Hanwha Ocean

데이터 분석을 위한 기초 통계 교육

강의자 : 이승한 (연세대학교)

일시 : 2023.07.19 (수)

Hanwha Ocean

Contents

1. 데이터 분석의 중요성

Hanwha Ocean

1. 데이터 분석의 중요성

빅데이터 (Big Data) 시대

- **데이터(Data)**의 양이 **방대(Big)**해지고 있음.
  - 일상 속 수 많은 곳을 통해서 방대한 데이터들은 수집이 되고 있음.
    - 예시) 컴퓨터, 스마트폰, 공장 내의 기계 등…
- 이러한 데이터를 기반으로 의사 결정을 효과적으로 수행할 수 있음

https://cdn.epnc.co.kr/news/photo/201907/91021_81259_3048.jpg

Hanwha Ocean

1. 데이터 분석의 중요성

빅데이터 (Big Data) 시대

- **데이터(Data)**의 양이 **방대(Big)**해지고 있음.
  - 일상 속 수 많은 곳을 통해서 방대한 데이터들은 수집이 되고 있음.
    - 예시) 컴퓨터, 스마트폰, 공장 내의 기계 등…
- 이러한 데이터를 기반으로 의사 결정을 효과적으로 수행할 수 있음

따라서, **주어진 데이터를 올바르게 분석하고 활용할 수 있는 능력**이 매우 중요해졌다!

https://cdn.epnc.co.kr/news/photo/201907/91021_81259_3048.jpg

Hanwha Ocean

1. 데이터 분석의 중요성

데이터의 종류

- 데이터는 **규모**에 있어서 뿐만이 아니라, **종류**에 있어서도 점점 다양해지고 있다.

https://blog.kakaocdn.net/dn/tBTGb/btrp96nSjDF/mY9GnqxqHBOQJDuoWDE2kk/img.jpg
https://m.segye.com/content/image/2022/05/23/20220523519355.jpg
https://www.danbinews.com/news/photo/202211/21976_31774_4546.jpg

Hanwha Ocean

1. 데이터 분석의 중요성

데이터의 종류

- 데이터는 **규모**에 있어서 뿐만이 아니라, **종류**에 있어서도 점점 다양해지고 있다.

정형 데이터

	A	B	C	D	E	F	G	H
1	일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
2	20180601	금	0	음식점-족	서울특별시	강남구	논현동	5
3	20180601	금	0	음식점-족	서울특별시	강동구	길동	5
4	20180601	금	0	음식점-족	서울특별시	강서구	내발산동	5
5	20180601	금	0	음식점-족	서울특별시	동대문구	제기동	5
6	20180601	금	0	음식점-족	서울특별시	서대문구	창천동	7
7	20180601	금	0	음식점-족	서울특별시	서초구	양재동	5
8	20180601	금	0	음식점-족	서울특별시	성동구	성수동2가	5
9	20180601	금	0	음식점-족	서울특별시	성북구	동선동2가	5
10	20180601	금	0	음식점-족	서울특별시	송파구	송파동	5
11	20180601	금	0	음식점-족	서울특별시	영등포구	문래동3가	5

테이블 (Table) 데이터

https://blog.kakaocdn.net/dn/tBTGb/btrp96nSjDF/mY9GnqxqHBOQJDuoWDE2kk/img.jpg
https://m.segye.com/content/image/2022/05/23/20220523519355.jpg
https://www.danbinews.com/news/photo/202211/21976_31774_4546.jpg

Hanwha Ocean

1. 데이터 분석의 중요성

데이터의 종류

- 데이터는 **규모**에 있어서 뿐만이 아니라, **종류**에 있어서도 점점 다양해지고 있다.

정형 데이터

	A	B	C	D	E	F	G	H
1	일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
2	20180601	금	0	음식점-족	서울특별시	강남구	논현동	5
3	20180601	금	0	음식점-족	서울특별시	강동구	길동	5
4	20180601	금	0	음식점-족	서울특별시	강서구	내발산동	5
5	20180601	금	0	음식점-족	서울특별시	동대문구	제기동	5
6	20180601	금	0	음식점-족	서울특별시	서대문구	창천동	7
7	20180601	금	0	음식점-족	서울특별시	서초구	양재동	5
8	20180601	금	0	음식점-족	서울특별시	성동구	성수동2가	5
9	20180601	금	0	음식점-족	서울특별시	성북구	동선동2가	5
10	20180601	금	0	음식점-족	서울특별시	송파구	송파동	5
11	20180601	금	0	음식점-족	서울특별시	영등포구	문래동3가	5

테이블 (Table) 데이터

비정형 데이터

이미지 (Image) 데이터

★★★★★ 5

배송 빨라요 | 성능 성능이 뛰어나요

뭔가 모르게 강아지 냄새가 나는거 같아서 탈취샴푸 구매해봤어요.
몇일간 꾸준히 사용해보니깐 조금씩 냄새가 사라지는거 같더라구요.
식물성분기반의 성분이라서 그런가 샴푸를 하면서도 강아지가 큰 거부감 없이 피부에 자극없이 잘 씻는데,
미국내 탈취제제품 전문회사인 [redacted] 사에서 만든 탈취샴푸라서 그런가..ㅎㅎ
스컹크 냄새까지 없애준다는 샴푸라고 하니 뭐 아주아주 좋은!
산뜻하니 레몬향으로 향이 좋더라구요.…

더보기

꾹 0 이 리뷰가 도움이 되었다면 꾹!

텍스트 (Text) 데이터

https://blog.kakaocdn.net/dn/tBTGb/btrp96nSjDF/mY9GnqxqHBOQJDuoWDE2kk/img.jpg
https://m.segye.com/content/image/2022/05/23/20220523519355.jpg
https://www.danbinews.com/news/photo/202211/21976_31774_4546.jpg

Hanwha Ocean

1. 데이터 분석의 중요성

데이터 분석을 위해 필요한 가장 기본적인 지식 = **“통계학”**

통계에 대한 기초적인 지식을 쌓음으로써,

주어진 **데이터에 대한 이해**를 할 수 있고,

더 나아가서 **의사결정에 도움**을 줄 수 있다!

Hanwha Ocean

1. 데이터 분석의 중요성

데이터 분석의 사례)

대형 마트에 “맥주”와 “기저귀”가 가까이 진열되어 있는 이유?

https://hr.wanted.co.kr/insights/pa_text/
https://dimg.donga.com/wps/NEWS/IMAGE/2008/06/14/7067261.1.jpg

Hanwha Ocean

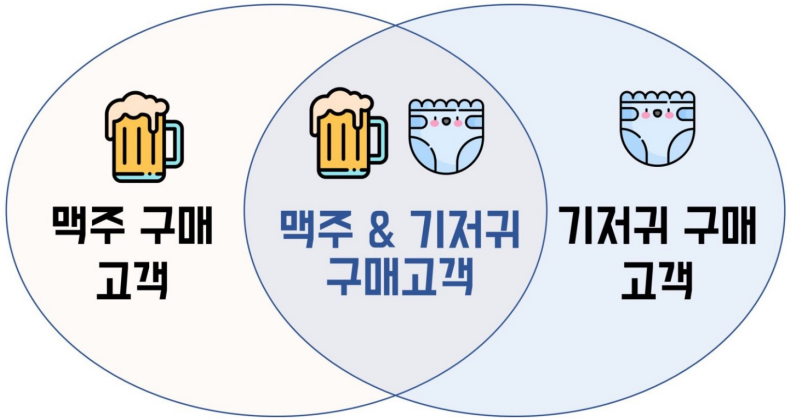
1. 데이터 분석의 중요성

데이터 분석의 사례)

대형 마트에 “맥주”와 “기저귀”가 가까이 진열되어 있는 이유?

- 맥주 구매자와 기저귀 구매자의 **소비 패턴 분석**
- 그 결과, 맥주와 기저귀를 함께 구매한 사람이 많은 것으로 나타남

이 둘을 가까운 진열대에 놓음으로써, 매출 증진 효과!



https://hr.wanted.co.kr/insights/pa_text/
https://dimg.donga.com/wps/NEWS/IMAGE/2008/06/14/7067261.1.jpg

2. 데이터의 형태

Hanwha Ocean

2. 데이터의 형태

데이터의 유형(형태)는 크게 2가지로 나뉜다

1. **범주형 (categorical)** : 몇 개의 **범주**로 나누어진 자료

   - 예시) 성별 (남/녀), 과일(바나나/사과/딸기)

Hanwha Ocean

2. 데이터의 형태

데이터의 유형(형태)는 크게 2가지로 나뉜다

1. **범주형 (categorical)** : 몇 개의 **범주**로 나누어진 자료

   - 예시) 성별 (남/녀), 과일(바나나/사과/딸기)

2. **수치형 (numerical)** : 숫자로 구성된 자료

   - 예시) 키 (170.6, 180.5, 165.5 ..), 사과의 개수 (15,20,17 ..)

Hanwha Ocean

2. 데이터의 형태

1. 범주형 (categorical) 데이터

범주형 데이터는 2가지로 구분된다.

(1) 명목형 : 단순히 분류된 자료 (순서 X)

- 예시) 과일 (사과, 바나나, 오렌지, 딸기), 혈액형 (A, B, AB, O)

Hanwha Ocean

2. 데이터의 형태

1. 범주형 (categorical) 데이터

범주형 데이터는 2가지로 구분된다.

(1) 명목형 : 단순히 분류된 자료 (순서 X)

- 예시) 과일 (사과, 바나나, 오렌지, 딸기), 혈액형 (A, B, AB, O)

(2) 순서형 : 순서가 존재하는 자료 (순서 O)

- 예시) 등급(상 > 중 > 하), 만족도(매우 만족 > 만족 > 보통 > 불만족 > 매우 불만족)

Hanwha Ocean

2. 데이터의 형태

2. 순서형 데이터

순서형 데이터는 2가지로 구분된다.

(1) 이산형 : 이산적인 값을 갖는 데이터

- 예시) 가족 구성원 수 (4,1,2,3 .. .)

Hanwha Ocean

2. 데이터의 형태

2. 순서형 데이터

순서형 데이터는 2가지로 구분된다.

(1) 이산형 : 이산적인 값을 갖는 데이터

- 예시) 가족 구성원 수 (4,1,2,3 .. .)

(2) 연속형 : 연속적인 값을 갖는 데이터

- 예시) 몸무게 (60.5kg, 71.3kg, …)

Hanwha Ocean

2. 데이터의 형태

Summary

- 범주형 : 몇 개의 범주로 나누어진 자료를 의미
  - 명목형 : 성별, 성공여부, 혈액형 등 단순히 분류된 자료
  - 순서형 : 개개의 값들이 이산적이며 그들 사이에 순서 관계가 존재하는 자료
- 수치형 : 이산형과 연속형으로 이루어진 자료를 의미
  - 이산형 : 이산적인 값을 갖는 데이터로 출산횟수 등을 의미
  - 연속형 : 연속적인 값을 갖는 데이터로 신장, 체중 등을 의미

Hanwha Ocean

2. 데이터의 형태

- 범주형 : 몇 개의 범주로 나누어진 자료를 의미
  - 명목형 : 성별, 성공여부, 혈액형 등 단순히 분류된 자료
  - 순서형 : 개개의 값들이 이산적이며 그들 사이에 순서 관계가 존재하는 자료
- 수치형 : 이산형과 연속형으로 이루어진 자료를 의미
  - 이산형 : 이산적인 값을 갖는 데이터로 출산횟수 등을 의미
  - 연속형 : 연속적인 값을 갖는 데이터로 신장, 체중 등을 의미

Quiz

Q. 만족도

매우 불만 (1) — 불만 (2) — 보통 (3) — 만족 (4) — 매우 만족 (5)

값이 커짐에 따라 만족도가 좋아지고 있다고 할 수 있는 경우 = ? 데이터!

EX) 만족도 조사 / 병의 단계 / 성적표

Q. 하루동안 은행에 방문하는 고객의 수

1명, 2명, 3명

3.5명 ?
소수점의 의미가 없음

수치적인 의미는 가지고 있으나 소수점으로 표현되지 않는 경우 = ? 데이터!

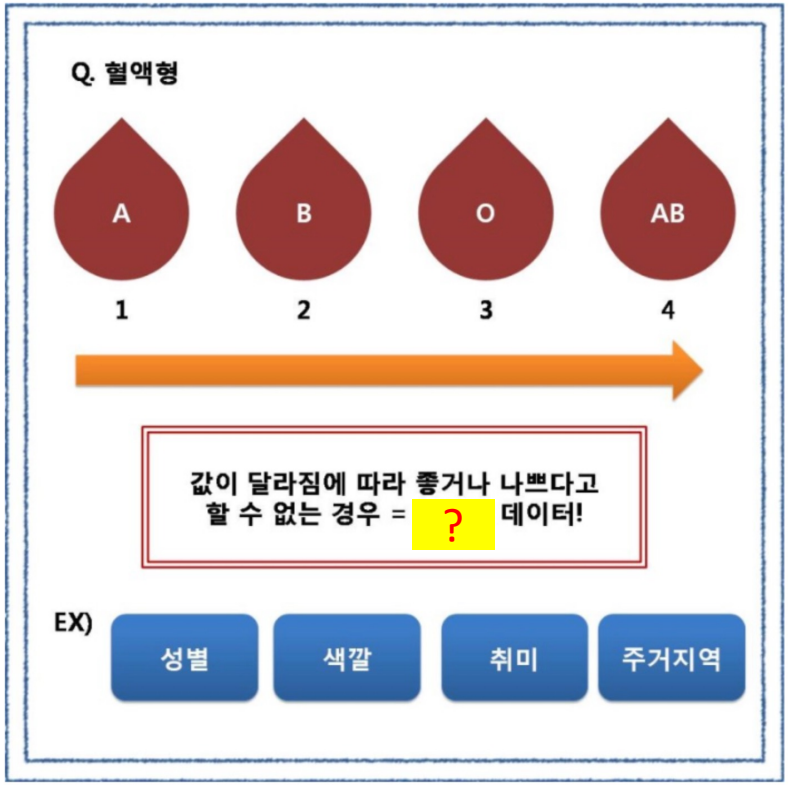
EX) 불량품 수 / 사고 건수 / 차량대수

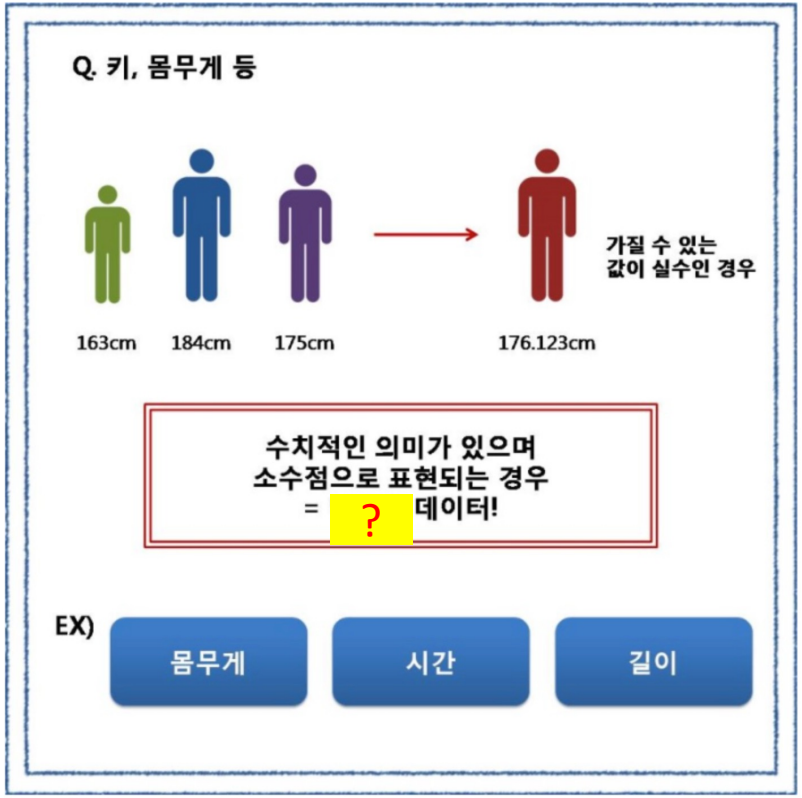
Hanwha Ocean

2. 데이터의 형태

- 범주형 : 몇 개의 범주로 나누어진 자료를 의미
  - 명목형 : 성별, 성공여부, 혈액형 등 단순히 분류된 자료
  - 순서형 : 개개의 값들이 이산적이며 그들 사이에 순서 관계가 존재하는 자료
- 수치형 : 이산형과 연속형으로 이루어진 자료를 의미
  - 이산형 : 이산적인 값을 갖는 데이터로 출산횟수 등을 의미
  - 연속형 : 연속적인 값을 갖는 데이터로 신장, 체중 등을 의미

Quiz





3. 통계학이란?

Hanwha Ocean

3. 통계학이란?

통계학 (Statistics) 이란?

데이터를 **수집, 분석, 및 해석**하는 학문

이를 통해, 현상을 이해하고 예측하는 데 도움을 얻을 수 있다!

Hanwha Ocean

3. 통계학이란?

통계학 (Statistics) 이란?

데이터를 **수집, 분석, 및 해석**하는 학문

이를 통해, 현상을 이해하고 예측하는 데 도움을 얻을 수 있다!

통계학의 두 분야

1. 기술 통계학
2. 추론 통계학

Hanwha Ocean

3. 통계학이란?

(1) 기술 통계학

데이터를 요약하고 설명하는 데 집중

예시) 한 학급에 있는 학생들의 몸무게의 평균은?

Hanwha Ocean

3. 통계학이란?

(1) 기술 통계학

데이터를 요약하고 설명하는 데 집중

예시) 한 학급에 있는 학생들의 몸무게의 평균은?

(2) 추론 통계학

데이터를 바탕으로, **미지의 집단에 대한 결론을 추론하고 예측**

예시) 100명의 사람을 대상으로, 특정 식품을 먹게 한 뒤 건강 증진 여부를 파악함.

과연, 이 식품은 전체 사람에게 건강에 도움을 주는 음식이라 할 수 있을까?

4. 통계학의 기초

Hanwha Ocean

4. 통계학의 기초

1. (정형) 데이터의 형태
2. 데이터의 요약값
3. 모집단과 표본
4. 확률
5. 조건부 확률
6. 사건의 독립과 종속

Hanwha Ocean

4. 통계학의 기초

1. (정형) 데이터의 형태

- 테이블(Table) 형식의 데이터

정형 데이터

	A	B	C	D	E	F	G	H
1	일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
2	20180601	금	0	음식점-족	서울특별시	강남구	논현동	5
3	20180601	금	0	음식점-족	서울특별시	강동구	길동	5
4	20180601	금	0	음식점-족	서울특별시	강서구	내발산동	5
5	20180601	금	0	음식점-족	서울특별시	동대문구	제기동	5
6	20180601	금	0	음식점-족	서울특별시	서대문구	창천동	7
7	20180601	금	0	음식점-족	서울특별시	서초구	양재동	5
8	20180601	금	0	음식점-족	서울특별시	성동구	성수동2가	5
9	20180601	금	0	음식점-족	서울특별시	성북구	동선동2가	5
10	20180601	금	0	음식점-족	서울특별시	송파구	송파동	5
11	20180601	금	0	음식점-족	서울특별시	영등포구	문래동3가	5

테이블 (Table) 데이터

https://blog.kakaocdn.net/dn/tBTGb/btrp96nSjDF/mY9GnqxqHBOQJDuoWDE2kk/img.jpg

Hanwha Ocean

4. 통계학의 기초

1. (정형) 데이터의 형태

- 테이블(Table) 형식의 데이터
- 테이블 데이터의 구성
  - **행 (row) : 각각의 데이터를 의미**
  - **열 (column) : 각각의 특징을 의미**

정형 데이터

	A	B	C	D	E	F	G	H
1	일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
2	20180601	금	0	음식점-족	서울특별시	강남구	논현동	5
3	20180601	금	0	음식점-족	서울특별시	강동구	길동	5
4	20180601	금	0	음식점-족	서울특별시	강서구	내발산동	5
5	20180601	금	0	음식점-족	서울특별시	동대문구	제기동	5
6	20180601	금	0	음식점-족	서울특별시	서대문구	창천동	7
7	20180601	금	0	음식점-족	서울특별시	서초구	양재동	5
8	20180601	금	0	음식점-족	서울특별시	성동구	성수동2가	5
9	20180601	금	0	음식점-족	서울특별시	성북구	동선동2가	5
10	20180601	금	0	음식점-족	서울특별시	송파구	송파동	5
11	20180601	금	0	음식점-족	서울특별시	영등포구	문래동3가	5

테이블 (Table) 데이터

https://blog.kakaocdn.net/dn/tBTGb/btrp96nSjDF/mY9GnqxqHBOQJDuoWDE2kk/img.jpg

Hanwha Ocean

4. 통계학의 기초

1. (정형) 데이터의 형태

- 테이블(Table) 형식의 데이터
- 테이블 데이터의 구성
  - **행 (row) : 각각의 데이터를 의미**
  - **열 (column) : 각각의 특징을 의미**

정형 데이터

	A	B	C	D	E	F	G	H
1	일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
2	20180601	금	0	음식점-족	서울특별시	강남구	논현동	5
3	20180601	금	0	음식점-족	서울특별시	강동구	길동	5
4	20180601	금	0	음식점-족	서울특별시	강서구	내발산동	5
5	20180601	금	0	음식점-족	서울특별시	동대문구	제기동	5
6	20180601	금	0	음식점-족	서울특별시	서대문구	창천동	7
7	20180601	금	0	음식점-족	서울특별시	서초구	양재동	5
8	20180601	금	0	음식점-족	서울특별시	성동구	성수동2가	5
9	20180601	금	0	음식점-족	서울특별시	성북구	동선동2가	5
10	20180601	금	0	음식점-족	서울특별시	송파구	송파동	5
11	20180601	금	0	음식점-족	서울특별시	영등포구	문래동3가	5

테이블 (Table) 데이터

* 열(column)은 **변수(variate)**이라고도 부른다

https://blog.kakaocdn.net/dn/tBTGb/btrp96nSjDF/mY9GnqxqHBOQJDuoWDE2kk/img.jpg

Hanwha Ocean

4. 통계학의 기초

1. (정형) 데이터의 형태

- 테이블(Table) 형식의 데이터
- 테이블 데이터의 구성
  - **행 (row) : 각각의 데이터를 의미**
  - **열 (column) : 각각의 특징을 의미**
- 예시
  - 하나의 **행** = 한 명의 **사람** (ex. 강성수, 이나영, 박상민 등)
  - 하나의 **열** = 하나의 **특징** (ex. 근무팀, 직급)

성명	계열사	근무팀	직급	목표 매출액	달성 매출액
강성수	성공산업	영업1팀	대리	5,500,000	3,000,000
이나영	성공산업	영업2팀	사원	4,000,000	2,200,000
박상민	성공산업	영업3팀	대리	4,000,000	2,800,000
태명우	성공산업	영업1팀	사원	3,000,000	2,200,000
김숙자	성공산업	영업2팀	대리	5,000,000	2,800,000
이규하	성공산업	영업3팀	대리	6,000,000	2,800,000
김민정	성공산업	영업1팀	사원	2,500,000	4,500,000
이명진	성공산업	영업2팀	사원	3,000,000	3,500,000
최민영	성공산업	영업3팀	대리	7,800,000	8,500,000
하상희	성공산업	영업2팀	사원	3,000,000	2,200,000
김신예	성공산업	영업3팀	사원	5,500,000	2,500,000

* 열(column)은 **변수(variate)**이라고도 부른다

https://help.hancom.com/hoffice/webhelp/9.0/ko_kr/hcell/images/7_pivot_table(main)_01.gif

Hanwha Ocean

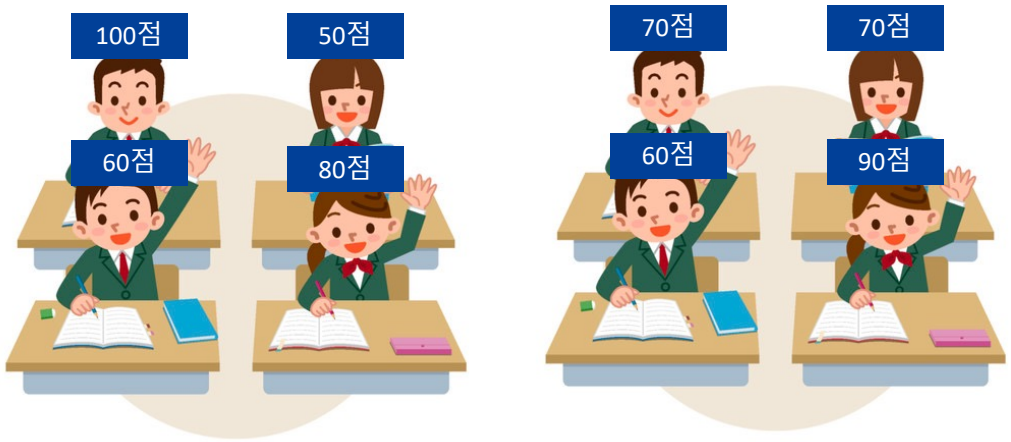
4. 통계학의 기초

2. 데이터의 요약값

데이터 요약의 필요성

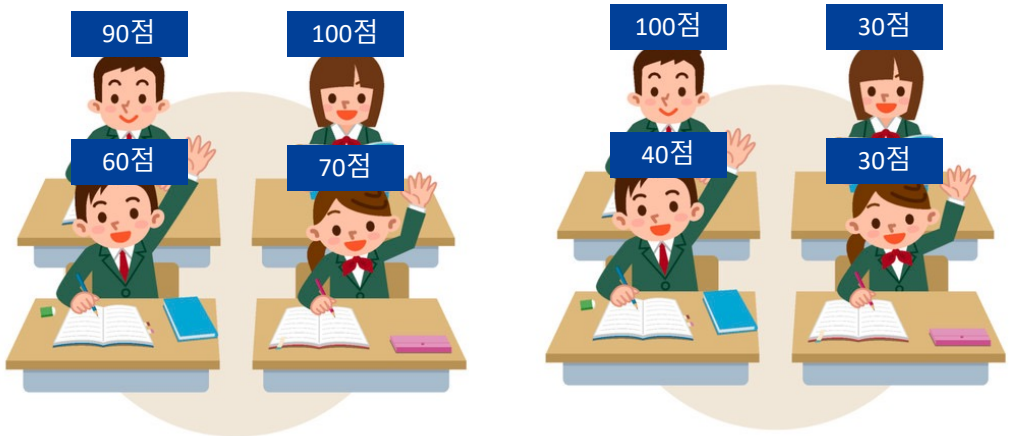
Q. 어느 학급이 더 공부를 잘한다고 할 수 있을까?

학급을 대표할 수 있는 "**하나의 (요약)점수**"를 만들어야 함!



학급 A

VS



학급 B

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 평균 (mean)

(2) 중위값 (median)

(3) 최빈값 (mode)

(4) 사분위수 (quartile)

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 평균 (mean)

총합을 전체 개수로 나눈 값

$$\mu = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Hanwha Ocean

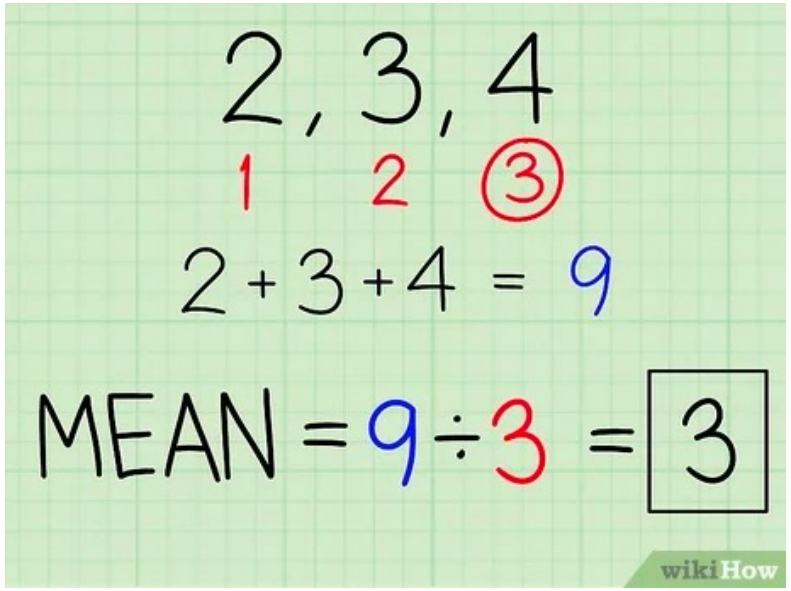
4. 통계학의 기초

2. 데이터의 요약값

(1) 평균 (mean)

총합을 전체 개수로 나눈 값

$$\mu = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$



https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(2) 중위값 (median)

중간 등수에 있는 값

https://www.wikihow.com/Find-Mean,-Median,-and-Mode

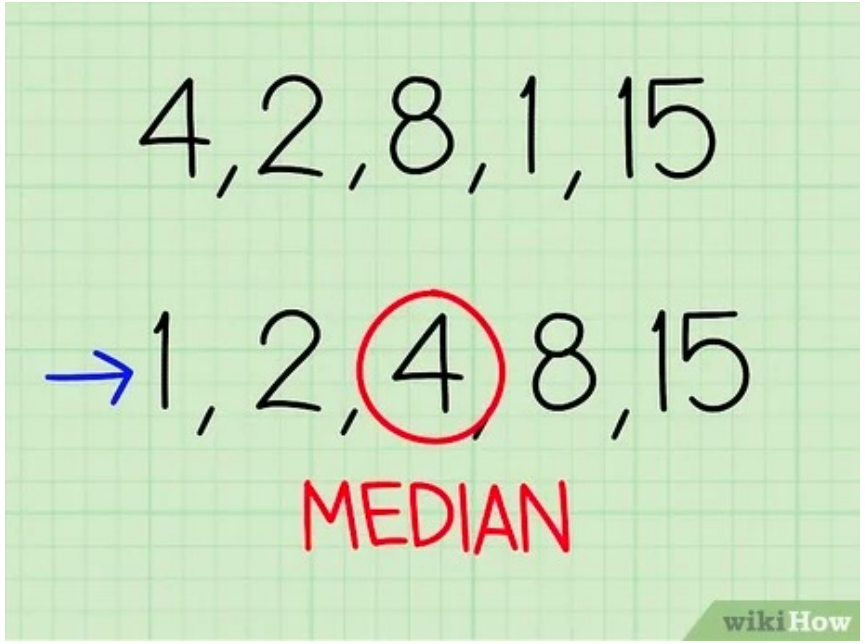
Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(2) 중위값 (median)

중간 등수에 있는 값



https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Hanwha Ocean

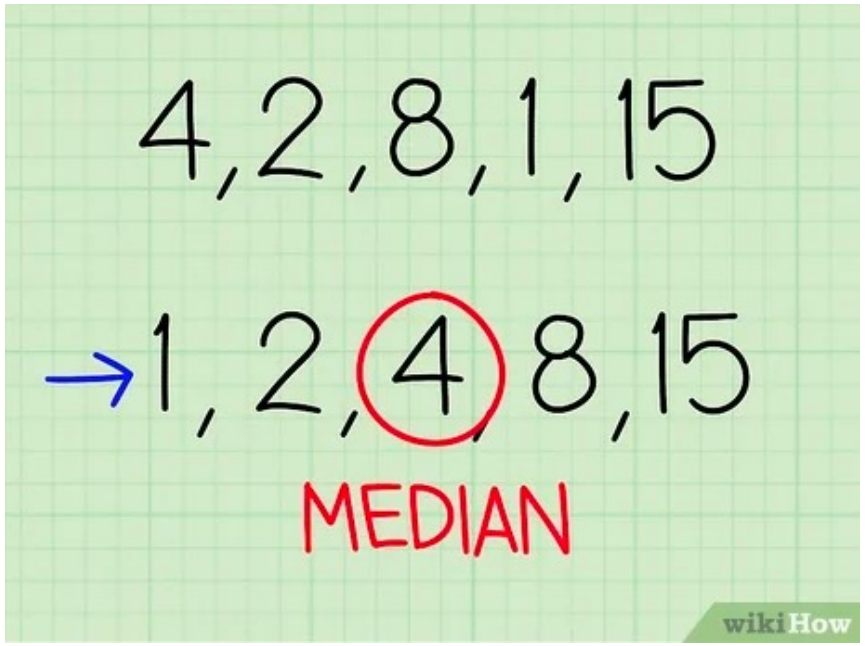
4. 통계학의 기초

2. 데이터의 요약값

(2) 중위값 (median)

중간 등수에 있는 값

*만약, 데이터의 개수가 **홀수 (2n+1) 개** 일 경우?*



https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(2) 중위값 (median)

중간 등수에 있는 값

데이터 개수(n)가

- 홀수인 경우 : (n+1) / 2 번째 값
- 짝수인 경우 : n/2 번째와 (n+2)/2번째 값의 평균

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

https://upload.wikimedia.org/wikipedia/commons/thumb/c/cf/Finding_the_median.png/640px-Finding_the_median.png

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(3) 최빈값 (mode)

가장 자주 등장하는 값

https://www.wikihow.com/Find-Mean,-Median,-and-Mode

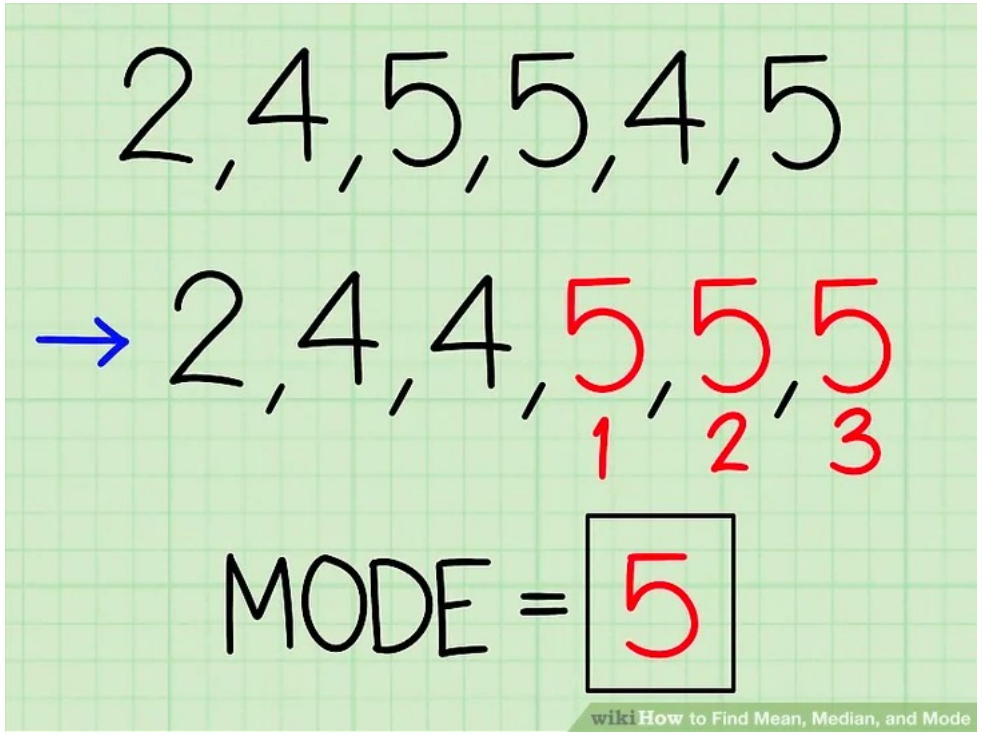
Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(3) 최빈값 (mode)

가장 자주 등장하는 값



https://www.wikihow.com/Find-Mean,-Median,-and-Mode

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

Question) 이상치에 영향을 가장 많이 받는 요약값은?

* 이상치 (outlier) : 다른 데이터들에 비해, 극단적으로 크거나/작은 값

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

예시) [160,170,175,180,180,185,250]

- 평균 : (160+170+175+180+185+250) / 6
- 중앙값 : 180
- 최빈값 : 180

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

예시) [160,170,175,180,180,185,**250**]

- 평균 : (160+170+175+180+185+**250**) / 6
- 중앙값 : 180
- 최빈값 : 180

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

예시) [160,170,175,180,180,185,**250**]

- 평균 : (160+170+175+180+185+**250**) / 6
- 중앙값 : 180
- 최빈값 : 180

하지만, 데이터의 수가 많으면 이상치의 효과도 감소하게 된다.

따라서, 일반적으로 **평균**을 가장 대표적인 요약값으로 사용한다.

Hanwha Ocean

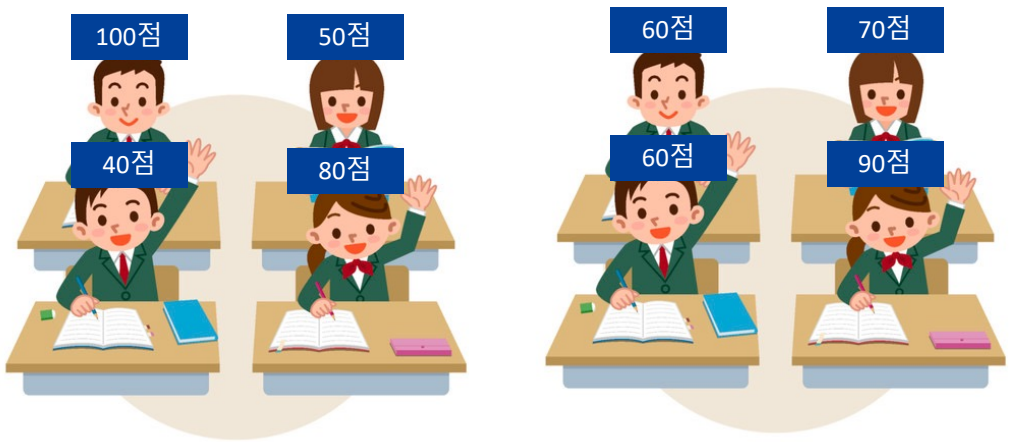
4. 통계학의 기초

2. 데이터의 요약값

평균 : 550/8 = 68.75점

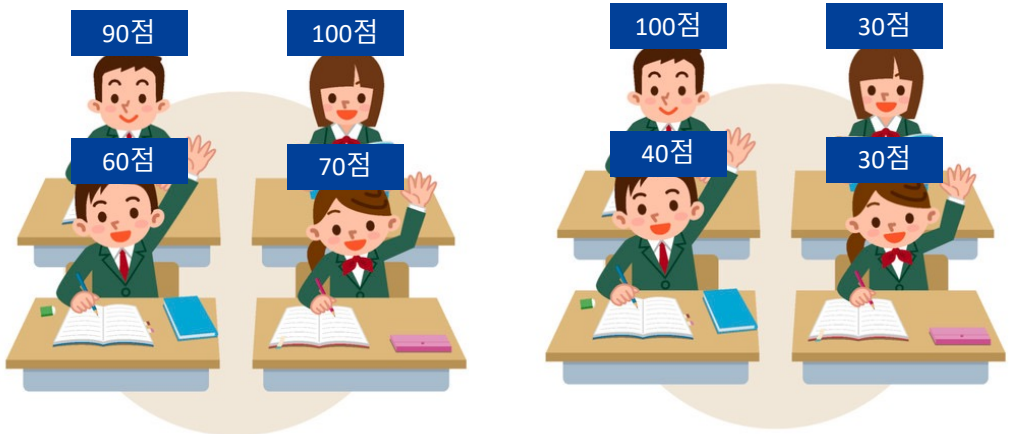
중앙값 : 40,50, 60,60,70,80,90,100 => 65점

최빈값 : 70점



학급 A

VS



학급 B

Hanwha Ocean

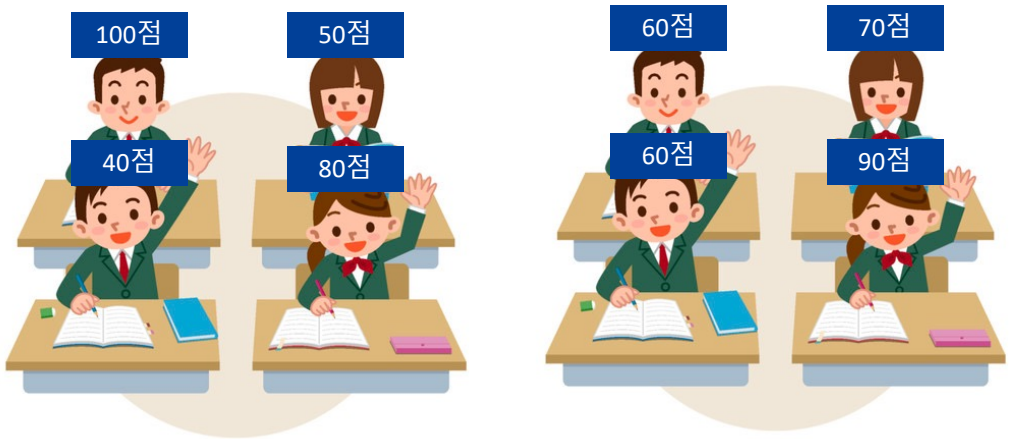
4. 통계학의 기초

2. 데이터의 요약값

평균 : 550/8 = 68.75점

중앙값 : 40,50, 60,60,70,80,90,100 => 65점

최빈값 : 70점



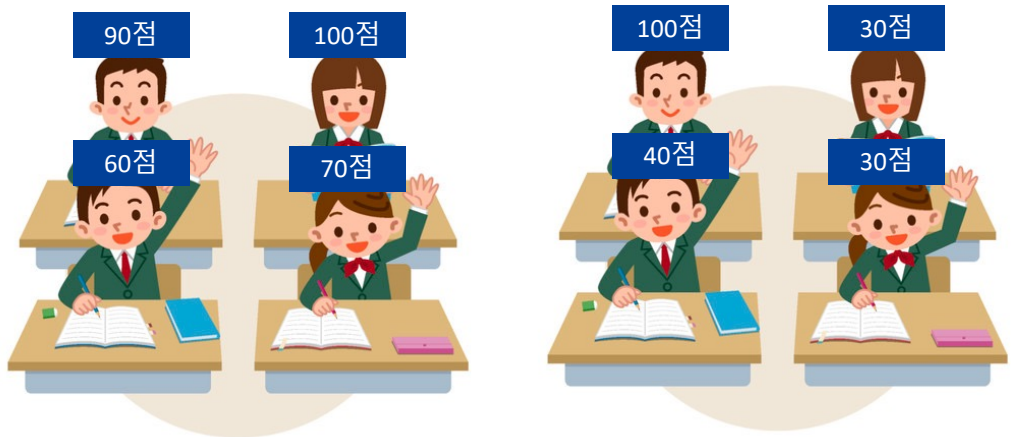
학급 A

VS

평균 : 520/8 = 65점

중앙값 : 30,30,40, 60,70,90,100,100 => 65점

최빈값 : 100점



학급 B

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(4) 사분위수 (quartile)

- **1분위수 (Q1)** : 하위 25%
- **2분위수 (Q2)** : 하위 50% (= 중앙값)
- **3분위수 (Q3)** : 하위 75% (= 상위 25%)

데이터의 **전체적인 모양(분포)**을 알 수 있다!

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(4) 사분위수 (quartile)

- **1분위수 (Q1)** : 하위 25%
- **2분위수 (Q2)** : 하위 50% (= 중앙값)
- **3분위수 (Q3)** : 하위 75% (= 상위 25%)

https://blog.kakaocdn.net/dn/dq4Bre/btqzkmQVgsF/ABK94lzsCWvTbbAkgjRUY0/img.jpg

데이터의 전체적인 모양(분포)울 알 수 있다!

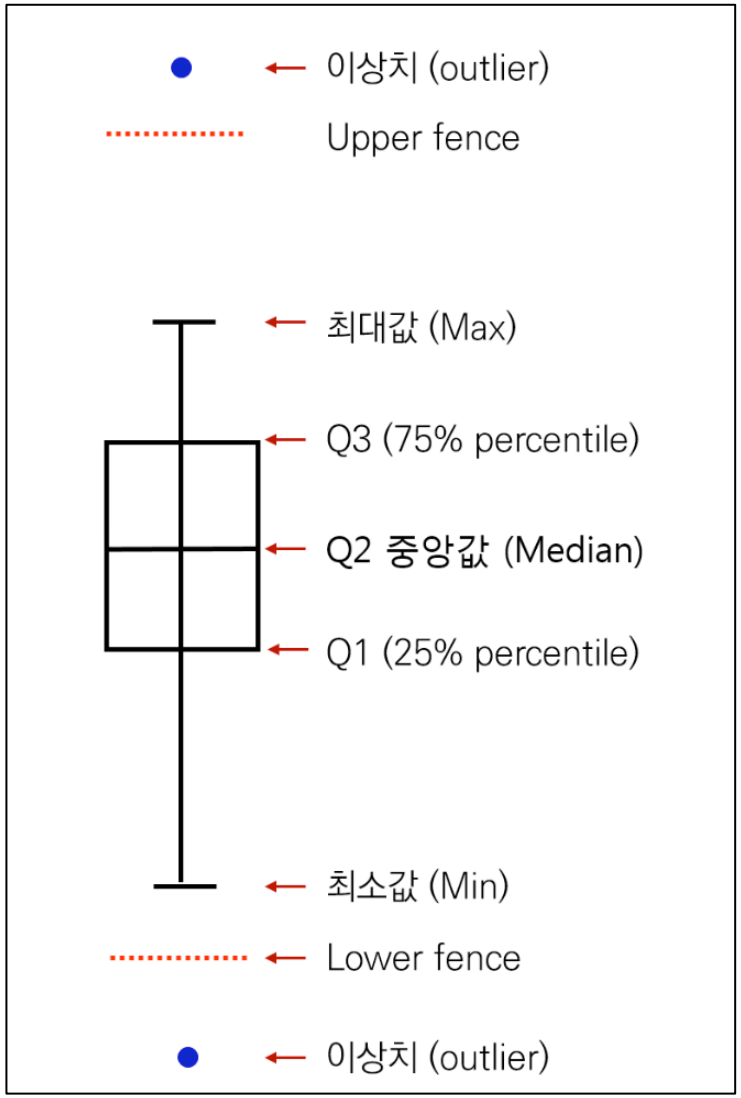
Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

상자 그림 (Box Plot)

- 데이터의 대략적인 분포를 **사분위수(quartile)**를 통해, **한 눈에 파악하기 용이하도록** 그린 그림



https://wikidocs.net/161870#outlier

Hanwha Ocean

4. 통계학의 기초

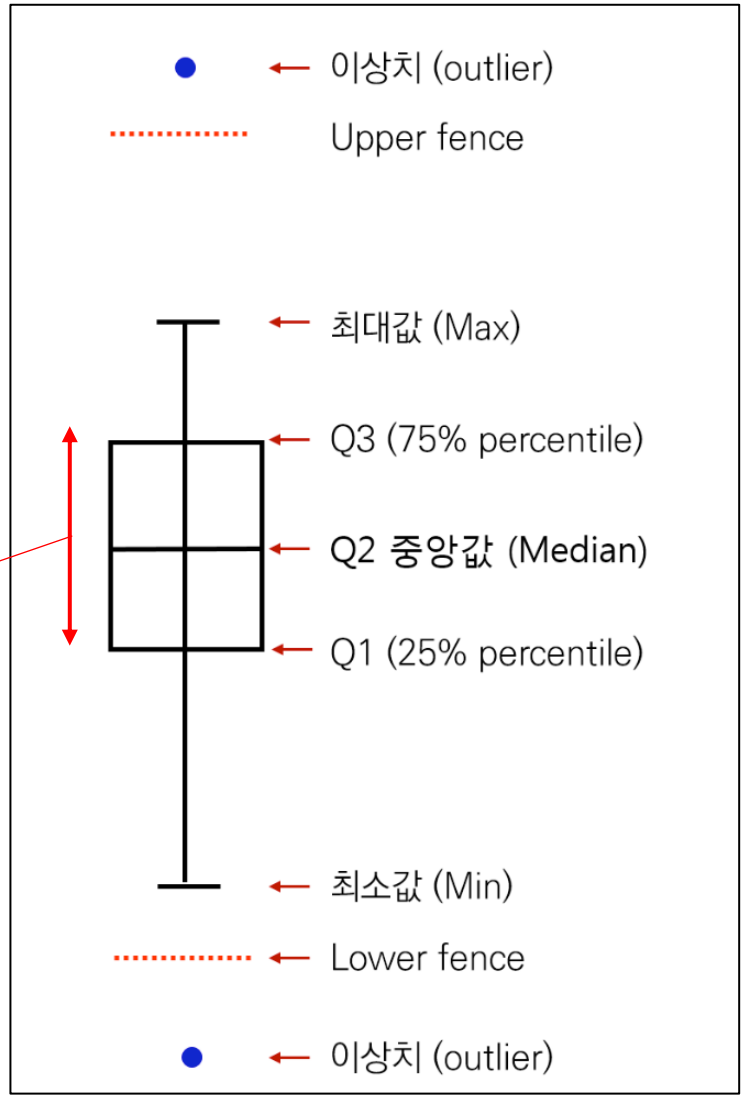
2. 데이터의 요약값

상자 그림 (Box Plot)

- 데이터의 대략적인 분포를 **사분위수(quartile)**를 통해, **한 눈에 파악하기 용이하도록** 그린 그림

IQR : Q3 & Q1 값의 차이

(= InterQuartile Range)



https://wikidocs.net/161870#outlier

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산

(2) 표준편차

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

- 데이터가 **평균**으로부터 **퍼져있는 정도**

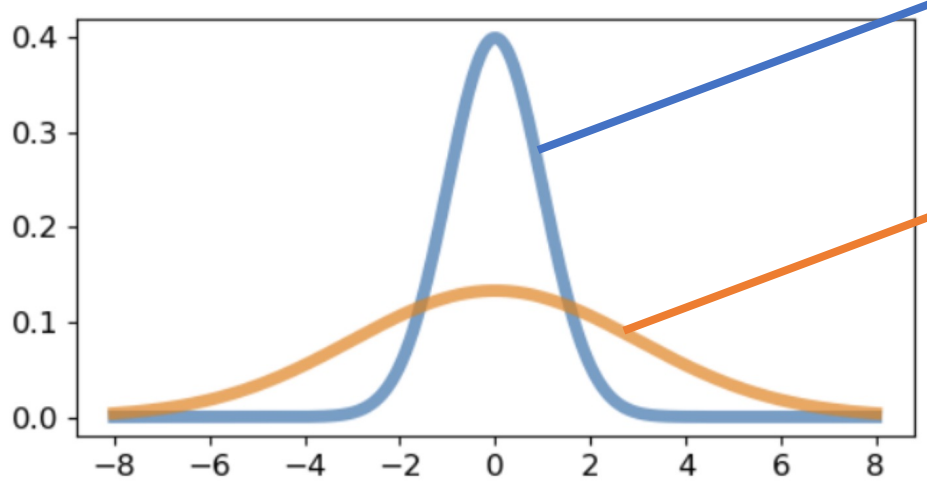
Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

- 데이터가 **평균**으로부터 **퍼져있는 정도**



분산이 **크다**

분산이 **작다**

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

- 예시)
  - Case 1) [10, 20, 30]
  - Case 2) [0, 20, 40]
  - Case 3) [19, 20, 21]

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

- 예시)
  - Case 1) [10, 20, 30]
  - Case 2) [0, 20, 40]
  - Case 3) [19, 20, 21]

세 경우 모두 **평균은 20**으로 동일하다.

하지만, 세 데이터들은 **서로 다른 모양**을 띈다.

이유 : **서로 다른 퍼져있는 정도 (= 분산)**

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

- 예시)
  - Case 1) [10, 20, 30]
  - Case 2) [0, 20, 40]
  - Case 3) [19, 20, 21]

세 경우 모두 **평균은 20**으로 동일하다.

하지만, 세 데이터들은 **서로 다른 모양**을 띈다.

이유 : **서로 다른 퍼져있는 정도 (= 분산)**

*따라서, (평균/중앙값 등의) **데이터의 중심값**만을 보는 것 뿐만 아니라,*

***데이터의 퍼져있는 정도**를 나타내는 **분산**을 함께 봐야 한다!*

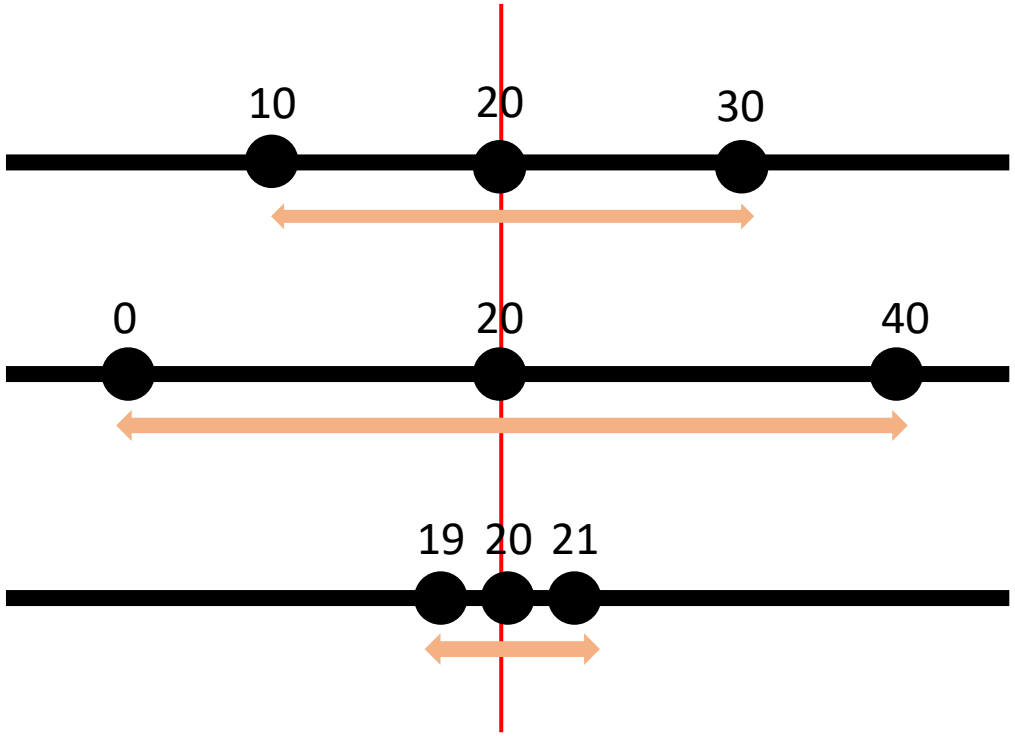
Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

- 예시)
  - Case 1) [10, 20, 30]
  - Case 2) [0, 20, 40]
  - Case 3) [19, 20, 21]



평균은 20으로 동일

분산(퍼져있는 정도)는 다르다

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) 편차를 제곱한다. (Why?)

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) 편차를 제곱한다

Step 4) 편차 제곱들의 평균을 구한다.

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) 편차를 제곱한다

Step 4) 편차 제곱들의 평균을 구한다.

평균 = 5

A 1, 3, 5, 7, 9

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) 편차를 제곱한다

Step 4) 편차 제곱들의 평균을 구한다.

평균 = 5

A	1	, 3	, 5	, 7	, 9	
편차	-4	, -2	, 0	, 2	, 4	

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) 편차를 제곱한다

Step 4) 편차 제곱들의 평균을 구한다.

평균 = 5

A	1	3	5	7	9
편차	-4	-2	0	2	4
편차의 제곱	16	4	0	4	16

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) 편차를 제곱한다

Step 4) 편차 제곱들의 평균을 구한다.

평균 = 5

A **1, 3, 5, 7, 9**

편차	**-4 , -2 , 0 , 2 , 4**
편차의 제곱	**16 , 4 , 0 , 4 , 16**
분산 (편차의 제곱의 평균)	$(16 + 4 + 0 + 4 + 16) \div 5 = 8$

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) **편차를 제곱한다**

Step 4) 편차 제곱들의 평균을 구한다.

Question) Step 3)에서 편차를 제곱하는 이유는?

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) **편차를 제곱한다** →

Step 4) 편차 제곱들의 평균을 구한다.

제곱을 하는 이유?

- 편차(=데이터-평균)은 음수일 수 있다.

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(1) 분산 (variance)

분산의 계산 방법은?

Step 1) 평균을 계산한다.

- 분산 = 평균으로부터 퍼져있는 정도

Step 2) 데이터 & 평균 사이의 차이를 계산한다 (= 편차)

Step 3) **편차를 제곱한다** →

Step 4) 편차 제곱들의 평균을 구한다.

제곱을 하는 이유?

- 편차(=데이터-평균)은 음수일 수 있다.
- 평균으로부터 떨어진 “거리”를 알고 싶은 것이므로, **양수로 변환해야 한다.**
- 따라서, 편차를 “제곱”한 뒤, 이의 평균을 계산한다!

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

2. 데이터의 요약값

(2) 표준편차 (standard deviation)

표준편차 = 분산의 제곱근

A **1, 3, 5, 7, 9**

편차	-4 , -2 , 0 , 2 , 4
편차의 제곱	16 , 4 , 0 , 4 , 16
분산 (편차의 제곱의 평균)	$(16 + 4 + 0 + 4 + 16) \div 5 = 8$
표준 편차 (분산의 제곱근)	$\sqrt{8}$

https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

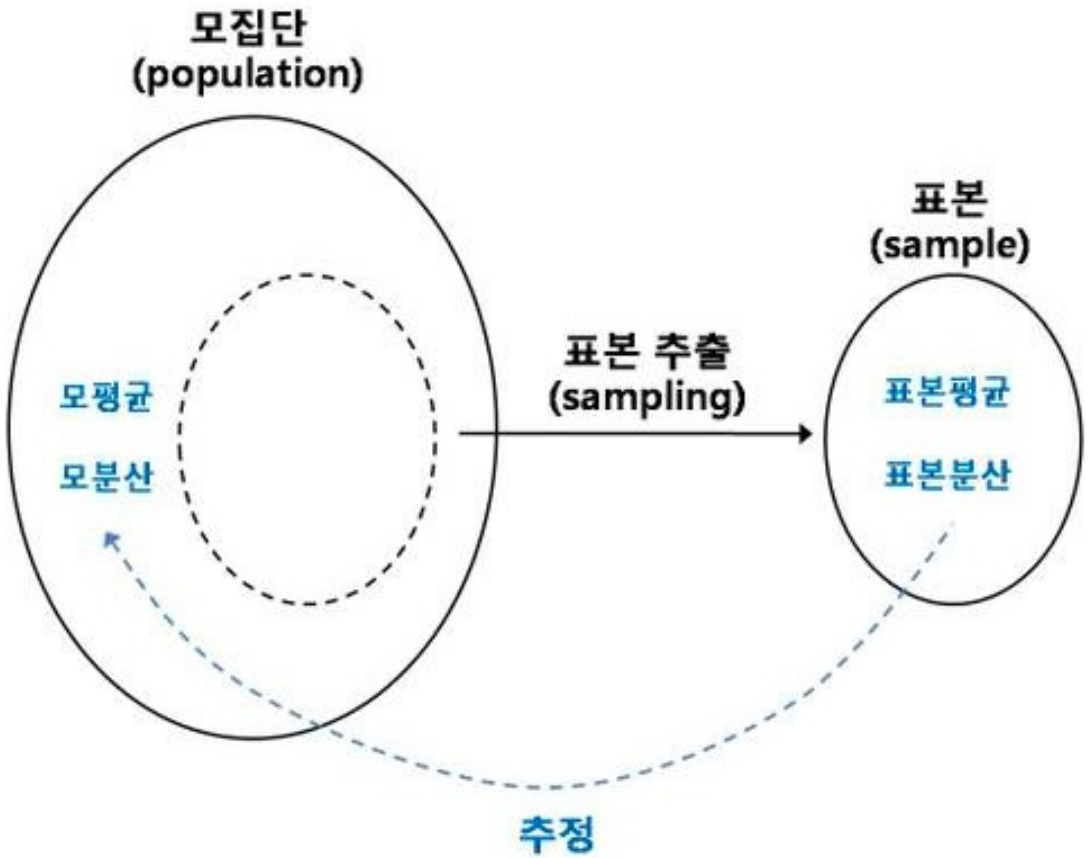
3. 모집단과 표본

모집단과 표본

관심 대상 전체를 전수 조사를 하면 좋겠으나,

시간/비용 등의 제약으로 인해 전체를 조사하기는 어려움.

따라서, **일부를 뽑아서 이들만**을 대상으로 조사를 한다.



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

모집단 & 표본

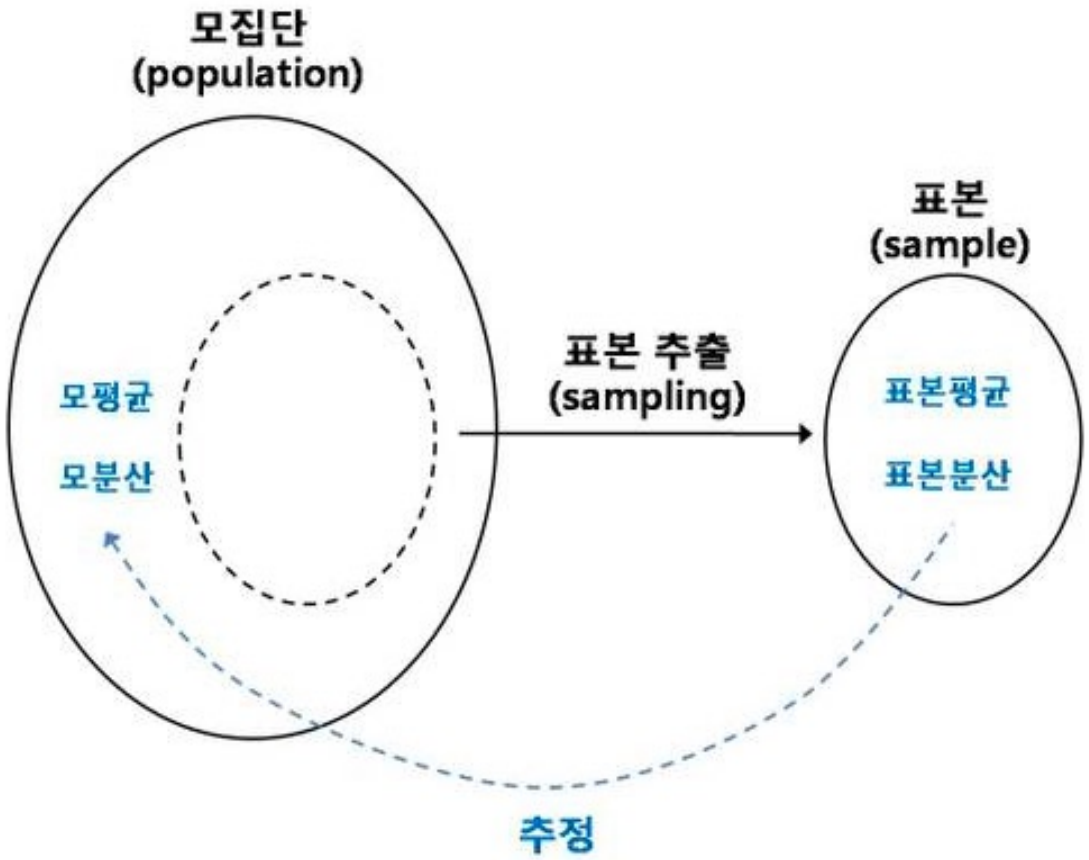
관심 대상 전체를 전수 조사를 하면 좋겠으나,

시간/비용 등의 제약으로 인해 전체를 조사하기는 어려움.

따라서, **일부를 뽑아서 이들만**을 대상으로 조사를 한다.

모집단 : 관심이 되는 전체 대상

표본 : 관심 대상 중, 일부 선택된(샘플된) 대상들



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

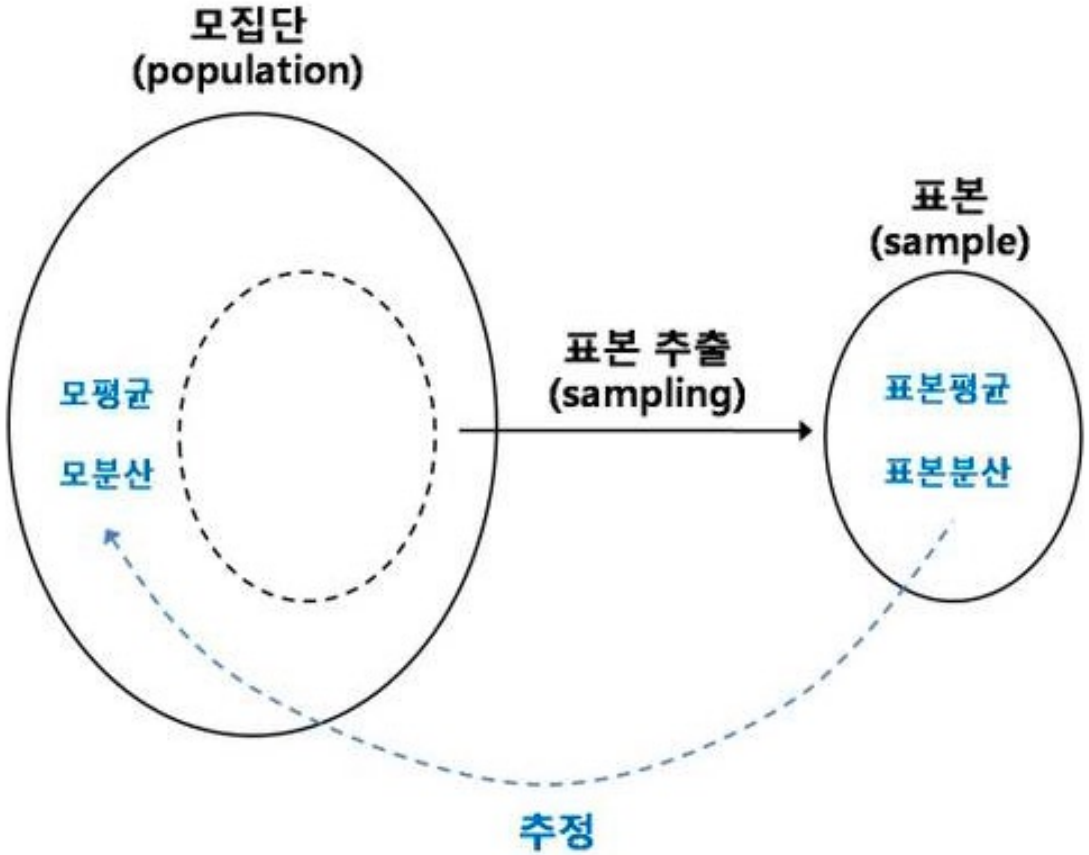
4. 통계학의 기초

3. 모집단과 표본

모수

모집단의 특성을 나타내는 값

- **모**평균 : **모집단**의 평균
- **모**분산 : **모집단**의 분산



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

모수

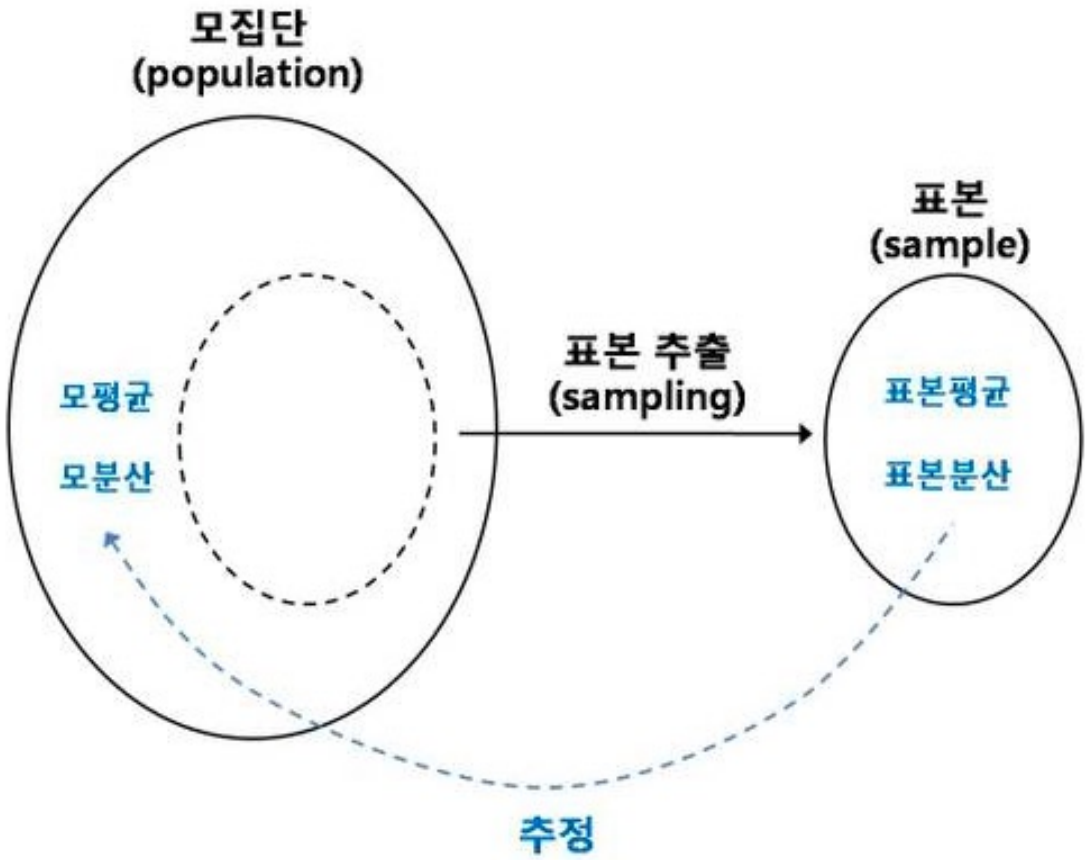
모집단의 특성을 나타내는 값

- **모**평균 : **모집단**의 평균
- **모**분산 : **모집단**의 분산

궁극적으로 알고자 하는 것은 “모수”

(+ 모수를 파악하기 위해서는 전수조사가 필요!)

현실적인 제약으로, 표본을 뽑는다 (= **표본 추출**)



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

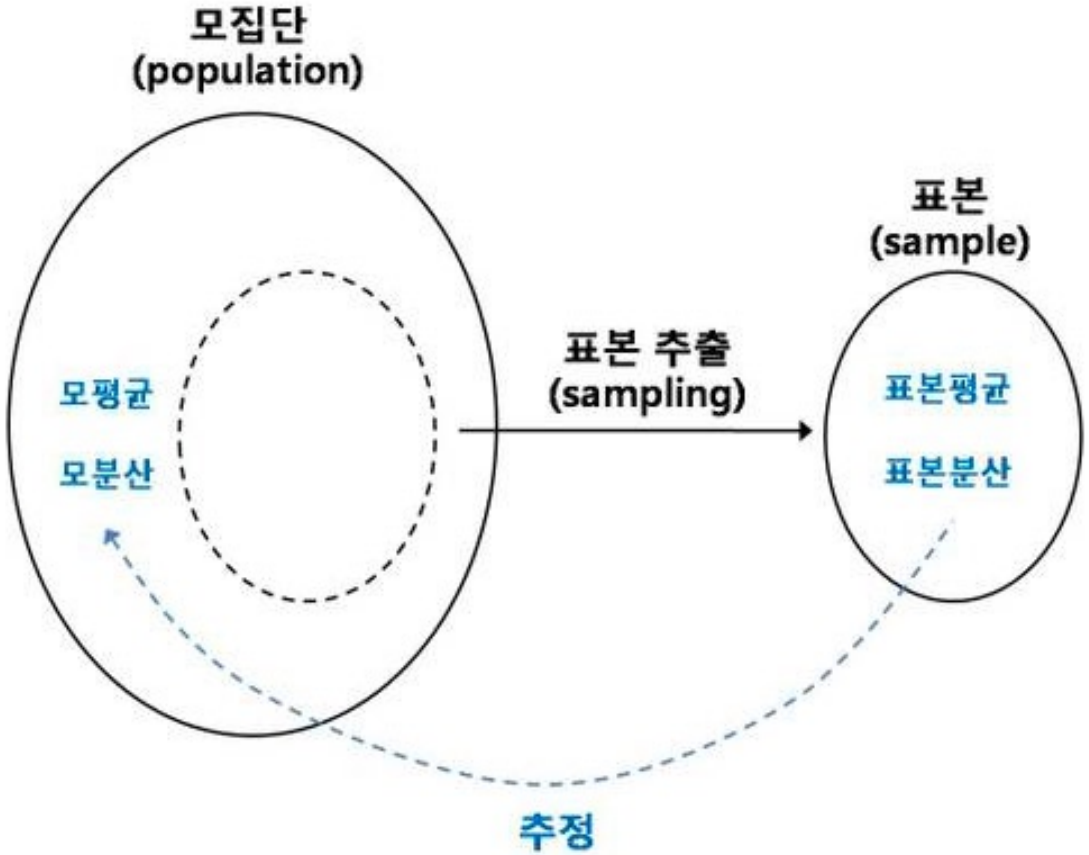
4. 통계학의 기초

3. 모집단과 표본

통계량

표본에서 얻어진 특성의 값

- **표본**평균 : **표본**의 평균
- **표본**분산 : **표본**의 분산



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

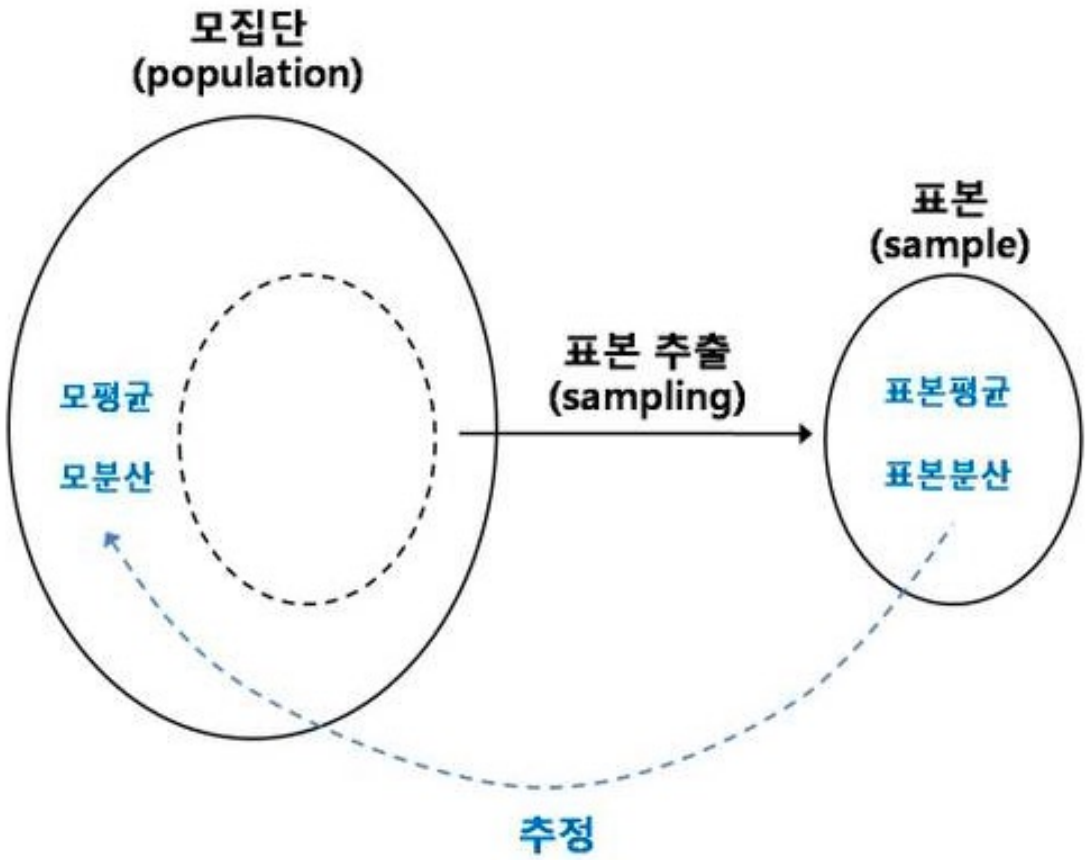
통계량

표본에서 얻어진 특성의 값

- **표본**평균 : **표본**의 평균
- **표본**분산 : **표본**의 분산

모수와 통계량 사이에는 차이가 존재하게 된다

(통계량의 경우, 전체를 대상으로 계산한 것이 아니므로)



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

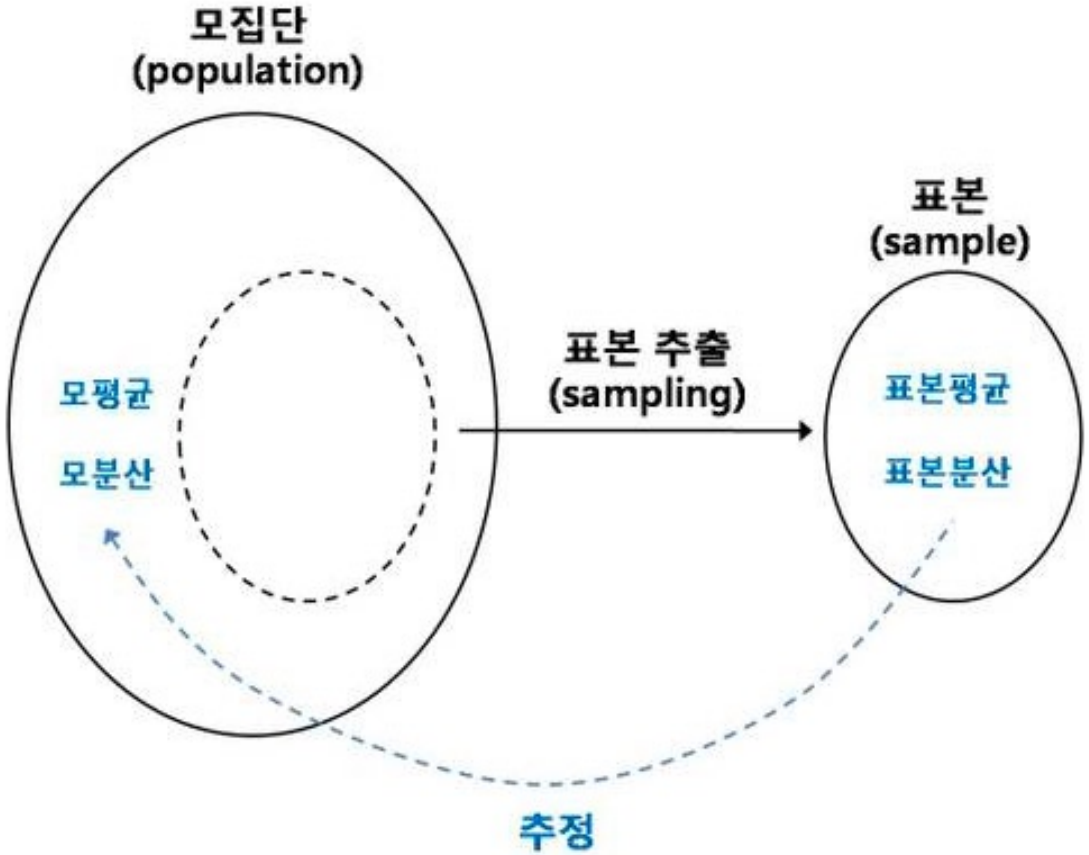
Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

추정 (Estimation)

표본에서 얻은 **통계량**을 사용하여 **모수**를 추측하는 과정



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

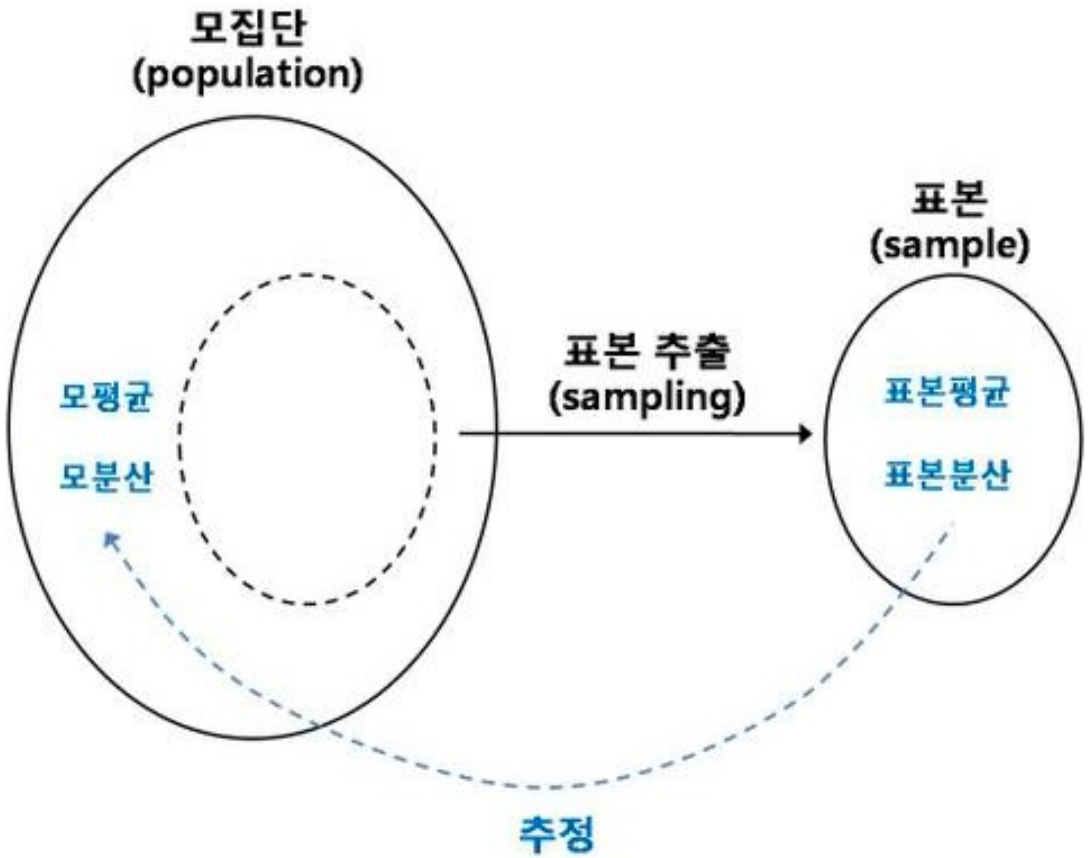
3. 모집단과 표본

추정 (Estimation)

표본에서 얻은 **통계량**을 사용하여 **모수**를 추측하는 과정

추정의 종류

1. **점** 추정 : 모수를 "하나의 값"으로 추정
2. **구간** 추정 : 모수를 포함하고 있는 "범위"를 추정



https://t1.daumcdn.net/cfile/tistory/2533803857C2FDE128

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

추정 (Estimation)

점 추정 vs 구간 추정

예시) 대한민국 성인 남성 키의 평균은?

Hanwha Ocean

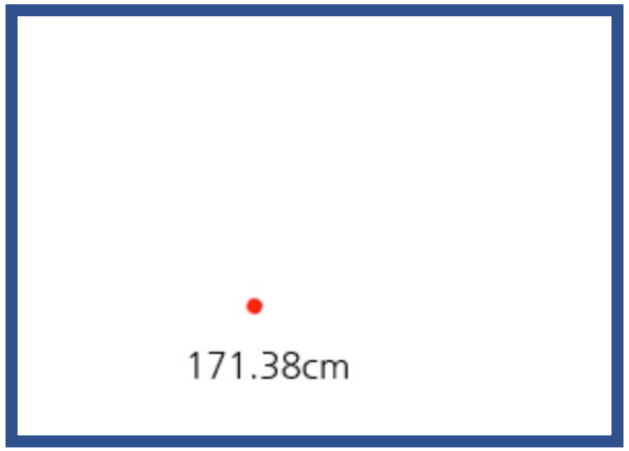
4. 통계학의 기초

3. 모집단과 표본

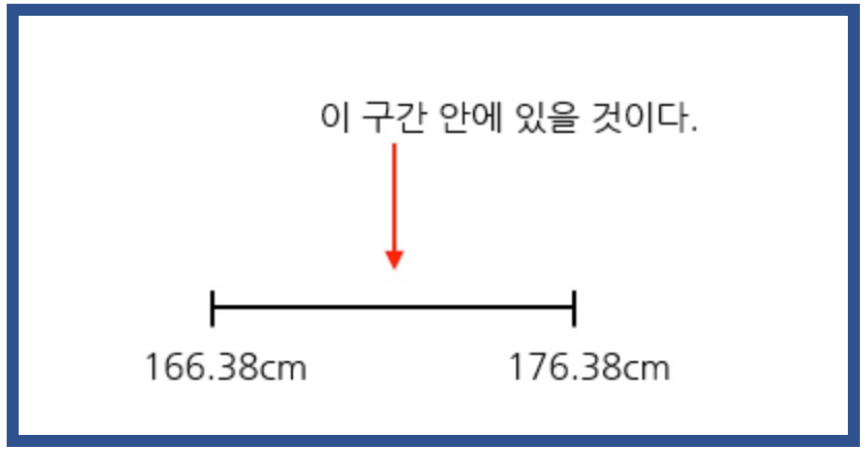
추정 (Estimation)

점 추정 vs 구간 추정

예시) 대한민국 성인 남성 키의 평균은?



점 추정



구간 추정

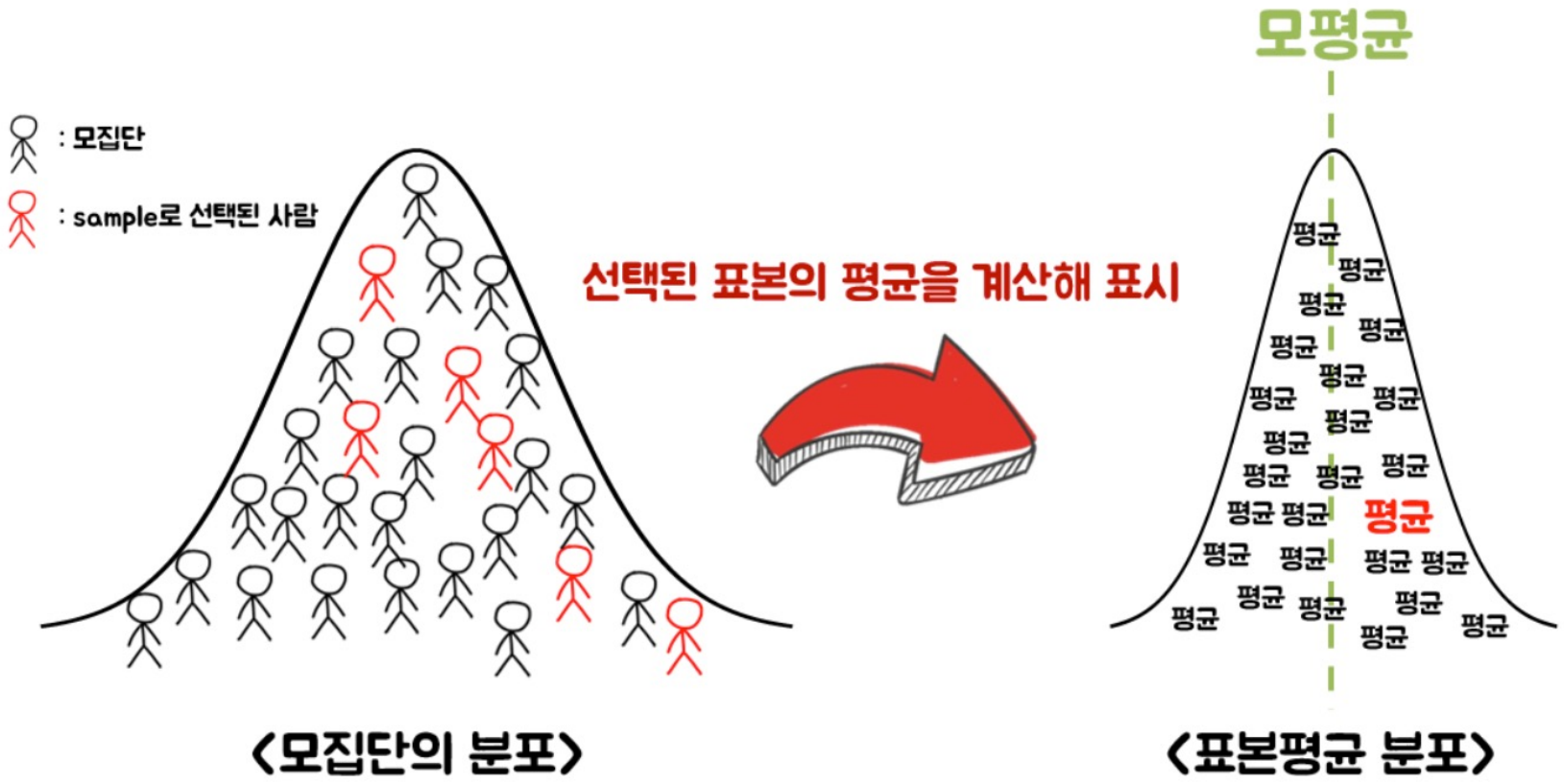
https://math100.tistory.com/49

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

추정 (Estimation)



https://angeloyeo.github.io/2021/01/05/confidence_interval.html

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

신뢰 구간

- 구간 추정 시 계산되는 범위
- 신뢰구간 = 통계량 ± 오차 범위

http://doc.mindscale.kr/ngv/stat01.pdf

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

신뢰 구간

- 구간 추정 시 계산되는 범위
- 신뢰구간 = 통계량 ± 오차 범위

신뢰 수준 : *계산한 구간이 얼마나 신뢰할만 한가?*

http://doc.mindscale.kr/ngv/stat01.pdf

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

신뢰 구간

- 구간 추정 시 계산되는 범위
- 신뢰구간 = 통계량 ± 오차 범위

신뢰 수준 : *계산한 구간이 얼마나 신뢰할만 한가?*

- 신뢰 구간은 모수를 포함할 수도, 안할 수도 있다.
  - 예시) 예상한 범위는 (169,171)이나, 실제 모수는 172

http://doc.mindscale.kr/ngv/stat01.pdf

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

신뢰 구간

- 구간 추정 시 계산되는 범위
- 신뢰구간 = 통계량 ± 오차 범위

신뢰 수준 : *계산한 구간이 얼마나 신뢰할만 한가?*

- 신뢰 구간은 모수를 포함할 수도, 안할 수도 있다.
  - 예시) 예상한 범위는 (169,171)이나, 실제 모수는 172
- 모수에서 뽑아낸 표본에 따라 통계량은 다름

  따라서, 신뢰 구간도 각각 다르게 계산된다.

http://doc.mindscale.kr/ngv/stat01.pdf

Hanwha Ocean

4. 통계학의 기초

3. 모집단과 표본

신뢰 구간

- 구간 추정 시 계산되는 범위
- 신뢰구간 = 통계량 ± 오차 범위

신뢰 수준 : *계산한 구간이 얼마나 신뢰할만 한가?*

- 신뢰 구간은 모수를 포함할 수도, 안할 수도 있다.
  - 예시) 예상한 범위는 (169,171)이나, 실제 모수는 172
- 모수에서 뽑아낸 표본에 따라 통계량은 다름

  따라서, 신뢰 구간도 각각 다르게 계산된다.
- **신뢰 수준 = 신뢰 구간이 모수를 포함하는 표본들의 비율**

http://doc.mindscale.kr/ngv/stat01.pdf

Hanwha Ocean

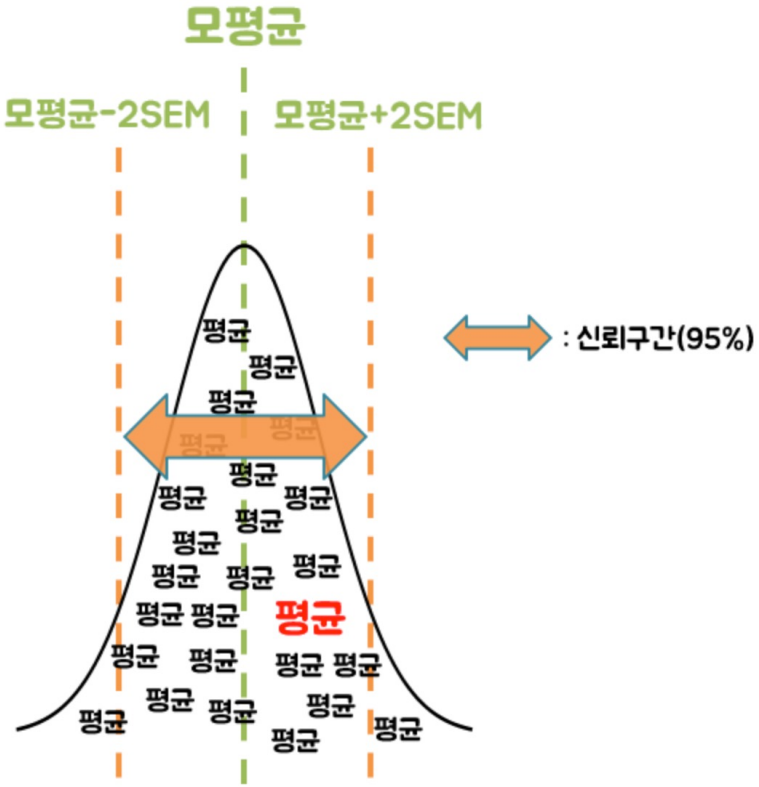
4. 통계학의 기초

3. 모집단과 표본

신뢰 구간 신뢰구간 = 통계량 ± 오차 범위

신뢰 수준 신뢰 수준 = 신뢰 구간이 모수를 포함하는 표본들의 비율

모평균 m에 대한 신뢰도 95%의 신뢰구간은, 크기가 n인 표본을 여러 번 추출하여 신뢰구간을 만들 때, 이러한 여러 표본 평균들의 95% 정도는 모평균 m을 포함할 것으로 기대된다.



〈표본평균 분포〉

https://angeloyeo.github.io/2021/01/05/confidence_interval.html

Hanwha Ocean

4. 통계학의 기초

4. 확률

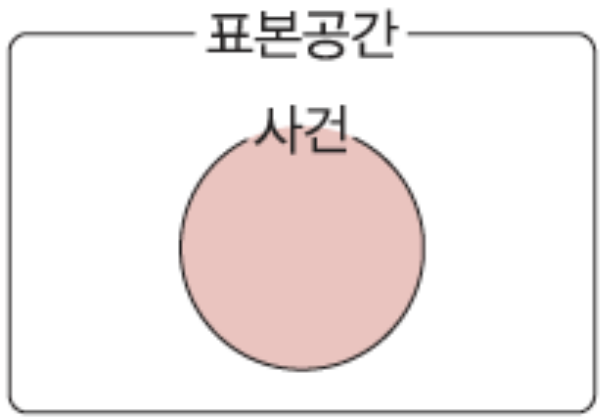
시행, 표본공간, 사건, 근원사건

시행 : 동전/주사위 던지는 것과 같이, 같은 조건에서 반복할 수 있고, 그 결과가 우연히 결정되는 실험

표본 공간 : 시행에서 일어날 수 있는 모든 가능한 결과의 집합

사건 : 표본 공간의 부분집합

근원 사건 : 한 개의 원소로 이루어진 사건



Hanwha Ocean

4. 통계학의 기초

4. 확률

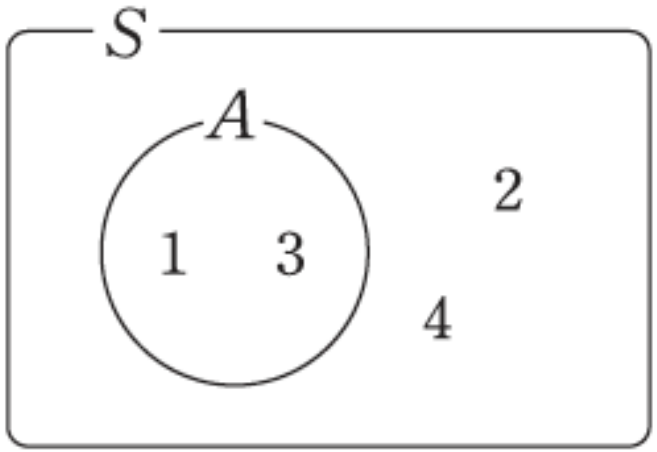
예시

4장의 카드 (1,2,3,4) 중, 임의로 1장의 코드를 뽑는 시행에서

(1) 표본공간 S : {1, 2, 3, 4}

(2) 뽑은 카드 숫자가 홀수인 사건 A : {1, 3}

(3) 근원 사건 : {1}, {2}, {3}, {4}



Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

1 ~ 20의 자연 수 중, 1장의 카드를 임의로 뽑는 시행에서 ..

(1) 표본공간 : ??

(2) 3의 배수를 뽑는 사건 : ??

(3) 20의 약수를 뽑는 사건 : ??

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

1 ~ 20의 자연 수 중, 1장의 카드를 임의로 뽑는 시행에서 ..

(1) 표본공간 : {1, 2, 3, .. ,19, 20}

(2) 3의 배수를 뽑는 사건 : {3, 6, 9, 12, 15, 18}

(3) 20의 약수를 뽑는 사건 : {1, 2, 4, 5, 10, 20}

Hanwha Ocean

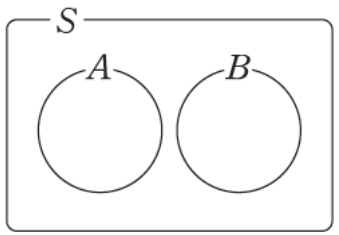
4. 통계학의 기초

4. 확률

배반사건 & 여사건

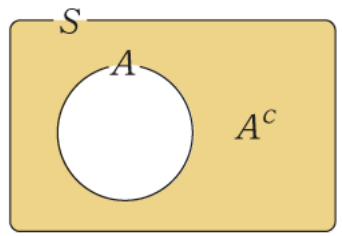
배반 사건 : $A \cap B = \varnothing$

- 사건 A와 사건 B가 동시에 일어나지 않을 때, A와 B는 배반사건이다.



여사건 : A^c

- 사건에 A에 대하여, A가 일어나지 않는 사건을 A의 여사건이라고 한다.



Hanwha Ocean

4. 통계학의 기초

4. 확률

예시

한 개의 주사귀를 던지는 시행에서,

- 표본공간 S : {1,2,3,4,5,6}
- 짝수의 눈이 나오는 사건 A : {2, 4, 6}
- 홀수의 눈이 나오는 사건 B : {1, 3, 5}

A와 B는 서로 배반사건이다.

A의 여사건은 {1, 3, 5}이다.

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

1~10 까지의 자연수가 적힌 10개의 공에서, 임의로 1개의 공을 꺼낸다.

- A : 꺼낸 공의 수가 홀수
- B : 꺼낸 공의 수가 4의 배수
- C : 꺼낸 공의 수가 9의 약수

(1) A&B, B&C, C&A 중, 서로 배반사건인 것은?

(2) $B \cup C$ 의 여사건은?

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

1~10 까지의 자연수가 적힌 10개의 공에서, 임의로 1개의 공을 꺼낸다.

- A : 꺼낸 공의 수가 홀수
- B : 꺼낸 공의 수가 4의 배수
- C : 꺼낸 공의 수가 9의 약수

(1) A&B, B&C, C&A 중, 서로 배반사건인 것은? (A&B), (B&C)

(2) $B \cup C$ 의 여사건은?

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

1~10 까지의 자연수가 적힌 10개의 공에서, 임의로 1개의 공을 꺼낸다.

- A : 꺼낸 공의 수가 홀수
- B : 꺼낸 공의 수가 4의 배수
- C : 꺼낸 공의 수가 9의 약수

(1) A&B, B&C, C&A 중, 서로 배반사건인 것은? (A&B), (B&C)

(2) $B \cup C$ 의 여사건은? {2,5,6,7,10}

Hanwha Ocean

4. 통계학의 기초

4. 확률

수학적 확률

확률: 어떤 시행에서 사건 A가 일어날 확률 $\mathrm{P}(A)$

어떤 시행에서 표본 공간 S의 각 근원사건이 일어날 가능성이 모두 같은 정도로 기대될 때,

이를 사건 A가 일어날 수학적 확률 이라고 함

$$\mathrm{P}(A)=\frac{n(A)}{n(S)}$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

각 면에 1,2,3,4가 하나씩 적힌 2개의 정사면체를 동시에 던질 때,

밑 면에 적힌 두 숫자의 합이 4일 확률은?

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

각 면에 1,2,3,4가 하나씩 적힌 2개의 정사면체를 동시에 던질 때,

밑 면에 적힌 두 숫자의 합이 4일 확률은?

$$S=\{(1,1),(1,2),\cdots,(4,3),(4,4)\}$$

$$A=\{(1,3),(2,2),(3,1)\}$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

각 면에 1,2,3,4가 하나씩 적힌 2개의 정사면체를 동시에 던질 때,

밑 면에 적힌 두 숫자의 합이 4일 확률은?

$$S=\{(1, 1), (1, 2), \cdots, (4, 3), (4, 4)\} \qquad n(S)=16$$

$$A=\{(1, 3), (2, 2), (3, 1)\} \qquad n(A)=3$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

각 면에 1,2,3,4가 하나씩 적힌 2개의 정사면체를 동시에 던질 때,

밑 면에 적힌 두 숫자의 합이 4일 확률은?

$$S=\{(1,1),(1,2),\cdots,(4,3),(4,4)\} \qquad n(S)=16$$

$$A=\{(1,3),(2,2),(3,1)\} \qquad n(A)=3$$

$$\mathrm{P}(A)=\frac{n(A)}{n(S)}=\frac{3}{16}$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

통계적 확률

여러 번의 반복적인 실험 및 경험을 통해서 얻어지는 상대도수를 통해서 추측한 확률

- ex) 야구 선수의 타율, 제품이 불량일 확률

Hanwha Ocean

4. 통계학의 기초

4. 확률

통계적 확률

여러 번의 반복적인 실험 및 경험을 통해서 얻어지는 상대도수를 통해서 추측한 확률

- ex) 야구 선수의 타율, 제품이 불량일 확률

야구선수의 타율이 0.3이라고 해서, 이 사람이 실제로 안타를 칠 확률이 과학적/절대적으로 30%?

NO! 단순히 과거의 경험 (ex. 300 타수 90안타)를 통해서, 경험적으로 추측한 확률일 뿐!

Hanwha Ocean

4. 통계학의 기초

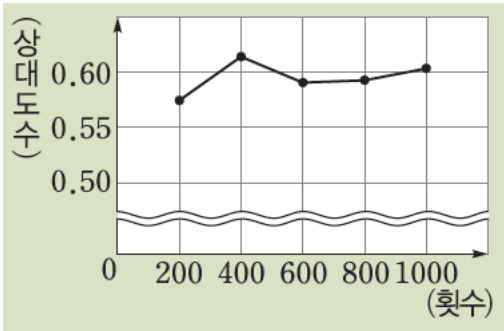
4. 확률

통계적 확률

하지만, 시행 횟수 (경험 횟수)가 많으면, 해당 통계적 확률은 보다 신뢰할 만 하다.

예시) 윷을 던져서 평평한 면이 나올 “통계적 확률”은?

윷짝을 던진 횟수	200	400	600	800	1000
평평한 면이 나온 횟수	115	246	354	475	601
상대도수	0.575	0.615	0.59	0.594	0.601



Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

어느 제약 회사에서 개발한 신약을, 1000명의 환자에게 투여하였더니 700명이 완치.

새로운 특정 환자에게, 해당 약을 투여했을 때, 독감이 완치될 “통계적 확률”은?

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 기본 성질

용어

- 표본 공간 S
- 표본 공간의 임의의 사건 A (= S의 부분 집합)

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 기본 성질

용어

- 표본 공간 S
- 표본 공간의 임의의 사건 A (= S의 부분 집합)

$$0 \leq n(A) \leq n(S)$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 기본 성질

용어

- 표본 공간 S
- 표본 공간의 임의의 사건 A (= S의 부분 집합)

$$0 \le n(A) \le n(S) \qquad 0 \le \frac{n(A)}{n(S)} \le 1$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 기본 성질

용어

- 표본 공간 S
- 표본 공간의 임의의 사건 A (= S의 부분 집합)

$$0 \le n(A) \le n(S) \qquad 0 \le \frac{n(A)}{n(S)} \le 1 \qquad \cdot 0 \le \mathrm{P}(A) \le 1$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 기본 성질

반드시 일어나는 사건 s

절대로 일어나지 않는 사건 $\varnothing$

$$P(S)=\frac{n(S)}{n(S)}=1 \qquad P(\varnothing)=\frac{n(\varnothing)}{n(S)}=0$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 기본 성질

요약) 확률의 기본 성질 3가지

표본공간이 S인 어떤 시행에서

1. 임의의 사건 A에 대하여 $0 \le P(A) \le 1$
2. 반드시 일어나는 사건 S에 대하여 $P(S)=1$
3. 절대로 일어나지 않는 사건 $\varnothing$에 대하여 $P(\varnothing)=0$

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

주머니 안에, 흰 공 2개 & 노란 공 3개가 있다.

임의로 3개의 공을 동시에 꺼낼 때…

Q1) 노란 공이 나올 확률

Q2) 3개 모두 흰 공이 나올 확률

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

주머니 안에, 흰 공 2개 & 노란 공 3개가 있다.

임의로 3개의 공을 동시에 꺼낼 때…

Q1) 노란 공이 나올 확률

= 1 – 노란 공이 하나도 안 나올 확률

= 1 – 0

= 1

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

주머니 안에, 흰 공 2개 & 노란 공 3개가 있다.

임의로 3개의 공을 동시에 꺼낼 때...

Q2) 3개 모두 흰 공이 나올 확률

= 0 (불가능)

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

두 사건 A와 B가 있을 때, A 또는 B가 일어날 확률은?

(= A 혹은 B 중 하나라도 발생할 확률)

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

두 사건 A와 B가 있을 때, A 또는 B가 일어날 확률은?

(= A 혹은 B 중 하나라도 발생할 확률)

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

예시) 숫자 1~10 중

- A 사건 = 2의 배수가 나오는 경우
- B 사건 = 8의 약수가 나오는 경우
- $A \cup B$ 사건 : 2의 배수 혹은 8의 약수가 나오는 경우

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

예시) 숫자 1~10 중

- A 사건 = 2의 배수가 나오는 경우 = { 2,4,6,8,10 }
- B 사건 = 8의 약수가 나오는 경우 = { 1,2,4,8 }
- $A \cup B$ 사건 : 2의 배수 혹은 8의 약수가 나오는 경우

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

예시) 숫자 1~10 중

- A 사건 = 2의 배수가 나오는 경우 = { 2,4,6,8,10 }
- B 사건 = 8의 약수가 나오는 경우 = { 1,2,4,8 }
- $A \cup B$ 사건 : 2의 배수 혹은 8의 약수가 나오는 경우 = { 1,2,4,6,8,10 }

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

5 4 3

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

예시) 숫자 1~10 중

- A 사건 = 2의 배수가 나오는 경우 = { 2,4,6,8,10 }
- B 사건 = 8의 약수가 나오는 경우 = { 1,2,4,8 }
- $A \cup B$ 사건 : 2의 배수 혹은 8의 약수가 나오는 경우 = { 1,2,4,6,8,10 }

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

$$\frac{n(A \cup B)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

$$\frac{n(A \cup B)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

$$\frac{n(A \cup B)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$ (A와 B가 배반사건일 경우)

Hanwha Ocean

4. 통계학의 기초

4. 확률

확률의 덧셈 정리 (요약)

표본공간 S의 두 사건 A, B에 대하여

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

특히 두 사건 A, B가 서로 배반사건이면

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

특정 학급에서,

- A 사이트에 가입한 학생은 3/5
- B 사이트에 가입한 학생은 1/2
- A와 B 사이트 모두에 가입한 사람은 1/5

한 명을 임의로 뽑았을 때, A 또는 B 사이트에 가입한 학생일 확률은?

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

특정 학급에서,

- A 사이트에 가입한 학생은 3/5
- B 사이트에 가입한 학생은 1/2
- A와 B 사이트 모두에 가입한 사람은 1/5

한 명을 임의로 뽑았을 때, A 또는 B 사이트에 가입한 학생일 확률은? 3/5 + 1/2 – 1/5 = 0.9

Hanwha Ocean

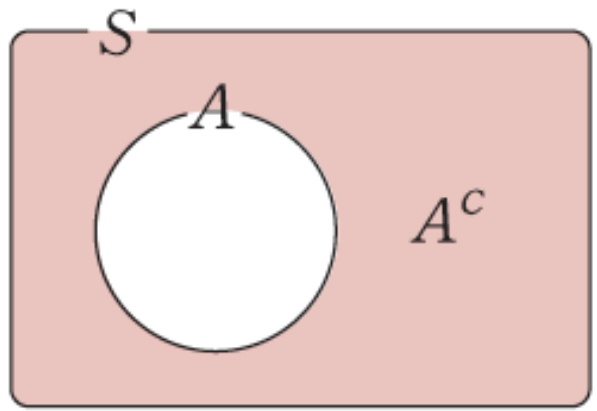
4. 통계학의 기초

4. 확률

여사건의 확률

사건 A와 그 여사건 A^C는 서로 배반사건이므로

$$\mathrm{P}(A \cup A^C) = \mathrm{P}(A) + \mathrm{P}(A^C)$$



Hanwha Ocean

4. 통계학의 기초

4. 확률

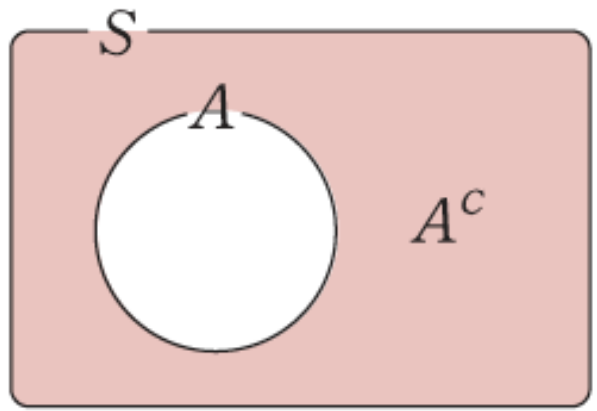
여사건의 확률

사건 A와 그 여사건 A^C는 서로 배반사건이므로

$$P(A \cup A^C) = P(A) + P(A^C)$$

$P(A \cup A^C) = P(S) = 1$이므로 $P(A) + P(A^C) = 1$,

$$\boxed{P(A^C) = 1 - P(A)}$$



Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

8개의 복권 중, 2개는 당첨, 6개는 꽝.

임의로 2개의 복권을 뽑을 때, 적어도 1개는 당첨일 확률은?

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

8개의 복권 중, 2개는 당첨, 6개는 꽝.

임의로 2개의 복권을 뽑을 때, 적어도 1개는 당첨일 확률은?

A) 1개 당첨일 확률 + 2개 당첨일 확률 (?)

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

8개의 복권 중, 2개는 당첨, 6개는 꽝.

임의로 2개의 복권을 뽑을 때, 적어도 1개는 당첨일 확률은?

A) 1개 당첨일 확률 + 2개 당첨일 확률 (?)

“1 – 0개 당첨일 확률”이 보다 수월하다!

Hanwha Ocean

4. 통계학의 기초

4. 확률

Quiz

8개의 복권 중, 2개는 당첨, 6개는 꽝.

임의로 2개의 복권을 뽑을 때, 적어도 1개는 당첨일 확률은?

A) 1개 당첨일 확률 + 2개 당첨일 확률 (?)

“1 – 0개 당첨일 확률”이 보다 수월하다!

1 – (6x5 / 8x7) = 1 - 30/56 = 26/56

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

조건부 확률이란?

일반적인 경우가 아니라, “사건 A가 일어났다고 가정할 때”, 사건 B가 일어날 확률

- P(B) : B가 일어날 확률
- P(B|A) : A가 일어났다고 가정할 때, B가 일어날 확률

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

조건부 확률이란?

일반적인 경우가 아니라, “사건 A가 일어났다고 가정할 때”, 사건 B가 일어날 확률

- P(B) : B가 일어날 확률
- P(B|A) : A가 일어났다고 가정할 때, B가 일어날 확률

위 두 경우는, 다를 수 있다!

예시) 키가 170을 넘을 확률 vs. (어렸을 때 농구를 많이 했다고 가정했을 때) 키가 170넘을 확률

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

예시

여행을 온 이용객 36명의 표

(단위: 명)

	어른	어린이	합계
남자	8	11	19
여자	7	10	17
합계	15	21	36

Q1) 전체 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

예시

여행을 온 이용객 36명의 표

(단위: 명)

	어른	어린이	합계
남자	8	11	19
여자	7	10	17
합계	15	21	36

$$n(S)=36,\ n(A)=19,\ n(A\cap B)=11$$

Q1) 전체 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

$$\frac{n(A\cap B)}{n(S)}=\frac{11}{36}$$

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

예시

여행을 온 이용객 36명의 표

(단위: 명)

	어른	어린이	합계
남자	8	11	19
여자	7	10	17
합계	15	21	36

Q1) 전체 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

Q2) 남자 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

예시

여행을 온 이용객 36명의 표

(단위: 명)

	어른	어린이	합계
남자	8	11	19
여자	7	10	17
합계	15	21	36

Q1) 전체 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

Q2) 남자 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

(조건이 추가적으로 붙었다)

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

예시

여행을 온 이용객 36명의 표

(단위: 명)

	어른	어린이	합계
남자	8	11	19
여자	7	10	17
합계	15	21	36

Q1) 전체 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

Q2) 남자 이용객 중 1명을 골랐을 때, 남자 어린이일 확률

(조건이 추가적으로 붙었다)

$$\frac{n(A \cap B)}{n(A)} = \frac{11}{19}$$

Hanwha Ocean

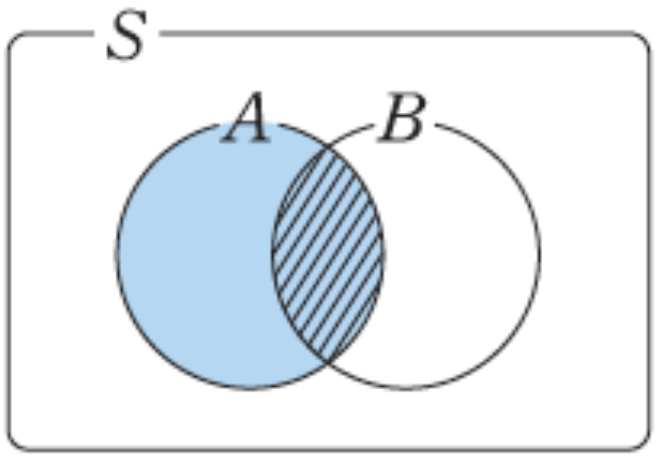
4. 통계학의 기초

5. 조건부 확률

계산 방법

사건 A가 일어났을 때, 사건 B의 조건부 확률은 :

$$P(B|A)=\frac{n(A\cap B)}{n(A)}$$



Hanwha Ocean

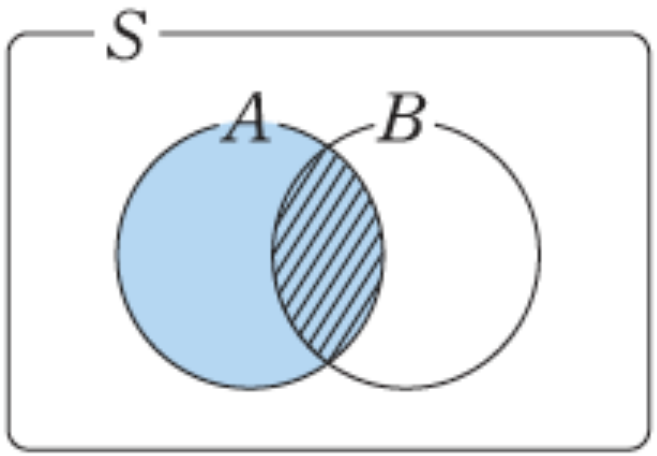
4. 통계학의 기초

5. 조건부 확률

계산 방법

사건 A가 일어났을 때, 사건 B의 조건부 확률은 :

$$P(B|A)=\frac{n(A\cap B)}{n(A)}$$



이 식의 우변의 분자와 분모를 각각 $n(S)$로 나누면

$$P(B|A)=\frac{\frac{n(A\cap B)}{n(S)}}{\frac{n(A)}{n(S)}}=\frac{P(A\cap B)}{P(A)}$$

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

한 개의 주사위를 던져서, 나온 눈이 소수라고 했을 때, 그 수가 홀수일 확률은?

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

한 개의 주사위를 던져서, **나온 눈이 소수라고 했을 때,** 그 수가 홀수일 확률은?

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

한 개의 주사위를 던져서, **나온 눈이 소수라고 했을 때,** 그 수가 홀수일 확률은?

$$S=\{1, 2, 3, 4, 5, 6\}$$

A : 소수인 사건

B : 홀수인 사건

$$A=\{2, 3, 5\},\ B=\{1, 3, 5\},\ A\cap B=\{3, 5\}$$

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

한 개의 주사위를 던져서, **나온 눈이 소수라고 했을 때,** 그 수가 홀수일 확률은?

$$S=\{1, 2, 3, 4, 5, 6\}$$

A : 소수인 사건

B : 홀수인 사건

$$A=\{2, 3, 5\},\ B=\{1, 3, 5\},\ A\cap B=\{3, 5\}$$

$$\mathrm{P}(A)=\frac{3}{6}=\frac{1}{2}$$

$$\mathrm{P}(A\cap B)=\frac{2}{6}=\frac{1}{3}$$

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

한 개의 주사위를 던져서, **나온 눈이 소수라고 했을 때,** 그 수가 홀수일 확률은?

$$S=\{1, 2, 3, 4, 5, 6\}$$

A : 소수인 사건

B : 홀수인 사건

$$A=\{2, 3, 5\},\ B=\{1, 3, 5\},\ A\cap B=\{3, 5\}$$

$$\mathrm{P}(A)=\frac{3}{6}=\frac{1}{2}$$

$$\mathrm{P}(A\cap B)=\frac{2}{6}=\frac{1}{3}$$

$$\mathrm{P}(B\mid A)=\frac{\mathrm{P}(A\cap B)}{\mathrm{P}(A)}$$

= 2/3

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

확률의 곱셈 정리

조건부 확률을 이용하여, 두 사건 A & B가 “동시”에 일어날 확률을 쉽게 계산할 수 있다!

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

확률의 곱셈 정리

조건부 확률을 이용하여, 두 사건 A & B가 “동시”에 일어날 확률을 쉽게 계산할 수 있다!

$$\mathrm{P}(B|A)=\frac{\mathrm{P}(A\cap B)}{\mathrm{P}(A)} \quad \Rightarrow \quad \boxed{\mathrm{P}(A\cap B)=\mathrm{P}(A)\mathrm{P}(B|A)}$$

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

확률의 곱셈 정리

조건부 확률을 이용하여, 두 사건 A & B가 “동시”에 일어날 확률을 쉽게 계산할 수 있다!

$$P(B|A)=\frac{P(A\cap B)}{P(A)} \rightarrow \boxed{P(A\cap B)=P(A)P(B|A)}$$

$$P(A|B)=\frac{P(A\cap B)}{P(B)} \rightarrow \boxed{P(A\cap B)=P(B)P(A|B)}$$

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

10개의 복권 중, 3개의 당첨 & 7개의 꽝이 있다. 철수와 영수가 차례로 1개씩 뽑을 때, 두 사람 모두 당첨될 확률은?

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

10개의 복권 중, 3개의 당첨 & 7개의 꽝이 있다. 철수와 영수가 차례로 1개씩 뽑을 때, 두 사람 모두 당첨될 확률은?

- A : 철수가 당첨되는 사건
- B : 영수가 당첨되는 사건

Hanwha Ocean

4. 통계학의 기초

5. 조건부 확률

Quiz

10개의 복권 중, 3개의 당첨 & 7개의 꽝이 있다. 철수와 영수가 차례로 1개씩 뽑을 때, 두 사람 모두 당첨될 확률은?

- A : 철수가 당첨되는 사건
- B : 영수가 당첨되는 사건

$$\mathrm{P}(A)=\frac{3}{10},\ \mathrm{P}(B|A)=\frac{2}{9}$$

$$\mathrm{P}(A\cap B)=\mathrm{P}(A)\mathrm{P}(B|A)=\frac{3}{10}\times\frac{2}{9}=\frac{1}{15}$$

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

두 사건이 독립이다 = 두 사건이 무관하다

두 사건이 종속이다 = 두 사건이 유관하다

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

두 사건이 독립이다 = 두 사건이 무관하다

두 사건이 종속이다 = 두 사건이 유관하다

관련이 있다

= ***서로가 발생할 확률에 영향을 끼친다!***

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

$$P(B|A)=P(B)$$

일때, 두 사건 A와 B는 서로 독립이다.

만약 위 식을 만족하지 않으면, 두 사건은 종속이다. (= 서로 연관이 있다)

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

두 사건이 독립이라고 하면, 확률의 곱셈 정리에 의해

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A) = \mathrm{P}(A)\mathrm{P}(B)$$

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

두 사건이 독립이라고 하면, 확률의 곱셈 정리에 의해

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

역으로 $P(A) \neq 0$이고 $P(A \cap B) = P(A)P(B)$이면

$$P(B) = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

=> 따라서, 두 사건은 독립이다!

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

두 사건이 독립이라고 하면, 확률의 곱셈 정리에 의해

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

역으로 $P(A) \neq 0$이고 $P(A \cap B) = P(A)P(B)$이면

$$P(B) = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

=> 따라서, 두 사건은 독립이다!

두 사건 A와 B가 서로 독립이기 위한 필요충분조건은

$$P(A \cap B) = P(A)P(B) \text{ (단, } P(A) \neq 0, P(B) \neq 0)$$

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

Quiz

주사위 1개를 던져서,

- A : 홀수가 나온 사건
- B : 2 이하가 나온 사건
- C : 소수인 사건

Q1) A & B는 서로 독립? 종속?

Q2) A & C는 서로 독립? 종속?

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

Quiz

주사위 1개를 던져서,

- A : 홀수가 나온 사건
- B : 2 이하가 나온 사건
- C : 소수인 사건

Q1) A & B는 서로 독립? 종속?

Q2) A & C는 서로 독립? 종속?

$A=\{1, 3, 5\}$, $B=\{1, 2\}$, $C=\{2, 3, 5\}$이므로

$$\mathrm{P}(A)=\frac{1}{2},\ \mathrm{P}(B)=\frac{1}{3},\ \mathrm{P}(C)=\frac{1}{2}$$

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

Quiz

주사위 1개를 던져서,

- A : 홀수가 나온 사건
- B : 2 이하가 나온 사건
- C : 소수인 사건

Q1) A & B는 서로 독립? 종속?

Q2) A & C는 서로 독립? 종속?

$A=\{1, 3, 5\}$, $B=\{1, 2\}$, $C=\{2, 3, 5\}$이므로

$$\mathrm{P}(A)=\frac{1}{2},\ \mathrm{P}(B)=\frac{1}{3},\ \mathrm{P}(C)=\frac{1}{2}$$

(1) $A\cap B=\{1\}$이므로 $\mathrm{P}(A\cap B)=\frac{1}{6}$

이때 $\mathrm{P}(A)\mathrm{P}(B)=\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$이므로 $\mathrm{P}(A\cap B)=\mathrm{P}(A)\mathrm{P}(B)$

따라서 두 사건 A와 B는 서로 독립이다.

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

Quiz

주사위 1개를 던져서,

- A : 홀수가 나온 사건
- B : 2 이하가 나온 사건
- C : 소수인 사건

Q1) A & B는 서로 독립? 종속?

Q2) A & C는 서로 독립? 종속?

$A=\{1, 3, 5\}$, $B=\{1, 2\}$, $C=\{2, 3, 5\}$이므로

$$\mathrm{P}(A)=\frac{1}{2},\ \mathrm{P}(B)=\frac{1}{3},\ \mathrm{P}(C)=\frac{1}{2}$$

(1) $A\cap B=\{1\}$이므로 $\mathrm{P}(A\cap B)=\frac{1}{6}$

이때 $\mathrm{P}(A)\mathrm{P}(B)=\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$이므로 $\mathrm{P}(A\cap B)=\mathrm{P}(A)\mathrm{P}(B)$

따라서 두 사건 A와 B는 서로 독립이다.

(2) $A\cap C=\{3, 5\}$이므로 $\mathrm{P}(A\cap C)=\frac{1}{3}$

이때 $\mathrm{P}(A)\mathrm{P}(C)=\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$이므로 $\mathrm{P}(A\cap C)\neq\mathrm{P}(A)\mathrm{P}(C)$

따라서 두 사건 A와 C는 서로 종속이다.

Hanwha Ocean

4. 통계학의 기초

6. 사건의 독립과 종속

독립 & 종속

$$P(B|A)=P(B)$$

일때, 두 사건 A와 B는 서로 독립이다.

만약 위 식을 만족하지 않으면, 두 사건은 종속이다. (= 서로 연관이 있다)

두 사건 A와 B가 서로 독립이기 위한 필요충분조건은

$$P(A\cap B)=P(A)P(B) \text{ (단, } P(A)\neq 0, P(B)\neq 0)$$

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

확률 변수

예시) 1개의 동전을 2번 던지는 시행

- 앞면 : H
- 뒷면 : T
- 앞면이 나오는 횟수 : X

(1) 표본 공간 S = { HH, HT, TH, TT }

(2) X가 가질 수 있는 값은? 0, 1, 2

Hanwha Ocean

4. 통계학의 기초

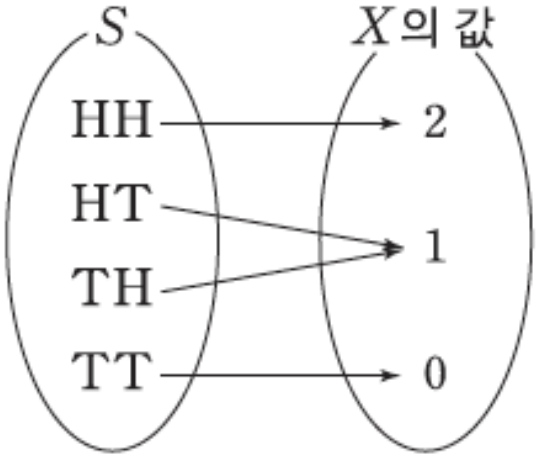
7. 확률 분포

확률 변수

(2) X가 가질 수 있는 값은? 0, 1, 2

- 각각의 확률은?

$$HH \longrightarrow 2,\ HT \longrightarrow 1,\ TH \longrightarrow 1,\ TT \longrightarrow 0$$



X가 0, 1, 2의 값을 가질 확률은 각각 $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

확률 변수

확률 변수 = 표본 공간 S의 각 원소에, 단 하나의 실수가 대응되는 함수 $\mathrm{P}(X=x)$

$$\mathrm{HH} \longrightarrow 2,\ \mathrm{HT} \longrightarrow 1,\ \mathrm{TH} \longrightarrow 1,\ \mathrm{TT} \longrightarrow 0$$

$$\mathrm{P}(X=0)=\frac{1}{4},\ \mathrm{P}(X=1)=\frac{1}{2},\ \mathrm{P}(X=2)=\frac{1}{4}$$

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

확률 변수의 종류

(1) 이산 확률변수

(2) 연속 확률변수

이산적(discrete) : 0, 1, 2, … 100, 150, ..

연속적(continuous) : 0.5,0.51,0.52, 0.521….

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

Quiz

다음 중 이산/연속 확률 변수는?

1. 학급 내 학생들의 키
2. 손목시계를 차고 온 학생의 수
3. 100미터 달리는데 걸린 시간

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

이산 확률변수

확률 변수가 가질 수 있는 값이 유한개 (or 셀 수 있을 때)

- ex) 지구의 인구 수
- ex) 이번 분기에 생산한 제품의 수

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

이산 확률변수

확률 변수가 가질 수 있는 값이 유한개 (or 셀 수 있을 때)

- ex) 지구의 인구 수
- ex) 이번 분기에 생산한 제품의 수

[확률 질량함수]

$$\mathrm{P}(X=x_i)=p_i \ (i=1,\ 2,\ \cdots,\ n)$$

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

이산 확률변수

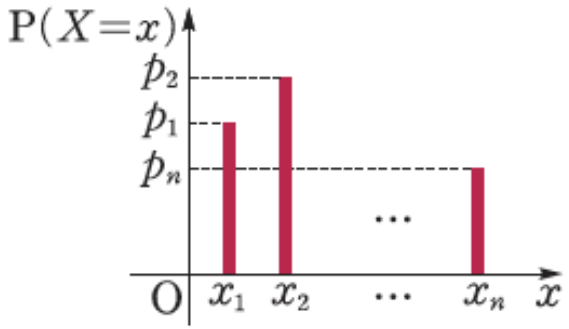
확률 변수가 가질 수 있는 값이 유한개 (or 셀 수 있을 때)

- ex) 지구의 인구 수
- ex) 이번 분기에 생산한 제품의 수

[확률 질량함수]

$$\mathrm{P}(X=x_i)=p_i \ (i=1, 2, \cdots, n)$$

X	x_1	x_2	$\cdots$	x_n	합계
$\mathrm{P}(X=x_i)$	p_1	p_2	$\cdots$	p_n	1



Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

이산 확률변수

확률 질량함수의 성질

이산확률변수 X의 확률질량함수 $\mathrm{P}(X=x_i)=p_i$ $(i=1, 2, \cdots, n)$에 대하여

❶ $0 \le p_i \le 1$

❷ $p_1+p_2+\cdots+p_n=1$

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

Quiz

어떤 이산확률변수 X의 확률 질량함수가 다음과 같을 때 …

$$P(X=x)=\begin{cases}\frac{1}{7} & (x=-2,\ 1,\ 2)\\ \frac{2}{7} & (x=-1,\ 0)\end{cases}$$

Q1) X의 확률분포를 표로 구하기

Q2) P(0<= X <= 1)은?

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

Quiz

어떤 이산확률변수 X의 확률 질량함수가 다음과 같을 때 …

$$P(X=x)=\begin{cases}\frac{1}{7} & (x=-2,\ 1,\ 2)\\ \frac{2}{7} & (x=-1,\ 0)\end{cases}$$

X	-2	-1	0	1	2	합계
$P(X=x)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	1

Q1) X의 확률분포를 표로 구하기

Q2) P(0<= X <= 1)은? 3/7

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

연속 확률변수

연속 확률변수 = 어떤 범위에 속하는 모든 실수의 값을 가지는 확률변수

예시) 배차 간격이 5분인 지하철을 기다리는 시간 = 확률 변수 X

- Q1) X가 가질 수 있는 범위는?
- Q2) X가 2 이하의 값을 가질 확률은?

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

연속 확률변수

연속 확률변수 = 어떤 범위에 속하는 모든 실수의 값을 가지는 확률변수

예시) 배차 간격이 5분인 지하철을 기다리는 시간 = 확률 변수 X

- Q1) X가 가질 수 있는 범위는? 0 ~ 5분
- Q2) X가 2 이하의 값을 가질 확률은?

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

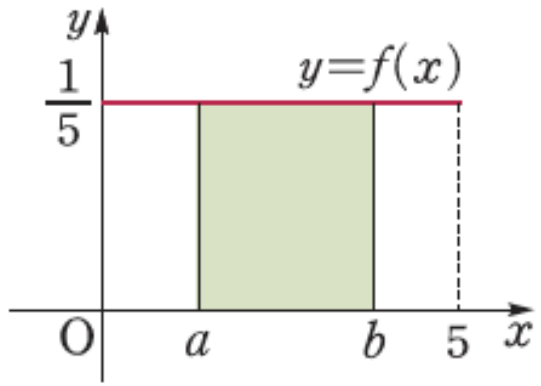
연속 확률변수

연속 확률변수 = 어떤 범위에 속하는 모든 실수의 값을 가지는 확률변수

예시) 배차 간격이 5분인 지하철을 기다리는 시간 = 확률 변수 X

- Q1) X가 가질 수 있는 범위는? 0 ~ 5분
- Q2) X가 2 이하의 값을 가질 확률은?

$$P(a \le X \le b) = \frac{b-a}{5} \quad (\text{단, } 0 \le a \le b \le 5)$$



넓이가 곧 확률이다!

즉, 넓이의 총 합은 1이다

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

연속 확률변수

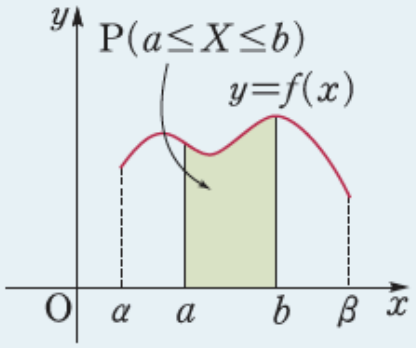
확률 밀도 함수

- 다음의 3가지 조건을 만족하는 함수

❶ $f(x) \geq 0$

❷ $y=f(x)$의 그래프와 x축 및 두 직선 $x=\alpha$, $x=\beta$로 둘러싸인 도형의 넓이는 1이다.

❸ $P(a \leq X \leq b)$는 $y=f(x)$의 그래프와 x축 및 두 직선 $x=a$, $x=b$로 둘러싸인 도형의 넓이와 같다. (단, $\alpha \leq a \leq b \leq \beta$)



Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

연속 확률변수

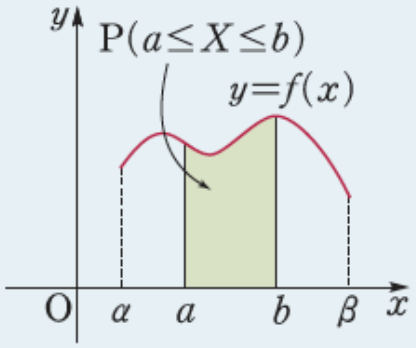
확률 밀도 함수

- 다음의 3가지 조건을 만족하는 함수

❶ $f(x) \geq 0$

❷ $y=f(x)$의 그래프와 x축 및 두 직선 $x=\alpha$, $x=\beta$로 둘러싸인 도형의 넓이는 1이다.

❸ $\mathrm{P}(a \leq X \leq b)$는 $y=f(x)$의 그래프와 x축 및 두 직선 $x=a$, $x=b$로 둘러싸인 도형의 넓이와 같다. (단, $\alpha \leq a \leq b \leq \beta$)



주의 :

- **이산** 확률변수의 확률 **질량** 함수에서는 **"높이"**가 확률
- **연속** 확률변수의 확률 **밀도** 함수에서는 **"넓이"**가 확률

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

Quiz

연속확률변수 X의 확률 밀도함수가 $f(x)=kx\ (0\leq x\leq 4)$ 일 때 ..

Q1) k의 값은?

Q2) P(2<=X<=4)는?

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

Quiz

연속확률변수 X의 확률 밀도함수가 $f(x)=kx\ (0\le x\le 4)$ 일 때 ..

Q1) k의 값은?

Q2) P(2<=X<=4)는?

(1) $f(x)=kx$의 그래프와 x축 및 직선 $x=4$로 둘러싸인 삼각형의 넓이가 1이므로

$$\frac{1}{2}\times 4\times 4k=1, \qquad k=\frac{1}{8}$$

Hanwha Ocean

4. 통계학의 기초

7. 확률 분포

Quiz

연속확률변수 X의 확률 밀도함수가 $f(x)=kx\ (0\le x\le 4)$ 일 때 ..

Q1) k의 값은?

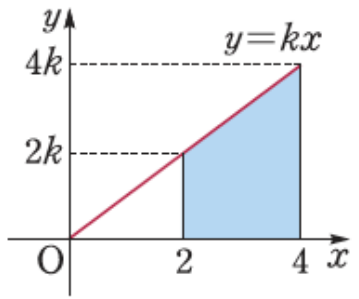
Q2) P(2<=X<=4)는?

(1) $f(x)=kx$의 그래프와 x축 및 직선 $x=4$로 둘러싸인 삼각형의 넓이가 1이므로

$$\frac{1}{2}\times 4\times 4k=1, \qquad k=\frac{1}{8}$$

(2) $\mathrm{P}(2\le X\le 4)$는 오른쪽 그림의 색칠한 사다리꼴의 넓이와 같으므로

$$\mathrm{P}(2\le X\le 4)=\frac{1}{2}\times\left(\frac{1}{4}+\frac{1}{2}\right)\times 2=\frac{3}{4}$$



5. 상관관계 분석

Hanwha Ocean

5. 상관관계 분석

상관관계 분석이란?

변수들 사이에 어떤 관계가 있는지 확인

Hanwha Ocean

5. 상관관계 분석

상관관계 분석이란?

변수들 사이에 어떤 관계가 있는지 확인

- ex) **키**와 **몸무게** 사이에는 어떤 관련이 있을까?
- ex) **성별**에 따른 **학업 성적**에는 어떤 차이가 있을까?
- ex) **직업**에 따라 **거주 지역**에 어떤 차이가 있을까?

Hanwha Ocean

5. 상관관계 분석

상관관계 분석이란?

변수들 사이에 어떤 관계가 있는지 확인

- ex) **키**와 **몸무게** 사이에는 어떤 관련이 있을까?
- ex) **성별**에 따른 **학업 성적**에는 어떤 차이가 있을까?
- ex) **직업**에 따라 **거주 지역**에 어떤 차이가 있을까?

빨간색 : 수치형 자료

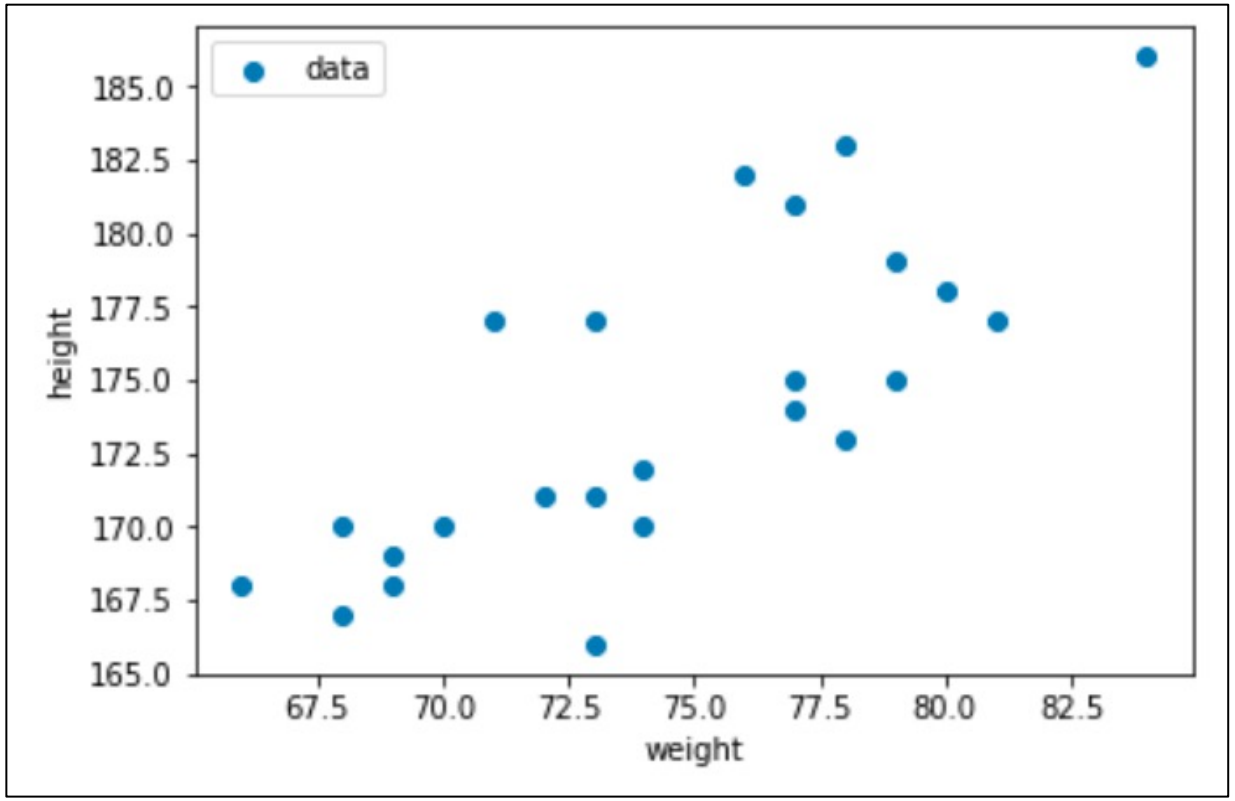
파란색 : 범주형 자료

Hanwha Ocean

5. 상관관계 분석

1. 수치형 & 수치형

- **산점도 (scatter plot)**를 통해 시각화
- 예시) 아이스크림 판매량과 기온의 관계



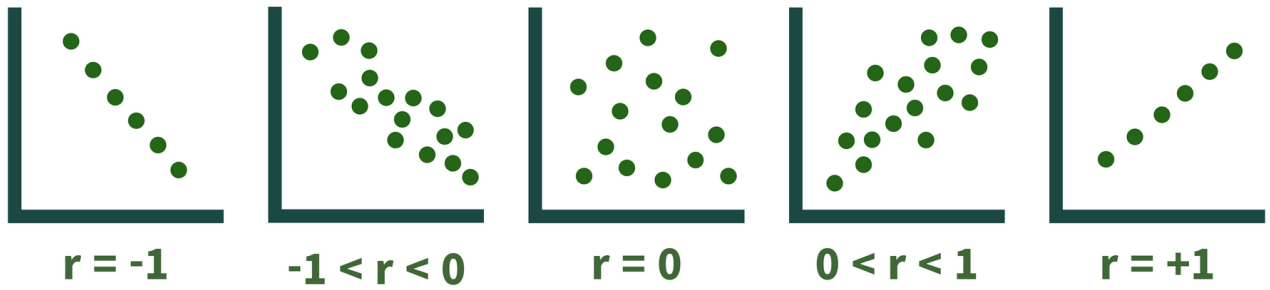
https://1.bp.blogspot.com/-FHcYVT-szGw/XIDIj5mFvpI/AAAAAAAAU2A/ThiIl6eiDXwQ52M8mgBHO0cOPKRxBxzxQCLcBGAs/s1600/%25E1%2584%2589%25E1%2585%25B3%25E1%2584%258F%25E1%2585%25B3%25E1%2584%2585%25E1%2585%25B5%25E1%2586%25AB%25E1%2584%2589%25E1%2585%25A3%25E1%2586%25BA%2B2019-03-07%2B%25E1%2584%258B%25E1%2585%25A9%25E1%2584%2592%25E1%2585%25AE%2B4.30.11.png

Hanwha Ocean

5. 상관관계 분석

1. 수치형 & 수치형

- **상관계수 (Correlation Coefficient)**
  - 범위 : -1 ~ 1
  - 1에 가까울수록 양의 상관관계
  - -1에 가까울수록 음의 상관관계
  - 0에 가까울수록 상관관계 없음



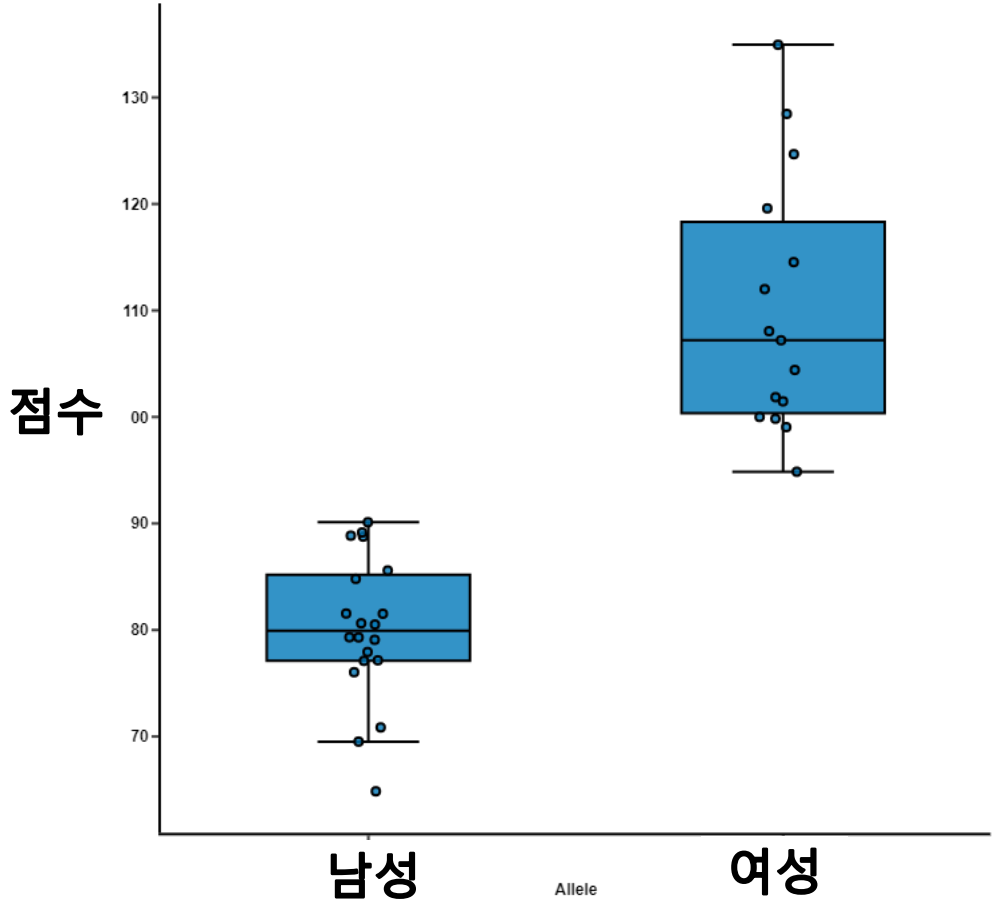
https://cdn.imweb.me/upload/S202101041a4e45576971e/a29a57bd847b1.png

Hanwha Ocean

5. 상관관계 분석

2. 수치형 & 범주형

- 범주 별로 수치형 변수 값들의 평균을 살펴보기
- 예시) 성별 상자그림



https://blog.bioturing.com/wp-content/uploads/2018/11/parallel-non-overlapped-box-plots.png

Hanwha Ocean

5. 상관관계 분석

3. 범주형 & 범주형

- 교차분석표를 통해 변수 간의 관계 파악
- 예시) 연령대 별로 주요 교통수단 비교

	10대	20대	30대	Total
지하철	57	45	32	134
버스	43	40	37	120
자차	0	15	31	46
Total	100	100	100	300

https://wikidocs.net/161870#central-tendency

Hanwha Ocean

5. 상관관계 분석

상관관계 vs 인과관계

상관관계가 인과관계를 보장하지는 않는다!

Hanwha Ocean

5. 상관관계 분석

상관관계 vs 인과관계

상관관계가 인과관계를 보장하지는 않는다!

예시) 아이스크림과 맥주의 판매량이 양의 상관관계

- 아이스크림을 많이 먹어서, 맥주가 많이 팔린것인가? (X)
- 맥주를 많이 먹어서, 아이스크림이 많이 팔린것인가? (X)

Hanwha Ocean

5. 상관관계 분석

상관관계 vs 인과관계

상관관계가 인과관계를 보장하지는 않는다!

예시) 아이스크림과 맥주의 판매량이 양의 상관관계

- 아이스크림을 많이 먹어서, 맥주가 많이 팔린것인가? (X)
- 맥주를 많이 먹어서, 아이스크림이 많이 팔린것인가? (X)

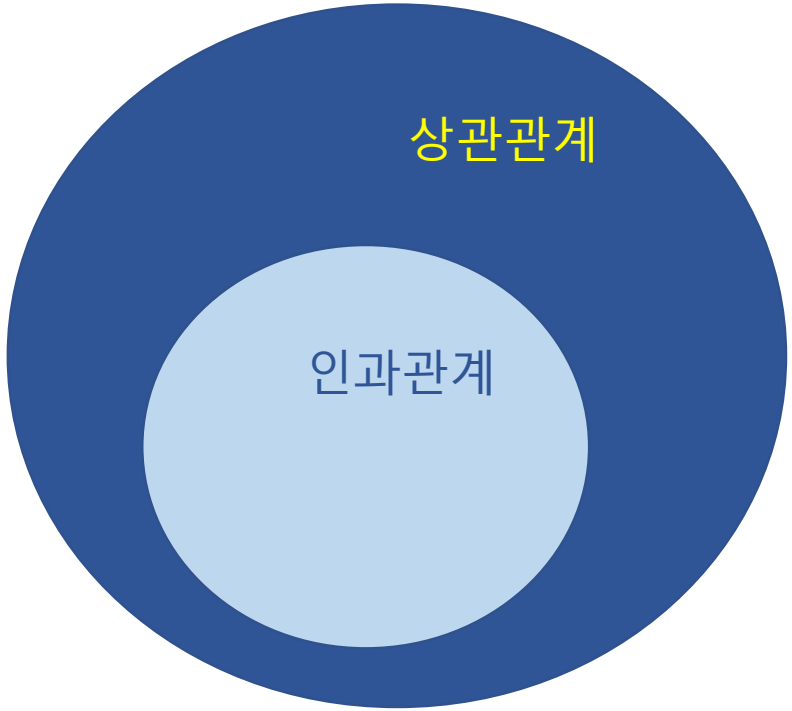
날이 더운 여름이라서, 아이스크림과 맥주가 모두 같이 많이 팔린것일 뿐, 인과성은 없다!

Hanwha Ocean

5. 상관관계 분석

상관관계 vs 인과관계

상관관계가 인과관계를 보장하지는 않는다!



예시) 아이스크림과 맥주의 판매량이 양의 상관관계

- 아이스크림을 많이 먹어서, 맥주가 많이 팔린것인가? (X)
- 맥주를 많이 먹어서, 아이스크림이 많이 팔린것인가? (X)

날이 더운 여름이라서, 아이스크림과 맥주가 모두 같이 많이 팔린것일 뿐, 인과성은 없다!

6. 회귀 분석

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

하나 혹은 그 이상의 변수를 사용하여, 다른 변수를 예측하는 분석법

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

하나 혹은 그 이상의 변수를 사용하여, 다른 변수를 예측하는 분석법

예시)

- 기온,습도,풍향 등을 사용하여, 미세먼지 농도 예측하기
- 학업 시간 및 성별을 사용하여, 시험 성적 예측하기

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

하나 혹은 그 이상의 변수를 사용하여, 다른 변수를 예측하는 분석법

예시)

- 기온,습도,풍향 등을 사용하여, 미세먼지 농도 예측하기
- 학업 시간 및 성별을 사용하여, 시험 성적 예측하기

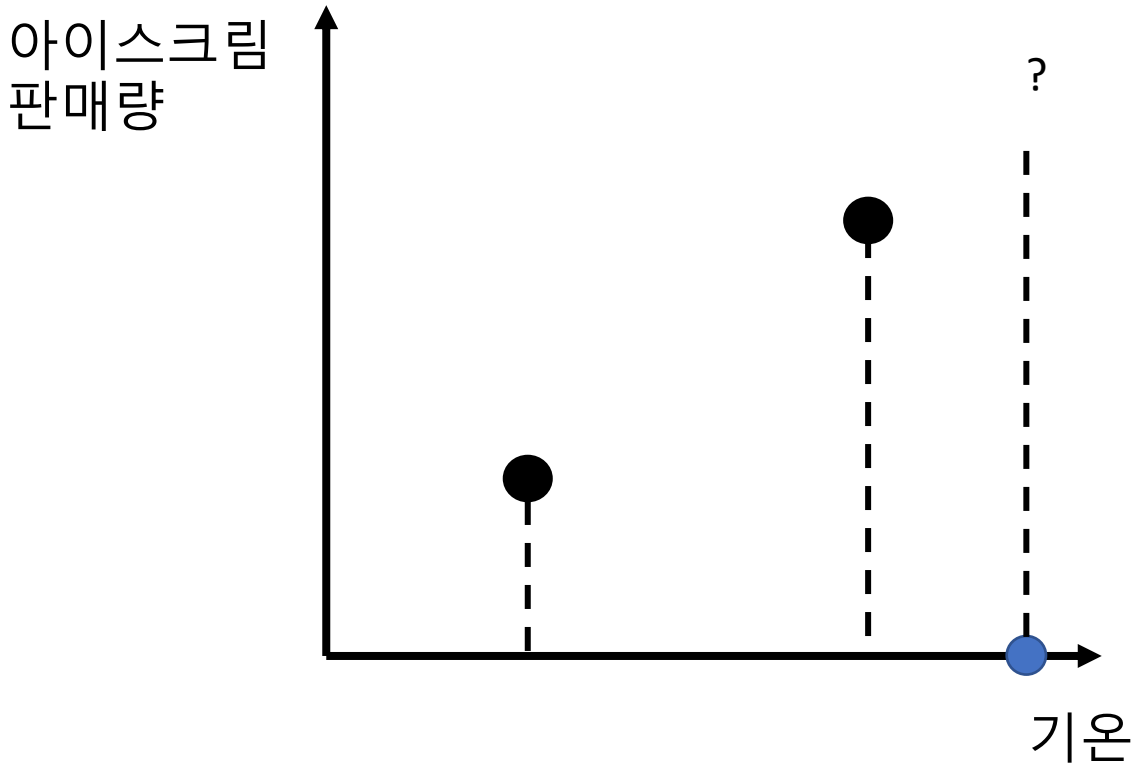
독립 변수 (X) : 예측을 위해 사용하는 변수

종속 변수 (Y) : 예측을 하고자 하는 변수

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

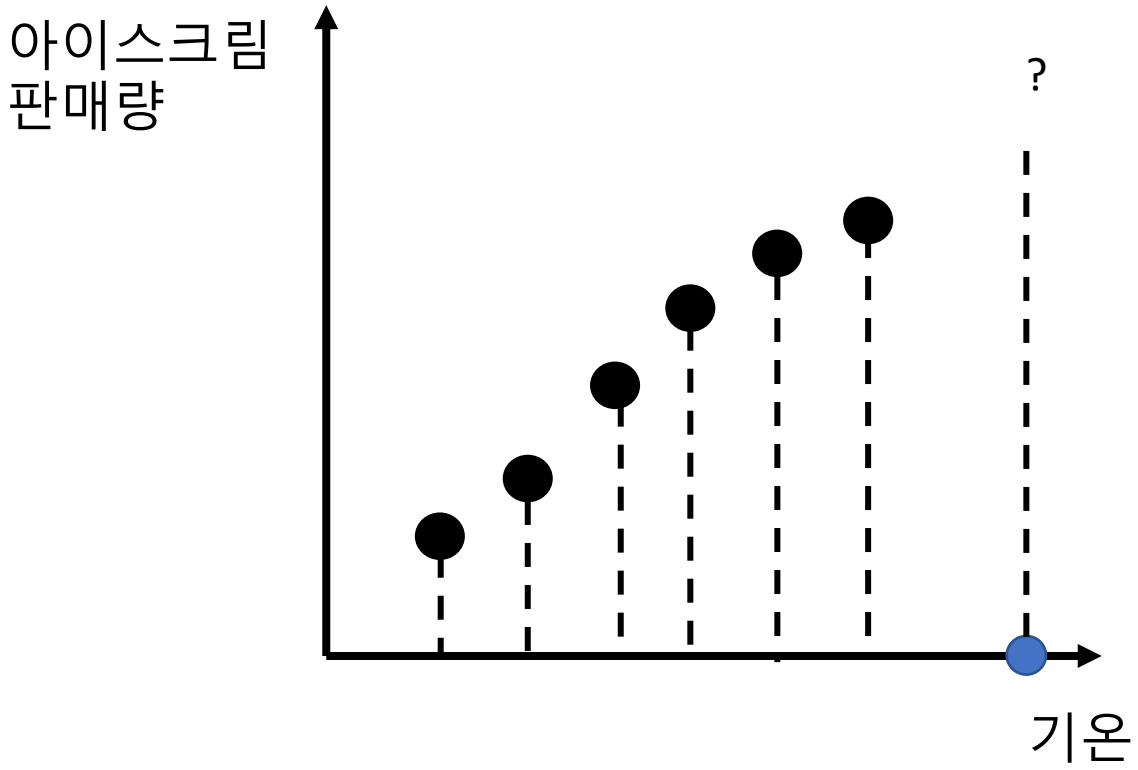


기온 (X)	아이스크림 판매량 (Y)
5	25
15	75
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

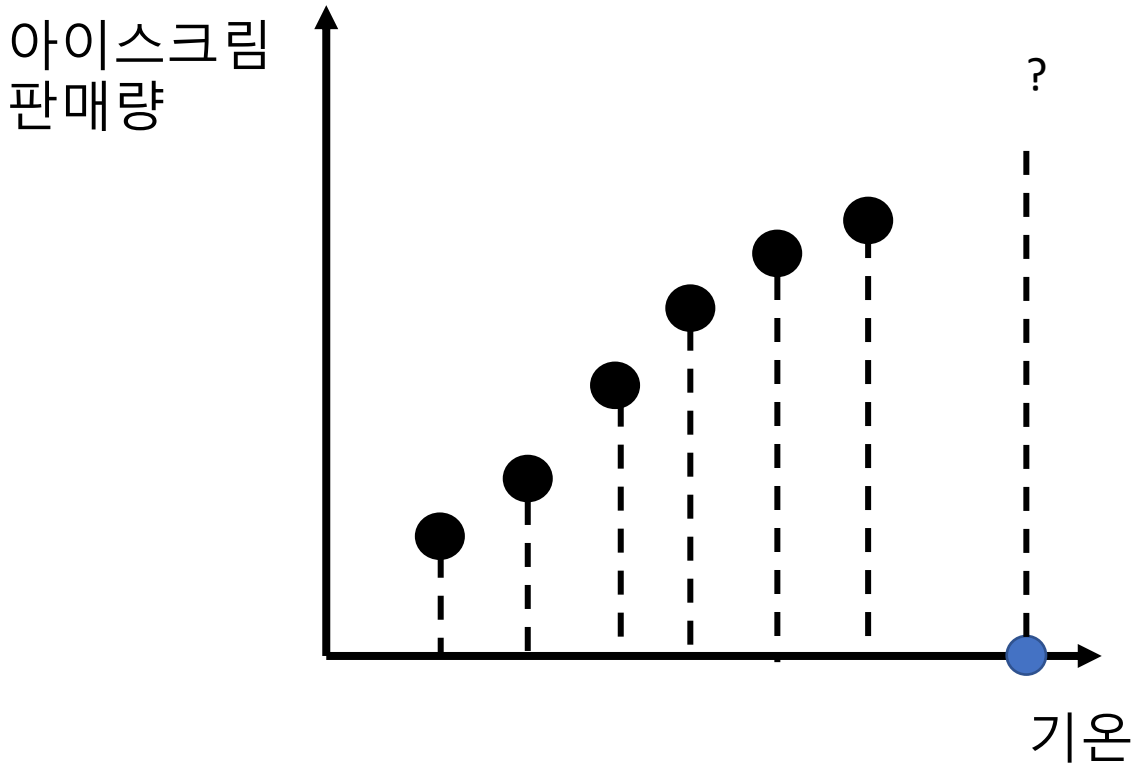


기온 (X)	아이스크림 판매량 (Y)
3	15
5	25
7	35
10	50
12	60
15	75
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?



기온 (X)	아이스크림 판매량 (Y)
3	15
5	25
7	35
10	50
12	60
15	75
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

아이스크림 판매량은 기온의 5 배구나!

$$Y = 5X$$

기온 (X)	아이스크림 판매량 (Y)
3	15
5	25
7	35
10	50
12	60
15	75
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

아이스크림 판매량은 기온의 5 배 +5 구나!

$$Y = 5X + 5$$

기온 (X)	아이스크림 판매량 (Y)
3	20
5	30
7	40
10	55
12	65
15	80
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

아이스크림 판매량은 기온의 5 배 +5 구나!

$$Y = 5X + 5$$

기울기 절편

이 둘을 통합하여, **계수 (coefficient)** 라고 부른다

회귀분석은, 이러한 계수를 찾는 분석이다!

기온 (X)	아이스크림 판매량 (Y)
3	20
5	30
7	40
10	55
12	65
15	80
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

아이스크림 판매량은 기온의 5 배 +5 구나!

$$Y = 5X + 5$$

기울기 절편

Question) 꼭 기온 만을 사용해야하나?

기온 (X)	아이스크림 판매량 (Y)
3	20
5	30
7	40
10	55
12	65
15	80
20	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

강수량 정보도 함께 사용하여 예측해보면?

기온 (X1)	강수량 (X2)	아이스크림 판매량 (Y)
3	2	16
5	1	28
7	4	32
10	3	49
12	1	63
15	2	76
20	1	??

Hanwha Ocean

6. 회귀 분석

회귀분석이란?

아이스크림 판매량은

기온의 5 배 – 강수량의 2 배 + 5 구나!

$$Y = 5X1 - 2X2 + 5$$

5: 기울기 / 2: 기울기 / 5: 절편

기온 (X1)	강수량 (X2)	아이스크림 판매량 (Y)
3	2	16
5	1	28
7	4	32
10	3	49
12	1	63
15	2	76
20	1	??

Hanwha Ocean

6. 회귀 분석

계수 추정

계수 (기울기, 절편)을 찾는 방법은?

Hanwha Ocean

6. 회귀 분석

계수 추정

계수 (기울기, 절편)을 찾는 방법은?

=> 오차를 최소화하는 계수를 찾는다.

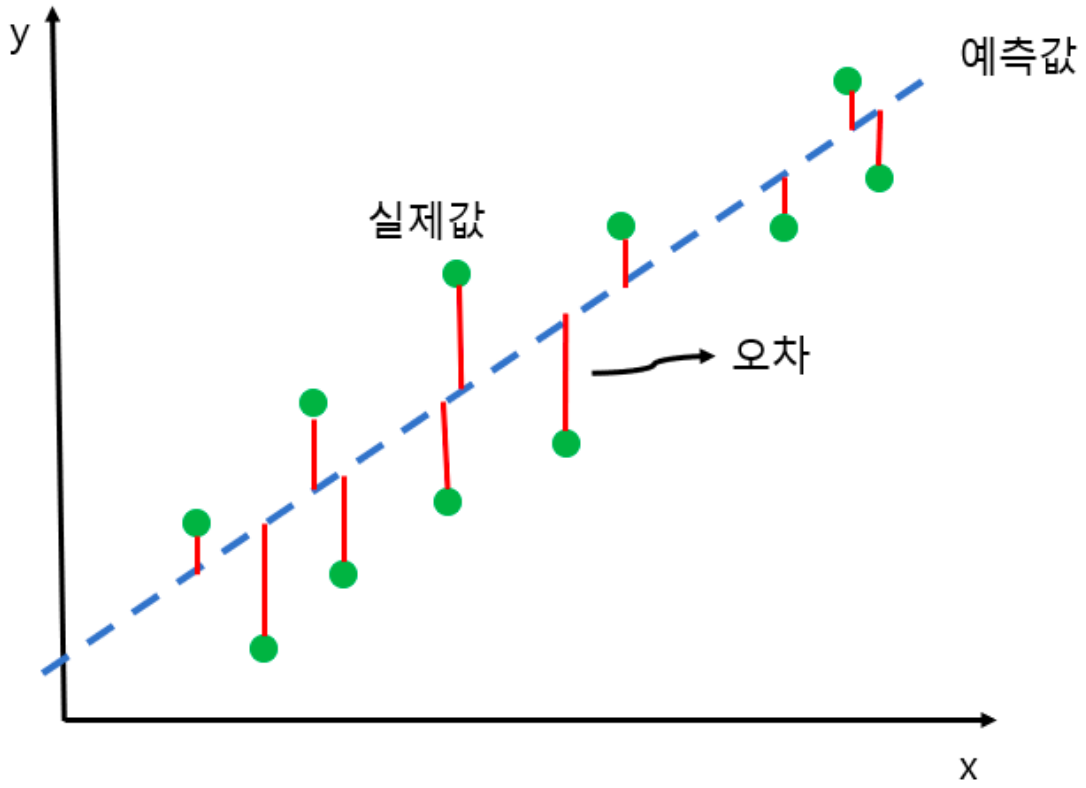
Hanwha Ocean

6. 회귀 분석

계수 추정

계수 (기울기, 절편)을 찾는 방법은?

=> 오차를 최소화하는 계수를 찾는다.



https://curriculum.cosadama.com/machine-learning/3-2/residual2.png

Hanwha Ocean

6. 회귀 분석

요약

회귀분석은 특정 변수(X)를 사용하여, 다른 특정 변수(Y)를 예측하는 분석 방법이다.

- X : 독립 변수
- Y : 종속 변수
- Y = AX + B
  - A : 기울기
  - B : 절편

  (A & B를 통틀어서 계수 라고 부른다)

Hanwha Ocean

6. 회귀 분석

요약

회귀분석은 특정 변수(X)를 사용하여, 다른 특정 변수(Y)를 예측하는 분석 방법이다.

- X : 독립 변수
- Y : 종속 변수
- Y = AX + B
  - A : 기울기
  - B : 절편

  (A & B를 통틀어서 계수 라고 부른다)

독립 변수가
- 1개인 경우 : 단순 회귀분석
- 2개 이상인 경우 : 다중 회귀분석

종속 변수가
- 1개인 경우 : 일변량 회귀분석
- 2개 이상인 경우 : 다변량 회귀분석

7. 시계열 분석

Hanwha Ocean

7. 시계열 분석

1. 시계열 데이터란
2. 시계열 데이터 분석의 목적
3. 시계열 분해
4. 이동평균법 & 지수평활법

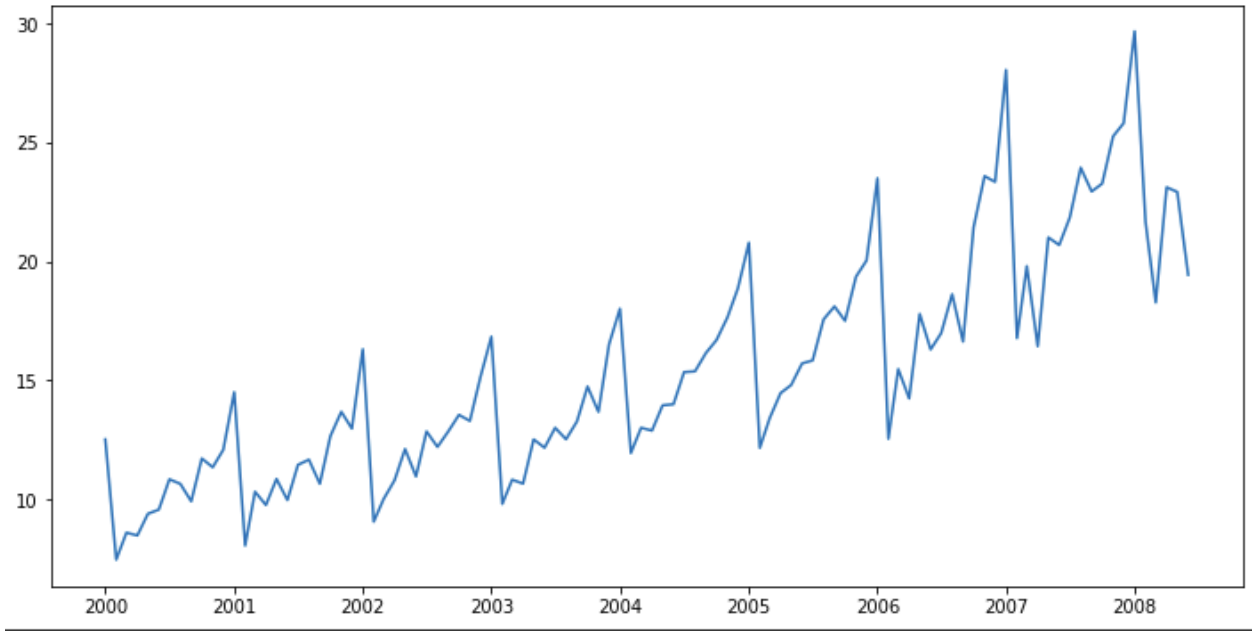
Hanwha Ocean

7. 시계열 분석

1. 시계열 (Time Series) 데이터란

시간에 따라 기록된 데이터

- 예시) 일별 평균 기온, 월별 생산량, 년별 매출 등



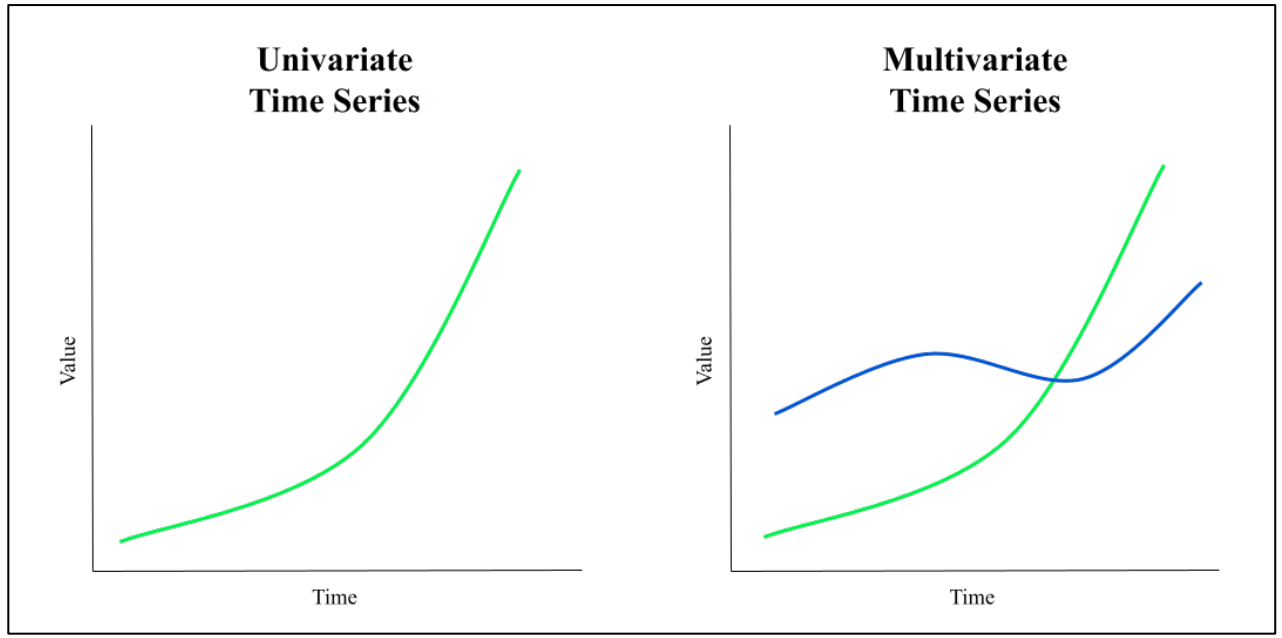
Hanwha Ocean

7. 시계열 분석

1. 시계열 (Time Series) 데이터란

시계열 데이터의 종류

- 1) **단변량 (Univariate)** 시계열
- 2) **다변량 (Multivariate)** 시계열



https://webimages.mongodb.com/_com_assets/cms/kz333w53m4bk2i8ma-image5.png?auto=format%252Ccompress

Hanwha Ocean

7. 시계열 분석

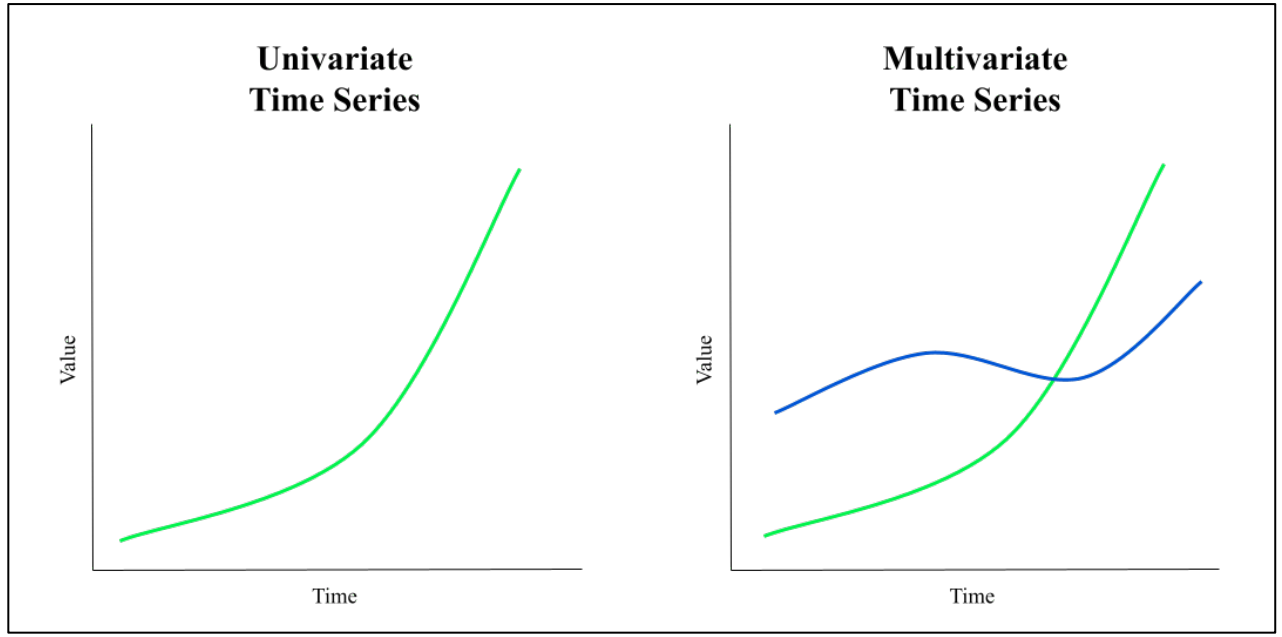
1. 시계열 (Time Series) 데이터란

시계열 데이터의 종류

- 1) **단변량 (Univariate)** 시계열
- 2) **다변량 (Multivariate)** 시계열

예시) 공장 내의 센서에서 매 시간마다 측정하는 대상이

- 1개일 경우 : 단변량
- 여러 개일 경우 : 다변량



https://webimages.mongodb.com/_com_assets/cms/kz333w53m4bk2i8ma-image5.png?auto=format%252Ccompress

Hanwha Ocean

7. 시계열 분석

2. 시계열 데이터 분석의 목적

크게 세 가지 태스크로 구분

(1) 시계열 예측 (Time Series Forecasting)

(2) 시계열 분류 (Time Series Forecasting)

(3) 시계열 이상치 탐지 (Time Series Anomaly Detection)

Hanwha Ocean

7. 시계열 분석

2. 시계열 데이터 분석의 목적

(1) 시계열 예측 (Time Series Forecasting)

- 과거 시점의 정보를 사용하여, 미래를 예측하는 것

Hanwha Ocean

7. 시계열 분석

2. 시계열 데이터 분석의 목적

(1) 시계열 예측 (Time Series Forecasting)

- 과거 시점의 정보를 사용하여, 미래를 예측하는 것
- 예시) 삼성전자의 과거 15일의 주가를 활용하여, 미래 5일의 주가를 예측

https://i.stack.imgur.com/I6hrO.png

Hanwha Ocean

7. 시계열 분석

2. 시계열 데이터 분석의 목적

(2) 시계열 분류 (Time Series Classification)

- 시계열 전체 정보를 활용하여, 속하는 클래스를 분류

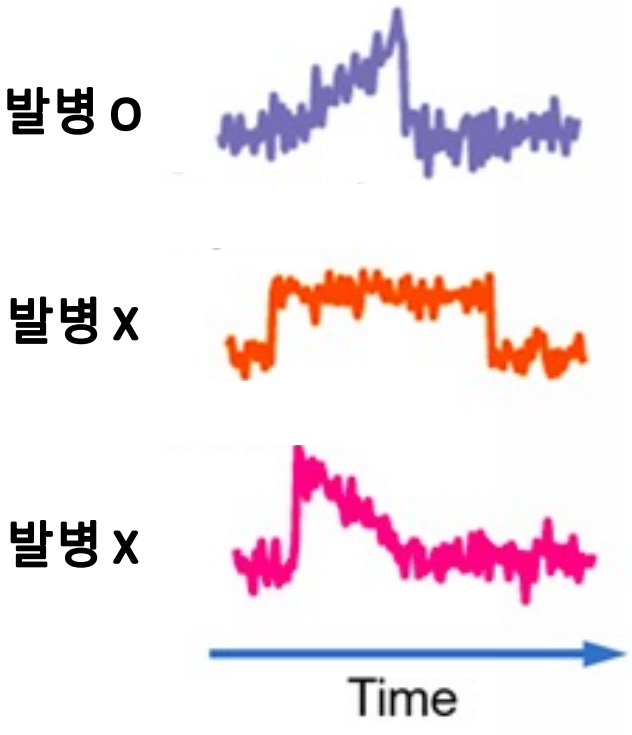
Hanwha Ocean

7. 시계열 분석

2. 시계열 데이터 분석의 목적

(2) 시계열 분류 (Time Series Classification)

- 시계열 전체 정보를 활용하여, 속하는 클래스를 분류
- 예시) 환자들의 생체 정보를 기록한 뒤, 발병 여부 (1/0) 예측



https://developersbay.se/wp-content/uploads/2020/11/image.png

Hanwha Ocean

7. 시계열 분석

2. 시계열 데이터 분석의 목적

(3) 시계열 이상치 탐지 (Time Series Anomaly Detection)

- 시계열 데이터의 특정 시점이 이상치인지 점검

https://miro.medium.com/v2/resize:fit:850/1*oxvTvzn0fztqoNlnZkK71g.png

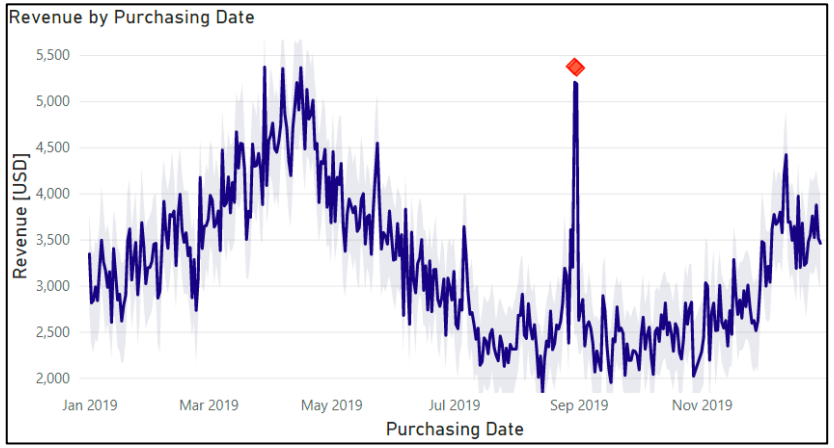
Hanwha Ocean

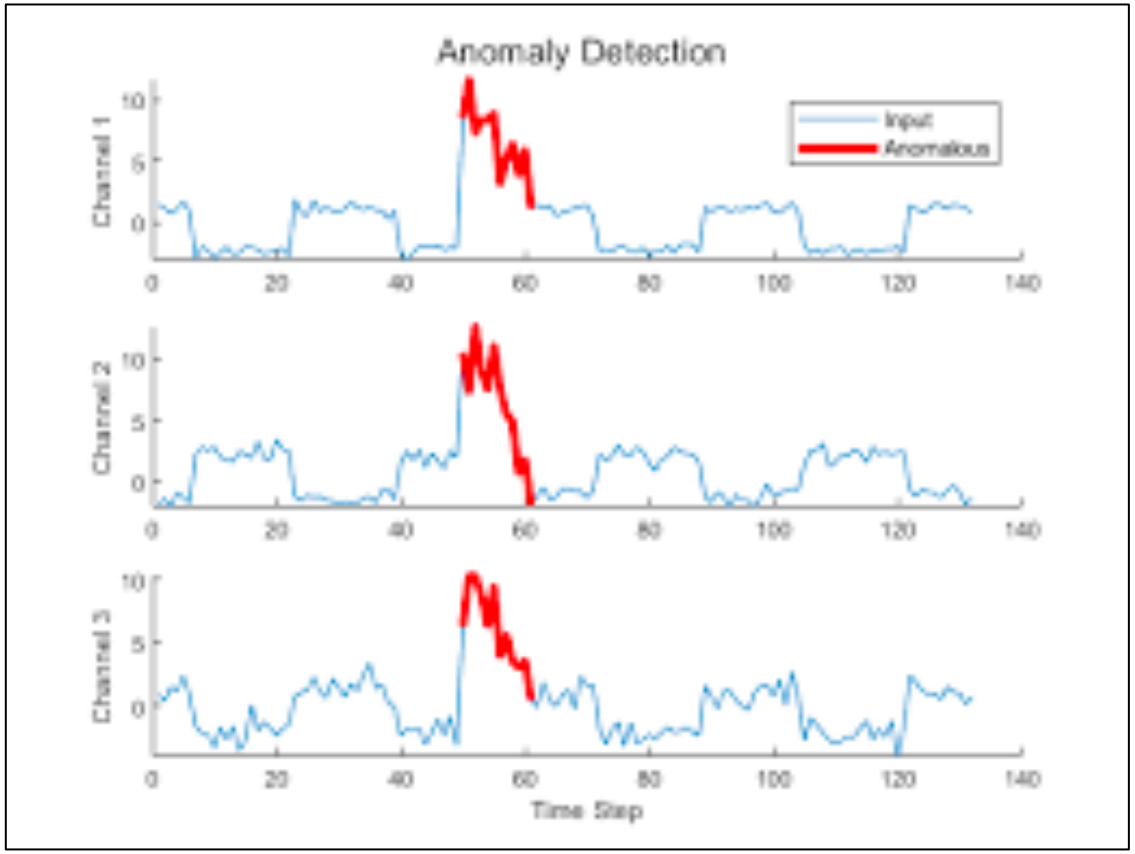
7. 시계열 분석

2. 시계열 데이터 분석의 목적

(3) 시계열 이상치 탐지 (Time Series Anomaly Detection)

- 시계열 데이터의 특정 시점이 이상치인지 점검
- 이상치는 특정 “시점” 일수도, “구간”일수도 있다





https://miro.medium.com/v2/resize:fit:850/1*oxvTvzn0fztqoNlnZkK71g.png

Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

시계열을 여러 개의 성분으로 분해함

Hanwha Ocean

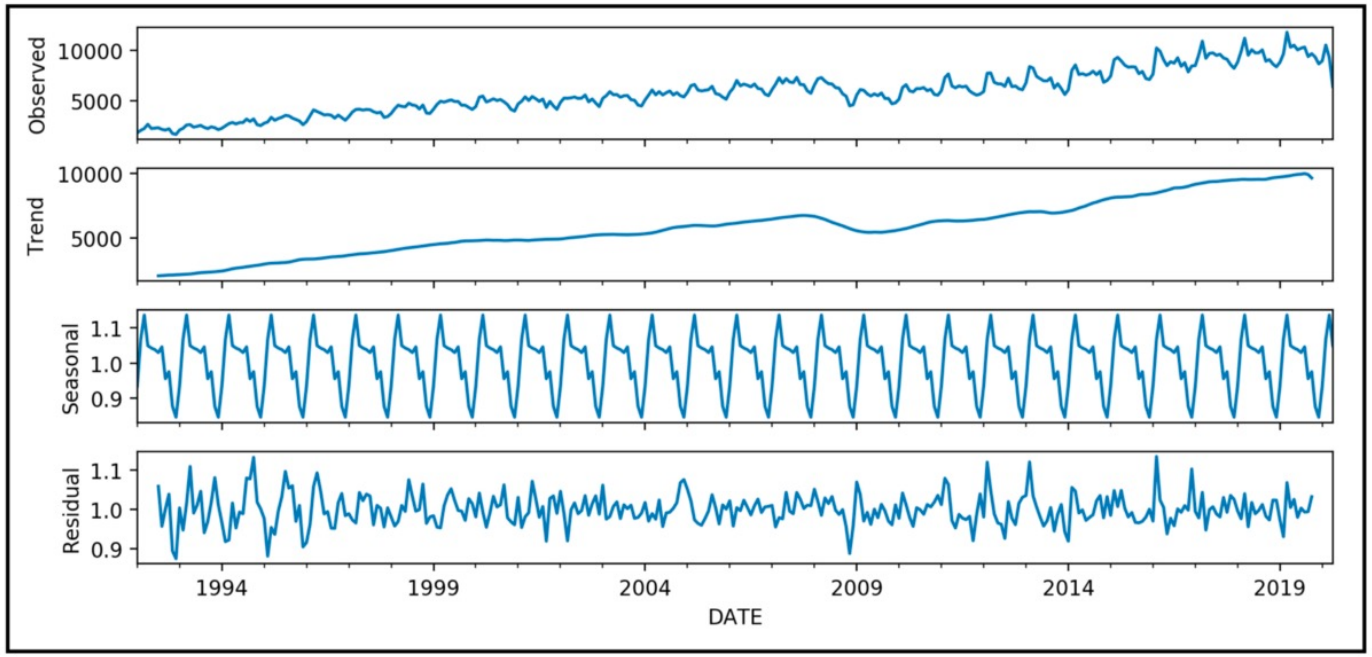
7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

시계열을 여러 개의 성분으로 분해함

시계열의 성분

- (1) 추세 (Trend)
- (2) 계절성 (Seasonality)
- (3) 잔차 (Residual)



Hanwha Ocean

7. 시계열 분석

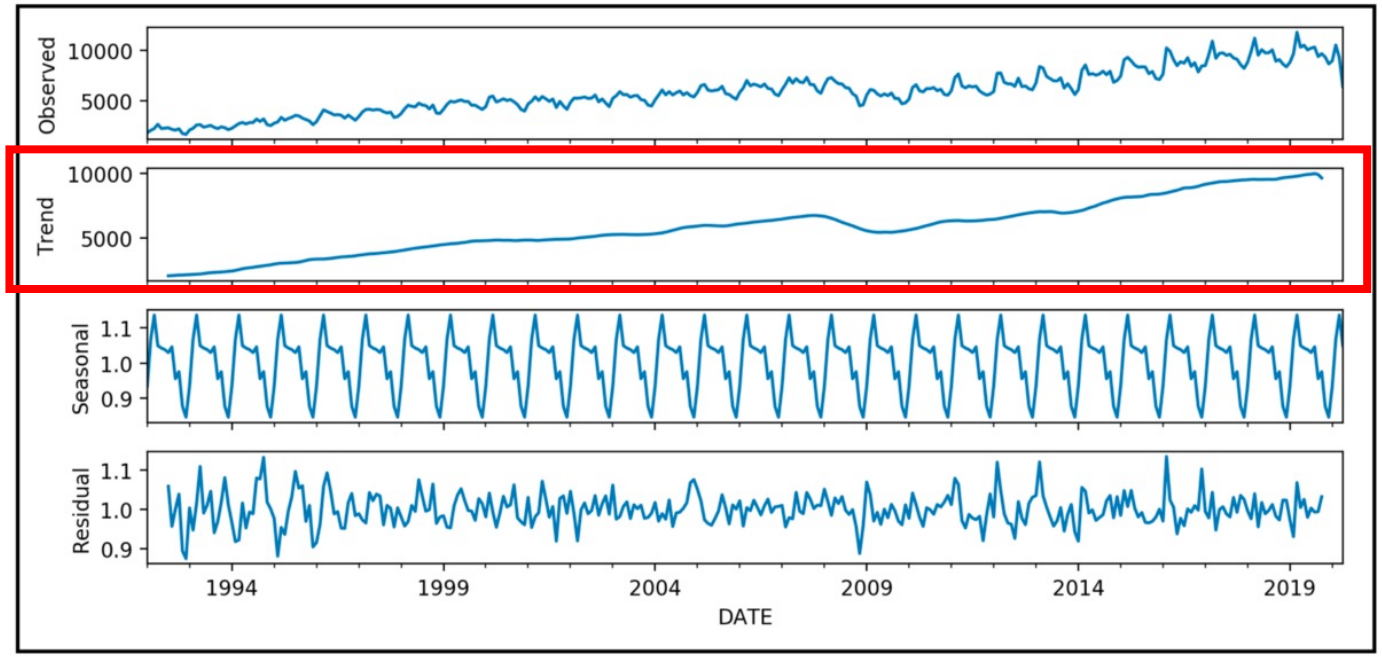
3. 시계열 분해 (Time Series Decomposition)

시계열을 여러 개의 성분으로 분해함

시계열의 성분

- **(1) 추세 (Trend)**
- (2) 계절성 (Seasonality)
- (3) 잔차 (Residual)

시계열의 전체적(전반적)인 트렌드



Hanwha Ocean

7. 시계열 분석

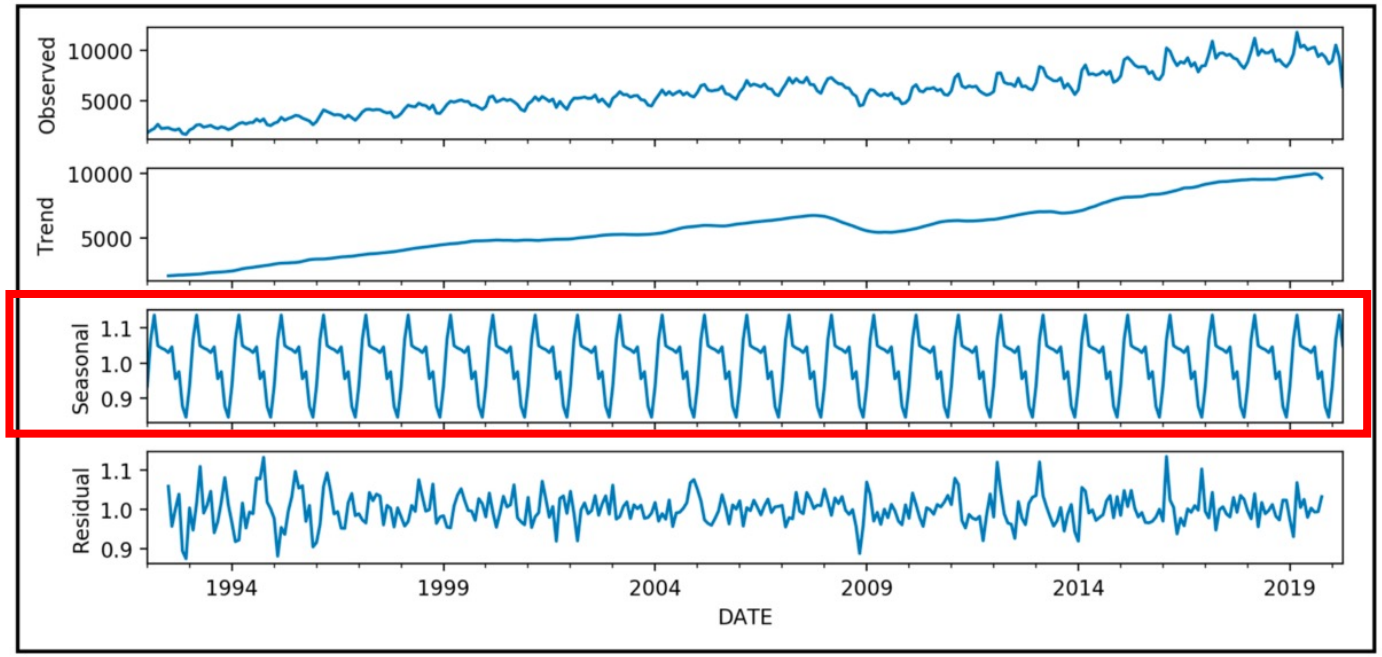
3. 시계열 분해 (Time Series Decomposition)

시계열을 여러 개의 성분으로 분해함

시계열의 성분

- (1) 추세 (Trend)
- **(2) 계절성 (Seasonality)**
- (3) 잔차 (Residual)

주기적으로 반복되는 싸이클



Hanwha Ocean

7. 시계열 분석

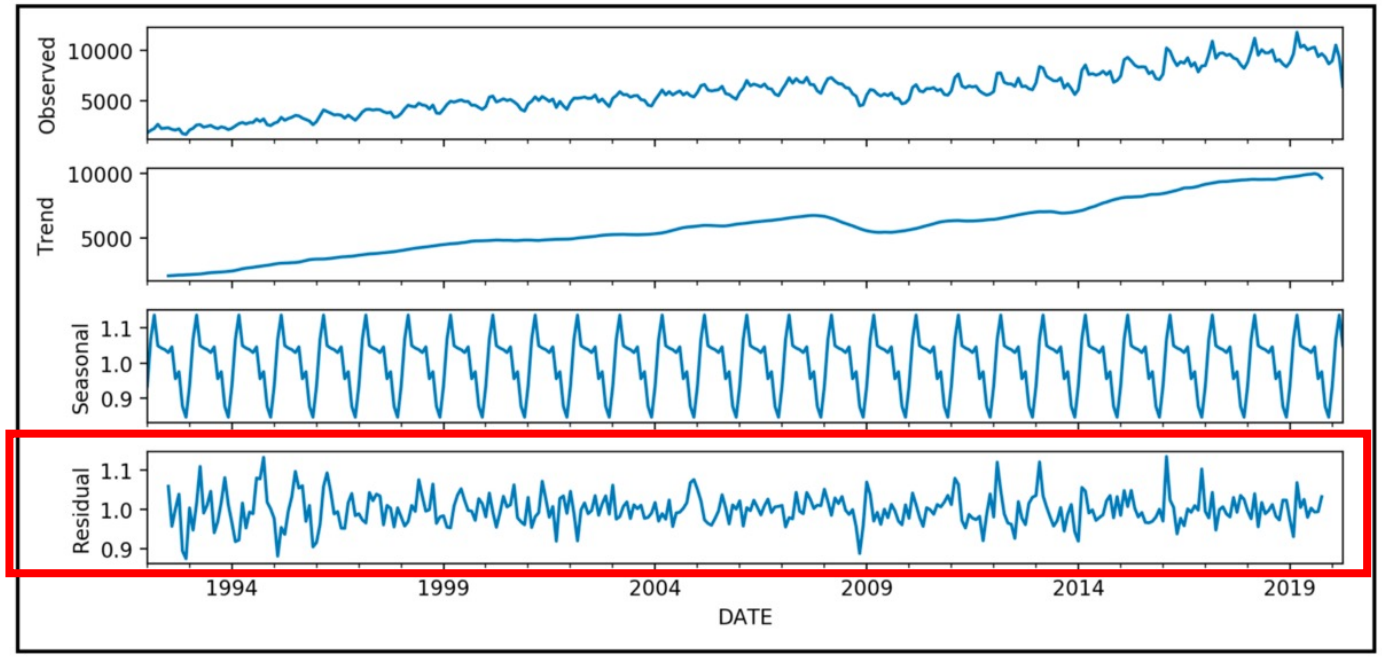
3. 시계열 분해 (Time Series Decomposition)

시계열을 여러 개의 성분으로 분해함

시계열의 성분

- (1) 추세 (Trend)
- (2) 계절성 (Seasonality)
- **(3) 잔차 (Residual)**

추세와 계절성을 제외한 나머지 부분



Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

추세와 계절성을 제외한 나머지 부분

추세(trend)

데이터가 장기적으로 증가하거나 감소할 때, *추세(trend)*가 존재합니다. 추세가 선형적일 필요는 없습니다. 때때로 어떤 추세가 증가에서 감소로 변화하는 경우에, 그것을 추세의 "방향이 변화했다"라고 언급할 것입니다. 그림 2.2의 당뇨병 약 매출 데이터에는 추세가 있습니다.

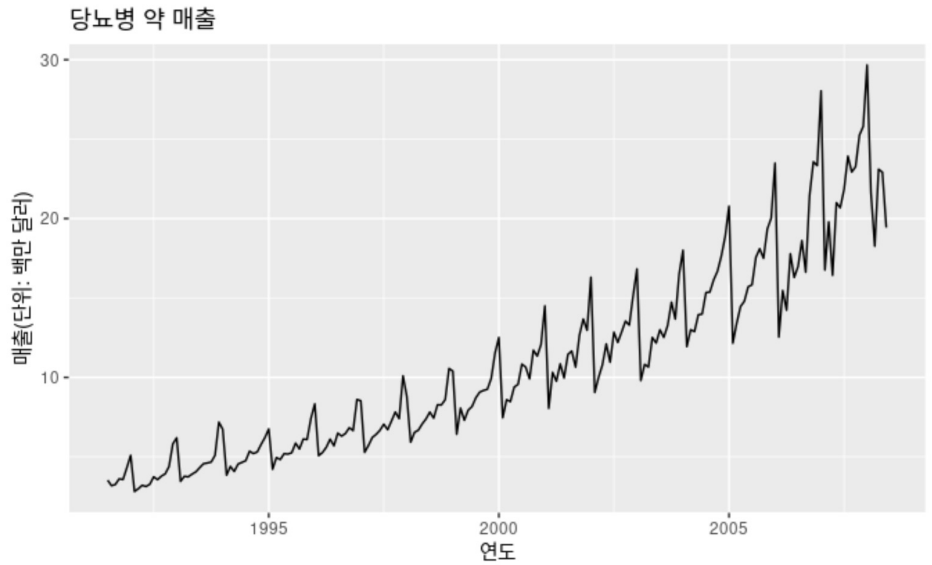


Figure 2.2: 호주 당뇨병 약 월별 매출.

https://otexts.com/fppkr/graphics-scatterplots.html

Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

추세와 계절성을 제외한 나머지 부분

계절성(seasonality)

해마다 어떤 특정한 때나 1주일마다 특정 요일에 나타나는 것 같은 계절성 요인이 시계열에 영향을 줄 때 *계절성 (seasonality)* 패턴이 나타납니다. 계절성은 빈도의 형태로 나타나는데, 그 빈도는 항상 일정하며 알려져 있습니다.
위의 당뇨병 약 월별 매출액에는 계절성이 나타나는데, 이 계절성은 부분적으로 연말에 발생하는 약품 가격 변동에 의한 것입니다.

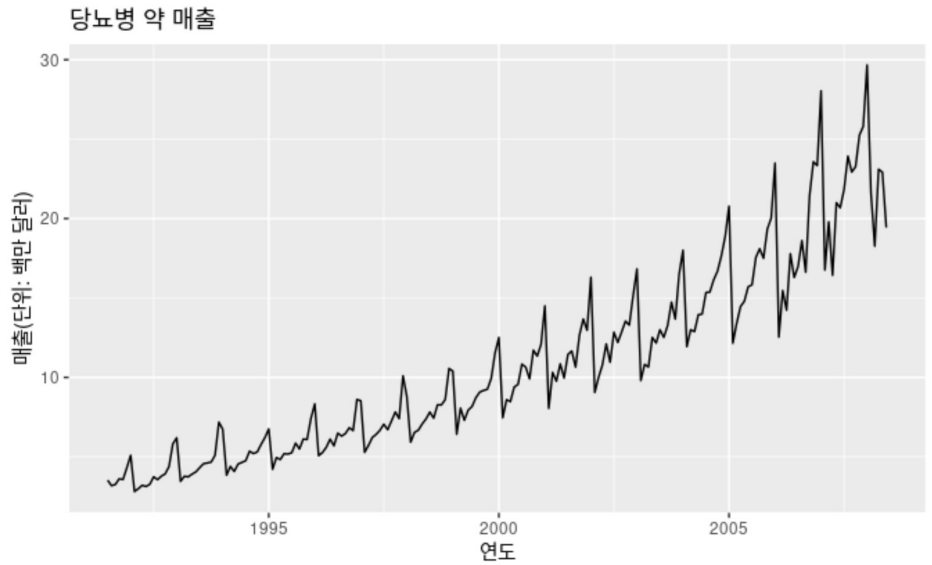


Figure 2.2: 호주 당뇨병 약 월별 매출.

https://otexts.com/fppkr/graphics-scatterplots.html

Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

Ex) 세 달에 전에 비해서 이번 달의 전략량이 크게 늘었다. 과연, 전력의 과소비가 이루어진 것인가?

- 전략량의 경우, 계절(기온)에 따라 크게 영향을 받을 수 있음.

Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

Ex) 세 달에 전에 비해서 이번 달의 전략량이 크게 늘었다. 과연, 전력의 과소비가 이루어진 것인가?

- 전략량의 경우, 계절(기온)에 따라 크게 영향을 받을 수 있음.
- 단순히 과거에 비해서 전력량이 늘었다고 안좋게 볼 것이 아니라, 계절성도 함께 고려해야함!

  (**여름이라서 에어컨 사용량이 늘어난 것일 수도**? 오히려 전체적인 추세는 감소했을 수도 있음)

Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

Ex) 세 달에 전에 비해서 이번 달의 전략량이 크게 늘었다. 과연, 전력의 과소비가 이루어진 것인가?

- 전략량의 경우, 계절(기온)에 따라 크게 영향을 받을 수 있음.
- 단순히 과거에 비해서 전력량이 늘었다고 안좋게 볼 것이 아니라, 계절성도 함께 고려해야함!

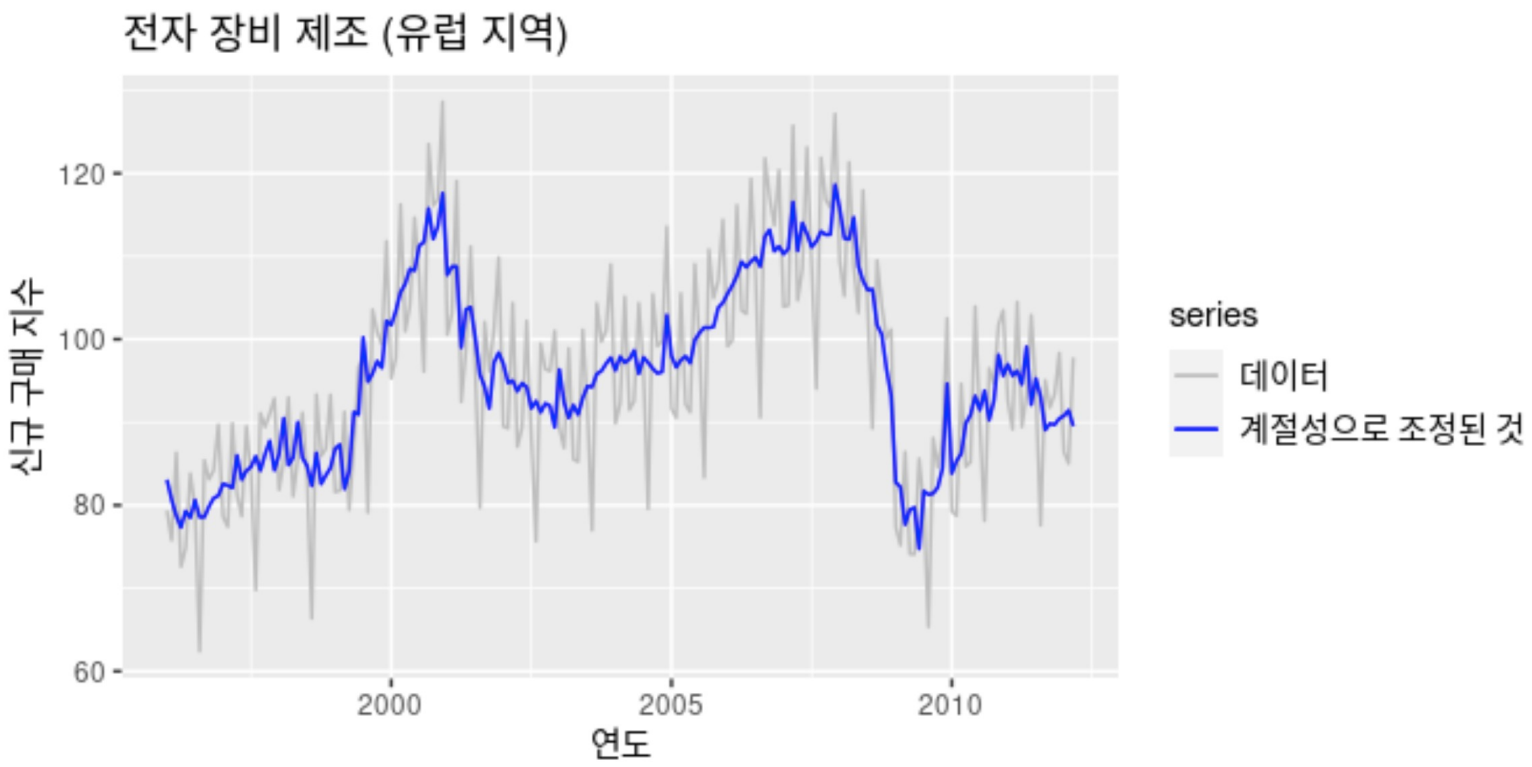
(**여름이라서 에어컨 사용량이 늘어난 것일 수도**? 오히려 전체적인 추세는 감소했을 수도 있음)

따라서, 시계열 분석 시, 추세와 계절성으로 분해해서 볼 필요가 있다!

Hanwha Ocean

7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)



https://otexts.com/fppkr/components.html

Hanwha Ocean

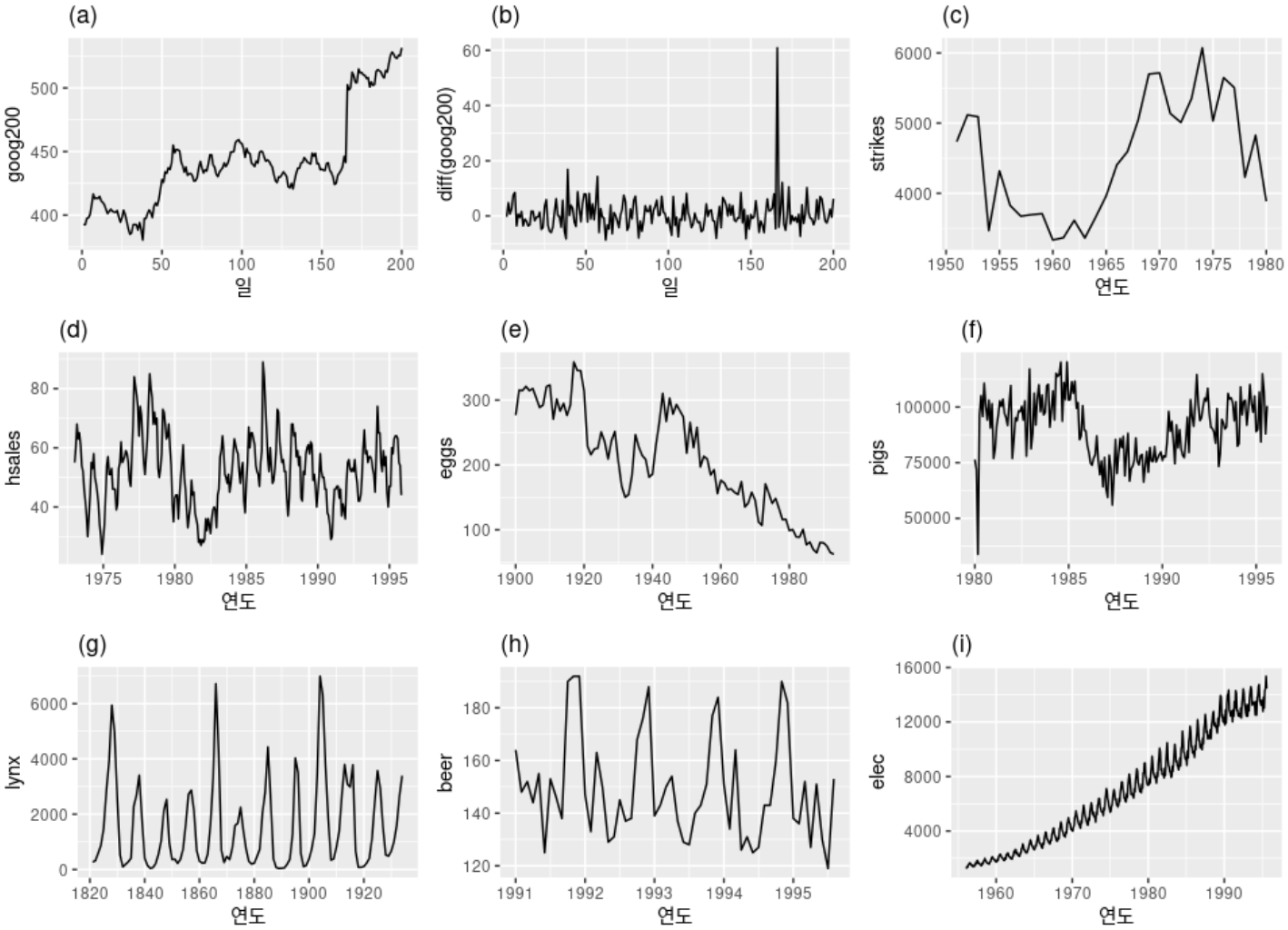
7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

Question)

다음 중, 계절성/추세가 존재하는 시계열은?

- 추세 : (hint = 3개)
- 계절성 : (hint = 4개)



Hanwha Ocean

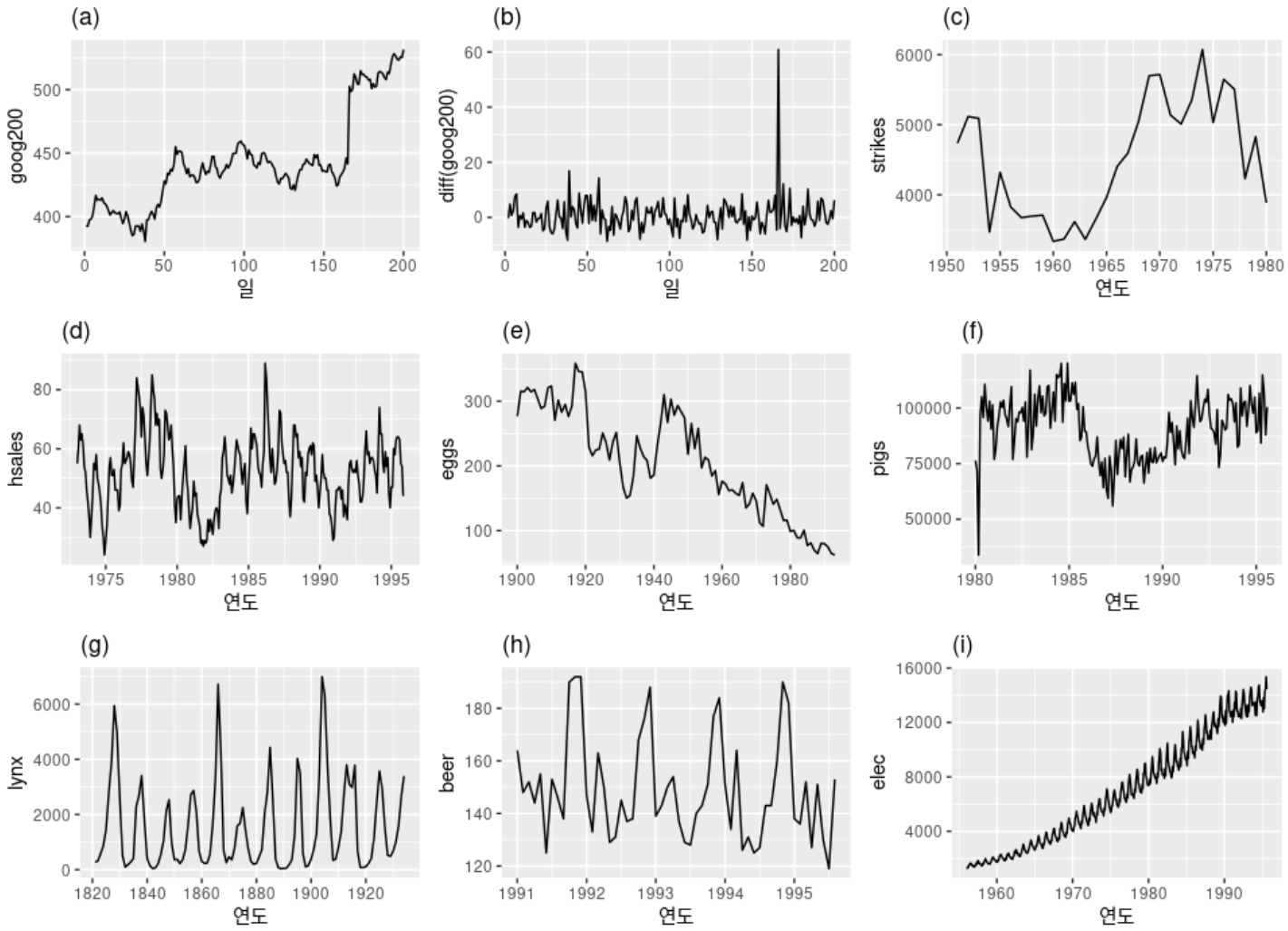
7. 시계열 분석

3. 시계열 분해 (Time Series Decomposition)

Question)

다음 중, 계절성/추세가 존재하는 시계열은?

- **추세 : (a), (e), (i)**
- **계절성 : (d), (g), (h), (i)**



Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

- Moving : 움직이는
- Average : 평균

=> 평균값을 움직이면서 계산한다!

Hanwha Ocean

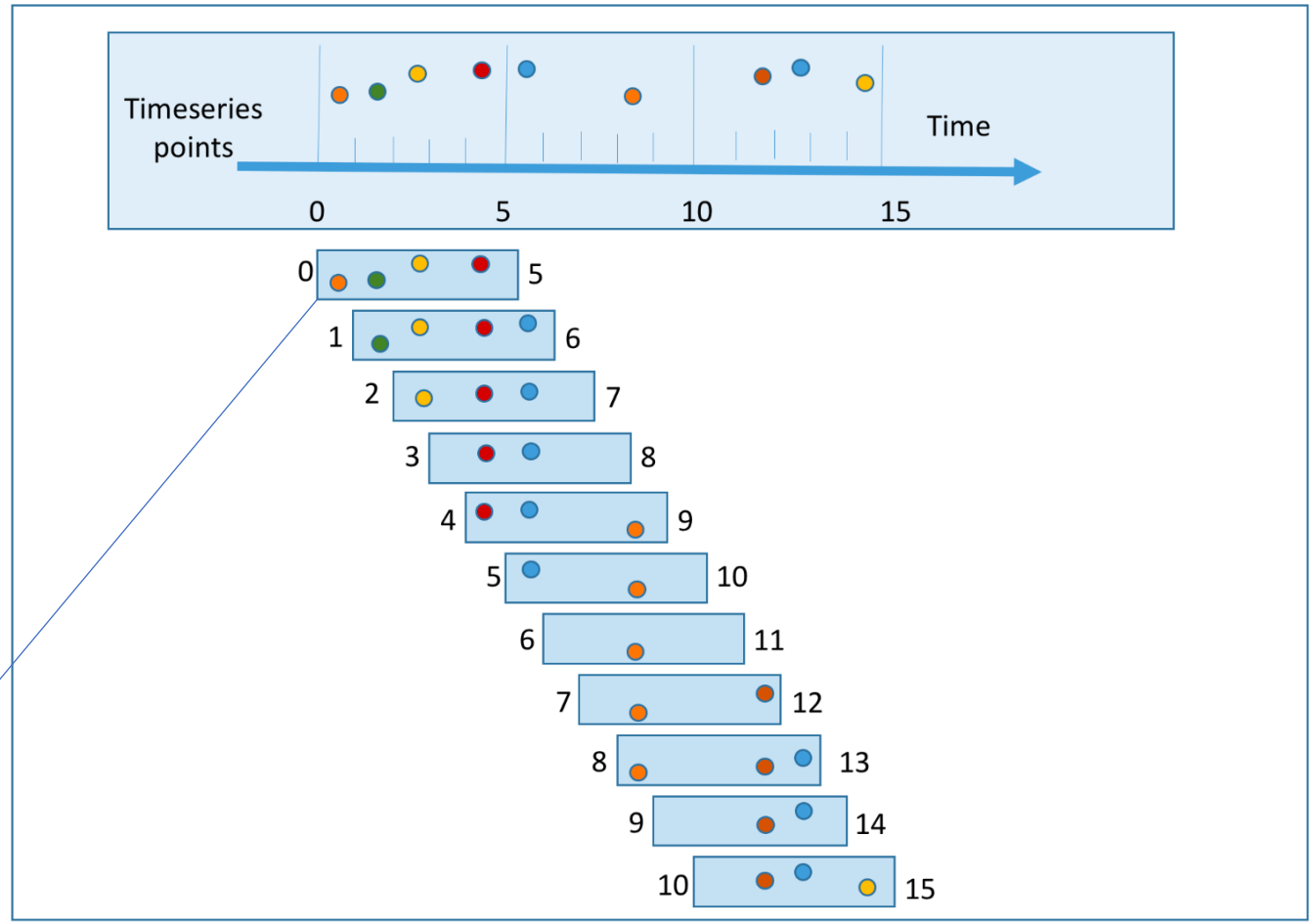
7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

- Moving : 움직이는
- Average : 평균

=> 평균값을 움직이면서 계산한다!



Window (size = 5)

https://docs.wavefront.com/images/5sec_moving_window.png

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

Ex) [10,15,20,25,20,15,20,30, 25, 15]

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

Ex) [10,15,20,25,20,15,20,30, 25, 15]

- Window size = 3
  - (10+15+20)/3 , (15+20+25)/3 , (20+25+20)/3 …

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

Ex) [10,15,20,25,20,15,20,30, 25, 15]

- Window size = 3
  - (10+15+20)/3 , (15+20+25)/3 , (20+25+20)/3 …
- Window size =4
  - (10+15+20+25)/4 , (15+20+25+20)/4, (20+25+20+15)/4, …

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

- 이동 평균을 계산한 값을 대상으로,

  다시 한번 이동 평균을 계산할 수 있다.

https://otexts.com/fppkr/moving-averages.html

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

- 이동 평균을 계산한 값을 대상으로,

  다시 한번 이동 평균을 계산할 수 있다.

$$\hat{T}_t = \frac{1}{2}\left[\frac{1}{4}(y_{t-2}+y_{t-1}+y_t+y_{t+1}) + \frac{1}{4}(y_{t-1}+y_t+y_{t+1}+y_{t+2})\right]$$
$$= \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1} + \frac{1}{8}y_{t+2}.$$

연도	분기	관측값	4-MA	2x4-MA
1992	Q1	443		
1992	Q2	410	451.25	
1992	Q3	420	448.75	450.00
1992	Q4	532	451.50	450.12
1993	Q1	433	449.00	450.25
1993	Q2	421	444.00	446.50
1993	Q3	410	448.00	446.00
1993	Q4	512	438.00	443.00
1994	Q1	449	441.25	439.62
1994	Q2	381	446.00	443.62
1994	Q3	423	440.25	443.12
1994	Q4	531	447.00	443.62
1995	Q1	426	445.25	446.12
1995	Q2	408	442.50	443.88
1995	Q3	416	438.25	440.38
1995	Q4	520	435.75	437.00
1996	Q1	409	431.25	433.50
1996	Q2	398	428.00	429.62
1996	Q3	398	433.75	430.88
1996	Q4	507	433.75	433.75

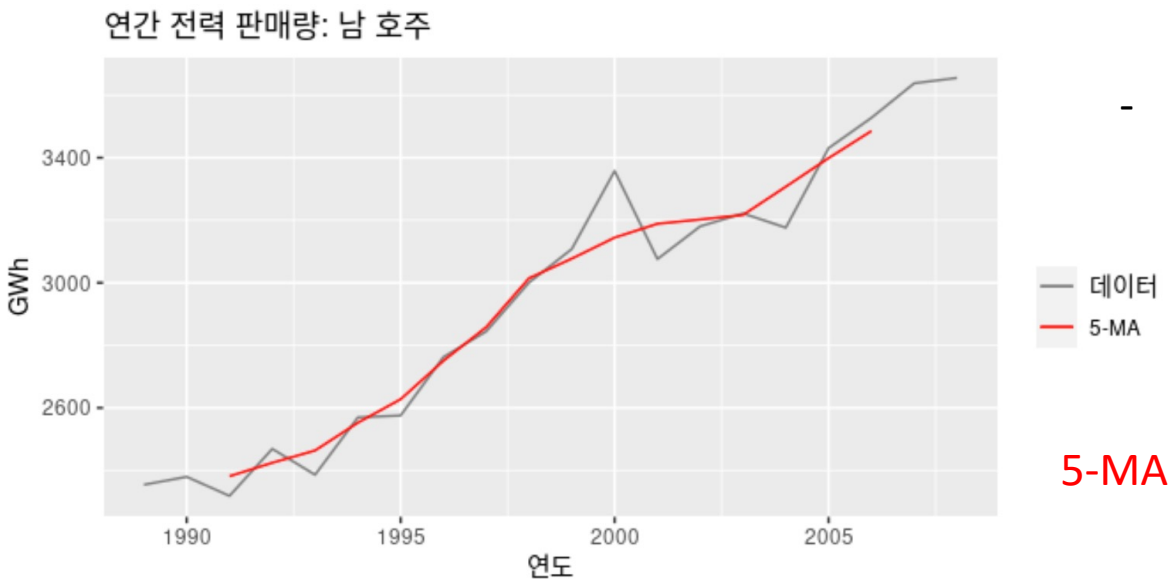
https://otexts.com/fppkr/moving-averages.html

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)



이동평균법의 효과 :

- 시계열이 **부드러워지는(smooth) 효과**가 있다.
- 따라서, 전체적인 **시계열의 추세**를 파악할 수 있다.

5-MA : Window size=5인 이동 평균법(Moving Average)

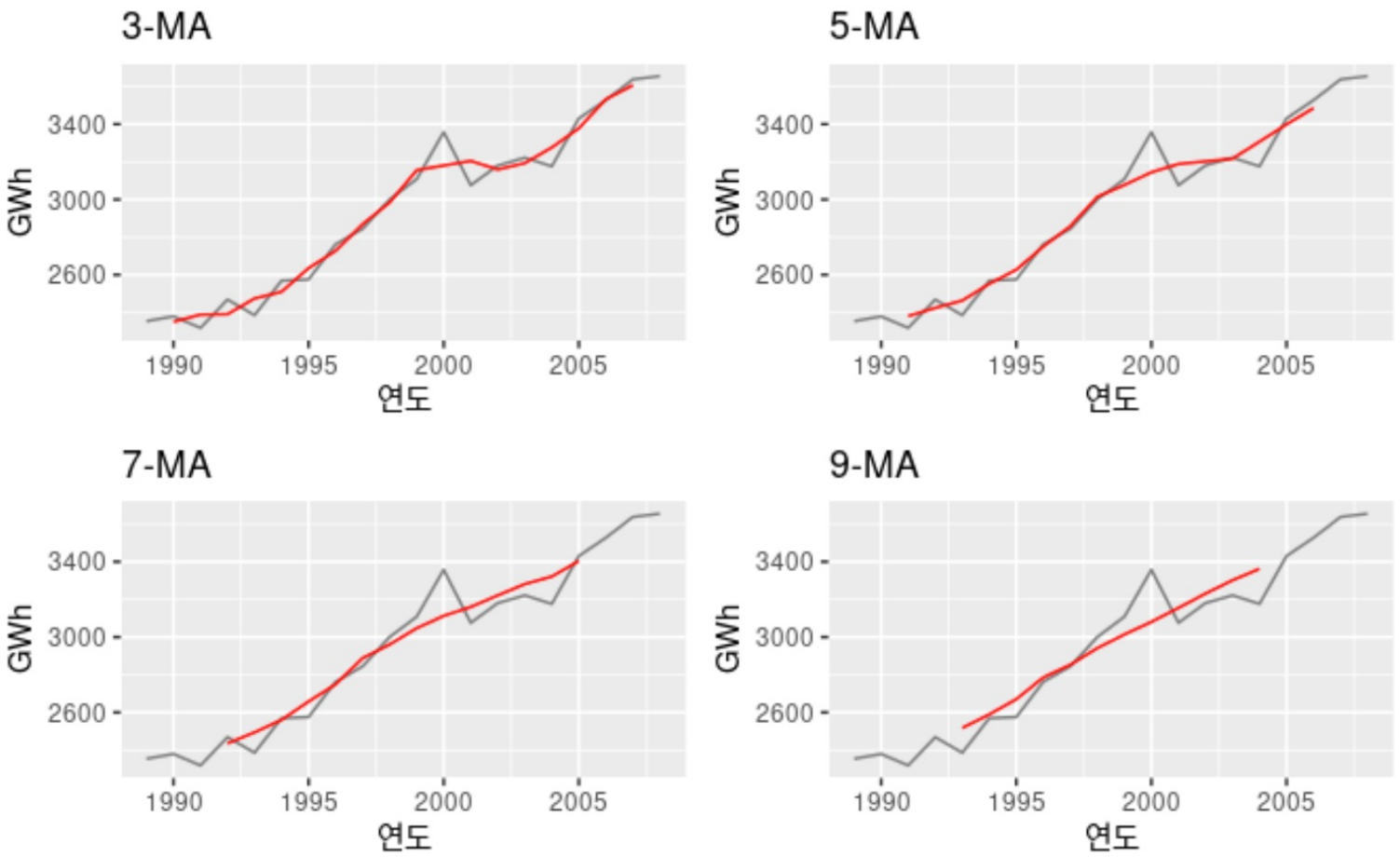
Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

Window Size 변화에 따른 추세의 차이



Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

Question) 이상치에 영향을 덜 받기 위해서는, window size를 늘려야/줄여야 한다?

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(1) 이동 평균법 (Moving Average)

Question) 이상치에 영향을 덜 받기 위해서는, window size를 늘려야/줄여야 한다?

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

- 추세나 계절성 패턴이 없는 시계열 데이터를 예측하는 데 좋다.
- 기본 아이디어) **미래 예측을 위해서는, 오래된 값보다 최근 값이 더 중요하다.**

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

- 추세나 계절성 패턴이 없는 시계열 데이터를 예측하는 데 좋다.
- 기본 아이디어) **미래 예측을 위해서는, 오래된 값보다 최근 값이 더 중요하다.**
- 예시) 2023.4분기 생산량 예측을 위해서,
  - (더 최근인) 2023.3분기 생산량이 더 유용?
  - (더 오래된) 2003.3분기 생산량이 더 유용?

미래 예측을 위해, 최근 값에 더 높은 가중치를 부여하자!

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

- 가중평균이란? (weighted average)
  - 여러 데이터에 대해서 평균을 계산할 것인데, 각각 다른 가중치를 부여하는 것!

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

- 가중평균 예시)

Assessment	Grade (%)	Weight (%)	Product
Assignment 01	60	10	600
Assignment 02	55	10	550
Practical Exam	60	30	1800
Theory Exam	74	50	3700
Sum			**6650**
Weighted Average			**66.5**

Product = Grade * Weight

Sum = Sum of Products

$$\text{Weighted Average} = \frac{\text{Sum of Products}}{\text{Total Weight}} = \frac{6650}{100} = 66.5$$

https://cdn.educba.com/academy/wp-content/uploads/2019/11/Weighted-Average-in-Excel.jpg

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

$0 \leq \alpha \leq 1$: 평활 매개변수

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

$0 \leq \alpha \leq 1$: 평활 매개변수

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

T+1 시점의 예측값

T 시점의 관측값

T-1 시점의 관측값

T-2 시점의 관측값

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

$0 \leq \alpha \leq 1$: 평활 매개변수

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

T 시점의 관측값을 얼마나 활용할지

T -1시점의 관측값을 얼마나 활용할지

T -2시점의 관측값을 얼마나 활용할지

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

$0 \leq \alpha \leq 1$: 평활 매개변수

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

T 시점의 관측값을 얼마나 활용할지

T -1시점의 관측값을 얼마나 활용할지

T -2시점의 관측값을 얼마나 활용할지

Question) 더 최근 값에 더 높은 가중치를 부여하려면, α를 키워야/줄여야 한다.

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

$0 \leq \alpha \leq 1$: 평활 매개변수

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

T 시점의 관측값을 얼마나 활용할지

T -1시점의 관측값을 얼마나 활용할지

T -2시점의 관측값을 얼마나 활용할지

Question) 더 최근 값에 더 높은 가중치를 부여하려면, α를 키워야/줄여야 한다.

Hanwha Ocean

7. 시계열 분석

4. 이동평균법 & 지수평활법

(2) 지수평활법 (Exponential Smoothing)

α 변화에 따른 예측값의 차이

	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
y_T	0.2000	0.4000	0.6000	0.8000
y_{T-1}	0.1600	0.2400	0.2400	0.1600
y_{T-2}	0.1280	0.1440	0.0960	0.0320
y_{T-3}	0.1024	0.0864	0.0384	0.0064
y_{T-4}	0.0819	0.0518	0.0154	0.0013
y_{T-5}	0.0655	0.0311	0.0061	0.0003

8. 데이터 시각화

Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

히스토그램 (Histogram)

선 그래프 (Line Plot)

산점도 (Scatter Plot)

막대 그래프 (Bar plot)

상자 그림 (Box plot)

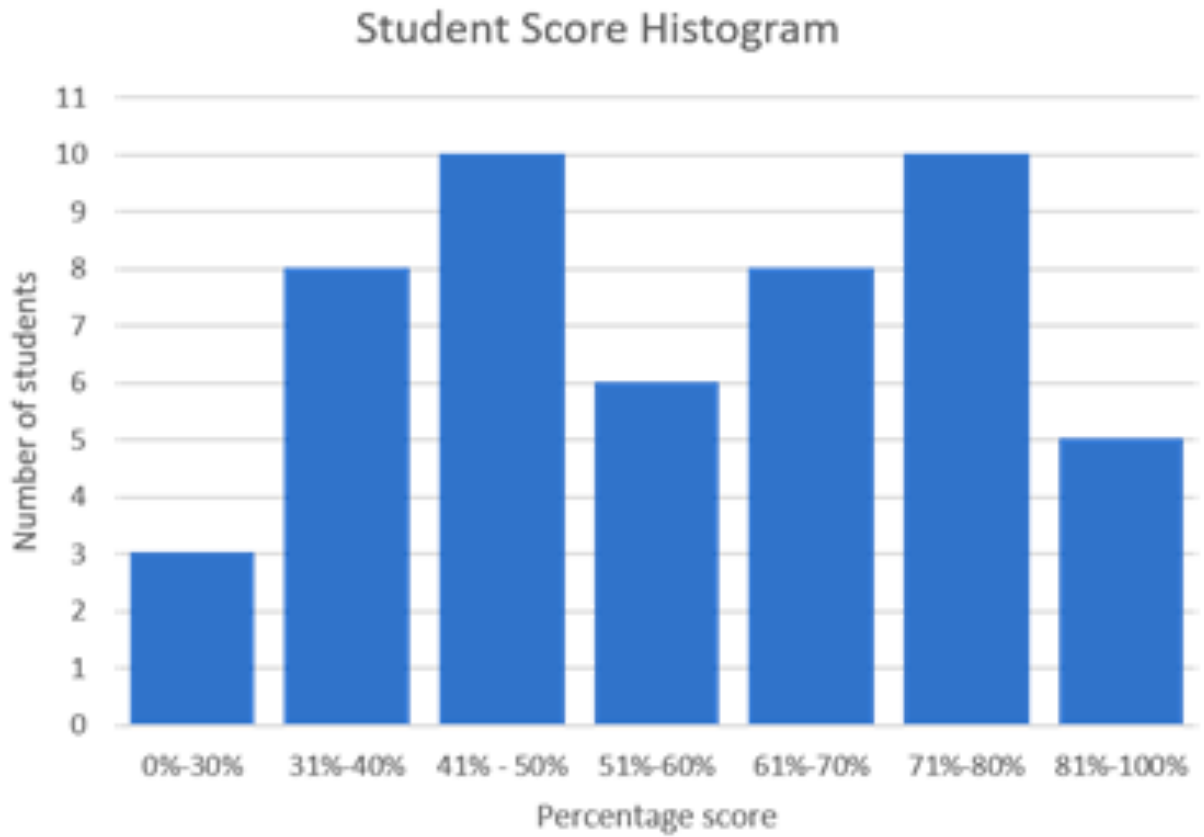
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

히스토그램 (Histogram)

- 도수(count) 분포를 그림으로 나타낸 것



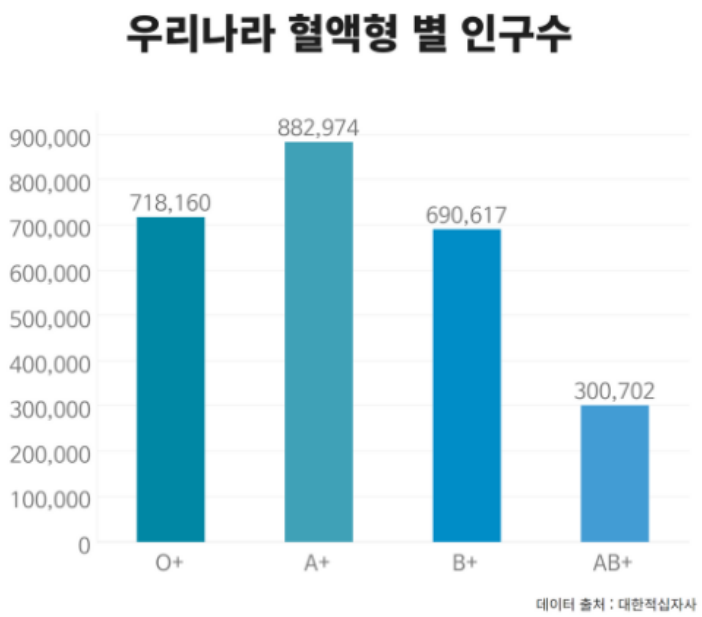
https://www.tibco.com/sites/tibco/files/media_entity/2022-01/histogram-example2.png

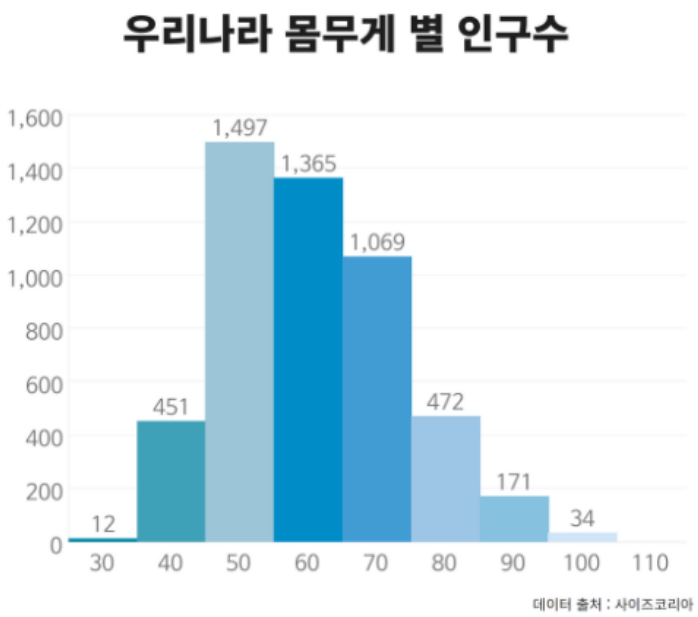
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

히스토그램 (Histogram)





https://cdn.imweb.me/upload/S202101041a4e45576971e/816736388e56d.png

Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

선 그래프 (Line Plot)

- 시계열 데이터를 시각화하는 데 주로 사용

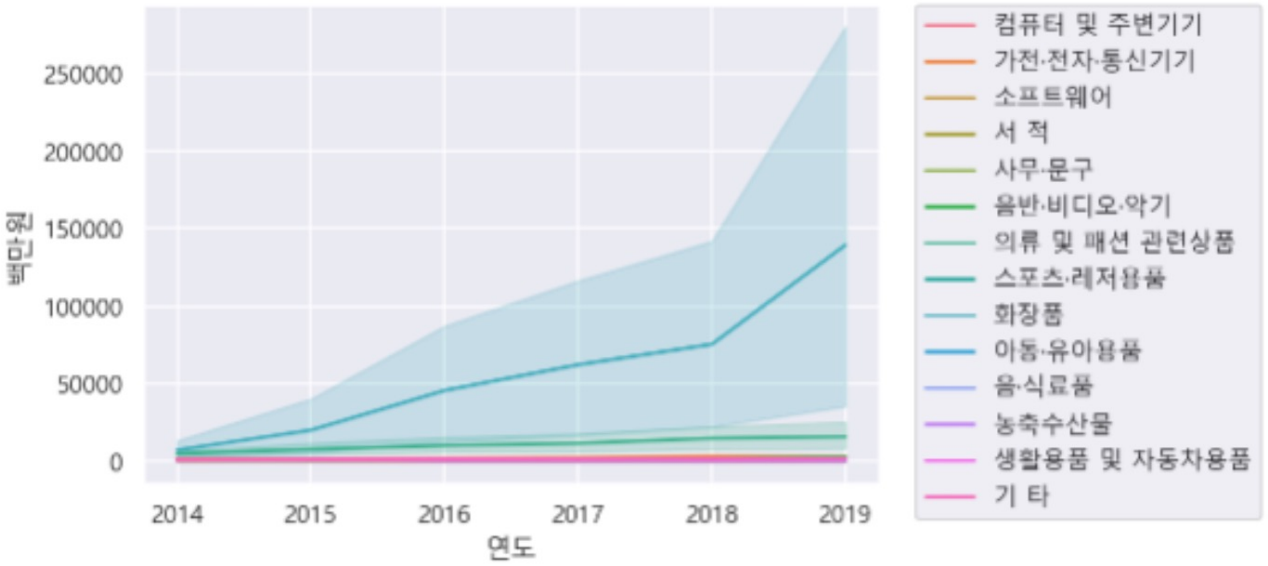
https://www.stockmarketeye.com/wp-content/uploads/2014/09/StockMarketEye-Stock-Chart-Line-Graph.png

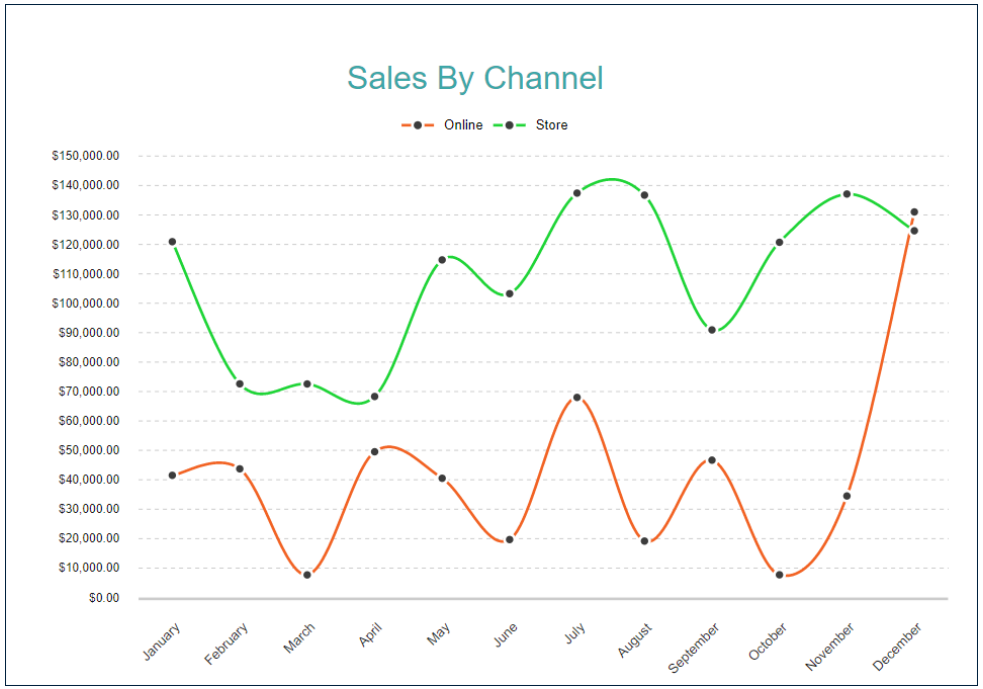
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

선 그래프 (Line Plot)





https://www.stockmarketeye.com/wp-content/uploads/2014/09/StockMarketEye-Stock-Chart-Line-Graph.png
https://ktae23.tistory.com/10

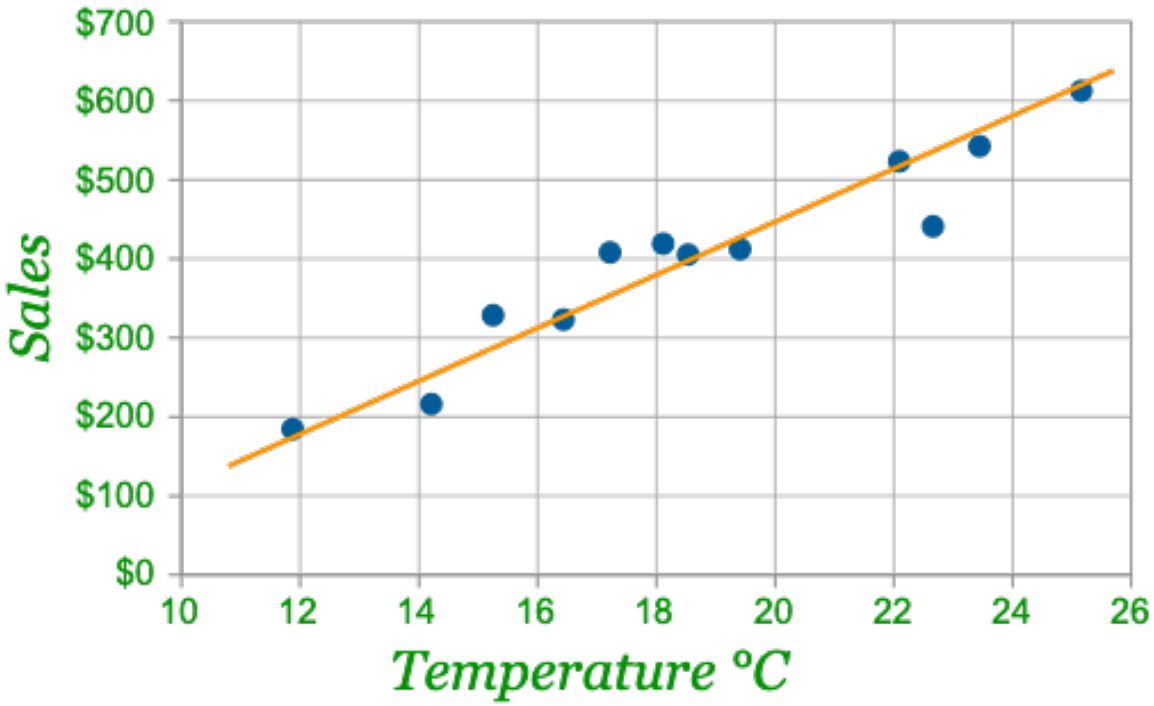
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

산점도 (Scatter Plot)

- 두 수치형 변수 간의 관계를 보기 위해 사용



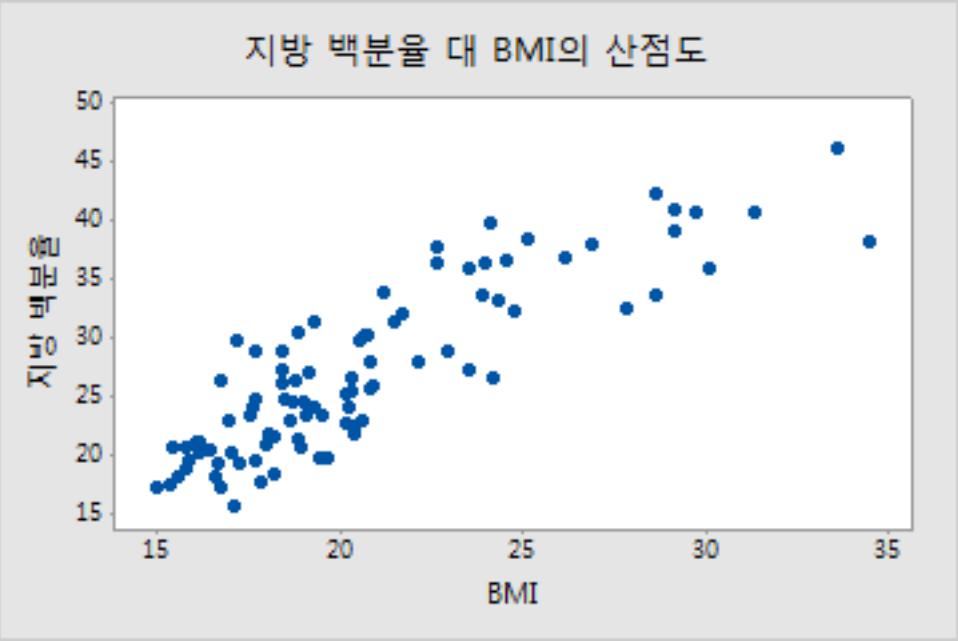
https://www.mathsisfun.com/data/images/scatter-ice-cream1a.svg

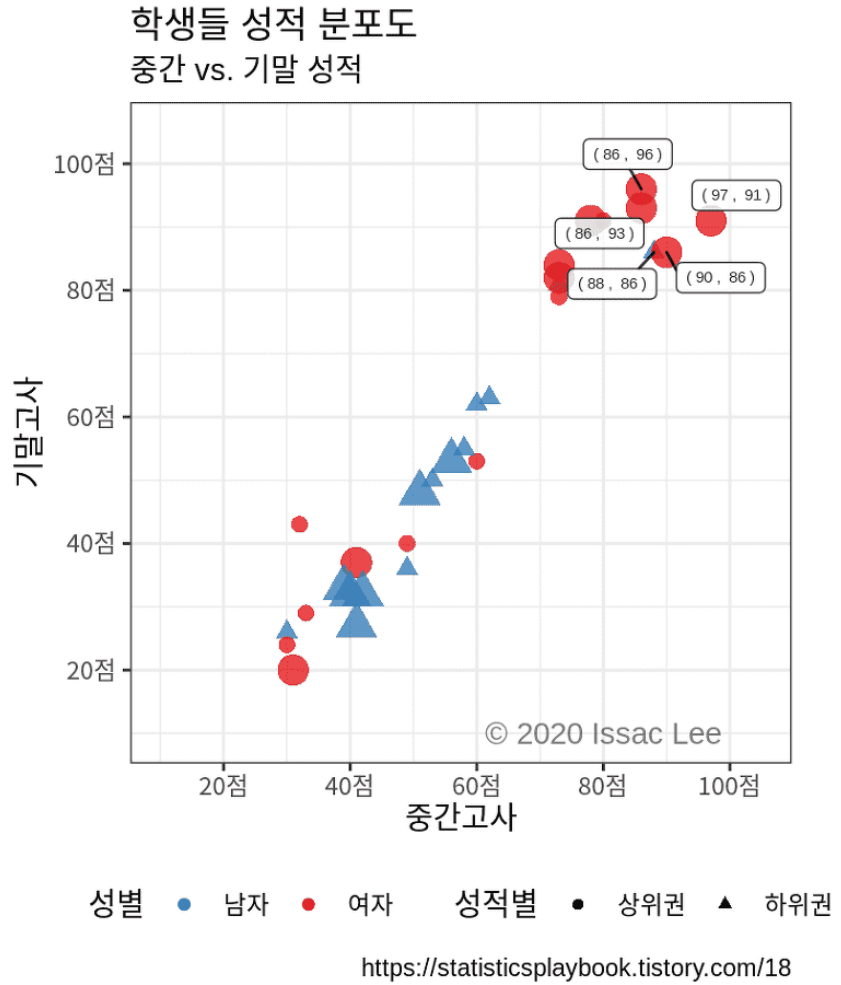
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

산점도 (Scatter Plot)





https://statisticsplaybook.tistory.com/18

https://support.minitab.com/ko-kr/minitab/21/media/generated-content/images/scatterplot_simple_example.png

Hanwha Ocean

8. 데이터 시각화

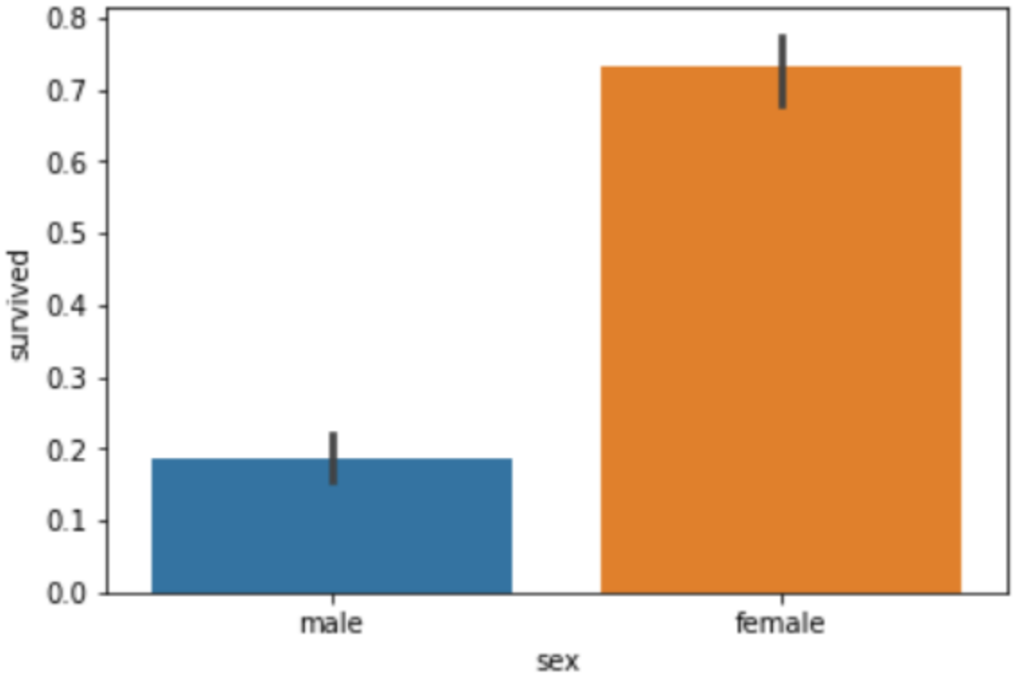
데이터 시각화의 종류

막대 그래프 (Bar plot)

- 범주형 변수와 수치형 변수의 관계를 보기 위해 사용

Q. 히스토그램과의 차이점은?

- 히스토그램 : 특정 값의 범위 대한 빈도 (count)
- 막대그래프 : 범위 (X)



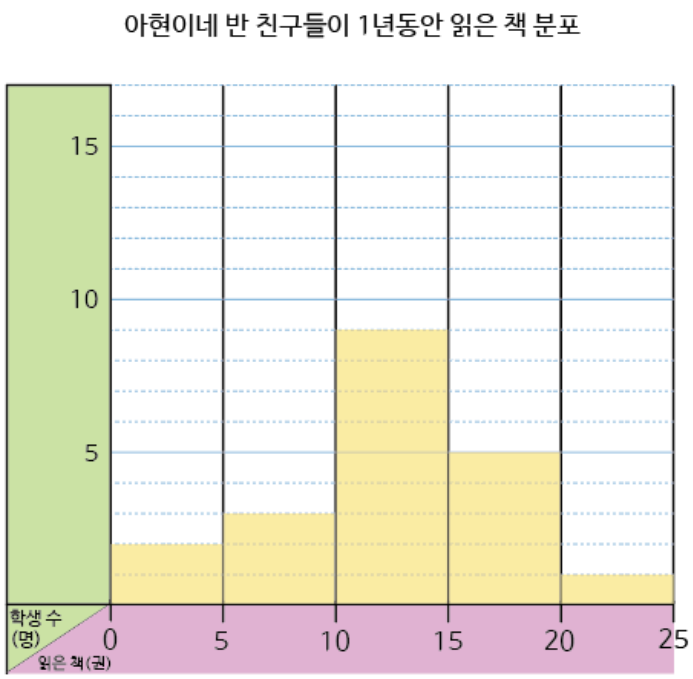
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

막대 그래프 (Bar plot)

막대 그래프



히스토그램

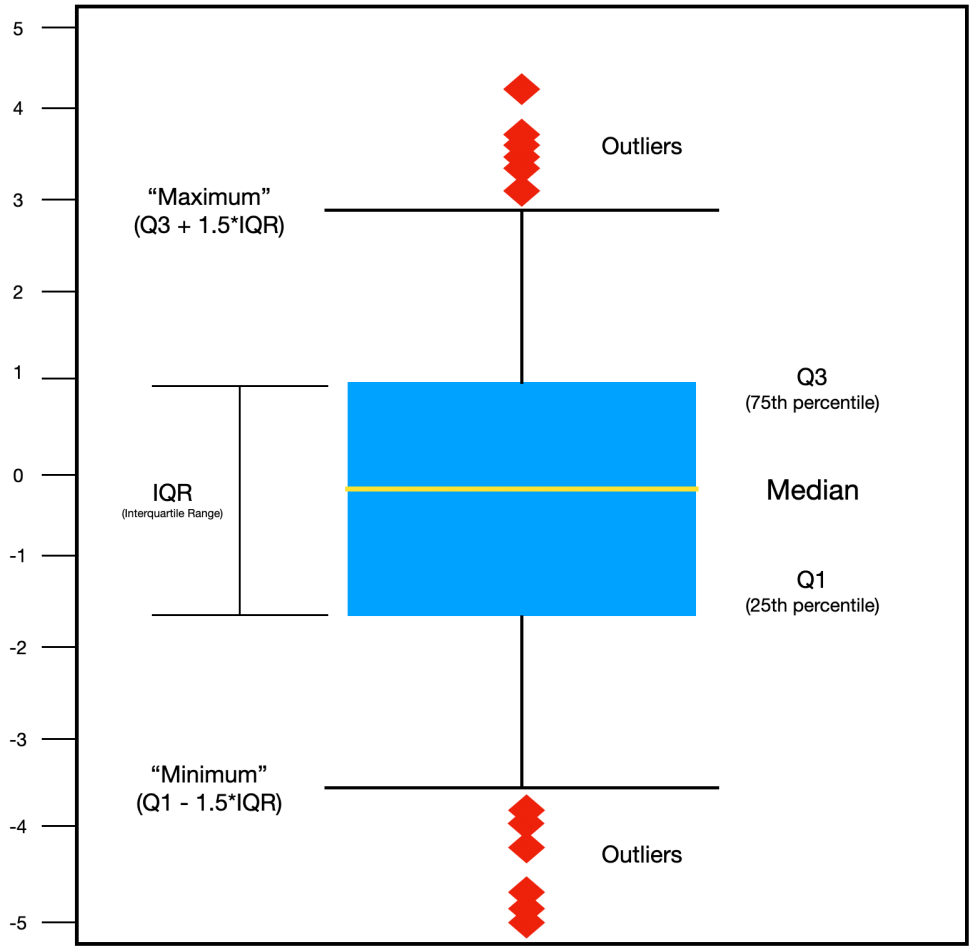
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

상자 그림 (Box Plot)

- 사분위수를 사용하여 시각화한 그림



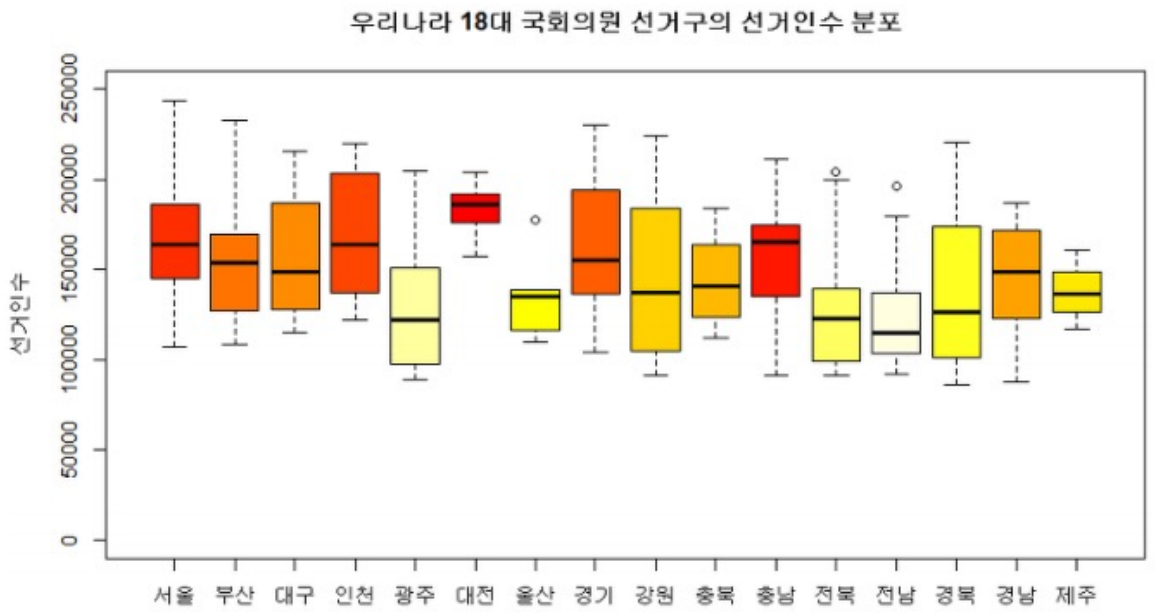
https://heytech.tistory.com/

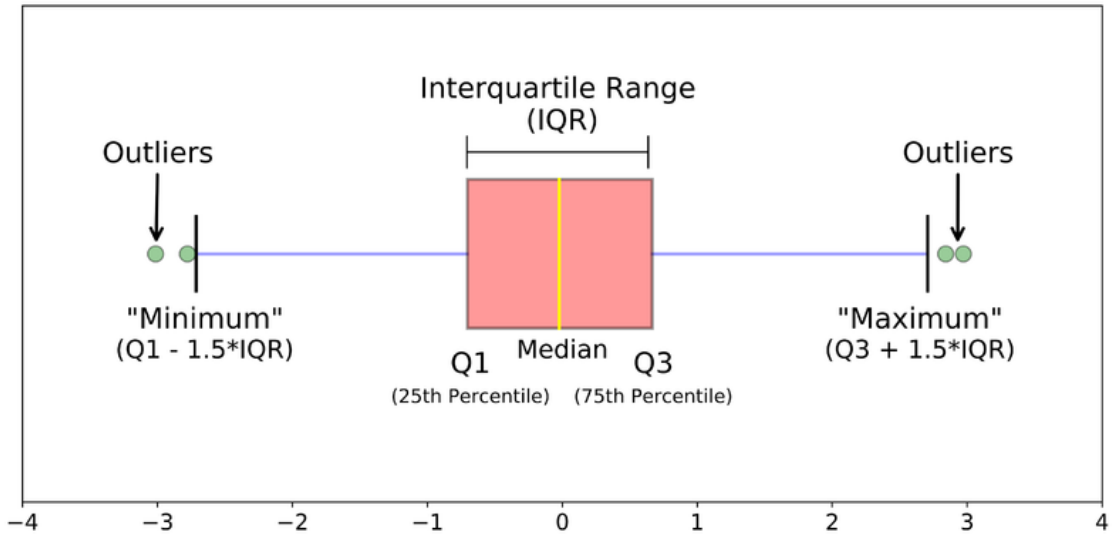
Hanwha Ocean

8. 데이터 시각화

데이터 시각화의 종류

상자 그림 (Box Plot)





https://t1.daumcdn.net/cfile/tistory/99E4E1435C49CB7B39
https://mblogthumb-phinf.pstatic.net/MjAxOTA2MTNfMjI5/MDAxNTYwMzk1MDM1NzA4.FLFUbY1ysVs70HUJhuidzMy1J5Pj3Y0--Ou6vSeTuTgg.rzKjGBjJOhmSMUS67Nu4shUbGGGHkoEDrWtoPV8ilzQg.PNG.narasemi/1_2c21SkzJMf3frPXPAR_gZA.png?type=w800

정리

Hanwha Ocean

1. 데이터 분석의 중요성

빅데이터 (Big Data) 시대

- **데이터(Data)**의 규모와 종류에서 그 정도가 **방대(Big)**해지고 있음.
  - 예시) 테이블, 이미지, 텍스트 데이터
- **주어진 데이터를 올바르게 분석하고 활용할 수 있는 능력**이 매우 중요해졌다
- 데이터를 기반으로 의사 결정을 효과적으로 수행할 수 있음

Hanwha Ocean

2. 데이터의 형태

범주형 데이터

- 명목형 : 순서 X
- 순서형 : 순서 O

수치형 데이터

- 이산형 : 이산적인 값
- 연속형 : 연속적인 값

Hanwha Ocean

3. 통계학이란?

기술 통계학 : 데이터를 요약 & 설명

추론 통계학 : 데이터를 바탕으로, 미지의 집단에 대한 결론을 추론

Hanwha Ocean

4. 통계학의 기초

데이터의 요약값 :

- 평균 / 중앙값 / 최빈값
- 사분위수
- 분산 / 표준편차

Hanwha Ocean

4. 통계학의 기초

모집단과 표본 :

- 모집단 / 표본집단
- 모수 / 통계량
- 추정 (점 추정 & 구간 추정)
- 신뢰구간 / 신뢰수준

Hanwha Ocean

4. 통계학의 기초

확률 :

- 시행, 표본공간, 사건
- 배반사건, 여사건
- 확률의 덧셈 정리

Hanwha Ocean

4. 통계학의 기초

조건부 확률 :

- 조건부 확률 : 특정 사건이 일어났을 때, 다른 사건이 발생할 확률
- 확률의 곱셈 정리

사건의 독립과 종속

Hanwha Ocean

4. 통계학의 기초

확률 변수:

- 이산형 확률변수 (& 확률 질량 함수)
- 연속형 확률변수 (& 확률 밀도 함수)

Hanwha Ocean

5. 상관관계 분석

변수들 사이에 어떤 관계가 있는지 확인

3가지 경우

- **수치형 & 수치형** : 산점도
- **수치형 & 명목형** : 범주 별 평균 계산 (& 상자 그림)
- **명목형 & 명목형** : 교차 분석표

상관관계 vs. 인과관계

Hanwha Ocean

6. 회귀 분석

하나 혹은 그 이상의 변수를 사용하여, 다른 변수를 예측하는 분석법

Y = AX + B

계수 = 기울기 & 절편

계수 추정 : **오차를 최소화**하는 계수 찾기

Hanwha Ocean

7. 시계열 분석

시계열 데이터 = 시간의 흐름에 따라 기록된 데이터

시계열 분해

- **추세 (Trend)**
- **계절성 (Seasonality)**
- 잔차 (Residual)

이동 평균법 & 지수 평활법

Hanwha Ocean

8. 데이터 시각화

히스토그램 (Histogram)

선 그래프 (Line Plot)

산점도 (Scatter Plot)

막대 그래프 (Bar plot)

상자 그림 (Box plot)

감사합니다!

연세대학교 이승한

010-8768-8472

seunghan9613@yonsei.ac.kr