ADsP 데이터 분석 준전문가

데이터 분석 개요

데이터 분석 순서

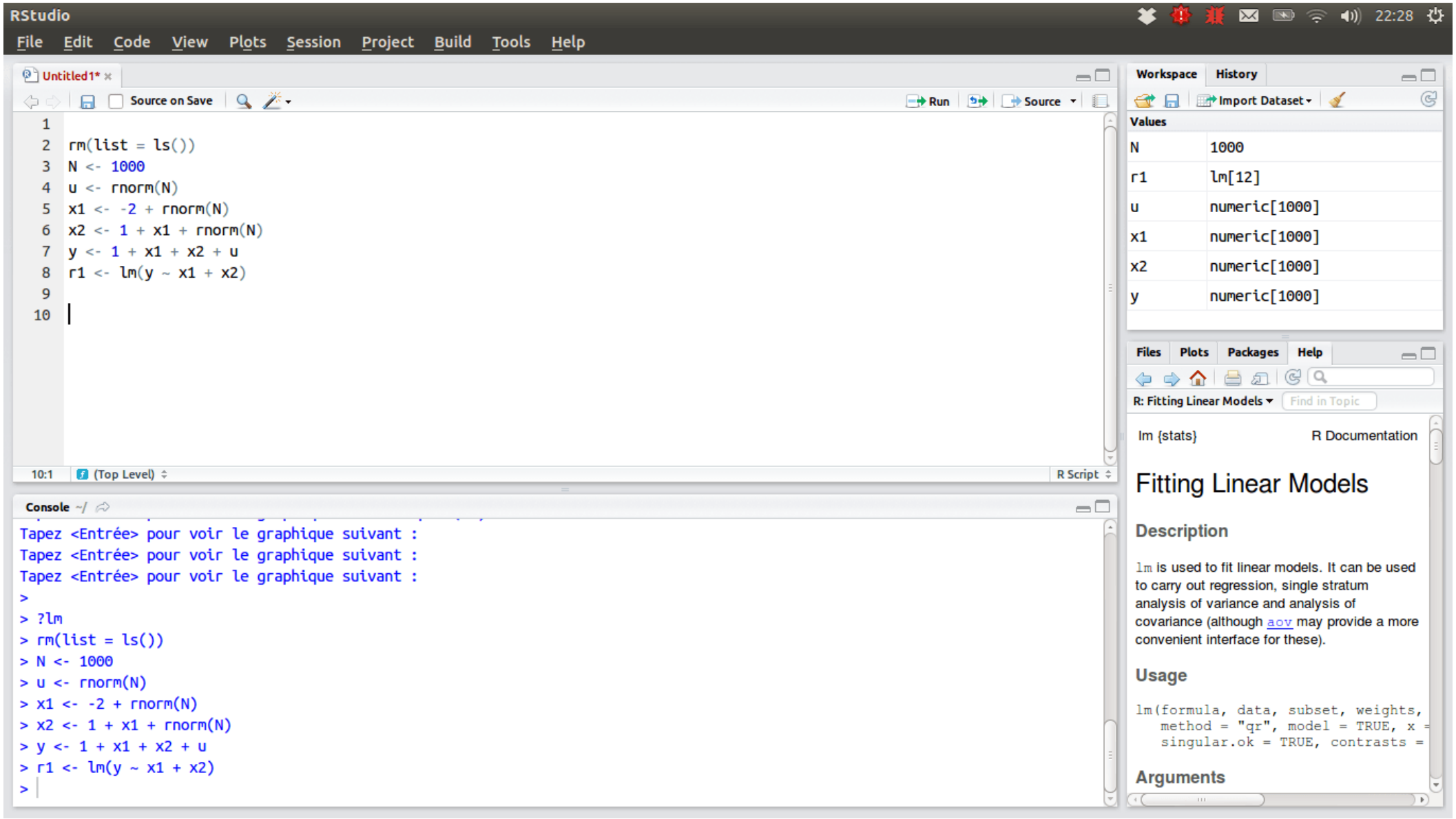
1. 데이터 획득 및 처리
    - 데이터 마이닝을 위한 분류 : 다양한 파생 변수 산출
    - 데이터 전처리 및 정형화 : 결측치 및 이상치 처리 -> 데이터 프레임 생성
2. 시각화
    - 그래프
    - 공간분석(GIS) : 공간과 관련된 속성들을 시각화
3. 데이터 분석
    - 탐색적 자료 분석(EDA) : 다양한 차원과 값을 조합해 특이점, 데이터의 구조적 관계를 알아내는 기법
    - 통계분석 : 기술통계, 추측(추론)통계
    - 데이터마이닝 : 데이터부터 유용한 지식을 추출하는 분석 방법

R언어

R 언어

- ❖ 오픈소스 프로그램, 통계, 데이터마이닝, 그래프를 위한 언어
- ❖ 주로 R Studio 프로그램을 이용
- ❖ 파이썬 코드와 매우 유사함

R 언어 vs 파이썬


RStudio
File Edit Code View Plots Session Project Build Tools Help
22:28
Untitled1*
Source on Save
Run
Source
1 rm(list = ls())
2 N <- 1000
3 u <- rnorm(N)
4 x1 <- -2 + rnorm(N)
5 x2 <- 1 + x1 + rnorm(N)
6 y <- 1 + x1 + x2 + u
7 r1 <- lm(y ~ x1 + x2)
10:1 (Top Level) R Script
Console ~/
Tapez <Entrée> pour voir le graphique suivant :
Tapez <Entrée> pour voir le graphique suivant :
Tapez <Entrée> pour voir le graphique suivant :
> ?lm
> rm(list = ls())
> N <- 1000
> u <- rnorm(N)
> x1 <- -2 + rnorm(N)
> x2 <- 1 + x1 + rnorm(N)
> y <- 1 + x1 + x2 + u
> r1 <- lm(y ~ x1 + x2)
Workspace History
Import Dataset
Values
N 1000
r1 lm[12]
u numeric[1000]
x1 numeric[1000]
x2 numeric[1000]
y numeric[1000]
Files Plots Packages Help
R: Fitting Linear Models
Find in Topic
lm {stats} R Documentation
Fitting Linear Models
Description
lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).
Usage
lm(formula, data, subset, weights,
method = "qr", model = TRUE, x =
singular.ok = TRUE, contrasts =
Arguments

데이터 형식과 사칙연산

❖ 숫자

```
10

[출력] 10
```

❖ 문자

```
"홍길동"

[출력] "홍길동"
```

❖ 논리값

```
TRUE

[출력] TRUE
```

❖ 사칙연산

```
2+2

[출력] 4
```

```
2-2

[출력] 0
```

```
2*2

[출력] 4
```

```
2/2

[출력] 1
```

❖ **몫**

```
2 %/% 2

[출력] 1
```

❖ **나머지**

```
2 %% 2

[출력] 0
```

❖ **거듭제곱**

```
2 ^ 2 ( 2**2 )

[출력] 4
```

R 기초문법

❖ 출력문

```
print(10)

[출력] 10
```

❖ 대입연산자(<-, =)

```
a <- 10
print(a)

[출력] 10
```

❖ 변수 목록보기(ls(), ls.str())

```
ls()

[출력] “a” “b”
```

❖ 변수 제거하기(rm())

```
rm(b)
b

[출력] Error in eval(expr, envir, enclos): 객체
       'b'를 찾을 수 없습니다
```

❖ 함수 생성하기

```
function( 매개변수 ) {

   a = 10        # 지역변수
   b <<- 20      # 전역변수
}
```

R 수학 함수

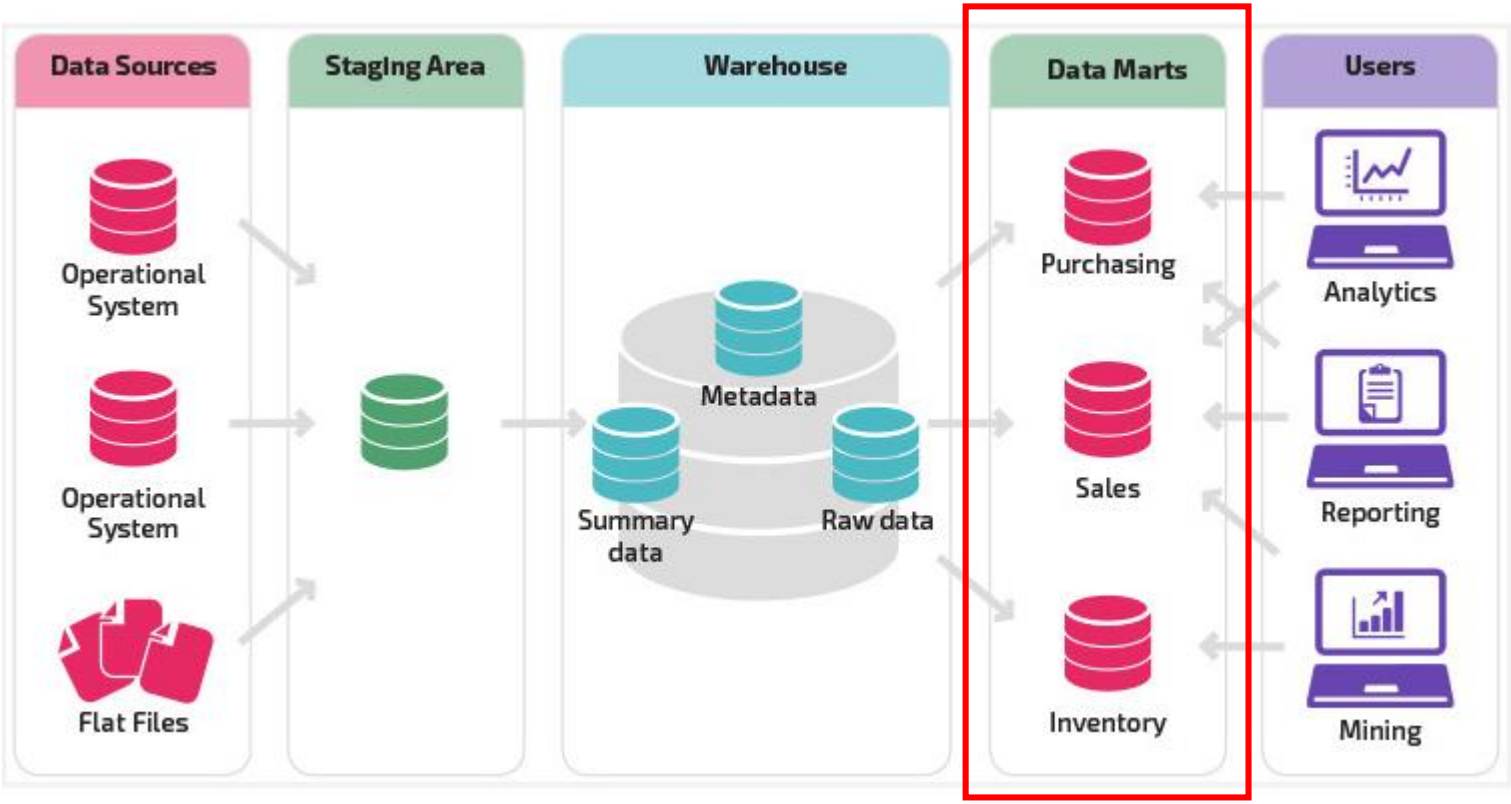
함수	기능
ceiling(x)	x의 값을 정수로 올림
floor(x)	x의 값을 정수로 내림
trunc(x)	x의 값의 소수점 자리를 버림
round(x)	x의 값을 반올림
round(x, digits=n)	x의 값을 소수점 n자리에서 반올림
sqrt(x)	x의 제곱근
exp(x)	x의 지수함수 값
log(x)	자연대수를 밑으로 하는 로그 값
log(x, base=a)	a를 밑으로 하는 로그 값
sin(x), cos(x), tan(x)	x의 삼각함수의 값
factorial(n)	n!=1×2×⋯×nn!=1×2×⋯×n
choose(n,k)	n 개 중 k를 뽑는 조합의 수

R 통계 함수

함수	기능
mean(x)	변수의 평균 산출
sum(x)	변수의 합계 산출
median(x)	변수의 중앙값 산출
max(x)	변수의 최대값 산출
min(x)	변수의 최소값 산출
log(x)	변수의 로그값 산출
sd(x)	변수의 표준편차 산출
var(x)	변수의 분산 산출
cov(x1, x2)	변수간 공분산 산출
cor(x1, x2)	변수간 상관계수 산출
length(x)	변수의 길이 산출
log(x)	변수의 로그값 산출

데이터 마트

- 데이터 웨어하우스와 사용자 사이의 중간층에 위치한 것
- 데이터 웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 함



출처 : https://images.app.goo.gl/ZuUvDPWRzxrocSLXA

통계분석

통계 조사 방법

❖ 전수조사 : 대상 집단 전체를 조사하는 것

↳ 시간과 비용이 많이 소요.

❖ 표본조사 : 모집단의 일부를 추출하여 분석

↳ 시간과 비용 절감

↳ 정확도 떨어질 수 있음

표본 추출 방법

❖ 단순랜덤 추출법 : 임의로 n개의 표본 추출

❖ 계통추출법 : 샘플에 번호를 부여한 후 일정한 간격별로 추출

❖ 집락추출법 : 군집을 나눈 후 군집별로 단순랜덤 추출법 적용

❖ 층화추출법 : 모집단을 계층으로 나누어, 계층을 대표하는 표본 추출

↳ 유사한 원소끼리 몇 개의 층을 나누어, 층 별로 랜덤 추출

측정 방법

❖ 명목척도 : 어느 집단에 속하는지 분류 (성별/출생지 등)

❖ 순서척도 : 측정 대상에 서열관계 관측 (만족도/학년/등수 등)

❖ 구간척도 : 속성의 양을 측정하는 것

↳ 구간이나 구간 사이의 간격이 의미가 있는 자료 (온도, 지수 등)

❖ 비율척도 : 간격에 대한 비율.

↳ 절대적 기준이 존재하고 사칙연산이 가능 (무게/ 나이 등등)

확률변수와 확률분포

❖ 확률변수 : 특정 사건이 일어날 가능성이 확률적으로 나타나는 변수

↳ E(X) = X라는 사건이 발생할 확률

❖ 이산형 : 0이 아닌 확률 값을 갖는 확률변수

↳ 베르누이 확률분포 · 이항분포 · 기하분포 · 다항분포 · 포아송분포

동전 던지기! 동전 던지기 n회 시공간

❖ 연속형 : 함수형태로 연속적으로 주어진 확률변수(실수 전체)

↳균일분포(일양분포) · 정규분포 · 지수분포 · t-분포 · 카이제곱-분포(x^2-분포)

· F-분포

추정

❖ 추정 : 표본으로부터 미지의 모수(모집단의 특징)을 추측

❖ 점추정 : 모수가 특정한 값일 것 이라고 추정

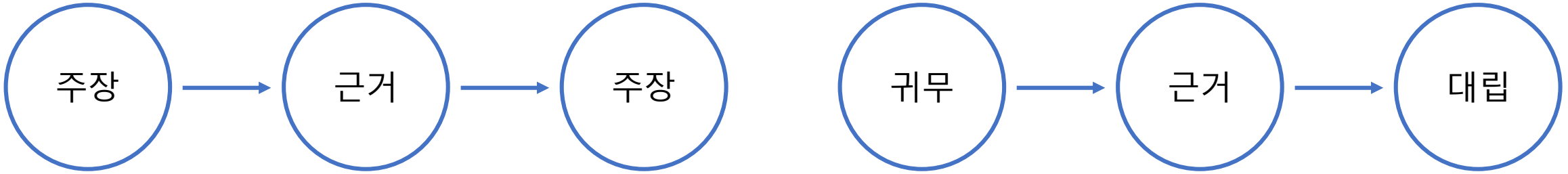
↳ 표본의 평균, 중위수, 최빈값 등을 사용

❖ 구간추정 : 모수가 특정한 구간에 있을 것이라고 선언

↳ 추정량의 분포에 대한 전제,구해진 구간 안에 모수가 있을 신뢰수준(가능성)

이 주어져야 함

가설검정



❖ 귀무가설 : '비교하는 값과 차이가 없다'를 기본개념으로 하는 가설

↳ ex) 용의자는 범인이 아니다.

❖ 대립가설 : 뚜렷한 증거가 있을 때 주장하는 가설 (채택되는 것이 목표)

↳ ex) 용의자는 범인이다.

가설검정

❖ 유의확률(p-value) : 귀무가설이 사실일 때, 관측된 통계량이 대립가설을 지지할 확률

↳ 어떤 사건이 우연히 발생할 확률

❖ 유의수준 : 귀무가설을 기각하게 되는 확률의 크기

↳ 우연히 발생하기 어렵다고 판단하는 기준 (0.05 = 5%)

❖ 제1종 오류 : 귀무가설이 옳은데 귀무가설을 기각하게 되는 오류

❖ 제2종 오류 : 귀무가설이 틀린데 귀무가설을 채택하는 오류

모수적 방법 & 비모수적 방법

❖ 모수적 방법

↳ 모집단의 분포에 대한 가정을 진행하고, 검정 실시

❖ 비모수적 방법

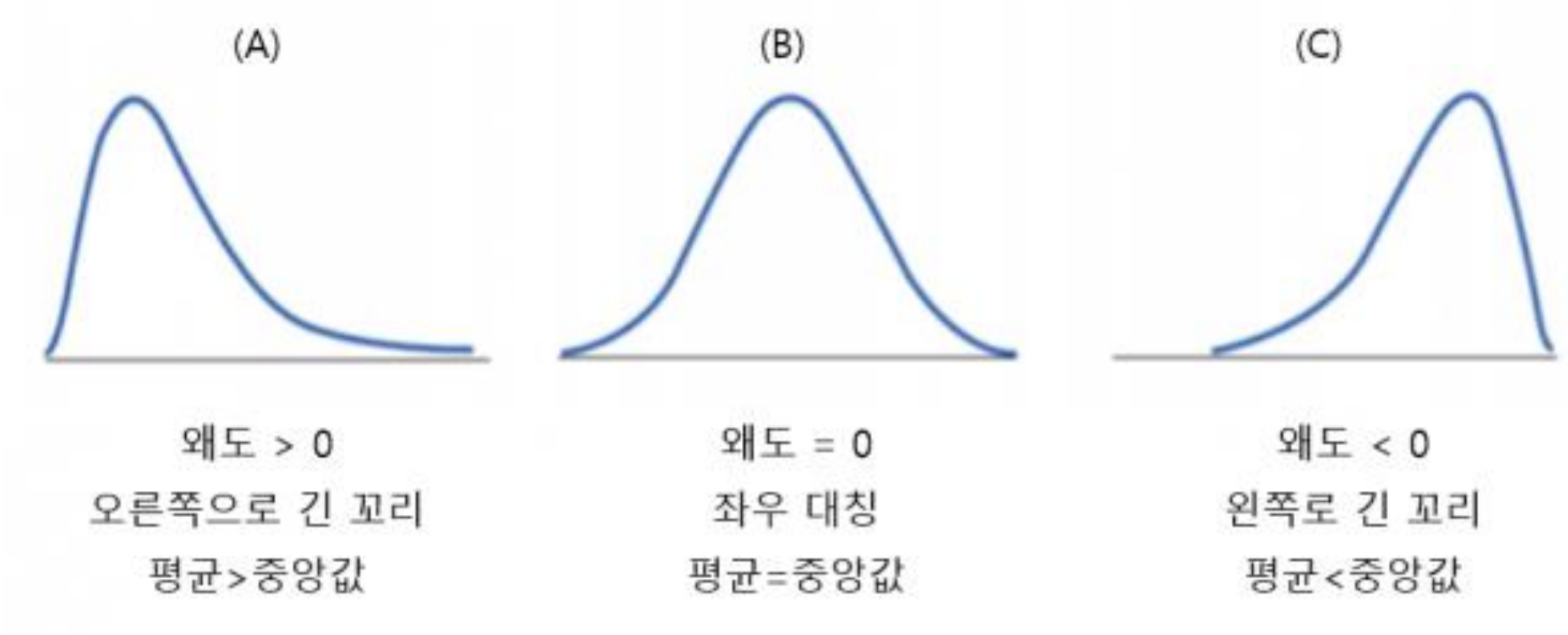
↳ 자료가 추출된 모집단의 분포에 아무 제약 않고 검정 실시

↳ 자료가 많지 않거나, 자료가 개체 간의 서열관계를 나타내는 경우에 사용

↳ 부호검정, 윌콕슨의 순위합 검정, 윌콕슨의 부호 순위 검정, 맨-휘트니의 U 검정, 런 검정, 스피어만의 순위상관계

왜도 값에 따른 분포

❖ 왜도 : 분포의 비대칭 정도를 나타내는 측도



출처 : https://images.app.goo.gl/5jzEnSYyGy9tizkZA

상관분석

❖ 데이터 안의 두 변수 간의 관계 정도를 알아보기 위한 분석 방법

상관 계수 범위	해설
0.7 < r <= 1	강한 양의 상관
0.3 < r <= 0.7	약한 양의 상관
0 < r <= 0.3	거의 상관 없음
r = 0	상관관계가 전혀 없음
-0.3 <= r < 0	거의 상관 없음
-0.7 <= r < -0.3	약한 음의 상관
-1 <= r < -0.7	강한 음의 상관

❖ p-value값이 0.05이하인 경우 변수 간 상관관계가 있다고 볼 수 있다.

↳ 우연히 발생하기 어렵다 = 두 변수 사이에 상관관계가 있다 = 통계적으로 유의미하다

회귀분석

❖ 하나 또는 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정하는 통계기법

❖ 선형회귀분석의 가정

선형성	입력변수와 출력변수의 관계가 선형이다.
등분산성	오차의 분산이 입력변수와 무관하게 일정하다.
독립성	입력변수와 오차는 관련이 없다.
비상관성	오차들끼리 상관이 없다.
정상성 (정규성)	오차의 분포가 정규분포를 따른다.

❖ **정**상성**성(선)**형성 **비**상관성**등**분산성한 **독립**성

❖ 데이터의 정상성(정규성)은 Q-Q plot, Shapiro-Wilks 검정 등을 사용해 확인

최적회귀방정식 변수 선택법

❖ 모든 후보 모형들에 대해 AIC 또는 BIC를 계산하고 그 값이 최소가 되는 모형 선택 -> 변수 개수가 적고, 확률이 높은 모델일 수록 AIC, BIC가 작음

❖ 전진선택법 : 절편만 있는 상수모형으로부터 시작해 중요하다고 생각되는 변수를 차례로 모형에 추가

❖ 후진제거법 : 모든 변수를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거

❖ 단계선택법 : 전진선택법으로 변수를 추가, 기존 변수가 영향을 받아 중요도가 약화되면 해당 변수를 다시 제거하는 등 단계별로 추가 및 제거 여부를 검토

시계열 분석

❖ 시간의 흐름에 따라 관찰된 값

❖ 정상시계열 (모든 시점에 대해 일정한 평균과 분산)

↳ 특정한 시차의 길이를 갖는 자기공분산을 측정하더라도 동일한 값을 갖는다.

❖ 비정상시계열을 정상시계열로 전환하는 방법

↳ 평균이 일정하지 않은 경우: 차분을 통해 정상화(현 시점에서 바로 전 시점의 자료 값을 뺌)

↳ 분산이 일정하지 않은 경우: 변환을 통해 정상화(자연로그를 취함)

시계열 모형

❖ 자기회귀 모형(AR 모형) : 이전 시점의 자료가 현재 자료에 영향을 줄 때

↳ 자기 자신의 이전 값에 가중치를 두어 현재의 상태를 표시

↳ 현재상태 = $\sum_{0}^{t-1} z_t \emptyset_t$

❖ 이동평균 모형(MA 모형)

↳ AR 모형과 반대로 자기 자신의 이전 값에 백색잡음의 결합으로 현재의 상태를 표시

↳ 현재상태 = $z_t - \sum_{0}^{t-1} z_t \theta_t$

시계열 모형

❖ 자기회귀누적이동평균 모형(ARIMA 모형)

↳AR 모형과 MA 모형의 결합

❖ 분해시계열

↳ 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

↳ 경향(추세)요인, 계절요인, 순환요인, 불규칙요인으로 이루어짐

다차원척도법(MDS)

❖ 객체간 근접성을 시각화하는 통계기법

❖ 개체들을 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사이의 집단화를 시각적으로 표현하는 분석방법

↳ 패턴 및 구조 찾기, 차원 축소에 용이

❖ 계량적 MDS → 데이터가 구간척도, 비율척도인 경우 활용

❖ 비계량적 MDS → 데이터가 순서척도인 경우 활용

주성분분석(PCA)

❖ 여러 변수들의 상관관계를 이용해 주성분이 높은 변수들을 요약 및 축소하여 소수의 주성분으로 차원을 축소하는 것

❖ 주성분분석 결과의 누적기여율이 85% 이상이면 주성분의 수로 결정할 수 있음

❖ Scree plot을 활용하여 고윳값이 수평을 유지하기 전단계로 주성분의 수 선택

문제풀이

3-1과목 문제풀이 정답

1	③	11	①	21	④	31	④	41	④	51	①	61	②	71	④
2	④	12	②	22	②	32	④	42	①	52	③	62	①	72	④
3	④	13	③	23	④	33	①	43	①	53	③	63	③	73	③
4	④	14	②	24	④	34	④	44	④	54	④	64	④	74	③
5	④	15	③	25	③	35	③	45	③	55	④	65	④	75	③
6	③	16	③	26	②	36	④	46	①	56	②	66	④		
7	②	17	②	27	①	37	④	47	②	57	③	67	③		
8	③	18	③	28	①	38	④	48	④	58	③	68	③		
9	②	19	①	29	①	39	②	49	②	59	③	69	④		
10	①	20	③	30	④	40	③	50	④	60	③	70	③		