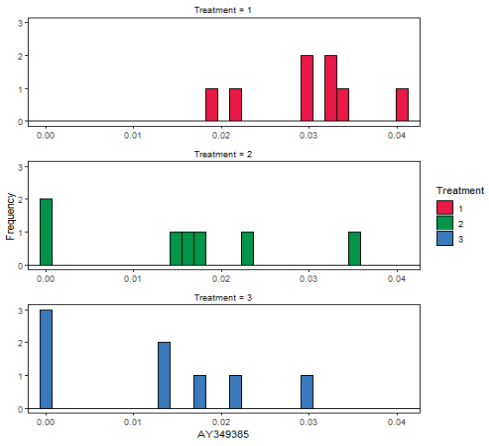
K-독립표본 비모수검정

실습

RexSoft /rexsoft.org

출 력





Var	Treatment 1 (N=8)	Treatment 2 (N=7)	Treatment 3 (N=8)	P-value*	Post-hoc†		
					1 vs 2	1 vs 3	2 vs 3
AY349385	0.03 (0.02,0.04)	0.02 (0.00,0.04)	0.01 (0.00,0.03)	0.0076	0.0738	0.0148	0.9044

Data was reported as median and interquartile range(IQR)
*P-value was calculated by Kruskal-Wallis test
†P-values were calculated Dunnet's method using Benjamini & Yekutieli's method

평균 비교 요약

- **두 그룹에서의 평균 비교**
  - 독립적으로 추출된 경우: 독립표본 T검정 (모수), 순위합 검정 (비모수)
  - 짝을 지어 추출된 경우: 짝표본 T검정 (모수), 부호 검정 (비모수), 부호순위 검정 (비모수)

- **셋 이상 그룹에서의 평균 비교**

분산 분석 (ANOVA)

- 정규성 가정이 만족되는 경우
- 두 개 이상의 모집단을 비교하기 위한 방법으로 F-test를 이용한다.
- F-test 로 모집단간 평균에 차이가 있다는 결론을 내리면 추가적으로 어떤 모집단이 가장 큰 차이를 보이는가를 다중비교을 통하여 추가적으로 검정

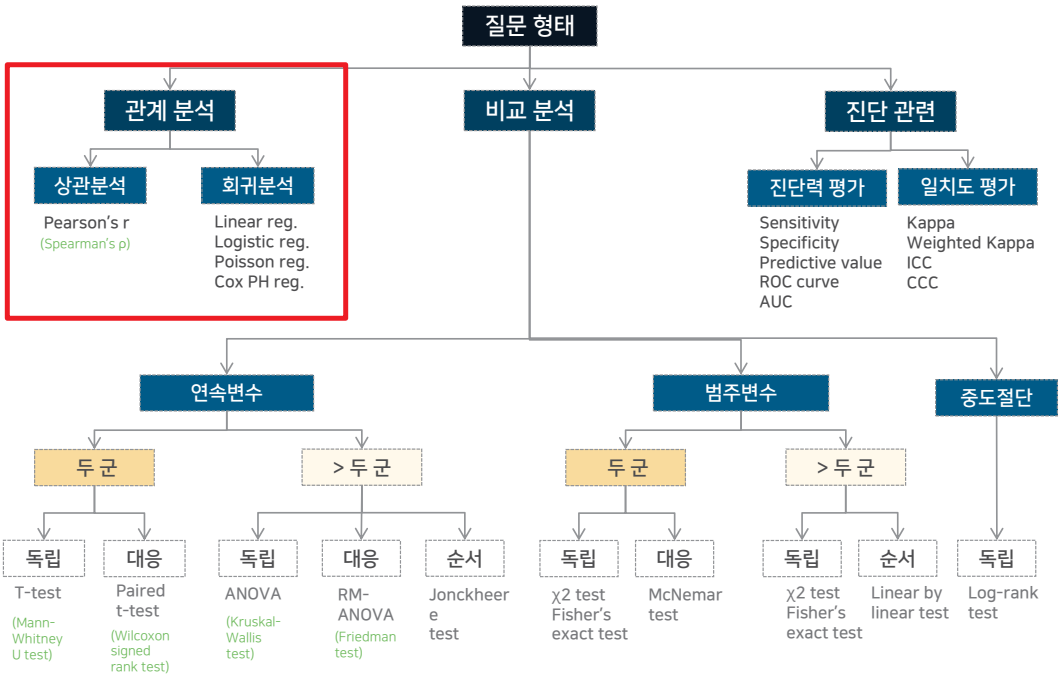
K-독립표본 비모수검정 (크루스칼-왈리스 검정)

- 정규성 가정이 만족되지 않는 경우
- 관측값의 순위를 이용한 비모수적 방법

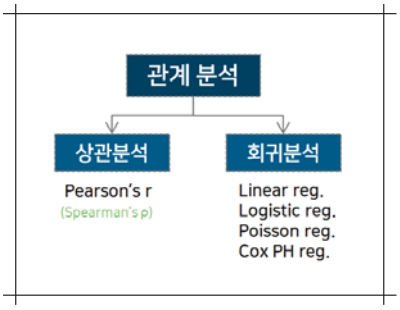
RexSoft
/rexsoft.org

03 상관분석 (Correlation Analysis)

통계분석 Flow chart



통계분석 Flow chart



관계 분석 = 연관성 분석

- 연령과 혈압은 어떤 관계가 있을까?
- 혈압에 영향을 미치는 인자는 무엇일까?
- 고혈압 유무에 영향을 미치는 인자는 무엇일까?
- 고혈압 발생률에 영향을 미치는 인자는 무엇일까?
- 고혈압 환자의 생존율에 영향을 미치는 인자는 무엇일까?

상관분석과 회귀분석

- (Non-directional association) 상관분석 : 두 연속형 변수의 선형적 상관관계의 정도를 나타내는 것
- (Directional association) 회귀분석 : 결과변수 (outcome, Y)가 위험인자 (risk factor) 나 예측변수 (predictor, Xs)들에 의해 어떻게 설명되거나 예측되는지 확인

1) Linear regression : 연속형 (continuous) 결과변수
2) Logistic regression : 이분형 (binary) 결과변수
3) Poisson regression : 단위당 계수형 (count per unit) 결과변수
4) Cox PH regression : 생존형 (survival or time-to-event) 결과변수

상관분석의 개념

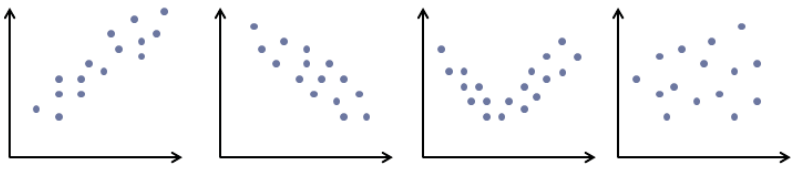
- **상관성 (correlation)**
  - 두 변수 사이의 선형적 연관성의 크기 또는 직선 관계의 강도(강/약/무) 및 패턴(정/부)를 의미

- **상관계수 (correlation coefficient)**
  - 상관성의 지표
  - 모상관계수 (ρ): 모집단에서의 상관성 지표
  - (표본)상관계수 (r): 자료로부터 추정한 상관성의 지표

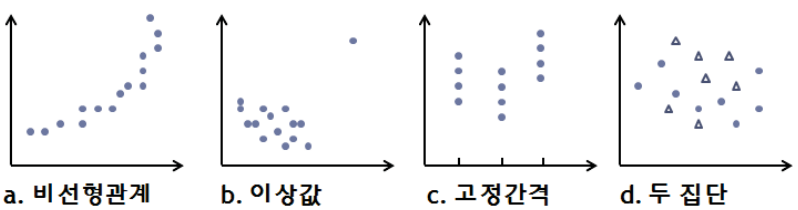
- **상관분석**
  - 자료로부터 상관계수 (r)를 추정하여 두 변수의 상관성에 대해 알아보는 과정
  - 정규성 가정을 만족할 때
    → Pearson's correlation coefficient
  - 정규성 가정을 만족하지 않을 때
    → Spearman's correlation coefficient

상관분석의 개념

- **상관관계의 종류**

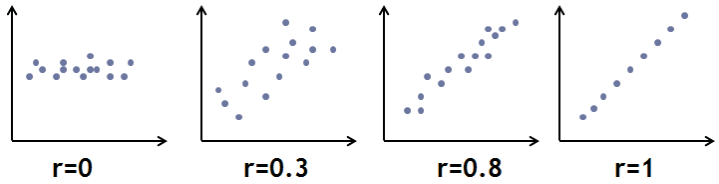


- **상관분석이 부적합한 경우**



상관분석의 개념

- **상관계수**



• Criteria

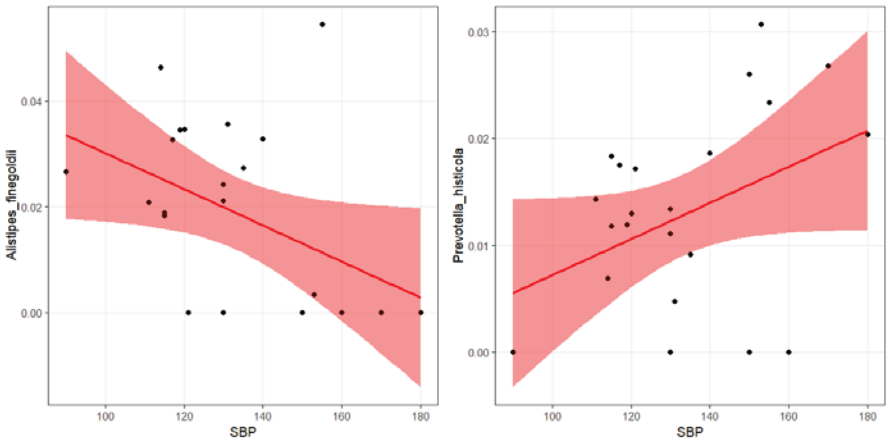
1) |r| ≤ 0.2 : 매우 약함 (very week)
   0.2 < |r| ≤ 0.4 : 약함 (week)
   0.4 < |r| ≤ 0.7 : 중간 (moderate)
   0.7 < |r| ≤ 0.9 : 강함 (strong)
   0.9 < |r| : 매우 강함 (very strong)

2) r > 0 : 양의 상관관계 (positive correlation)
   r < 0 : 음의 상관관계 (negative correlation)
   r = 0 : 상관관계 없음 (no correlation)

상관분석

실습

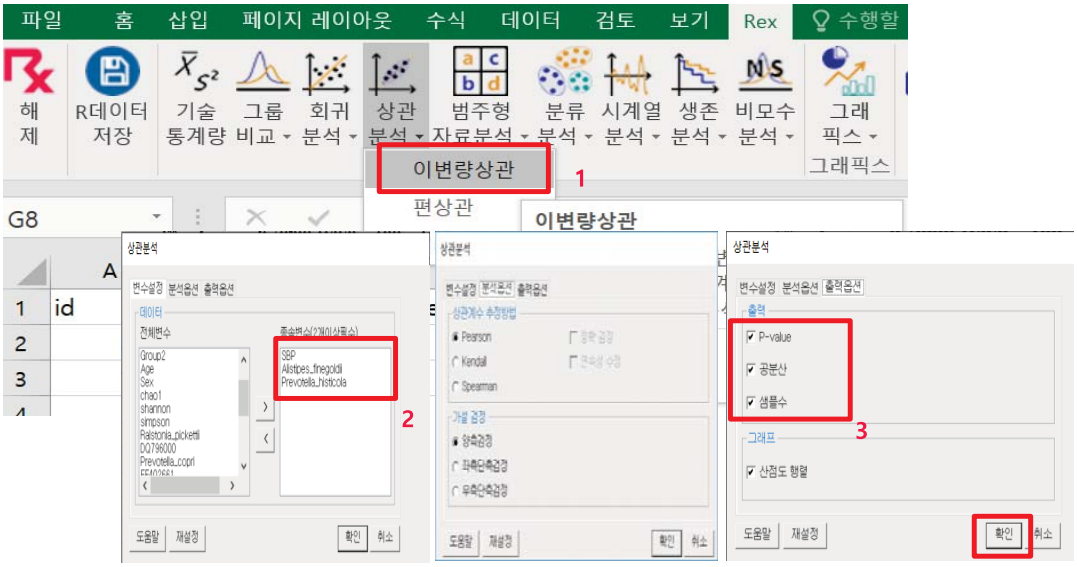
- **SBP와 Alistipes_finegoldii, Prevotella_histicola 에 대하여 상관분석을 수행하시오.**



실습

RexSoft /rexsoft.org

입 력 Rex ► 상관분석 ► 이변량상관



실습

RexSoft /rexsoft.org

출 력

데이터구조

Category	Variable	N	N.valid	(%.valid)	N.miss	(%.miss)
Dependent variable	SBP	23	23	(100.00%)	0	(0.00%)
Dependent variable	Alistipes_finegoldii	23	23	(100.00%)	0	(0.00%)
Dependent variable	Prevotella_histicola	23	23	(100.00%)	0	(0.00%)

상관분석결과

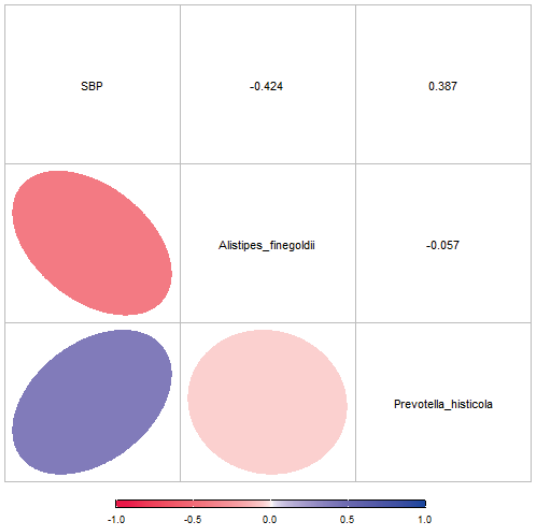
	SBP	Alistipes_finegoldii	Prevotella_histicola
SBP		-0.4242 0.0436 -0.1547 23	0.3867 0.0683 0.0765 23
Alistipes_finegoldii	-0.4242 0.0436 -0.1547 23		-0.0571 0.796 0 23
Prevotella_histicola	0.3867 0.0683 0.0765 23	-0.0571 0.796 0 23	

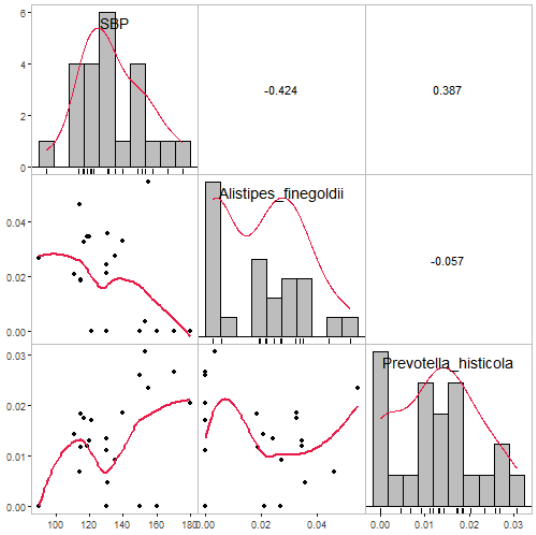
Cross Correlation Table
- Correlation Coefficients
- p-value
- Covariance
- No. of Observations

실습

RexSoft
/rexsoft.org

출력





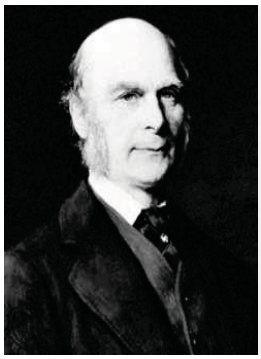
RexSoft
/rexsoft.org

04 선형회귀분석
(Linear Regression Analysis)

회귀분석 개요

▪ 회귀분석의 어원

- 유전학자 F. Galton이 아버지와 아들의 신장 (height) 간의 관계 연구
- 직선형 관계 (linear relationship) 가정
- 아들의 신장이 평균값으로 회귀(regress)되는 현상을 발견한 것에서부터 유래



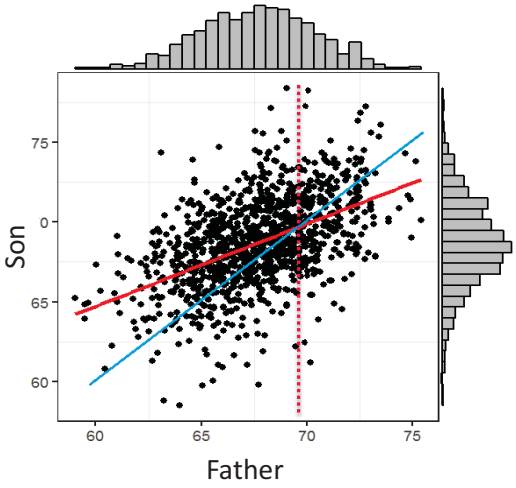
Francis Galton

BORN
February 16, 1822, England

DIED
January 17, 1911 (aged 88)

SUBJECTS OF STUDY
human intelligence

부모의 신장에 비해 2세의 신장이 일반 평균치에 복귀 (regress toward population mean)하는 특성을 발견



회귀분석 개요

▪ 회귀분석의 목표

- 결과변수 (Y)와 설명변수 (Xs) 간의 인과관계 $Y = f(X)$ 를 모형화하여 추론
- 결과변수를 설명변수의 선형조합 (linear combination) 즉 가중합 (weighted sum)의 함수로 모형화
- 예. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

▪ 세부목적

- 모형에 대한 추론 (모형의 추정 → 유의성 확인 → 설명력 평가)
- 결과변수의 설명/예측에 대한 각 설명변수의 상대적 기여도 추정
- 설명변수의 특정한 값 x_0에 대한 결과변수 값 예측

회귀분석 개요: Y = f(X) 에서 역할에 따른 변수의 종류

RexSoft /rexsoft.org

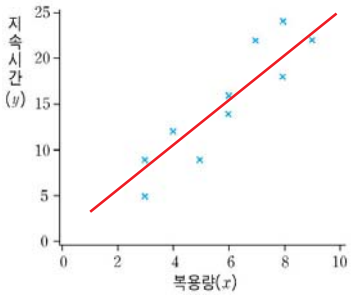
- **Y: 종속변수 (dependent variable)**
  - 관심의 대상이 되는 특정 현상을 나타내는 변수
  - 일반적으로 기호 Y 로 표시
  - 출력변수 (output variable),
    반응변수 (response variable),
    결과변수 (outcome variable, endpoint)
    라고도 함

- **X: 독립변수 (independent variable)**
  - 종속변수에 영향을 미칠 수 있거나 종속변수를 설명할 수 있는 변수
  - 일반적으로 기호 X로 표시
  - k 개의 독립변수는 $X_1, \ldots, X_k$로 표시
  - 입력변수 (input variable),
    설명변수 (explanatory variable),
    위험인자 (risk factor),
    예측변수 (predictor) 라고도 함

회귀분석 개요: X 와 Y 의 관계

RexSoft /rexsoft.org

- **결정적인 관계 (Deterministic relationship)**
  - $Y = f(X)$
  - Y 의 값이 X 의 값에 따라 유일하게 정의되는 관계
  - (선형) 결정 모형
  - $Y = \beta_0 + \beta_1 X$



- **통계적인 관계 (Statistical relationship)**
  - $Y = f(X) + \varepsilon$ (ε: 오차항)
  - Y 의 값이 X 의 값뿐 아니라 오차를 포함하여 확률적으로 예측되는 관계
  - (선형) 확률 모형
  - $Y = \beta_0 + \beta_1 X + \varepsilon$
  - Probabilistic model =
    Deterministic model + random error

회귀분석의 종류 – 결과변수의 유형별 (1)

RexSoft /rexsoft.org

- **선형모형 (linear model)**
  - 연속형 결과변수
  - 가정: 결과변수가 예측변수의 가중합과 직접 관련됨
  - 예. $Y = \beta_0 + \beta_1 X + \varepsilon$
  - (선형)회귀분석 (linear regression)

회귀분석의 종류 – 결과변수의 유형별 (2)

RexSoft /rexsoft.org

- **일반화 선형모형 (generalized linear model)**
  - 범주형 및 이산형 결과변수
  - 가정: 결과변수가 예측변수의 가중합과 연결함수 (link function)을 통해 관련됨
  - 예. $g(Y) = \beta_0 + \beta_1 X + \varepsilon$
  - 로지스틱회귀분석 (logistic regression)
    - ✓ 이진형 결과변수 (binary outcome)
    - ✓ 예. 당뇨병 유병 여부
    - ✓ $g(Y)$ = logit link function
  - 포아송회귀분석 (Poisson regression)
    - ✓ 이산형 결과변수 (count outcome)
    - ✓ 예. 강원지역의 10년간 폐쇄성무호흡증 발생건수
    - ✓ $g(Y)$ = log link function

회귀분석의 종류 – 결과변수의 유형별 (3)

RexSoft /rexsoft.org

- **Cox 비례위험 모형 (Cox's proportional hazard model)**
  - 생존형 결과변수 (survival outcome)
  - 가정: 결과변수가 예측변수의 가중합과 로그위험함수 (log hazard function)을 통해 관련됨
  - 예. $ln(h(Y)) = \beta_0 + \beta_1 X + \varepsilon$
  - Cox 비례위험회귀분석 (Cox's proportional hazard regression)

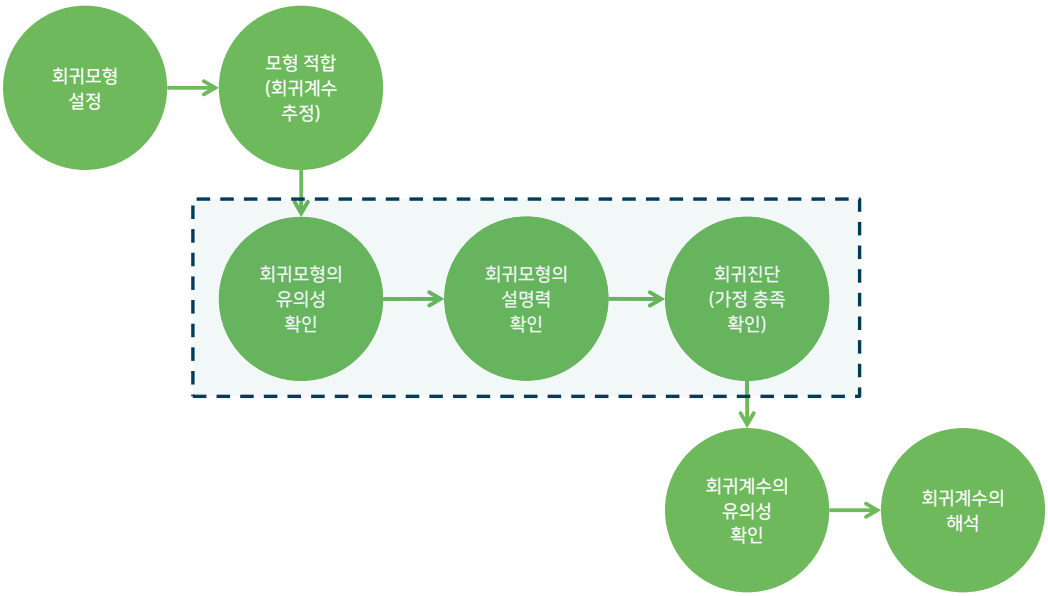
회귀분석의 종류 – 설명변수의 갯수별

RexSoft /rexsoft.org

- **단순회귀 (simple regression)**
  - 1개의 설명변수
  - 단변수분석 (univariable analysis)
  - 예. $Y = \beta_0 + \beta_1 X + \varepsilon$

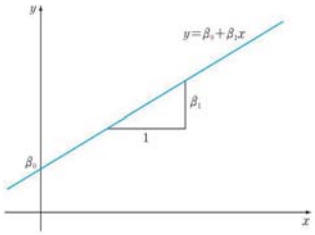
- **다중회귀 (multiple regression)**
  - 2개 이상의 설명변수
  - 다변수분석 (multivariable analysis)
  - 예. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

회귀분석의 절차



(선형)회귀분석 개요

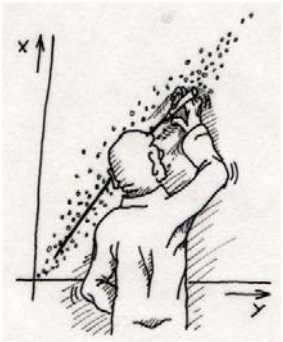
- 목적 : 연속형 결과변수 (y)에 대한 설명변수 (x)의 관계를 선형 회귀식으로 추정
- 회귀식/회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim i.i.d.N(0, \sigma^2)$
- 가정 : ① 선형성 (y 와 x가 직선관계에 있음)
  ② 오차항의 정규성 (오차가 정규분포를 따름)
  ③ 오차항의 독립성 (오차들이 서로 영향을 받지 않고 독립적임)
  ④ 오차항의 등분산성 (설명변수의 모든 값에서 오차의 분산은 동일함)
- Y 절편 (intercept) β_0
  - $x = 0$ 일 때 y 값들의 평균
- 기울기 (slope) β_1
  - 회귀계수 (regression coefficient): 설명변수의 상대적 기여도
  - x가 1단위 증가할 때 y 의 평균적인 증가량
  - 주로 최소제곱법에 의해 추정됨 (오차항의 제곱합이 최소가 되는 직선)



- 오차항 (error term) ε
  - 회귀식에서 x 에 의해 설명되지 않는 남은 변동
  - 오차항에 대한 가정 확인 시, 오차항의 추정치인 잔차 (residual, e)를 이용 (잔차분석)

(선형)회귀분석 개요

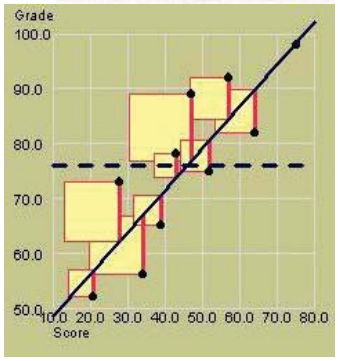
- **최적합 직선 (best fitting line)**
  - 자료를 가장 잘 설명하는 직선
  - 자료와의 오차가 가장 작은 직선



- **최소제곱법 (least square method)**
  - 오차제곱합 (SSE, sum of squares due to error)

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

  - 오차제곱합을 최소로 하는 (β_0 , β_1) 추정하자!



(선형)회귀분석: 분산분석 (ANOVA, Analysis of Variance)

- **결과변수 값의 변동 (variation)을 두 부분으로 분해**

총변동 Total variation = 회귀변동: 설명변수에 의한 변동 Variation due to regression line + 오차에 의한 변동 Variation due to error

회귀직선이 설명하는 변동

회귀직선으로 설명되지 않는 변동

- **두 변동의 상대적인 비율로 설명변수의 효과를 판단!**

(선형)회귀분석: 총변동의 분해

RexSoft /rexsoft.org

- **SST (total sum of squares)**
  - 총제곱합
  - 결과변수의 총변동을 측정한 값

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy}$$

- **SSR (sum of squares for regression)**
  - 회귀제곱합
  - 회귀직선으로 설명되는 결과변수의 변동을 측정한 값

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{y})^2$$

- **SSE (sum of squares for error)**
  - 잔차제곱합
  - 회귀직선으로 설명하고 남은 설명되지 않는 결과변수 값 (오차)의 변동을 측정한 값

$$SSE = \sum_{i=1}^{k}(y_i - \hat{y}_i)^2 = SST - SSR$$

(선형)회귀분석: 분산분석표 (ANOVA table)

RexSoft /rexsoft.org

Source	df	SS	MS	F
Regression	1	SSR	MSR = SSR	MSR/MSE
Error	$n-2$	SSE	MSE = SSE/$(n-2)$	
Total	$n-1$	SST		

- **SST = SSR + SSE**
- **F 검정 통계량: F= MSR/MSE~ F(1, n – 2)**

(선형)회귀분석: 모형 평가 (1)

RexSoft /rexsoft.org

- **회귀식의 유의성 확인**
  - 회귀직선이 의미가 있는가?
  - 회귀직선이 설명하는 변동 (SSR)이 충분히 큰가?

- **분산분석표의 F-test의 p-value 로 판정**
  - H_0: regression model/line is not significant (i.e., $\beta_1 = 0$)
  - Test statistic $F = \frac{MSR}{MSE} \sim F(1, n-2)$
  - 유의확률 (p-value) < 유의수준 (α) → 귀무가설 기각
    → 회귀식이 통계적으로 유의하다고 판단

(선형)회귀분석: 모형 평가 (2)

RexSoft /rexsoft.org

- **회귀식의 설명력 확인**
  - 회귀직선이 얼마나 선형 인과관계를 설명하는가?
  - 결정계수 (coefficient of determination) R^2 를 사용하여 판단
  - 총변동에서 직선이 설명하는 변동의 비율

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SSR}{SST}$$

  - $0 \le R^2 \le 1$
  - $R^2 \approx 1$ → 설명력 높음 → good 모형
  - $R^2 \approx 0$ → 설명력 낮음 → poor 모형

(선형)회귀분석: 모형 평가 (3)

RexSoft /rexsoft.org

- **회귀진단 (가정 충족 확인)**
  - 회귀분석 결과는 모형에 대한 가정이 만족될 때 신뢰할 수 있다!

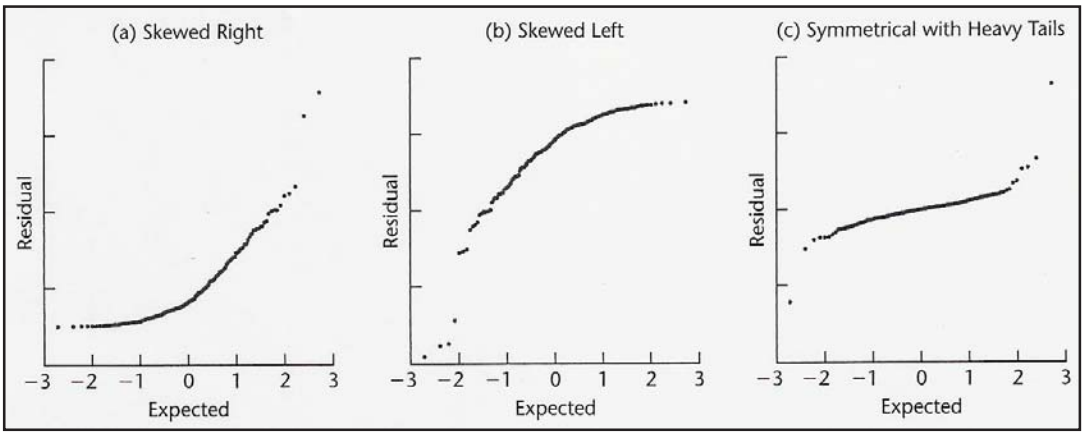
- **인과성 형태에 대한 가정**
  - 직선 형태 (linearity) → 산점도와 상관계수로 확인

- **오차에 대한 가정 → 잔차 (residual) 분석!**
  - 정규성 (normality) → 정규확률그림
  - 등분산성 (equal variance) → 잔차그림
  - 독립성 (independence) → 잔차그림 및 Durbin-Watson 통계량으로 확인
    - ✓ DW가 2 안팎이면 이웃하는 잔차 간에 상관성이 없다고 판단
    - ✓ DW가 0이나 4에 가까우면 독립성 위배로 판단

(선형)회귀분석: 모형 평가 (3) 정규확률그림

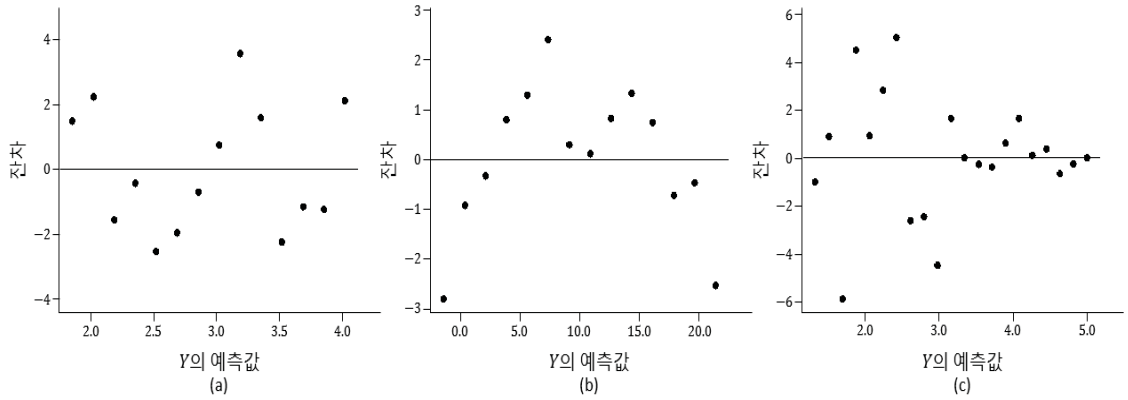
RexSoft /rexsoft.org

- **정규확률그림**
  - 대부분의 점들이 기준선 근처에 있다면 정규성이 만족한다고 판정함
  - 기준선에서 떨어져 있는 점들이 (특히 양 극단에서) 관찰되면 정규성이 만족되지 않는다고 판단



(선형)회귀분석: 모형 평가 (3) 잔차그림

- **잔차그림**
  - 기본 가정이 만족될 경우, 점들이 기준선을 중심으로 random하게 (즉, 패턴없이) 분포
  - 특정 패턴이 관찰될 경우 기본 가정 중 일부 또는 전부가 위배되는 것으로 판단함



문제없음

독립성 가정이
만족되지 않음

등분산성 가정이
만족되지 않음

(선형)회귀분석: 모형 평가 (3)

- **영향력 있는 관측값 (influential observation)**
  - 분석에서 제외할 경우 회귀계수 추정에 상당한 변화가 생기는 경우
  - 이상값이거나 지렛값
  - Rex 에서는 쿡의 거리, 지렛값으로 확인
- **이상값 (outlier)**
  - 잔차가 매우 큰 관측값
  - 예. sample peculiarity or a data entry error
- **지렛값 (leverage point)**
  - 설명변수 값이 평균에서 멀리 떨어져 있는 관측값

(선형)회귀분석: 회귀계수의 유의성 확인 및 해석

RexSoft /rexsoft.org

▪ 설명변수의 효과 추정

- 회귀계수 = 직선의 기울기의 추정값
- 회귀분석표에서 Coefficient 사용

▪ 회귀계수의 유의성 확인

- 회귀분석표에서 t-검정의 p-value 사용하여 판단
  - ✓ H_0: $\beta_1 = 0$ (즉, 설명변수의 효과가 없다)
  - ✓ Test statistic $t = \frac{b_1}{\sqrt{MSE/S_{xx}}} \sim t(n-2)$ under H_0

▪ 회귀계수의 해석

- 설명변수 값이 1단위 증가할 때 결과변수 값의 평균적인 증가량

예제. Alistipes와 SBP의 연관성 규명 연구

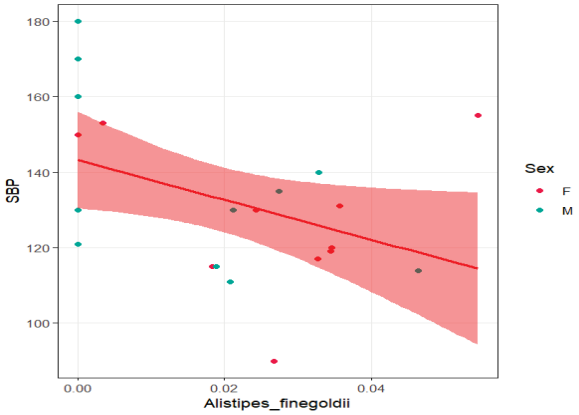
RexSoft /rexsoft.org

▪ 데이터 소개

- Alistipes가 SBP에 미치는 효과를 규명하고자 한다.
- SBP는 성별, 나이에 영향을 받으므로, 성별, 나이 효과를 보정해야할 필요가 있다.
- 일반적으로 성별, 나이는 교란변수로 고려되기 때문에 메타데이터를 분석하는 경우 대부분의 분석에 공변량으로 포함할 필요가 있다.

▪ 데이터 및 산점도

obs	SBP	Alistipes
1	120	0.0346
2	115	0.0183
3	130	0.0212
4	131	0.0357
5	111	0.0209
⋮	⋮	⋮

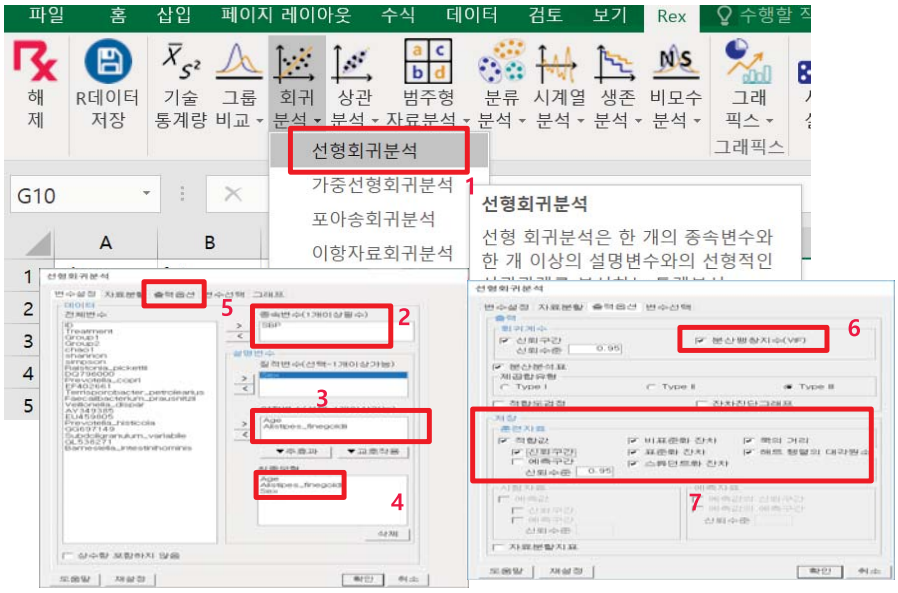


예제. Alistipes와 SBP의 연관성 규명 연구

실습

RexSoft /rexsoft.org

입 력 Rex ► 회귀분석 ► 선형회귀분석



예제. Alistipes와 SBP의 연관성 규명 연구

실습

RexSoft /rexsoft.org

분산분석표

출 력

Model Effect (Goodness of Fit Test)

	SS	DF	MS	F-value	P-value	R2	adj.R2
Regression	2476.5000	3	825.5	2.094	0.1349	0.2485	0.1298
Residual	7490.3696	19	394.23				
Total	9966.8696	22					

P>0.05, 회귀식은 통계적으로 유의하지 않음.

R2=0.2363, 제시한 모형은 이 회귀식에 의해 23.6%만큼 설명됨

T검정

Coefficient Estimates from a regression

	Estimate	SE	T-value	P-value	Lower bound of 95% CI for Estimate*	Upper bound of 95% CI for Estimate*
(Intercept)	152.1760	13.5379	11.2408	$7.75x10^{-10}$	123.8410	180.5111
Age	-0.3030	0.2552	-1.1875	0.2497	-0.8370	0.2310
Alistipes_finegoldii	-409.4945	266.4252	-1.5370	0.1408	-967.1288	148.1398
SexM	4.8224	8.6942	0.5547	0.5856	-13.3747	23.0194

B=4.8224 (95% CI, -13.3747 to 23.0194, p<0.0001)
허리둘레가 1cm 증가하면 복부지방양은 평균적으로 0.1942g만큼 증가하며
이는 통계적으로 유의하다

예제. Alistipes와 SBP의 연관성 규명 연구

실습

RexSoft /rexsoft.org

출 력

모형적합도

Model Fitness Measurements

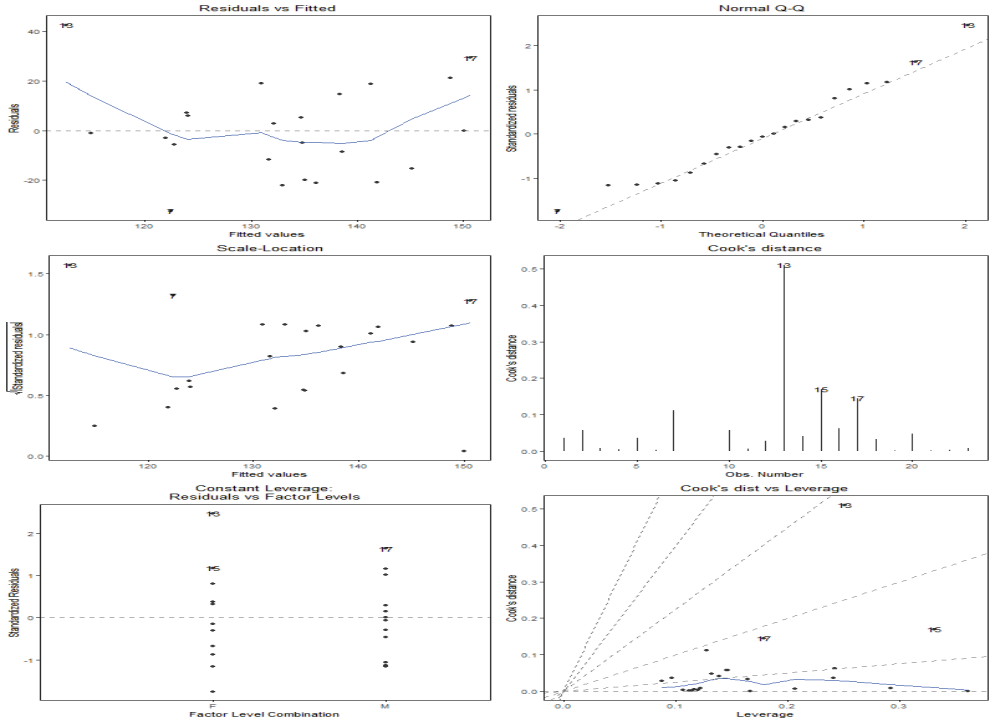
	Value	DF
Deviance	7490.3696	19
Pearson's chi-square	7490.3696	19
-2*log-likelihood	198.3464	5
AIC	208.3464	
BIC	214.0239	

예제. Alistipes와 SBP의 연관성 규명 연구

실습

RexSoft /rexsoft.org

출 력



주의사항

▪ 정규성 가정 불만족시

- 회귀모형에서 y 에 관한 추론(추정 및 가설검정)은 y의 정규성을 기초로 하기 때문에 이 조건이 만족되는지를 확인하는 작업이 필요

- 만약 y 의 정규성이 만족되지 않는다면
  1. 변수변환(variable transformation)
     로그변환, 제곱근변환, 제곱변환 등
  2. 비모수적인(nonparametric) 방법

▪ 변수변환 예시

- 허리둘레과 심복부 지방 조직의 승법적인 관계식이 성립하는 경우에는 심복부지방을 로그변환하여 선형회귀모형으로 적합시킬 수 있음

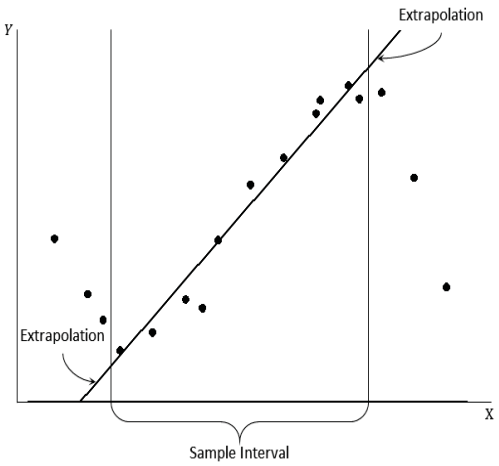
obs	fat	waist	log_fat
1	75	26	1.875061263
2	73	26	1.86332286

주의사항

▪ 외삽법 (extrapolation)

- 설명변수가 관측되지 않은 영역에서 반응변수를 예측하는 것을 외삽법이라고 함.

- 외삽법은 잘못된 결과를 도출할 가능성이 매우 크다. 예를 들어 관측된 설명변수의 영역에서 두 변수 사이에 선형관계가 존재할 수 있으나, 관측되지 않은 영역에서는 곡선의 관계가 있을 수 있음.



다중회귀분석: 다중공선성 (multicollinearity)

▪ 다중공선성의 정의

- 설명변수들이 너무 밀접한 관련이 있어서 각각의 영향을 보정하여 분석하기 어려운 상태
- 다중 공선성이 있을 경우 모형 추정 시 오차 증가 → 회귀분석 결과를 신뢰하기 어려움

▪ 다중공선성의 징후

- 모형의 설명력(R2)은 높은데 유의한 설명변수가 없을 경우
- 특정 설명변수를 포함한 모형과 포함하지 않은 모형을 비교할 때, 다른 설명변수의 (편)회귀계수 값이 모형에 따라 차이가 나는 경우
- 설명변수 사이의 상관계수가 높을 경우
- 분산팽창요인 (VIF, variance inflation factor)이 큰 설명변수들이 존재하는 경우 (기준: 10 이상)
- 공차한계 (Tolerance)가 작은 설명변수가 존재하는 경우 (기준: 0.1 이하)

▪ 다중공선성의 해결

- 다중공선성을 보이는 여러 개의 설명변수 중 1개만 모형에 포함하는 방법
  - ✓ 임상적으로 중요 또는 유용한 변수 선택
  - ✓ 단변수분석 결과에서 가장 효과가 큰 (=회귀계수 절대값이 가장 큰) 변수나 가장 유의한 (=p-value가 가장 작은) 변수 선택
- 변수선택법 적용
- 정규화(regularization) 회귀분석 사용 - Ridge regression, LASSO 등

다중회귀분석: 변수선택법

▪ 모든 부분집합 선택법 (all possible subsets selection)

▪ 순차적 변수선택법 (sequential variable selection procedure)

- 전진선택법 (forward selection)
  - ✓ 영모형 (null model)에서 출발
  - ✓ 설명변수 중에서 가장 크게 영향을 미칠 것 같은 변수부터 하나씩 포함하면서 변수선택 진행
  - ✓ 모형 적합도가 더 이상 개선되지 않으면 종료
- 후진제거법 (backward elimination)
  - ✓ 모든 설명변수를 포함하는 포화모형 (full model)에서 출발
  - ✓ 설명변수 중에서 가장 적게 영향을 미칠 것 같은 변수부터 하나씩 제거하면서 변수선택 진행
  - ✓ 모형 적합도가 상당히 감소할 것 같으면 종료
- 단계별선택법 (stepwise selection)
  - ✓ 전진선택법과 후진제거법을 결합한 형태

▪ 정규화 기반 선택법

- 예. LASSO, Elastic-net

다중회귀분석: 변수선택의 기준

- **모형의 적합도**
  - 수정된 결정계수 adj. R^2
  - F-검정 통계량
  - Deviance
  - -2*로그우도값 (-2*log-likelihood)
  - Akaike information criterion (AIC) → Rex 변수선택의 기준
  - Bayesian information criterion (BIC)

- **모형의 예측력**
  - 오차제곱합 SSE
  - Mallow's C_p
  - 예측오차제곱합 PRESS

다중회귀분석: 회귀계수의 해석

- **회귀계수**
  - 설명변수 값이 1단위 증가할 때 결과변수 값의 평균적인 증가량
  - 각 설명변수들이 결과변수에 미치는 영향력의 크기를 비교하고자 할 때, 설명변수들의 단위가 다른 경우 회귀계수로 직접적인 비교가 불가능

- **표준화 회귀계수 (standardized regression coefficient)**
  - 설명변수들의 값을 평균 0, 표준편차 1로 표준화하여 산출한 값
  - 표준화 회귀계수의 절댓값으로 여러 설명변수들 가운데 가장 영향력이 큰 변수를 확인할 수 있음

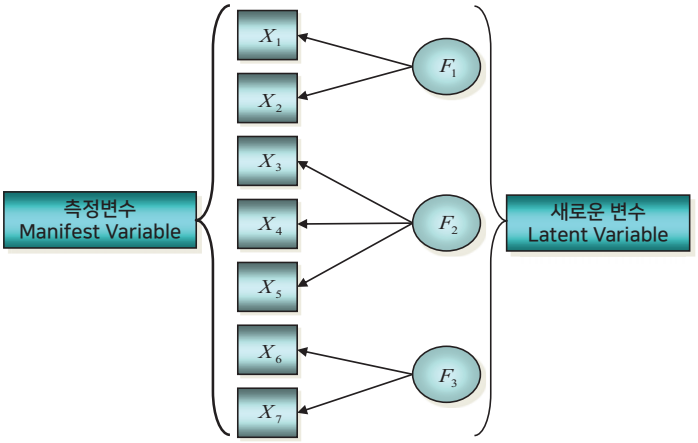
05 차원축소

Principal Component Analysis: PCA
Principal Coordinate Analysis: PCoA

차원축소 개념

▪ 차원축소의 필요성

- 여러 개의 변수사이의 관련성을 분석하여 그 변수들에 공통적으로 부여 가능한 인자를 파악하여 해석 가능한 적은 개수의 새로운 변수로 차원을 축소

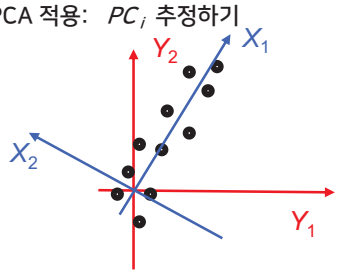


- 여러 개의 변수로부터 적은 수의 공통인자를 찾아내어 이 인자들을 회귀분석 또는 판별분석 등 차후의 분석에 이용할 수 있음.
- 측정(관측)된 변수들은 가상의 인자들의 선형결합으로 표현

차원축소 개념

▪ 주성분분석

- PCA 적용: PC_i 추정하기

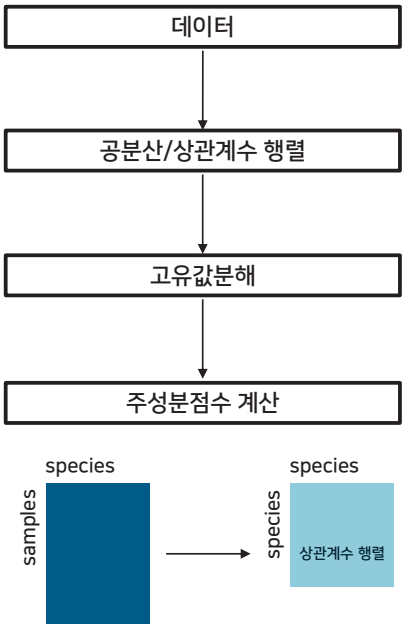


$$x_1 = \text{argmax}_{|w|=1} \; E\left(w^T Y\right)^2$$
$$x_2 = \text{argmax}_{|w|=1} \; E\left(w^T \hat{Y}_2\right)^2,$$
$$\text{where} \quad \hat{Y}_2 = Y - e_1 e_1^T Y$$

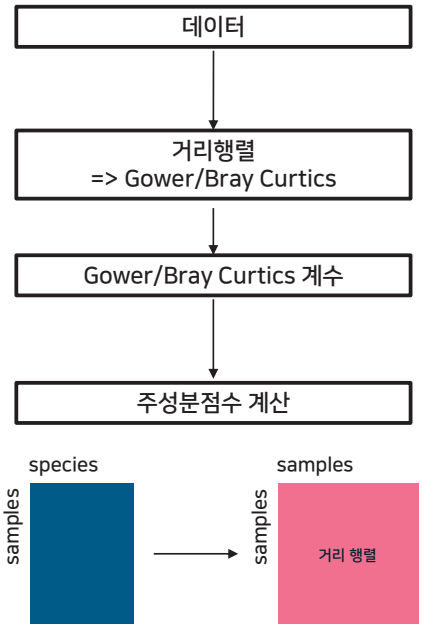
- x_1 와 x_2 는 고유값분해(eigenvalue decomposition) 기법을 활용하여 계산할 수 있다
  - ➔ x_1 는 제1 고유벡터이고, x_2 는 제2 고유벡터가 된다.
  - ➔ x_i: i번째 주성분점수 (principal component score)
  - ➔ $x_i \cdot Y$: i번째 주성분 (principal component)
  - ➔ var($e_i \cdot Y$) = λ_i

PCA vs PCOA

▪ PCA



▪ PCOA



▪ 주성분분석

- 고유값분해

$$A=\begin{pmatrix}1 & 2\\ 2 & 5\end{pmatrix} \Rightarrow (A-\lambda I)\underline{x}=0$$

$$\Rightarrow |A-\lambda I|=\begin{vmatrix}1-\lambda & 2\\ 2 & 5-\lambda\end{vmatrix}=0$$

$$\Rightarrow \lambda^2-6\lambda+1=0$$

$$\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}=\begin{pmatrix}3+2\sqrt{2}\\ 3-2\sqrt{2}\end{pmatrix},\quad \begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}1\\ 1+\sqrt{2}\end{pmatrix},\ \begin{pmatrix}1\\ 1-\sqrt{2}\end{pmatrix}$$

즉,

$$\begin{pmatrix}1\\ 1+\sqrt{2}\end{pmatrix}^T\begin{pmatrix}1\\ 1-\sqrt{2}\end{pmatrix}=0$$

$$\begin{pmatrix}1 & 2\\ 2 & 5\end{pmatrix}\begin{pmatrix}1\\ 1+\sqrt{2}\end{pmatrix}=(3+2\sqrt{2})\begin{pmatrix}1\\ 1+\sqrt{2}\end{pmatrix},$$

$$\begin{pmatrix}1 & 2\\ 2 & 5\end{pmatrix}\begin{pmatrix}1\\ 1-\sqrt{2}\end{pmatrix}=(3-2\sqrt{2})\begin{pmatrix}1\\ 1-\sqrt{2}\end{pmatrix}$$

▪ 질환 여부 따라 유의한 차이가 있는 균주를 찾으시오 (데이터 : stat_example.xls)

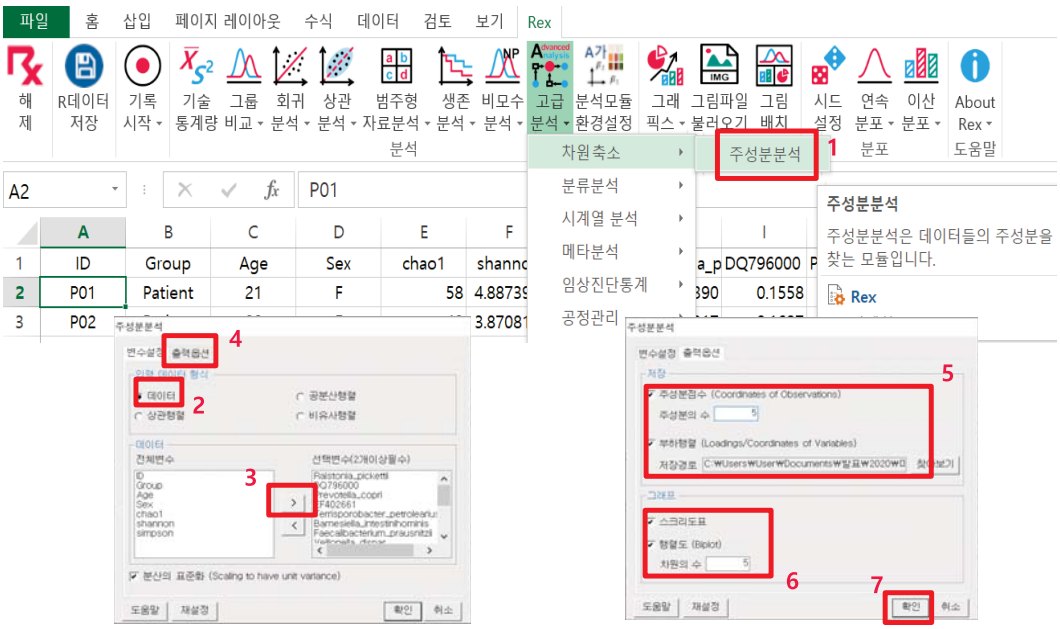
- meta+species sheet (샘플별 메타정보 및 균주 정보)
  - ✓ Option
  - ✓ beta
  - ✓ ID : 일련번호
  - ✓ Group : 질환여부 (patient, treated, control)
  - ✓ Age : 각 샘플의 나이
  - ✓ Sex : 각 샘플의 성별 정보
- UniFrac sheet (샘플 간의 UniFrac 거리)

PCA 실습

실습

RexSoft /rexsoft.org

입 력 Rex ► 고급분석 ► 차원축소 ► 주성분분석

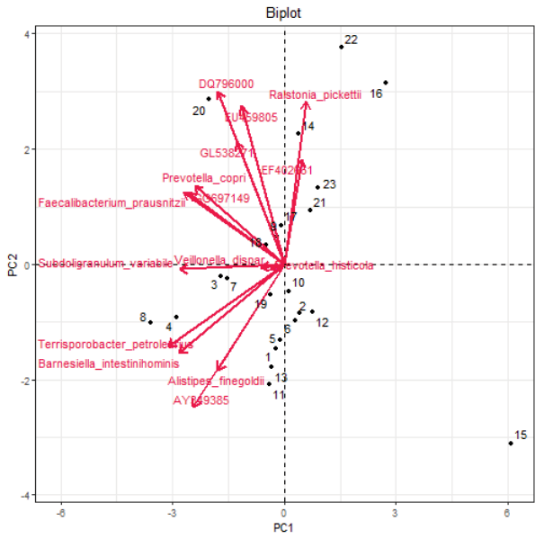


PCA 실습

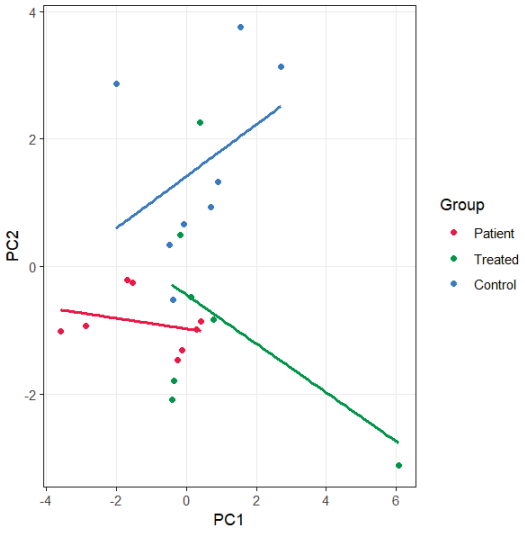
실습

RexSoft /rexsoft.org

BIPLOT



출 력

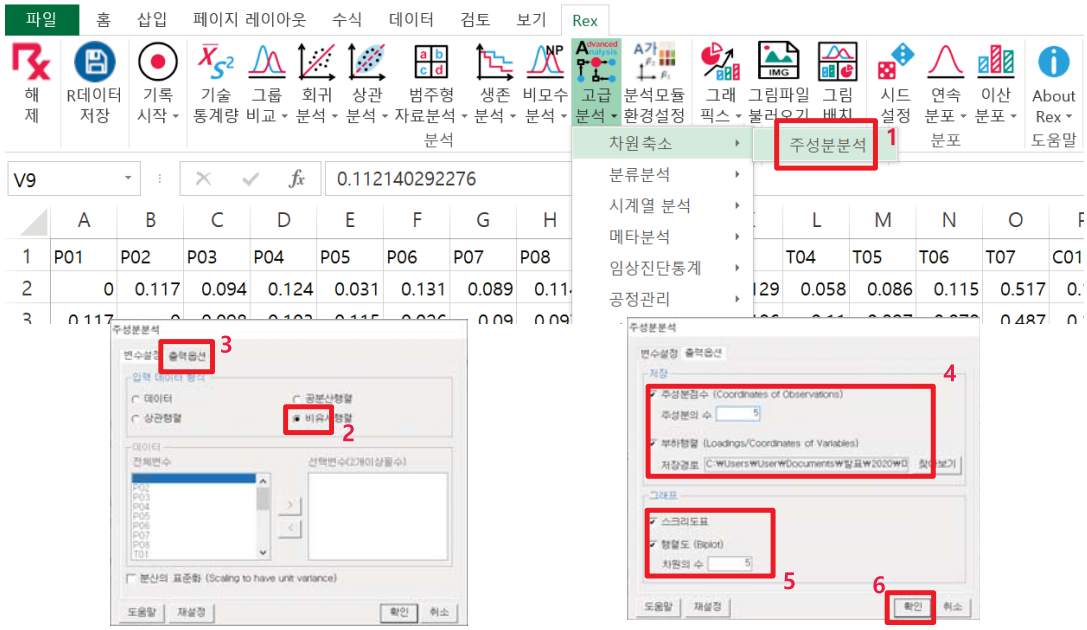


PCOA 실습

실습

RexSoft /rexsoft.org

입 력 Rex ► 고급분석 ► 차원축소 ► 주성분분석



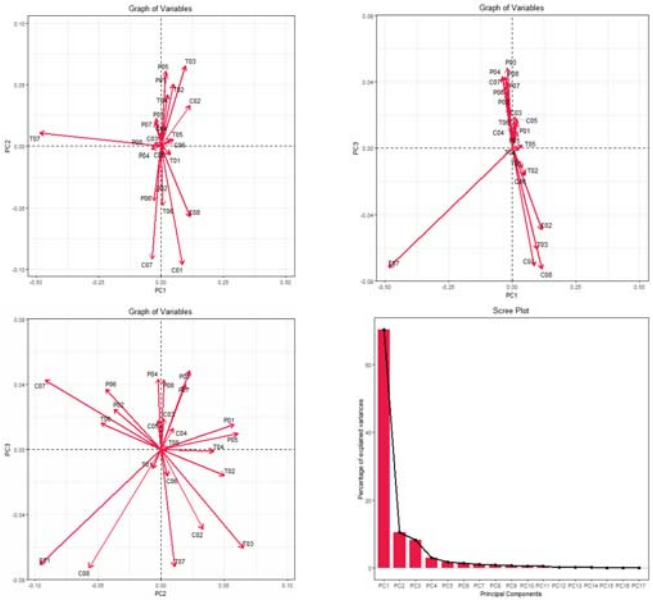
PCOA 실습

실습

RexSoft /rexsoft.org

BIPLOT

출 력



RexSoft

감사합니다

93