

No.

date

A

/

title

영어 노트

[illegible]

혼자 공부하는 R 데이터 분석

혼자 공부하며 함께 만드는

혼공 용어 노트

목차

가나다순

가설 검정 statistical hypothesis test	21	마크다운 markdown	22
검정통계량 test statistic	20	막대 그래프 bar chart	15
결측치 missing value	17	매개변수 parameter	11
관측치 observations	08	반복문 iteration	12
귀무 가설 null hypothesis	21	배열 array	09
그래픽 사용자 인터페이스		범주형 자료 categorical data	09
GUI; Graphical User Interface	06	벡터 vector	08
기술통계량 descriptive statistic	13	변수 variable	10
기울기 slope	19	분산 variance	14
네이티브 파이프 연산자 native pipe operator	21	분산분석 ANOVA; analysis of variance	22
누적 막대 그래프 stacked bar chart	18	분위수 quantile	13
대립 가설 alternative hypothesis	21	빅데이터 big data	06
데이터 과학 data science	06	빈도분석 frequency analysis	14
데이터 분석 data analysis	07	사분위수 quartile	13
데이터 세트 data set	07	산점도 scatter plot	16
데이터 시각화 data visualization	18	상관분석 correlation analysis	20
데이터 유형 data type	08	상자 그림 boxplot	15
데이터 재구조화 reshaping data	17	선버스트 차트 sunburst chart	18
데이터 전처리 data preprocessing	16	선형성 linearity	20
데이터 프레임 data frame	09	스크립트 script	06
독립변수 independent variable	19	연산자 operator	11
등분산성 homoscedasticity	21	왜도 skewness	14
리스트 list	09	원시 데이터 raw data	12

웹 크롤링 web crawling	07
유의 확률 p-value	20
응용 프로그램 프로그래밍 인터페이스	
API; Application Programming Interface	20
이상치 outlier	18
인덱스 index	09
인코딩 encoding	06
인터랙티브 웹 interactive web	22
절편 intercept	18
정규분포 normal distribution	14
제이슨 JSON; JavaScript Object Notation	12
조건문 conditional	11
종속변수 dependent variable	19
줄기 잎 그림 stem-and-leaf plot	16
첨도 kurtosis	14
컬럼 column	13
키 key	17
테이블 table	08
통합 개발 환경	
IDE; Integrated Development Environment	06
파이차트 pie chart	15
파이프 연산자 pipe operator	16
패키지 package	11

표준편차 standard deviation	14
할당 연산자 assignment operator	10
함수 function	10
행렬 matrix	09
회귀분석 regression analysis	19
히스토그램 histogram	15
f 검정 f-test	21
t 검정 t-test	21
XML eXtensible Markup Language	12

ABC 순

alternative hypothesis 대립 가설	21	eXtensible Markup Language XML	12
ANOVA: analysis of variance 분산분석	22	f-test f 검정	21
API: Application Programming Interface		frequency analysis 빈도분석	14
응용 프로그램 프로그래밍 인터페이스	20	function 함수	10
array 배열	09	GUI: Graphical User Interface	
assignment operators 할당 연산자	10	그래픽 사용자 인터페이스	06
bar chart 막대 그래프	15	histogram 히스토그램	15
big data 빅데이터	06	homoscedasticity 등분산성	21
boxplot 상자 그림	15	IDE: Integrated Development Environment	
categorical data 범주형 자료	09	통합 개발 환경	06
column 컬럼	13	independent variable 독립변수	19
conditional 조건문	11	index 인덱스	09
correlation analysis 상관분석	20	interactive web 인터랙티브 웹	22
data analysis 데이터 분석	07	intercept 절편	18
data frame 데이터 프레임	09	iteration 반복문	12
data preprocessing 데이터 전처리	16	JSON: JavaScript Object Notation 제이슨	12
data science 데이터 과학	06	key 키	17
data set 데이터 세트	07	kurtosis 첨도	14
data type 데이터 유형	08	linearity 선형성	20
data visualization 데이터 시각화	18	list 리스트	09
dependent variable 종속변수	19	markdown 마크다운	22
descriptive statistic 기술통계량	13	matrix 행렬	09
encoding 인코딩	06	missing value 결측치	17

native pipe operator 네이티브 파이프 연산자	21
normal distribution 정규분포	14
null hypothesis 귀무 가설	21
observations 관측치	08
operator 연산자	11
outlier 이상치	18
p-value 유의 확률	20
package 패키지	11
parameter 매개변수	11
pie chart 파이차트	15
pipe operator 파이프 연산자	16
quantile 분위수	13
quartile 사분위수	13
raw data 원시 데이터	12
regression analysis 회귀분석	19
reshaping data 데이터 재구조화	17
scatter plot 산점도	16
script 스크립트	06
skewness 왜도	14
slope 기울기	19
stacked bar chart 누적 막대 그래프	18
standard deviation 표준편차	14
statistical hypothesis test 가설 검정	21

stem-and-leaf plot 줄기 잎 그림	16
sunburst chart 선버스트 차트	18
t-test t 검정	21
table 테이블	08
test statistic 검정통계량	20
variable 변수	10
variance 분산	14
vector 벡터	08
web crawling 웹 크롤링	07

01장

□ 빅데이터 **big data** [01장 026쪽]


여러 종류의 데이터가 결합한 대규모 데이터.

□ 데이터 과학 **data science** [01장 026쪽]

데이터를 수집하고 가공하여 데이터에서 의미를 찾는 다양한 방법을 말한다.

□ 통합 개발 환경 **IDE; Integrated Development Environment** [01장 032쪽]

개발을 편하게 할 수 있도록 도와주는 개발도구. 코딩, 디버그, 컴파일, 배포 등을 모두 처리할 수 있다.


 R에서는 RGui나 R 스튜디오를 사용한다.

□ 그래픽 사용자 **GUI; Graphical User Interface** [01장 032쪽]

인터페이스 사용자와 컴퓨터가 서로 상호작용할 수 있도록 알기 쉬운 아이콘이나 그림으로 나타낸 인터페이스.

□ 스크립트 **script** [01장 051쪽]

코드를 작성한 문서.

 R 스튜디오에서는 Script 탭에서 코드를 편하게 작성할 수 있다.

□ 인코딩 **encoding** [01장 061쪽]

컴퓨터 정보의 어떤 형식을 다른 형식으로 변환하는 것.

그것이 알고싶다 **UTF-8**

유니코드를 표현하는 문자 인코딩 방식 중 하나. 전 세계의 모든 문자를 컴퓨터나 웹 페이지에 표현할 수 있다.

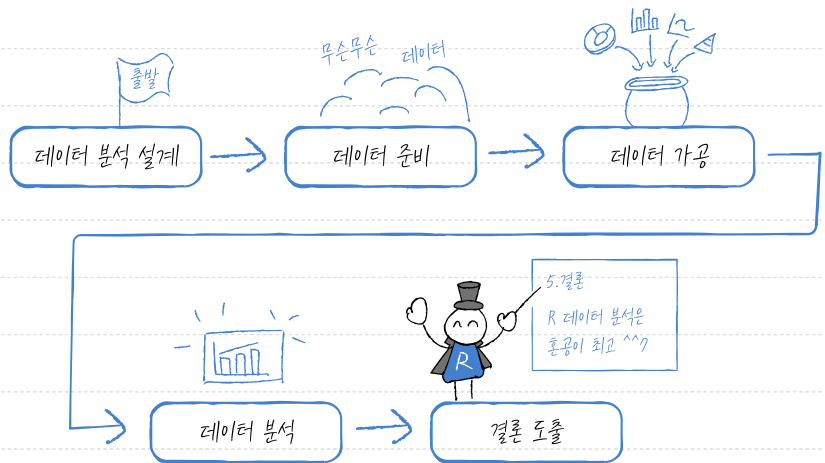
02장

□ 데이터 분석

data analysis

[02장 074쪽]

데이터를 활용하여 수치적으로 검증하고, 의사 결정을 합리적으로 하기 위해 전달하는 정보를 만드는 과정.



□ 웹 크롤링

web crawling

[02장 076쪽]

프로그램으로 웹사이트에서 원하는 정보를 가져오는 행위.

→ 웹 크롤러라고 한다.

□ 데이터 세트

data set

[02장 082쪽]

각각의 속성을 가진 여러 관측치가 모인 데이터의 집합.

□ 테이블

table

[02장 082쪽]

행과 열로 이루어진 데이터 세트.

행	열		
	번호	이름	나이
	1	혼양이	비밀
	2	혼공양이	1,200
	3	혼공스	22

□ 관측치

observations

[02장 082쪽]

데이터 세트를 이루는 행을 말한다.

□ 데이터 유형

data type

[02장 083쪽]

숫자로만 이루어진 숫자형, 문자로만 이루어진 문자형, TRUE 혹은 FALSE로 이루어진 논리형 등이 있다.

- 단일형: 숫자형, 문자형처럼 한 가지 데이터 유형으로 구성된 데이터
- 다중형: 여러 가지 데이터 유형으로 구성된 데이터

□ 벡터

vector

[02장 084쪽]

데이터 구조의 가장 기본 형태. 1차원으로 구성된 단일형 데이터.

문자형 벡터

U	A	N	D	I	♥
---	---	---	---	---	---

숫자형 벡터

8	8	2	8	8	2
---	---	---	---	---	---

논리형 벡터

TRUE	FALSE	TRUE
------	-------	------


전화는 안 받아줘!
문자만 받아줘!

저 드립 맞나?

맞나?

맞나?

□ 범주형 자료

categorical data 

[02장 089쪽]

명목형 자료를 범주화한 특수한 형태의 벡터.

그것이 알고싶다 명목형 자료

- 명목형 자료: 과일, 나라명, 도서명... 순서가 없는 자료.
- 수치형 자료: 1, 2, 3... 정수형, 실수형 자료.

□ 행렬

matrix

[02장 090쪽]

행과 열로 구성된 2차원의 단일형 데이터.

□ 배열

array

[02장 092쪽]

행렬을 n차원으로 확대한 단일형 데이터.

□ 리스트

list

[02장 093쪽]

1차원 데이터인 벡터나 서로 다른 구조의 데이터를 그룹으로 묶은 다중형 데이터 세트.

□ 인덱스

index

[02장 94쪽]

리스트 안에 있는 값의 위치를 의미한다. [] 대괄호로 위치를 가리키는데 이를 인덱싱이라고 한다.

□ 데이터 프레임

data frame

[02장 095쪽]

리스트를 2차원으로 확대한 다중형 데이터.

03장

□ 변수

variable

[03장 103쪽]


이름 그대로 ‘변하는 값’. 특정 범위 안에서 어떠한 값이라도 저장할 수 있다. 분석 편의를 위해 임시 값을 저장할 수도 있다.

- 첫 문자는 반드시 영문자(알파벳) 또는 마침표(.)를 사용한다.
- 첫 문자에는 숫자, 밑줄 문자(_)를 사용할 수 없다.
- 마침표(.)와 밑줄 문자(_)를 제외한 특수 문자는 사용할 수 없다.
- 대문자와 소문자를 구분한다.
- 변수명 중간에 빈칸을 넣을 수 없다. 빈칸은 밑줄 문자(_)를 활용하여 표현한다.

□ 할당 연산자

assignment operator

[03장 104쪽]

변수를 생성할 때 사용하는 연산자로  기호를 사용한다.

```
> x <- 10  
> x  
[1] 10
```

□ 함수

function  매개변수

[03장 105쪽]

특정 기능을 하도록 만들어진 프로그래밍 구문을 묶어 놓은 것.

- 내장 함수

함수명(인자)

• 사용자 정의 함수

```
함수명 <- function(매개변수1, 매개변수2, ....) {
  함수가 구현할 내용
  ...
  return(결괏값)
}
```

→ 사용자가 함수 기능을 정의한다.

□ 매개변수	parameter	[03장 107쪽]
	함수의 변수. 함수가 호출될 때 전달되는 어떠한 값. 없을 수도 있고 여러 개가 있을 수도 있다.	
□ 패키지	package	[03장 114쪽]
	여러 함수를 기능에 따라 묶어서 제공하는 것. 패키지에 있는 함수를 사용하려면 패키지를 설치하고 로드해야 한다.	
□ 연산자	operator	[03장 125쪽]
	프로그램에서 데이터를 처리하는 연산 기호.	
	<ul style="list-style-type: none"> • 할당 연산자: 특정한 값을 변수에 저장한다. • 산술 연산자: 숫자를 계산하는 연산자. • 관계 연산자: 변수 간의 혹은 변수와 값을 비교하여 관계를 TRUE와 FALSE 진릿값으로 알려준다. • 논리 연산자: 진릿값을 연산한다. 	
□ 조건문	conditional	[03장 130쪽]
	조건이 TRUE면 실행되는 코드 구문.	

```
if(조건) {
  조건이 TRUE(참)일 때 실행되는 구문1
} else {
  조건이 FALSE(거짓)일 때 실행되는 구문2
}
```

← 조건을 더 추가하려면 else if를 추가한다.

□ 반복문

iteration

[03장 132쪽]

정해진 조건만큼 반복 실행하는 구문.

- `apply()` 함수: 행렬을 연산한다.
- `lapply()` 함수: 벡터, 행렬, 리스트, 데이터 프레임을 연산한다. 실행 결과를 리스트로 반환한다.
- `sapply()` 함수: 벡터, 행렬, 리스트, 데이터 프레임을 연산한다. 실행 결과를 데이터 프레임으로 반환한다.

04장 ✓

□ 원시 데이터

raw data

[04장 146쪽]

가공하지 않은 처음의 데이터.

□ XML

eXtensible Markup Language

[04장 160쪽]

사용자가 <> 괄호로 직접 정의한 태그에 데이터 내용이 들어있는 파일.

□ JSON

JavaScript Object Notation

[04장 161쪽]

데이터 속성과 값이 쌍으로 이루어진 중첩 데이터 구조의 데이터 파일.

그것이 알고싶다 중첩 데이터

"가족관계": {"#": 2, "아버지": "홍판서", "어머니": "춘섬"}

↓
속성

↓
값 안에 다시 속성과 값이 있다.

□ 컬럼

column

[04장 168쪽]

데이터 프레임에서 열을 말한다. 변수에 해당한다.

5개 컬럼

Num	Size	weight	Tail	Species
1	45	6	30	cat
2	30	3	22	cat
...				
149	30	10	22	dog
150	53	17	22	dog

150개 관측치

□ 기술통계량

descriptive statistic

[04장 173쪽]

데이터를 요약한 대푯값, 데이터를 의미 있는 수치로 요약하여 데이터 특성을 파악할 수 있다.
평균, 중앙값, 최솟값, 최댓값 등

□ 분위수

quantile

[04장 175쪽]

전체 데이터를 크기 순으로 정렬하여 n개로 나누었을 때 그 경계에 해당하는 값.

□ 사분위수

quartile

[04장 175쪽]

데이터를 4등분 한 지점을 관측한 값.

- 제1사분위수: 제0.25분위수, 하위 25%에 해당하는 값.
- 제2사분위수: 제0.50분위수, 50%에 해당하는 값.
- 제3사분위수: 제0.75분위수, 하위 75% 혹은 상위 25%에 해당하는 값.
- 제4사분위수: 제1분위수, 100%에 해당하는 값.

□ 분산

variance

[04장 177쪽]

데이터가 평균으로부터 퍼진 정도를 설명하는 값.

□ 표준편차

standard deviation

[04장 177쪽]

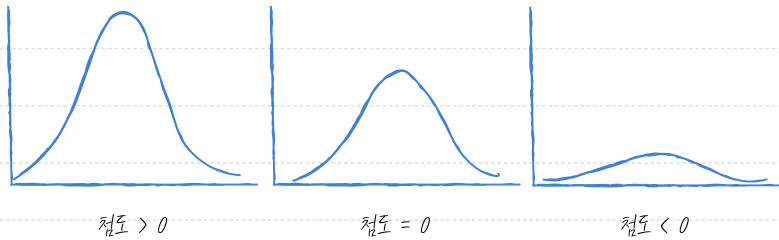
데이터 값이 퍼진 정도를 설명하는 값.

□ 첨도

kurtosis

[04장 178쪽]

데이터 분포가 정규분포 대비 뾰족한 정도를 설명하는 값.

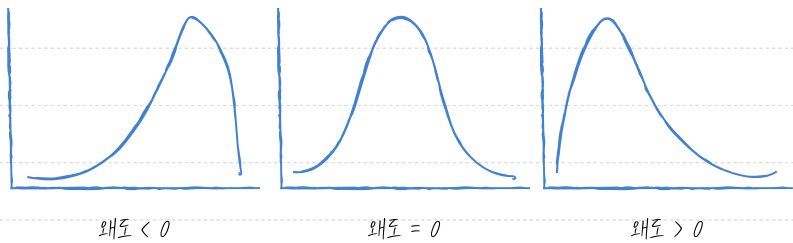


□ 왜도

skewness

[04장 178쪽]

데이터 분포의 비대칭성을 설명하는 값.



□ 정규분포

normal distribution

[04장 178쪽]

평균을 중심으로 좌우가 대칭이며 하나의 꼭지를 갖는 종 모양의 분포 형태를 말한다.

□ 빈도분석

frequency analysis

[04장 180쪽]

데이터 항목별로 빈도와 빈도 비율을 구하는 분석 방법. 데이터 분포를 파악할 때 가장 많이 사용한다.

□ 막대 그래프

bar chart

[04장 185쪽]

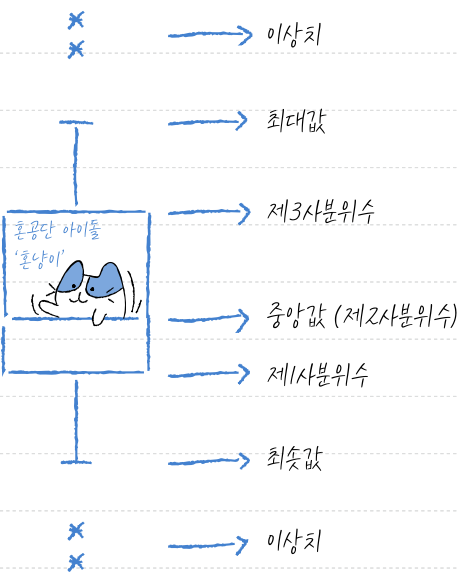
범주형 데이터의 수량이 많고 적음을 나타낼 때 적합한 그래프.

□ 상자 그림

boxplot

[04장 191쪽]

데이터 분포를 확인하고, 데이터 분포에서 벗어난 극단의 데이터를 판단할 때 적합한 그래프.



□ 히스토그램

histogram

[04장 193쪽]

연속형 데이터를 일정하게 구간을 나누어 각 구간에 해당하는 데이터 빈도를 그릴 때 적합한 그래프.

□ 파이차트

pie chart

[04장 195쪽]

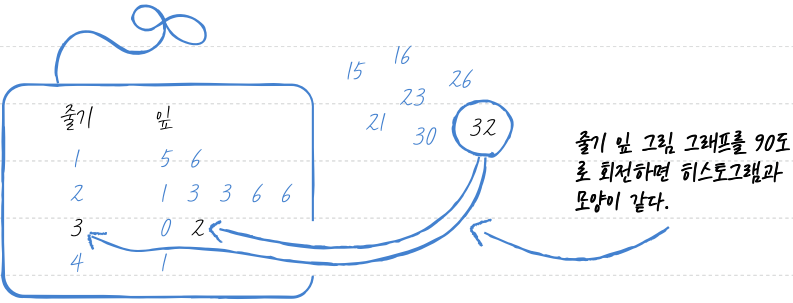
원을 데이터 범주 구성 비례에 따라 파이 조각 모양처럼 표현할 때 적합한 그래프.

□ 줄기 잎 그림

stem-and-leaf plot

[04장 196쪽]

변수 값을 자릿수로 분류한 것을 시각화하여 데이터 전체 형태를 파악할 때 적합한 그래프.



□ 산점도

scatter plot

[04장 199쪽]

두 변수 간의 관계를 점으로 나타낼 때 적합한 그래프.

05장 ✓

□ 파이프 연산자

pipe operator

[05장 219쪽]

dplyr 패키지

%>% 기호를 사용하여 데이터나 결과값을 변수로 저장하는 과정을 거치지 않고 데이터와 함수를 연결하여 사용할 수 있다. 파생변수를 만들지 않아도 된다.

□ 데이터 전처리

data preprocessing

[05장 224쪽]

변수를 생성하거나 변수명을 변경하고, 조건에 맞는 데이터를 추출하거나 변경하고, 데이터를 정렬하고 병합하는 일련의 과정.

유사 용어 데이터 가공, 데이터 핸들링, 데이터 마트

그것이 알고싶다 **dplyr 패키지**

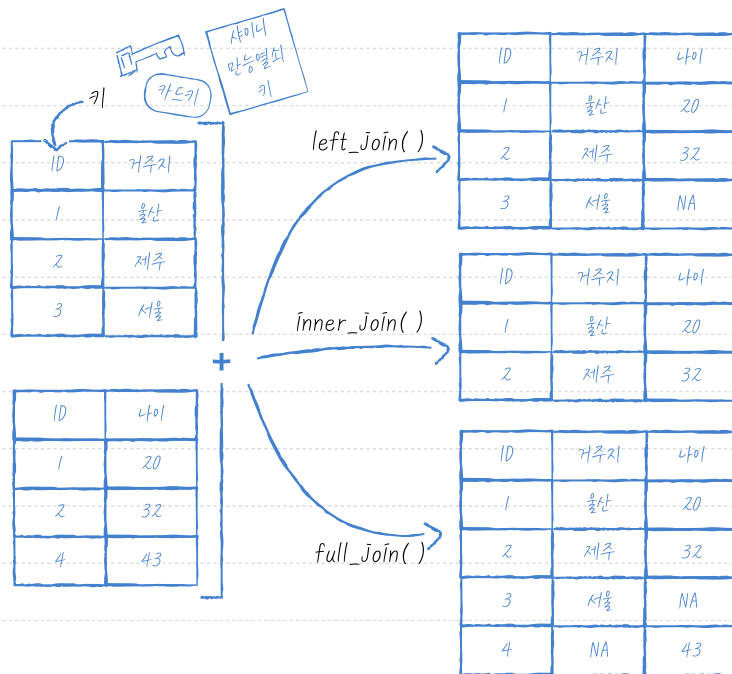
데이터 가공 필수 패키지. 사용자 친화적인 방식으로 설계되어, R에서 가장 많이 다루는 데이터 유형인 데이터 프레임을 직관적으로 조작할 수 있다.

□ 키

key

[05장 236쪽]

데이터를 정렬할 때 다른 데이터와 구별할 수 있는 고유한 식별자. 데이터를 결합할 때 기준이 된다.



□ 데이터 재구조화

reshaping data

[05장 246쪽]

동일한 데이터를 목적에 따라, 분석 기준에 따라 데이터 구조를 변형하는 것.

- melt(): 데이터의 열을 행으로 바꾼다.
- acast(): 데이터의 행을 열로 바꾼다. 결괏값을 벡터, 행렬, 배열로 반환한다.
- dcast(): 데이터의 행을 열로 바꾼다. 데이터 프레임으로 반환한다.

□ 결측치

missing value

[05장 263쪽]

데이터가 없는 것. 값이 누락된 것을 의미한다.

NA로 표기한다.



정상적인 데이터 분포에서 벗어난 값을 의미한다. 극단치라고도 한다.

06장

복잡해 보이는 수치 데이터를 이미지화하여 누구나 쉽게 내용을 이해할 수 있도록 시각적으로 전달하는 것을 말한다. 데이터의 특성을 파악할 때, 분석할 때, 공유할 때 등 데이터 분석 전 과정에서 활용할 수 있다.

그것이 알고싶다 ggplot 2 패키지

R의 내장 함수로도 그래프를 그릴 수 있지만, ggplot2 패키지에 있는 함수를 사용하면 더 다양한 기능을 활용할 수 있다.

전체적인 빈도와 각 변수의 범주의 빈도를 같이 보여줄 때 적합한 그래프. 막대 그래프 안에  색상으로 비율을 표시할 수 있다.
 함수에 fill 옵션을 지정한다.

계층 구조의 데이터를 범주별로 비율을 나타낼 때 적합한 그래프. 누적 막대 그래프와 변환이 가능하다.

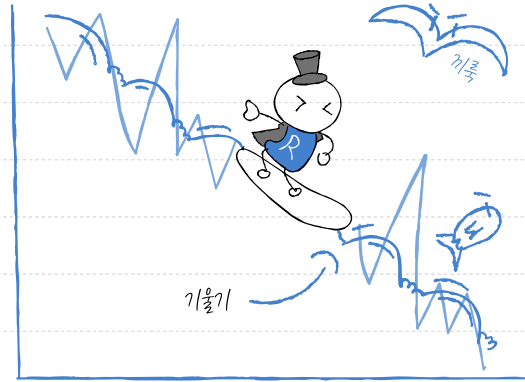
직선이 x축이나 y축과 만나는 좌표.

□ 기울기

slope

[06장 292쪽]

직선의 경사도, 값이 클수록 경사지게 표현된다.



□ 회귀분석

regression analysis

[06장 303쪽]

독립변수와 종속변수 간의 인과관계를 구하는 분석 기법.

- 단순회귀분석: 독립변수가 1개일 때
- 다중회귀분석: 독립변수가 2개 이상일 때

□ 독립변수

independent variable

[06장 303쪽]

다른 변수의 변화에 영향을 받지 않는 독립적인 변수.

□ 종속변수

dependent variable

[06장 303쪽]

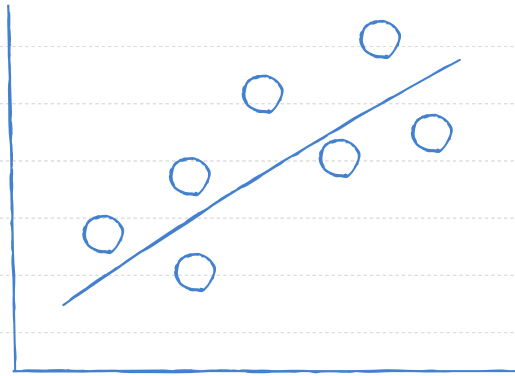
독립변수에 영향을 받아 변하는 변수.

□ 선형성

linearity

[06장 303쪽]

직선 형태를 가지는 것. 독립변수가 종속변수에 영향을 주는 경우 선형관계가 발생하며 그래프에 직선 형태로 나타난다.



□ 상관분석

correlation analysis

[06장 303쪽]

연속적인 두 변수 간의 연관성을 구하는 분석 기법.
e 상관관계를 확인한다.

□ 검정통계량

test statistic

[06장 304쪽]

가설을 검정할 때 표출에서 산출한 통계량.

□ 유의 확률

p-value

[06장 304쪽]

두 변수 간 상관관계가 통계적으로 의미가 있는지 판단하는 검정통계량. 일반적으로 0.05보다 작으면 '통계적으로 유의하다'고 해석한다.
e 통계적으로 의미가 있다!

□ 응용 프로그램

API; Application Programming Interface

[06장 308쪽]

프로그래밍

응용 프로그램들이 서로 상호작용할 수 있도록 도와주는 매개체.

인터페이스

07 장

□ 네이티브	native pipe operator	[07장 347쪽]
파이프 연산자	파이프 연산자와 동일하며 패키지를 설치하지 않아도 사용 가능하다.  기호를 사용한다. → R 버전 4.1.0이상	
□ 가설 검정	statistical hypothesis test	[07장 380쪽]
	가설이 통계적으로 유의한지 판단하는 검정. 가설을 세우고 그 가설이 맞는지 입증한다.	
□ 귀무 가설	null hypothesis	[07장 380쪽]
	<u>기준에 알려진 사실</u> 을 기준으로 설정하는 가설.	
□ 대립 가설	alternative hypothesis	[07장 380쪽]
	귀무 가설과는 반대로 새롭게 주장하려는 가설. <u>입증하고자 하는 가설</u> 이다.	
□ 등분산성	homoscedasticity	[07장 381쪽]
	비교하는 집단 간의 분산이 서로 같다는 것을 의미한다.	
□ f 검정	f-test	[07장 381쪽]
	두 집단의 <u>분산에 차이</u> 가 있는지 검정하는 기법.	
□ t 검정	t-test	[07장 381쪽]
	두 집단 간 <u>평균 차이</u> 가 있는지 검정하는 기법.	

세 개 이상 집단 간 평균 차이가 있는지 검정하는 기법.

08장

일반 텍스트 내용과 서식을 함께 작성하여 웹에 공유할 수 있는 마크업 언어.

사용자가 입력한 데이터에 따라 웹이 상호 작용하며 동작하는 웹 애플리케이션.



MEMO

Handwriting practice lines consisting of 20 horizontal dashed lines.

MEMO

Handwriting practice area with horizontal dashed lines.

혼자 공부하는 사람들을 위한 용어 노트

[illegible]