

Module2

EDA(탐색적 자료분석)



◆ 학습목표

범주형 자료 분석방법을 학습하고, 수치형 자료 분석방법을 학습한다.

I. 범주형 자료 분석

II. 수치형 자료 분석

I. 범주형 자료 분석

자료란

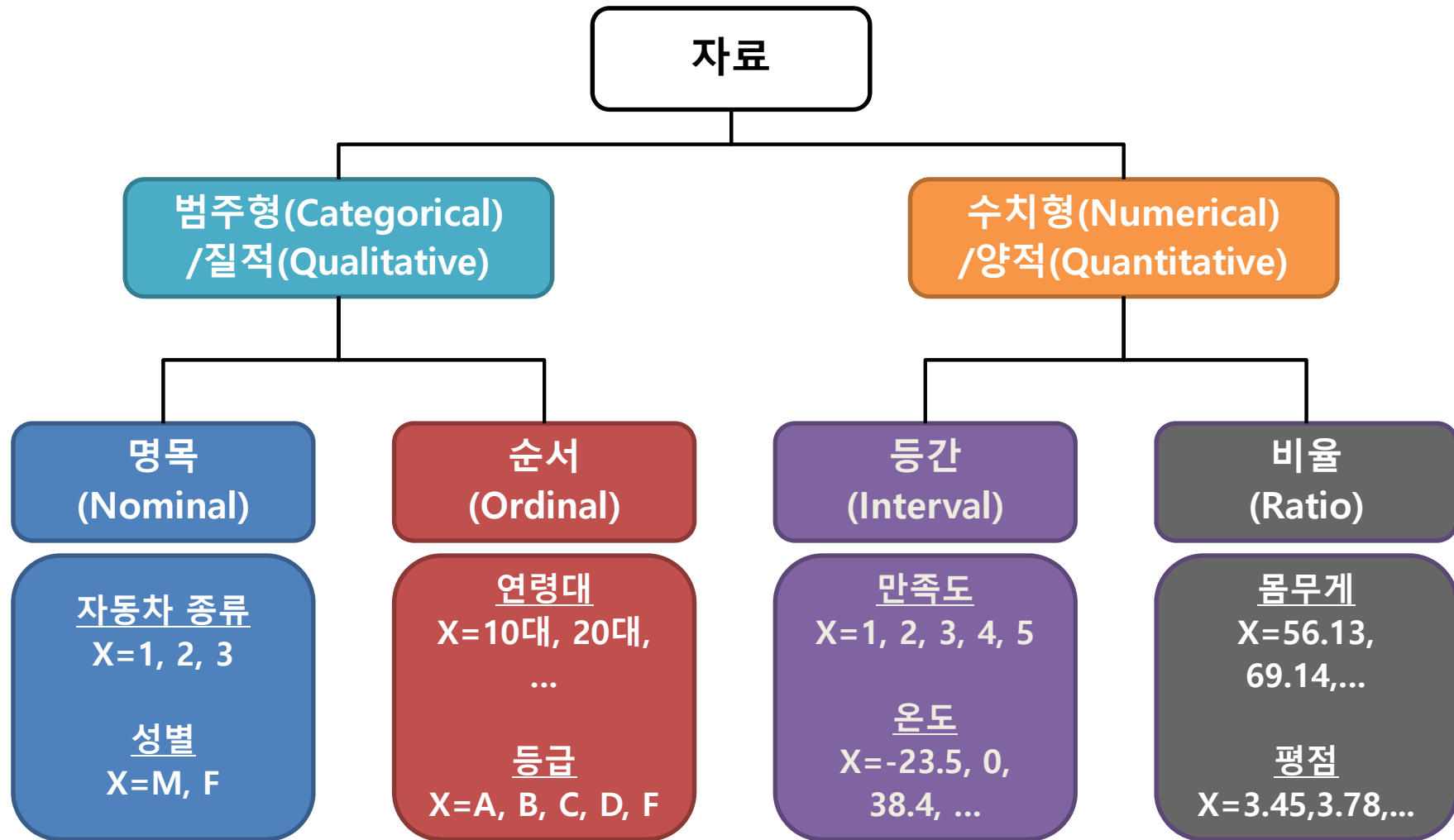
자료와 분석방법

❖ 자료의 형태가 중요한 이유

- 자료의 형태와 분석목적에 따라 통계분석이 결정

형태		요약방법	자료정리	그래프	분석방법
범주형	범주형	도표 그래프	도수분포표 분할표	막대도표 원도표	교차분석
범주형	수치형	도표+ 수치	그룹별 평균	그룹별 막대도표 그룹별 상자도표	t-test ANOVA
수치형	수치형	수치 그래프	산술평균 중앙값 조화평균	히스토그램 상자도표 산점도	상관분석 회귀분석 등

자료와 척도



출처 : Doane and Seward(2011), Applied Statistics in Business & Economics

자료와 척도

❖ 척도(Measurement)의 종류

- 척도: 측정을 하기 위해서 사용한 측정도구
- 예) 키-cm, 몸무게-kg, 성별-M,F, 속도-km/h, 실업률-%

❖ 범주형 자료

- 명목자료/척도 (Nominal measurement)

측정 대상의 특성을 분류하거나 확인

예) 성별, 혈액형, 직업구분

- 순서자료/척도(Ordinal measurement)

측정대상의 특성을 몇 개의 범주로 구분할 뿐만 아니라 그 범주들 사이에 순서관계가 성립하는 경우

예) 학력, 학점, 나이대 등

예) 좋아하는 과목의 순서: 통계학=1위, 간호학=2위, 수학=3위

예) 먹고 싶은 순서대로 순위를 정하세요

아이스크림(3), 초콜렛(1), 솜사탕(4), 오렌지 (2)

자료와 척도

❖ 수치형 자료

– 등간자료/척도 (Interval measurement)

측정 대상의 양적인 차이를 나타내주는 변수

절대영점이 존재하지는 않지만 균일한 간격을 두고 분할하여 측정

예) 설문지의 설문문항(리커르트 5점), 온도, 아이큐지수

– 비율자료/척도(Ratio measurement)

측정 대상의 양적인 차이를 나타내주는 변수

절대영점이 존재하며, 비율계산이 가능

예) 시험 점수, 스트레스 점수, 키, 몸무게

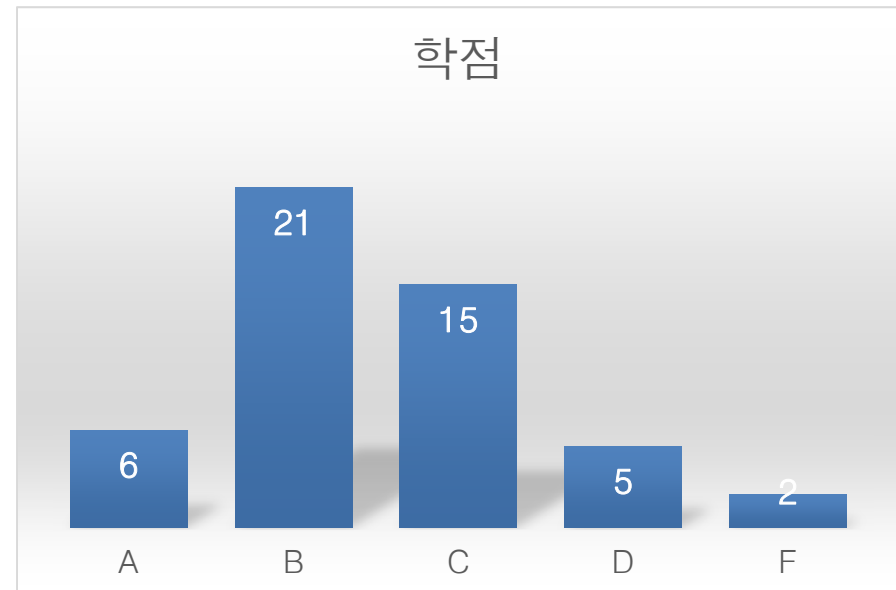
이름	연령	연령대
홍길동	21	20대
이길동	29	30대
백두산	35	30대

도수분포표

❖ 도수분포표와 막대그래프 (Bar chart)

- 누적비율: 순서형일 경우에 사용하면 편리

학점	빈도	비율(%)	누적비율(%)
A	6	12	12
B	21	42	54
C	15	30	84
D	6	12	96
F	2	4	100
합계	50	100	100



❖ 분할표(Contingency table)

- 관측치를 몇 개의 범주로 분할하여 그 해당도수로 자료를 정리해 놓은 표
- 다변량 자료: 범주형 변수가 2개 일 때
- 비율을 표시하는 방법이 중요(행, 열, 전체): 분석목적에 따라 비율표시
- 분할표를 이용한 통계분석 : 교차분석(chi-square)

일원분할표

대조군	처리군	합계
60	40	100

일원분할표

생존	사망	합계
50	50	100

이원분할표

	대조군	처리군	계
생존	40	10	50
사망	20	30	50
계	60	40	100

❖ 사전(실험)설계일 때

- 사전에 그룹의 수를 결정해서 연구할 때 -> 해석: 그룹에 따른 차이
- 예) 비타민과 감기에 대한 연구를 하기 위해, 비타민을 투여할 실험군과 가짜약을 투여할 대조군으로 사전에 구분하여 연구
- 실험군과 대조군에 따른 차이를 검정: 교차분석(동질성검정)
- 비율기준: 그룹별 자료수

사후

사전

그룹	감기발병		합계
	유	무	
실험군 (비타민)	17 (34.0%)	33 (66.0%)	50 (100.0%)
대조군 (Placebo)	38 (76.0%)	12 (24.0%)	50 (100.0%)
합계	55 (55.0%)	45 (45.0%)	100 (100.0%)

사후설계 분할표(독립성)

❖ 사후설계 분할표

- 사전에 그룹의 인원수를 정하지 못하고 사후의 결과를 토대로 연구할 때 → 해석: 두 변수간의
관련성
- 예) 흡연이 폐암과 연관이 있는지를 연구하기 위해 흡연자와 비흡연자를 대상으로 폐암발생여부를 사후에 조사
- 폐암과 흡연간의 관련성을 검정: 교차분석(독립성검정)
- 비율기준: 전체 자료수

사전

사후

폐암	흡연					합계
	비흡연군	장기금연군	단기금연군	재흡연군	흡연군	
무	170,867 (52.0%)	51,690 (15.7%)	46,598 (14.2%)	29,178 (8.9%)	27,784 (8.5%)	326,117 (99.3%)
유	723 (0.2%)	370 (0.1%)	497 (0.2%)	319 (0.1%)	504 (0.2%)	2,413 (0.7%)
합계	171,590 (52.2%)	52,060 (15.8%)	47,095 (14.4%)	29,497 (9.0%)	28,288 (8.6%)	328,530 (100.0%)

출처: 한국인에서 흡연과 폐암의 상관관계 및 폐암의 위험인자 분석, 국민건강보험 일산병원 연구소 (2016)

범주형 자료 분석

▽ 1.package 설치

```
[1] # 1.기본
import numpy as np # numpy 패키지 가져오기
import matplotlib.pyplot as plt # 시각화 패키지 가져오기
import seaborn as sns # 시각화

# 2.데이터 가져오기
import pandas as pd # csv -> dataframe으로 전환
```

2.데이터 불러오기

2.1 데이터 프레임으로 저장

인버테이션(inv)은 dataframe 형태로 가져오기(pandas)
✓ 1초 오후 6:07에 완료됨

2.데이터 불러오기

2.데이터 불러오기

2.1 데이터 프레임으로 저장

- 원본데이터(csv)를 dataframe 형태로 가져오기(pandas)

0초

[3]

```
url = "https://raw.githubusercontent.com/leecho-bigdata/statistics-python/main/01_1.EDA.csv"
eda_df = pd.read_csv(url, encoding="cp949")
eda_df.head(10)
```

	id	성별	문반	학년	몸무게	출석	중간	기말
0	1	남자	1	1	40	100	87	80
1	2	여자	2	2	50	100	83	60
2	3	남자	1	3	56	100	84	60
3	4	여자	2	4	51	100	73	60
4	5	남자	1	1	55	100	68	60
5	6	남자	2	2	61	100	77	50
6	7	여자	1	3	69	100	40	80
7	8	여자	2	2	44	100	73	30
8	9	여자	1	2	66	80	64	40
9	10	남자	2	2	60	100	66	40

Next steps: [View recommended plots](#)

2.2 자료구조 살펴보기

0초


[4]

```
eda_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
```

✓ 1초

오후 6:07에 완료됨


LG
 Life's Good

80/862

0초 [4] eda_df.info()

0초 [5] eda_df.shape

0 票 [6] eda_df.dtypes

2.3 범주형 변수 처리

- ✓ 1초 오후 6:07에 완료됨

2.데이터 불러오기

2.3 범주형 변수 처리

- 가변수 처리시 문자로 처리를 해야 변수명 구분이 쉬움

```
[7] eda_df['성별'] = eda_df['성별'].replace([1,2], ['남자', '여자'])
eda_df['분반'].replace({1:'A반', 2:'B반'}, inplace=True)
eda_df['학년'].replace({1:'1학년', 2:'2학년', 3:'3학년', 4:'4학년'}, inplace=True)

eda_df.head()
```

	id	성별	분반	학년	몸무게	출석	중간	기말
0	1	남자	A반	1학년	40	100	87	80
1	2	여자	B반	2학년	50	100	83	60
2	3	남자	A반	3학년	56	100	84	60
3	4	여자	B반	4학년	51	100	73	60
4	5	남자	A반	1학년	55	100	68	60

Next steps: [View recommended plots](#)

```
[8] eda_df.dtypes
```

```
id      int64
성별     object
분반     object
학년     object
몸무게   int64
출석     int64
중간     int64
기말     int64
dtype: object
```

```
[9] # datatype을 category로 변경
eda_df['성별'] = eda_df['성별'].astype('category')
eda_df['분반'] = eda_df['분반'].astype('category')
eda_df['학년'] = eda_df['학년'].astype('category')
```

✓ 1초 오후 6:07에 완료됨

2.데이터 불러오기

```

0초 10초
기말      int64
dtype: object

[9] # datatype을 category로 변경
eda_df['성별'] = eda_df['성별'].astype('category')
eda_df['분반'] = eda_df['분반'].astype('category')
eda_df['학년'] = eda_df['학년'].astype('category')

[10] eda_df.dtypes

id      int64
성별    category
분반    category
학년    category
몸무게  int64
출석    int64
중간    int64
기말    int64
dtype: object

  3.범주형 변수(1개) (one categorical)

  3.1 돛수분포표(freq_table)

0초 # value_counts()
eda_df['성별'].value_counts()

남자    54
여자    48
Name: 성별, dtype: int64

[12] freq_table = pd.DataFrame(eda_df['성별'].value_counts())
freq_table.columns = ['count']
freq_table

count
남자    54
  
```

1초 오후 6:07에 완료됨

3.범주형 변수(1개) (one categorical)

3.범주형 변수(1개) (one categorical)

3.1 돗수분포표(freq_table)

```
[11] # value_counts()
eda_df['성별'].value_counts()
```

```
남자    54
여자    48
Name: 성별, dtype: int64
```

```
[12] freq_table = pd.DataFrame(eda_df['성별'].value_counts())
freq_table.columns = ['count']
freq_table
```

	count
남자	54
여자	48

```
[13] ### crosstab이용
freq_table = pd.crosstab(index = eda_df["성별"],
                        columns = ['count'])
freq_table
```

col_0	count
성별	
남자	54
여자	48

```
[14] # 비율 추가
freq_table['prop'] = np.round(freq_table['count']/sum(freq_table['count']), 2)
```

✓ 1초 오후 6:07에 완료됨

3. 범주형 변수(1개) (one categorical)

✓ 0초

[13]

남자	54
여자	48

✓ 0초

[14]

비율 추가

freq_table['prop'] = np.round(freq_table['count']/sum(freq_table['count']), 2)

freq_table

col_0	count	prop
성별		
남자	54	0.53
여자	48	0.47

Next steps:

View recommended plots

✓ 0초

[15]

누적 비율

freq_table['cum_prop'] = np.cumsum(freq_table['prop'])

freq_table

col_0	count	prop	cum_prop
성별			
남자	54	0.53	0.53
여자	48	0.47	1.00

Next steps:

View recommended plots

✓ 0초

[16]

범주형 변수(막대그래프)

sns.catplot(x = "성별",

kind = "count",

data = eda_df)

plt.show()

✓ 1초

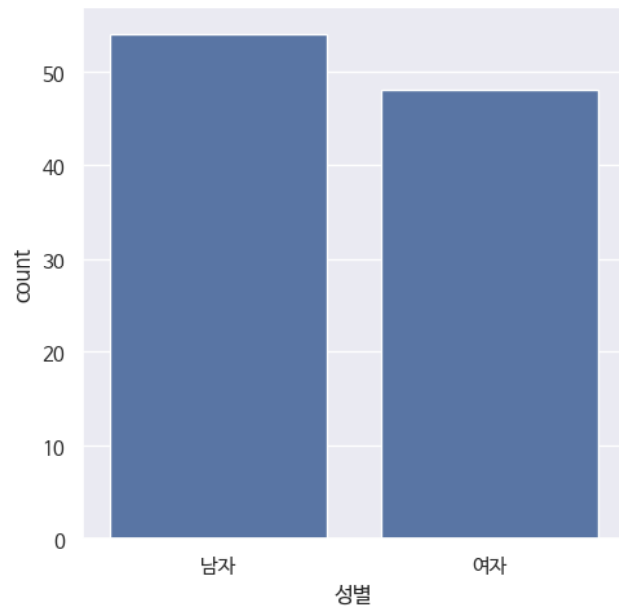
오후 6:07에 완료됨

3. 범주형 변수(1개) (one categorical)

3.2 그래프 그리기(막대 그래프)

```
[16] # 범주형 변수(막대그래프)
sns.catplot(x = "성별",
            kind = "count",
            data = eda_df)

plt.show()
```



```
[17] #pd_plot 이용
ax = freq_table["count"].plot(figsize = (8, 6),
                              kind = "bar")

ax.set(title = '막대그래프',
       xlabel = '성별',
```

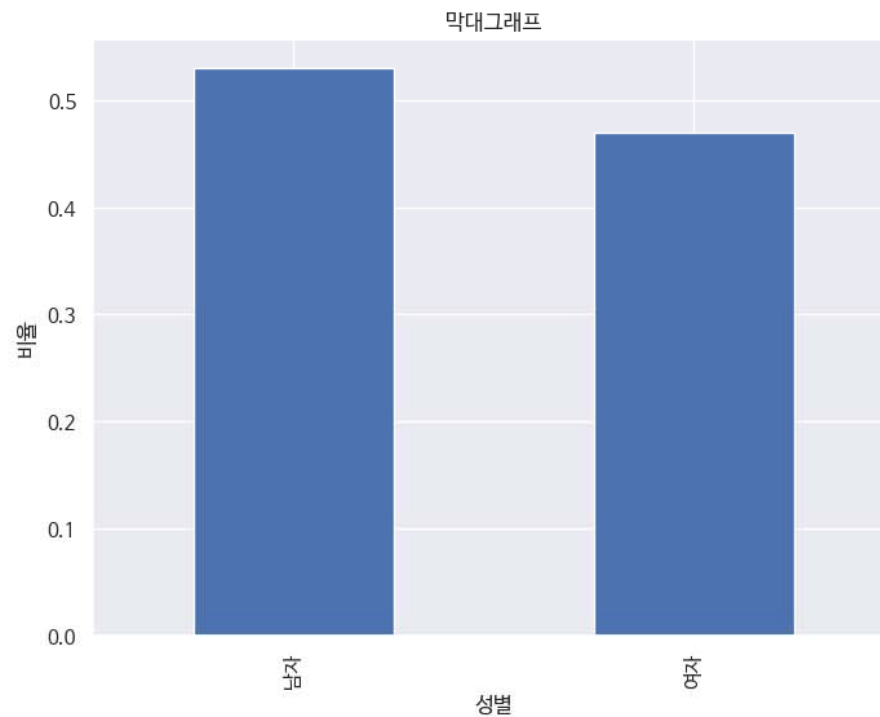
✓ 1초 오후 6:07에 완료됨

3. 범주형 변수(1개) (one categorical)



3. 범주형 변수(1개) (one categorical)

```
[18] #pd.plot 이용(비율)
ax = freq_table["prop"].plot(figsize = (8, 6),
                             kind = "bar")
ax.set(title = '막대그래프',
       xlabel = '성별',
       ylabel = "비율")
plt.show()
```



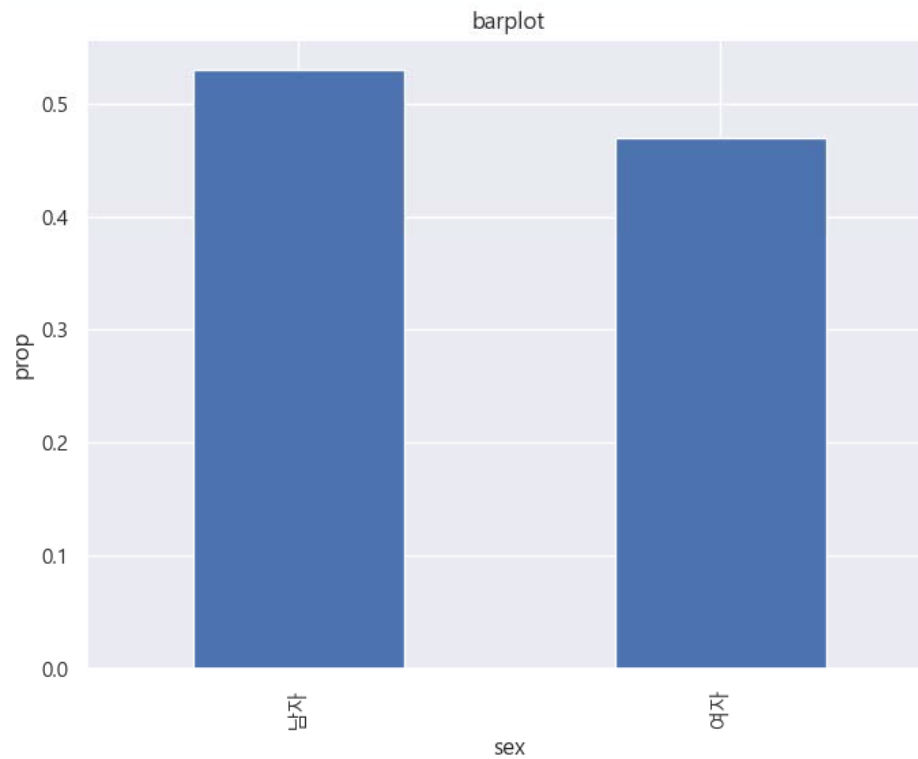
4. 범주형 변수(2개)

✓ 1초 오후 6:07에 완료됨

3. 범주형 변수(1개) (one categorical)

```
[19]: #pd_plot 이용(비율)
ax = freq_table["prop"].plot(figsize = (8, 6),
                             kind = "bar")

ax.set(title = 'barplot',
       xlabel = 'sex',
       ylabel = "prop")
plt.show()
```



4. 범주형 변수(2개)

4. 범주형 변수(2개)

4.1 분할표(Cross-tabulation)

```
[19] cross_table = pd.crosstab(index = eda_df["성별"],
                           columns = eda_df["분반"])
cross_table
```

	분반	A반	B반
성별			
남자	21	33	
여자	31	17	

Next steps: [View recommended plots](#)

```
[20] # margins
cross_table = pd.crosstab(index = eda_df["성별"],
                           columns = eda_df["분반"],
                           margins = True)
# cross_table.index = ["남자", "여자"]
cross_table
```

	분반	A반	B반	All
성별				
남자	21	33	54	
여자	31	17	48	
All	52	50	102	

Next steps: [View recommended plots](#)

✓ 1초 오후 6:07에 완료됨

4. 범주형 변수(2개)

Next steps: [View recommended plots](#)

✓

0초

[21] # 정리

```
cross_table.index = ["남자", "여자", "열전체"]
cross_table.columns = ["A반", "B반", "행전체"]
```

```
cross_table
```

	A반	B반	행전체
남자	21	33	54
여자	31	17	48
열전체	52	50	102

Next steps: [View recommended plots](#)

✓

0초

[22] # 전체비율

```
cross_table/cross_table.loc["열전체", "행전체"]
```

	A반	B반	행전체
남자	0.205882	0.323529	0.529412
여자	0.303922	0.166667	0.470588
열전체	0.509804	0.490196	1.000000

✓

0초

[23] # 열비율

```
cross_table/cross_table.loc["열전체"]
```

	A반	B반	행전체
남자	0.403846	0.66	0.529412
여자	0.596154	0.34	0.470588
열전체	1.000000	1.00	1.000000

✓

0초

[24] # 행비율

✓ 1초

오후 6:07에 완료됨

4. 범주형 변수(2개)

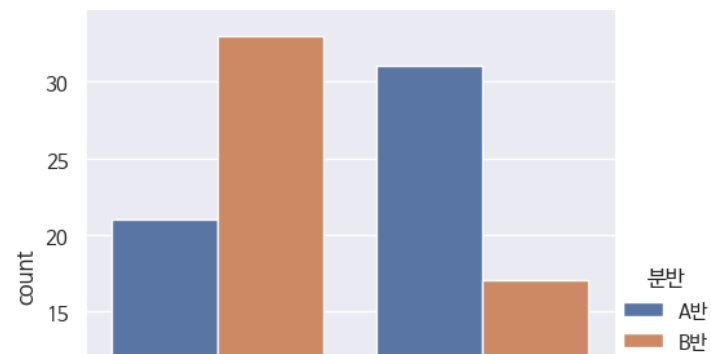
✓ 0초 [24] # 행비율
`cross_table.div(cross_table["행전체"], axis=0)`

	A반	B반	행전체
남자	0.388889	0.611111	1.0
여자	0.645833	0.354167	1.0
열전체	0.509804	0.490196	1.0

✓ 0초 [25] # 전체비율 저장
`cross_table_prop = cross_table/cross_table.loc["열전체", "행전체"]`

4.2 그래프 그리기(막대 그래프)

✓ 1초 [26] # 누적 막대그래프(count)
`sns.catplot(x = "성별",
hue = "분반",
kind = "count",
data = eda_df)
plt.show()`



✓ 1초 오후 6:07에 완료됨

4.범주형 변수(2개)



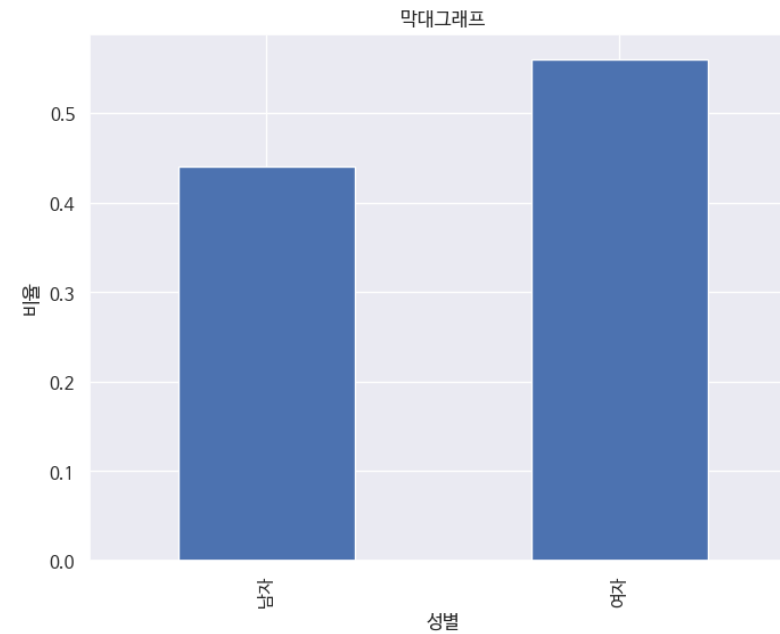
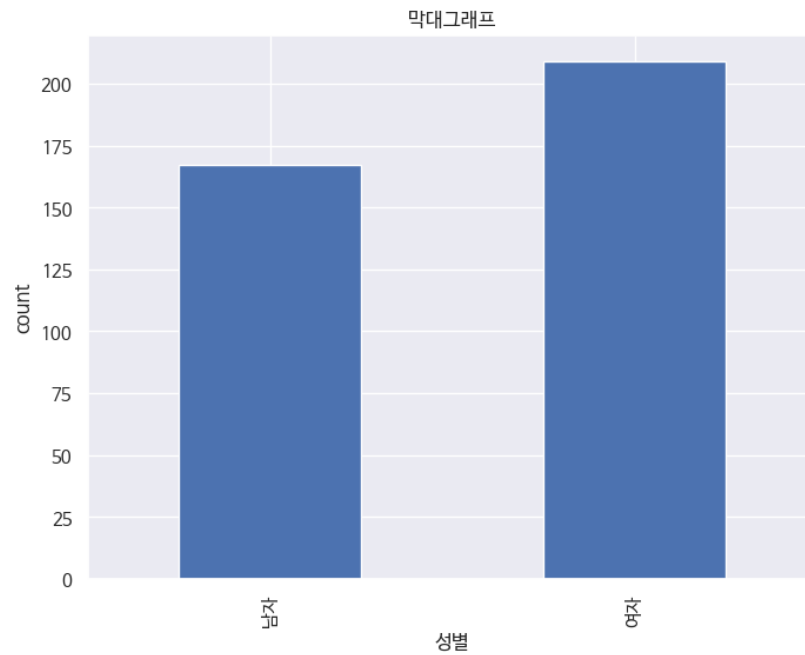
연습문제

연습문제

- ❖ 01_2.OnlineGame.csv를 이용하여 범주형 자료를 분석하세요.
- ❖ 성별: 1=남자, 2=여자
- ❖ 결혼: 1=결혼, 2=미혼
- ❖ 학력: 1=초중고생, 2=고졸, 3=대학생, 4=대졸
- ❖ 성별의 돛수분포표 및 막대그래프를 만들어 보세요
- ❖ 결혼과 학력의 교차분석표 및 막대그래프를 만들어 보세요. 비율은 전체 %로 계산하세요.

- ❖ 성별(sex)의 돛수분포표, 막대그래프(돛수), 막대그래프(비율),

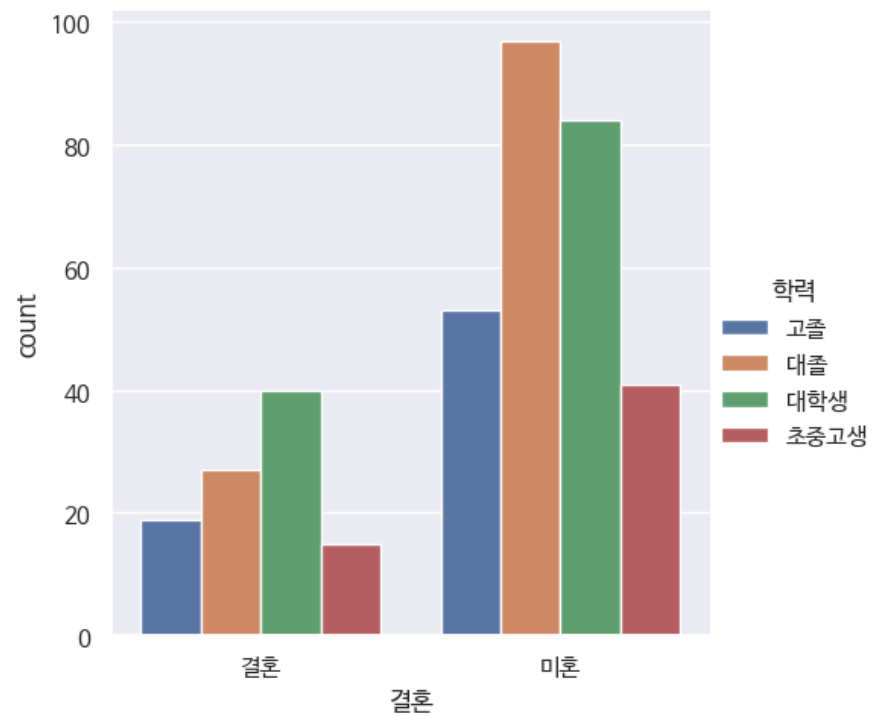
col_0	count	prop	cum_prop
성별			
남자	167	0.44	0.44
여자	209	0.56	1.00



❖ 결혼(mrg)과 학력(school)의 이원분할표, 막대그래프

	고졸	대졸	대학생	초중고생	행전체
결혼	19	27	40	15	101
미혼	53	97	84	41	275
열전체	72	124	124	56	376

	고졸	대졸	대학생	초중고생	행전체
결혼	0.050532	0.071809	0.106383	0.039894	0.268617
미혼	0.140957	0.257979	0.223404	0.109043	0.731383
열전체	0.191489	0.329787	0.329787	0.148936	1.000000



II. 수치형 자료 분석

중심위치[평균]

수치형 자료

❖ 수치형 자료 정리

- 일반적으로 연속자료의 특성을 시각적으로 파악하기 보다는 숫자로 기술
- 분포의 특성을 숫자로 표현하는 법
- 지능지수 : IQ, 경제현상:GDP, 불쾌지수 등
- 중심위치와 산포경향

❖ 중심위치(central location)

- 관찰된 자료들이 어디에 집중되어 있는가를 나타냄
- 종류 : 산술평균, 중앙값, 최빈값, 기하평균, 조화평균, 가중평균

❖ 산포경향 (변동)

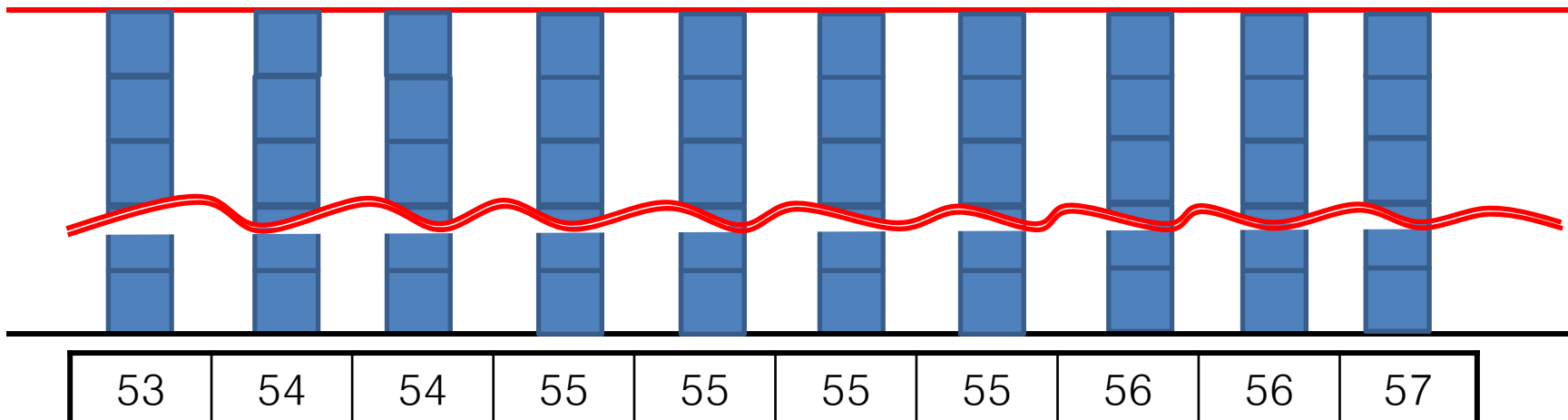
- 자료가 중심위치로부터 어느 정도 흩어져 있는가를 나타냄
- 자료가 평균으로 부터 떨어진 평균 차이(거리)
- 종류 : 범위, 편차, 분산, 표준편차

(산술)평균 (Mean)

❖ 평균

- 균등하게 나누다
- G대학 경영통계 수강생의 몸무게 분석

$$\begin{aligned}\bar{x} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\left(\sum_{i=1}^n x_i\right) \\ &= \frac{1}{10}(53 + 54 + \cdots + 56 + 57) \\ &= 55\end{aligned}$$



(산술)평균 (Mean)

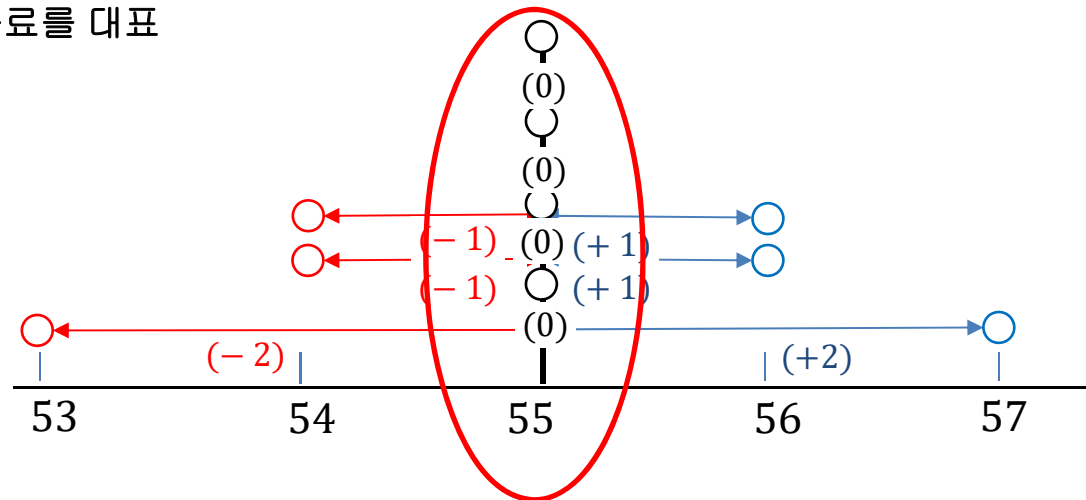
❖ 평균

- 자료의 중심적인 경향을 나타내는 수치(무게중심)
- 중심위치(central location): 분포상의 무게중심
- 중심위치: 평균을 중심으로 왼쪽 자료와 오른쪽 자료를 다 더하면 0

- 편차: $\sum (x_i - \bar{x}) = 0$

$$(53 - 55) + (54 - 55) + \dots + (56 - 55) + (57 - 55) = 0$$

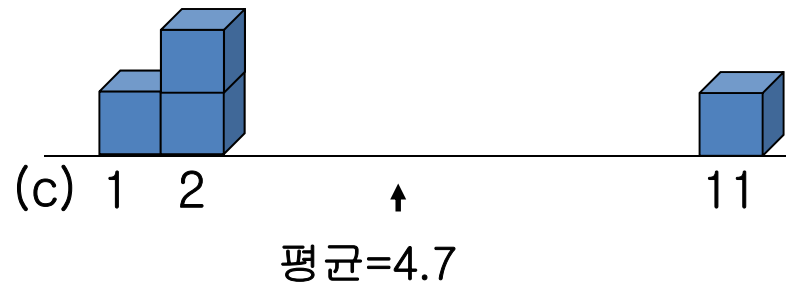
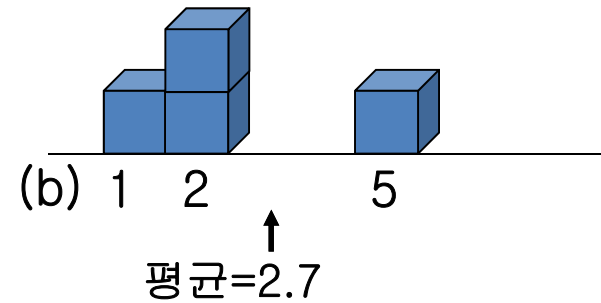
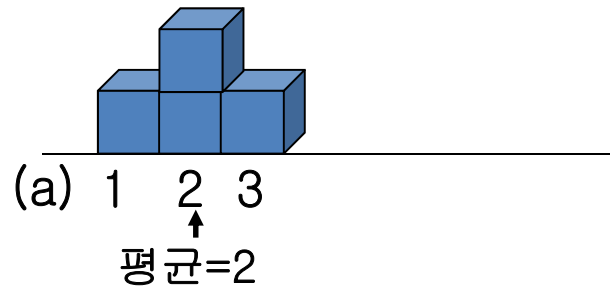
- 관찰된 자료들이 어디에 집중되어 있는가를 나타냄
- 전체 자료를 대표



(산술)평균 (Mean)

❖ 산술평균의 문제

- 이상치(outlier)에 민감하게 반응함
- 보완: 중앙값, 최빈값, 절사평균



중앙값 (Median)

- ❖ 자료를 크기 순으로 나열할 때 가장 가운데 오는 값

$$\tilde{x} = \frac{x_{n+1}}{2}$$

- ❖ 특징

- 이상치의 영향을 받지 않음
- 중앙값을 중심으로 좌우 분포면적이 같음
- 원데이터를 크기순서대로 재 배열
- 짝수일 경우에는 가운데 두개 값의 평균

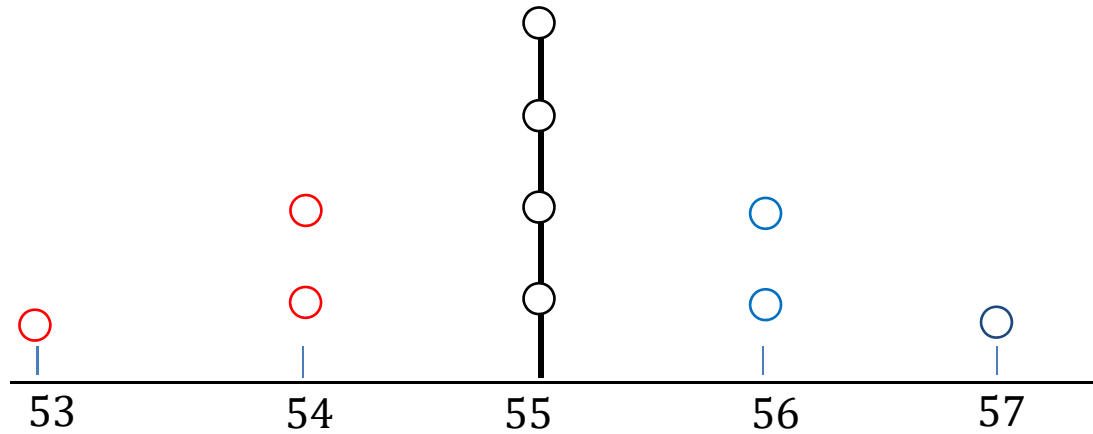
53,54, ..., 55, 55, ..., 56, 57
1 2 6 7 9 10

$$\tilde{x} = \frac{55 + 55}{2} = 55.5$$

최빈값 (Mode)

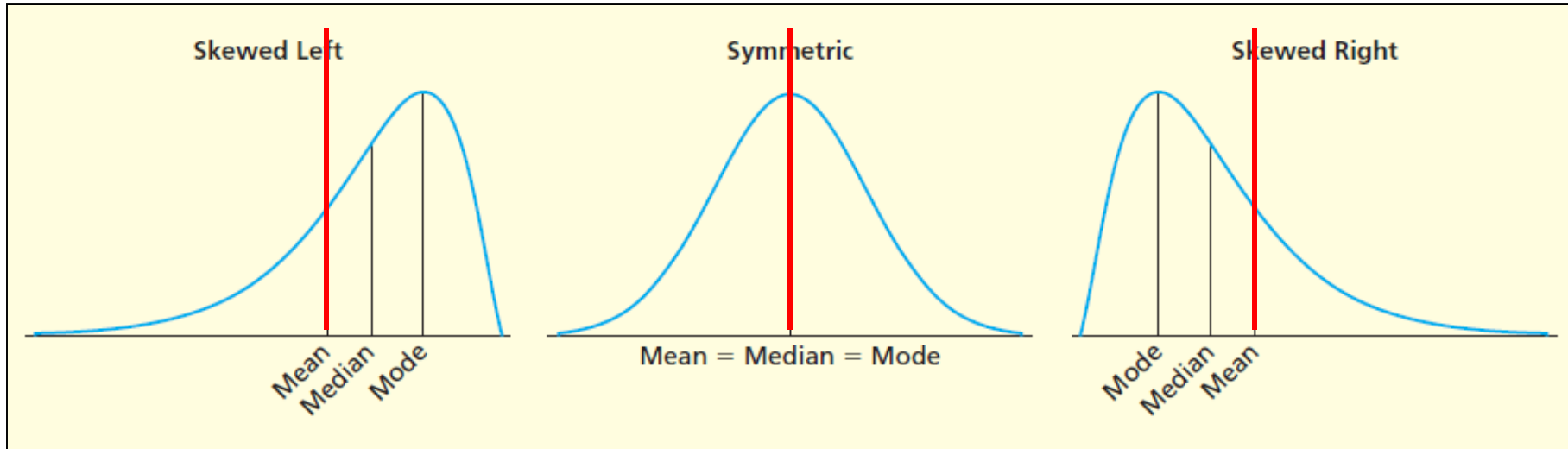
- ❖ 자료 중 발생빈도가 가장 높은 값
 - 빈도수에 의해 산출
 - 유일하지 않을 수도 있음

몸무게	빈도
53	1
54	2
55	4
56	2
57	1



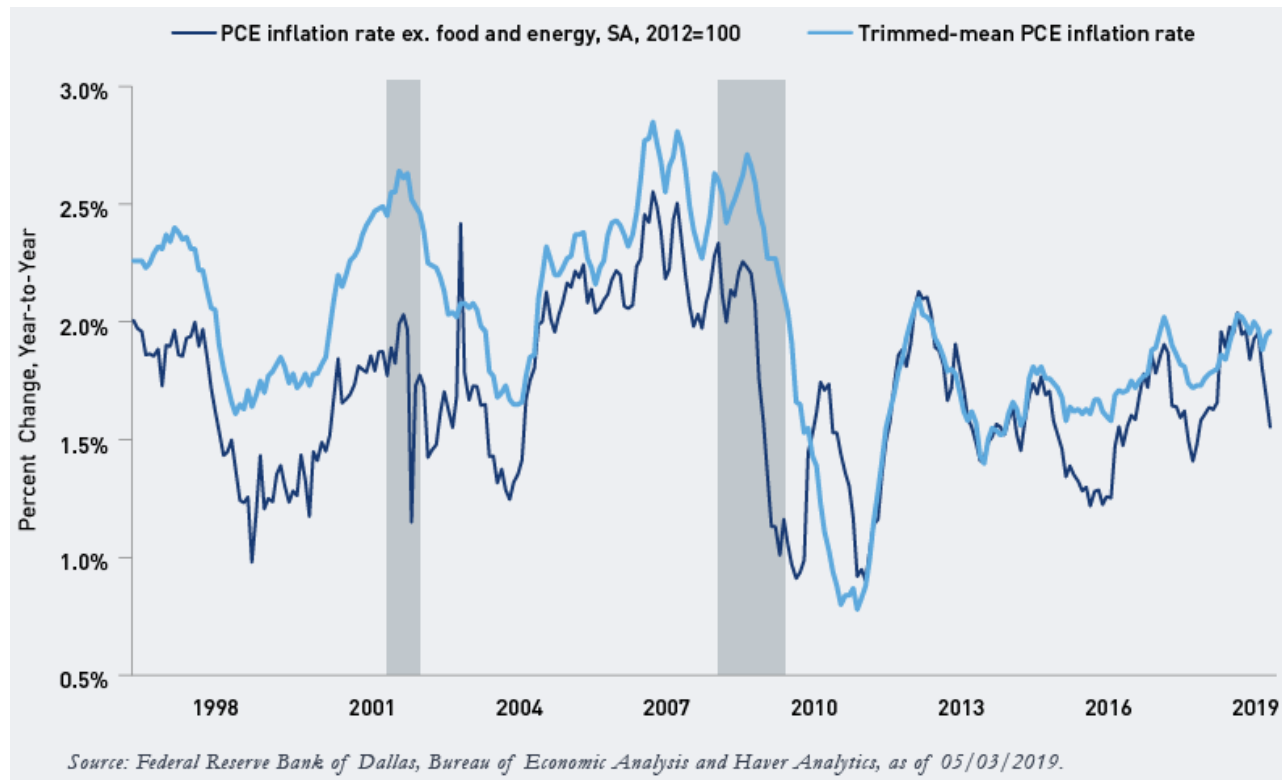
중심위치의 모양

- ❖ 자료의 분포와 평균(Mean), 중앙값(Median), 최빈값(Mode)과의 관계



절단평균(Trimmed Mean)

- ❖ 최대, 최소값 중 K개를 제외한 평균
- ❖ 스포츠경기에서 많이 사용 (극단치가 있는 경우)
 - 스포츠 경기에서는 최대값과 최소값을 빼고 계산



출처 : <https://blog.loomissayles.com/maybe-inflation-didnt-ease-a-look-at-trimmed-mean-inflation>

가중평균(Weighted Mean)

- ❖ 각 항의 수치에 그 중요도에 비례하는 계수를 곱한 다음 산출
 - 정밀도나 들어온 양이 같지 않은 물품의 평균 가격처럼 원래의 수치가 동등하지 않다고 생각되는 경우에 주로 사용

$$\bar{x}_A = \frac{\sum w_i x_i}{\sum w_i}, w_i = i\text{번째 관찰치의 가중치}$$

- ❖ 사례) 중간고사(30)와 기말고사(70)
 - A: 중간고사 95, 기말고사 80 B: 중간고사 80, 기말고사 95

$$\begin{aligned}\bar{x}_A &= \frac{(30 \times 95) + (70 \times 80)}{30 + 70} \\ &= 84.5\end{aligned}$$

$$\begin{aligned}\bar{x}_B &= \frac{(30 \times 80) + (70 \times 95)}{30 + 70} \\ &= 90.5\end{aligned}$$

가중평균(Weighted Mean)

LGE Internal Use Only

❖ 수강생이 다른 세 반을 통합한 전체 평균은?

반	수강생	반평균
A	10	60
B	50	70
C	40	80

$$\bar{x} = \frac{60 + 70 + 80}{3} = 70$$

$$\bar{x}_w = \frac{10 \times 60 + 50 \times 70 + 40 \times 80}{10 + 50 + 40} = 73$$

기하평균(Geometric Mean)

❖ 곱의 형태로 변화하는 자료

- 비율의 평균계산에 많이 사용
- 물가상승률, 인구변동률...
- 연평균 성장률
- 수학적 대칭관계(뒤장 참고)

$$\begin{aligned}
 CAGR &= \sqrt[n-1]{\frac{x_n}{x_1}} - 1 \\
 &= \sqrt[4]{\left(\frac{998}{635}\right)\left(\frac{1,265}{998}\right)\left(\frac{1,701}{1,265}\right)\left(\frac{2,363}{1,701}\right)} - 1 \\
 &= \sqrt[4]{\left(\frac{2,363}{635}\right)} - 1 \\
 &= 0.389
 \end{aligned}$$

- 평균의 의미: 매년 똑같이 38.9%씩 증가

$$\sqrt[4]{0.389} \leftrightarrow 0.389^4 = 0.389 \times 0.389 \times 0.389 \times 0.389$$

년도	수익	증가율
2010	635	
2011	998	0.572
2012	1,265	0.268
2013	1,701	0.345
2014	2,363	0.389
산술평균		0.394

년도	수익	증가율
2010	635	
2011	882	0.389
2012	1225	0.389
2013	1702	0.389
2014	2364	0.389
CAGR		0.389

- ❖ 속도 등과 같이 여러 단위가 결합되어 있을 때 계산

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

- ❖ 예제) 서울과 부산 (400km)를 왕복하는데, 가는데 시속 400km/h로 가고, 오는데 시속 100km/h로 왔다면 왕복하는데 걸린 평균 시속은?

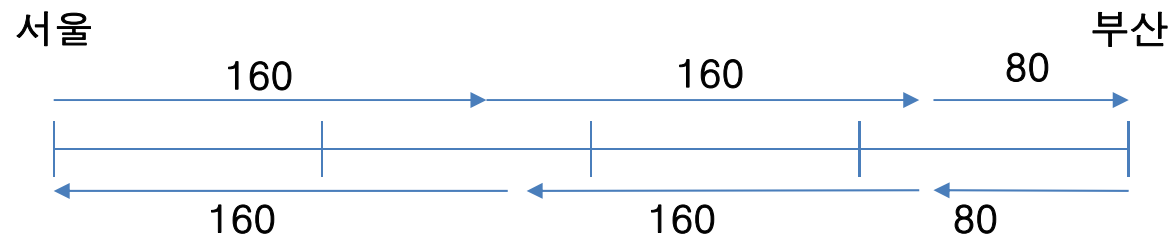
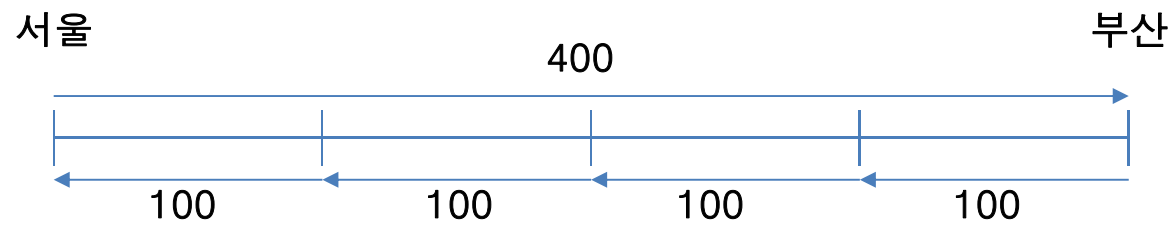
$$\bar{x}_H = \frac{2}{\frac{1}{400} + \frac{1}{100}} = \frac{2}{\frac{5}{400}} = \frac{800}{5} = 160km/h$$

평균의 의미: 5시간 동안 똑같이 나눈 속도는

조화평균(Harmonic mean)

LGE Internal Use Only

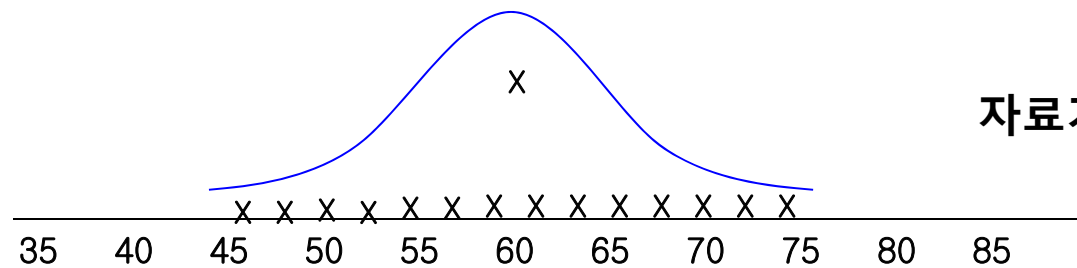
$$\bar{x}_H = \frac{2A}{\frac{A}{400} + \frac{A}{100}} = \frac{800}{\frac{400}{400} + \frac{400}{100}} = \frac{800}{5} = 160km/h$$



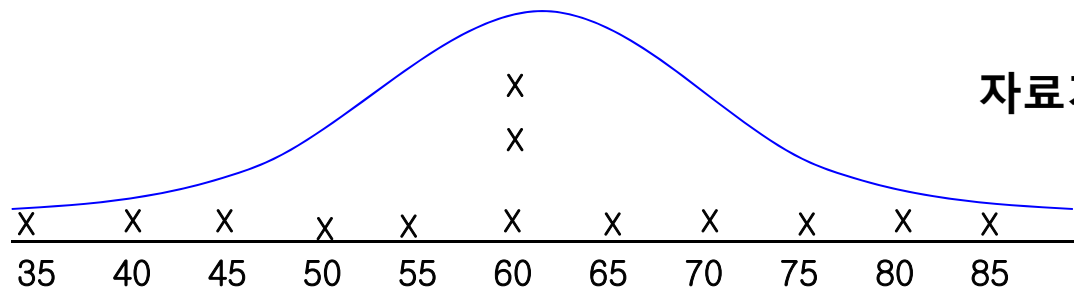
산포(변동)

❖ 산포 (Dispersion, 변동)

- 자료가 중심위치로부터 어느 정도 흩어져 있는가를 나타냄
- 자료가 평균으로 부터 떨어진 평균 차이(거리)
- 수치 자료의 특징을 정리할 때 평균과 같이 제공
- 종류 : 범위, 4분위, 편차, 분산, 표준편차
- 중심위치(평균)이 얼마나 안정적인지에 대한 정보



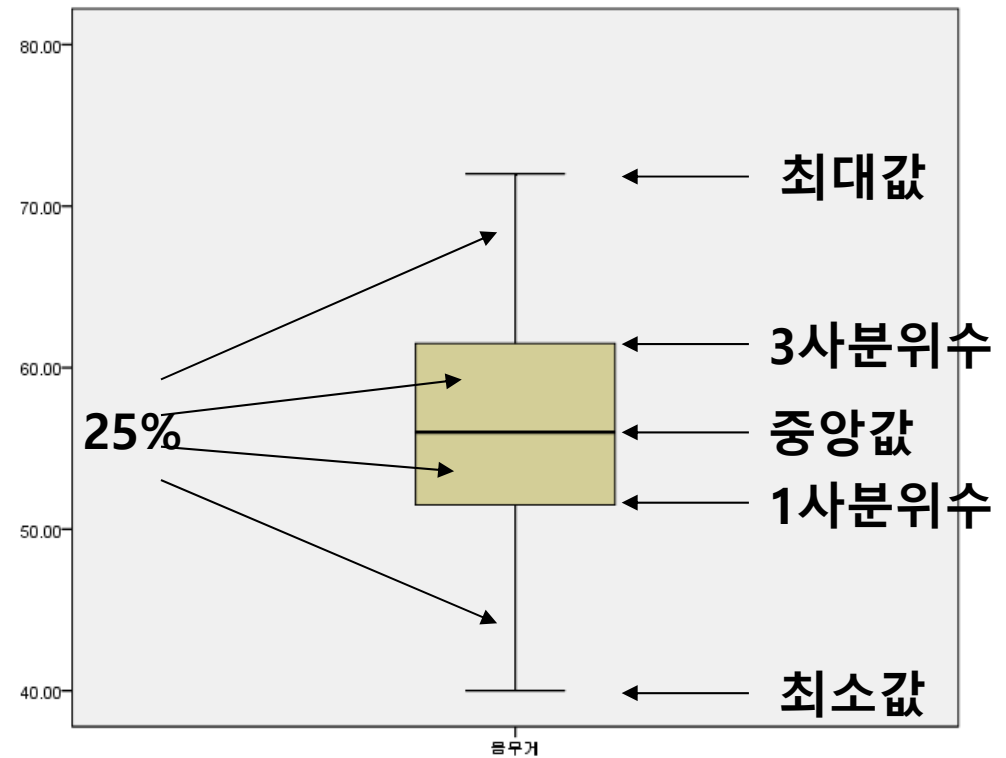
자료가 조밀하게 퍼져 있음: 안정적



자료가 넓게 퍼져 있음: 덜 안정적

❖ 사분위(Interquartile-Range)

- 자료를 동일한 비율로 4등분
- 자료를 순서대로 정렬
- 제1사분위수(Q_1): 25%
- 제2사분위수(Q_2): 50%
- 제3사분위수(Q_3): 75%
- 제4사분위수(Q_4): 100%
- 상자도표(Box Plots)에서 사용



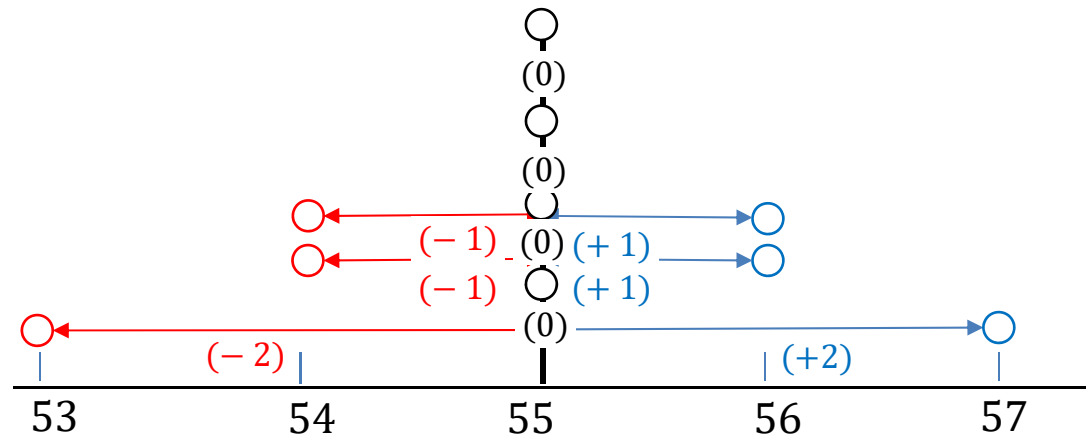
편차

❖ 편차

- 평균으로부터 데이터들이 떨어져 있는 거리
- 평균을 중심으로 왼쪽 자료와 오른쪽 자료를 다 더하면 0

$$\sum (x_i - \bar{x}) = 0$$

$$(53 - 55) + (54 - 55) + \dots + (56 - 55) + (57 - 55) = 0$$



분산과 표준편차

❖ 분산(variance)

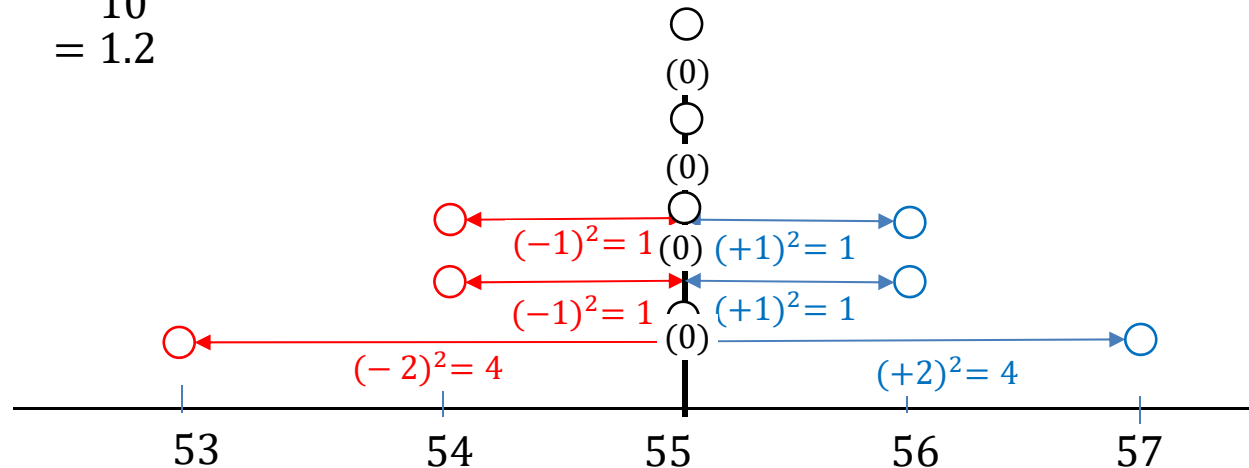
- 편차 → 분산: "-"를 없애주기 위해
- 자료가 평균을 중심으로 얼마나 광범위하게 분포하고 있는 가를 하나의 수치로 나타낸 통계량

$$var(\bar{x}) = \sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{10} (53 - 55)^2 + (54 - 55)^2 + \dots + (56 - 55)^2 + (57 - 55)^2$$

$$= \frac{1}{10} (4 + 1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 + 4)$$

$$= 1.2$$



분산과 표준편차

❖ 표준편차(Standard Deviation)

- 분산을 원 자료의 측정단위로 다시 전환하기 위해
- 자료가 평균으로 부터 떨어진 평균 차이(거리)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

$$\sigma = \sqrt{1.2} = 1.1$$

- 편차 → 분산 → 표준편차
- $\sum (x_i - \bar{x}) = 0 \rightarrow \frac{1}{n} \sum (x_i - \bar{x})^2 = \sigma^2 \rightarrow \sigma = \sqrt{\sigma^2}$

표본분산과 표본표준편차

❖ 분산(variance)

- 모분산(σ^2), 표본 분산(s^2),

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- 자유도(degree of freedom)

자유롭게 가질 수 있는 편차의 수

(n-1)의 의미: 평균은 고정

❖ 표준편차(Standard Deviation)

- 자료가 평균으로 부터 떨어진 평균 차이(거리)
- 모표준편차(σ), 표본표준편차(s)

$$s = \sqrt{s^2}$$

통계에서 표준편차가 왜 중요한가요?

분산과 표준편차

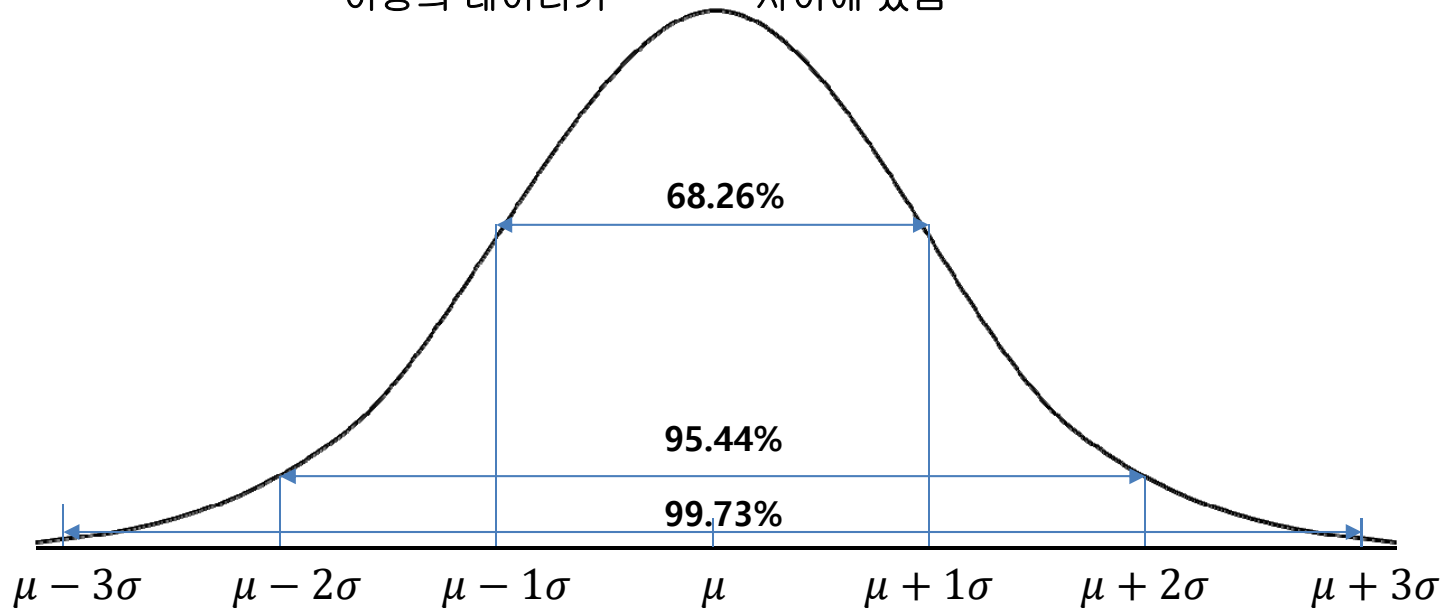
❖ 표준편차의 중요성

- 자료의 분포와 변동에 대한 중요한 정보를 제공
- 통계학의 중요한 규칙과 연결
- Empirical Rule (경험적 법칙)

$k = 1$, 68.26% 이상의 데이터가 $\mu \pm 1\sigma$ 사이에 있음

$k = 2$, 95.44% 이상의 데이터가 $\mu \pm 2\sigma$ 사이에 있음

$k = 3$, 99.73% 이상의 데이터가 $\mu \pm 3\sigma$ 사이에 있음



표준화

❖ 표준화(standardization)

- 측정단위 등과 관계없이 자료를 표준화 시킨 값
- z값으로 변환된 자료

$$z_i = \frac{x_i - \bar{x}}{s}$$

- 사례) $\bar{x} = 50kg$, $s = 5kg$ 일 때 40, 65kg인 자료의 표준화된 값은?

$$z = \frac{40 - 50}{5} = -2 \quad z = \frac{65 - 50}{5} = 3$$

- 모든 자료가 $\bar{z} = 0$, $s_z = 1$ 로 표준화 됨 → 절대 비교가 가능

변동계수(Coefficient of Variation)

- ❖ 측정단위가 다르거나 자료 값의 차이가 큰 경우에 사용

$$CV = \frac{s}{\bar{x}} \times 100$$

- (예) 유치원 여자 어린이들의 몸무게와 50대 주부의 몸무게 분포 비교

	\bar{x}	s	CV
어린이	20	8	0.40
50대 주부	55	13	0.23

- 키(cm)와 몸무게(kg)

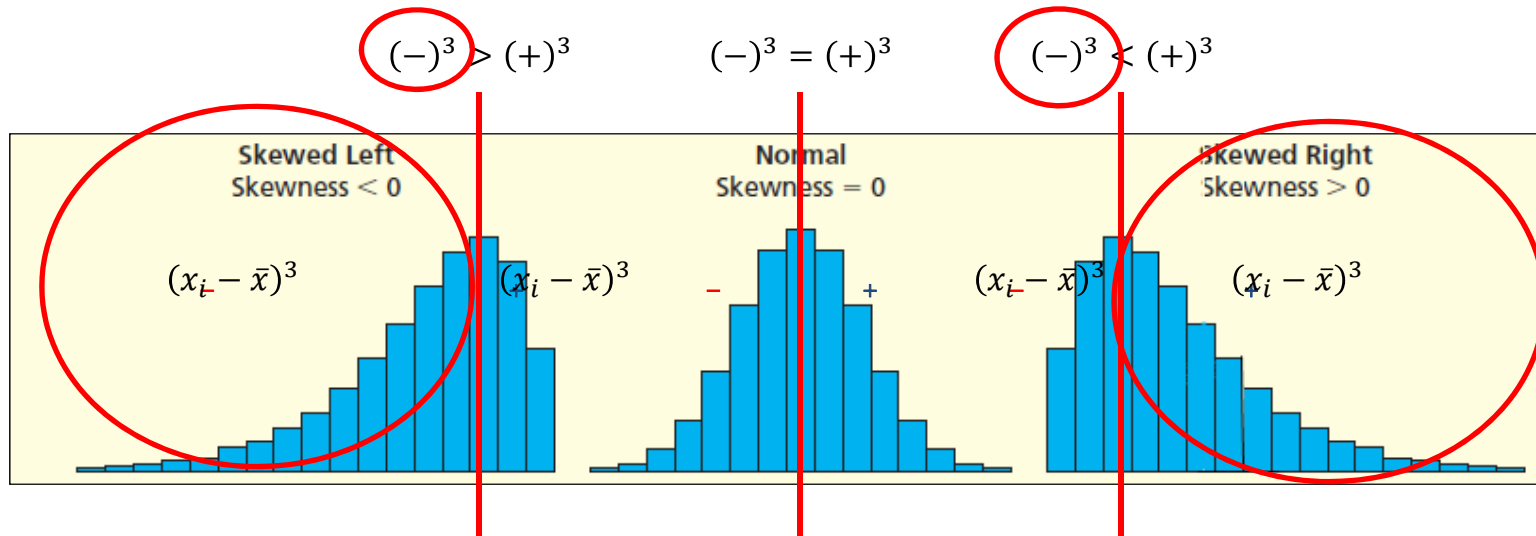
	\bar{x}	s	CV
키	175	15	0.09
몸무게	73	9	0.12

분포 형태

❖ 자료의 분포 형태

- 왜도(skewed): 자료가 평균을 중심으로 대칭인지 → 정규분포인지 확인
- 자료에 이상점이 있는지 점검

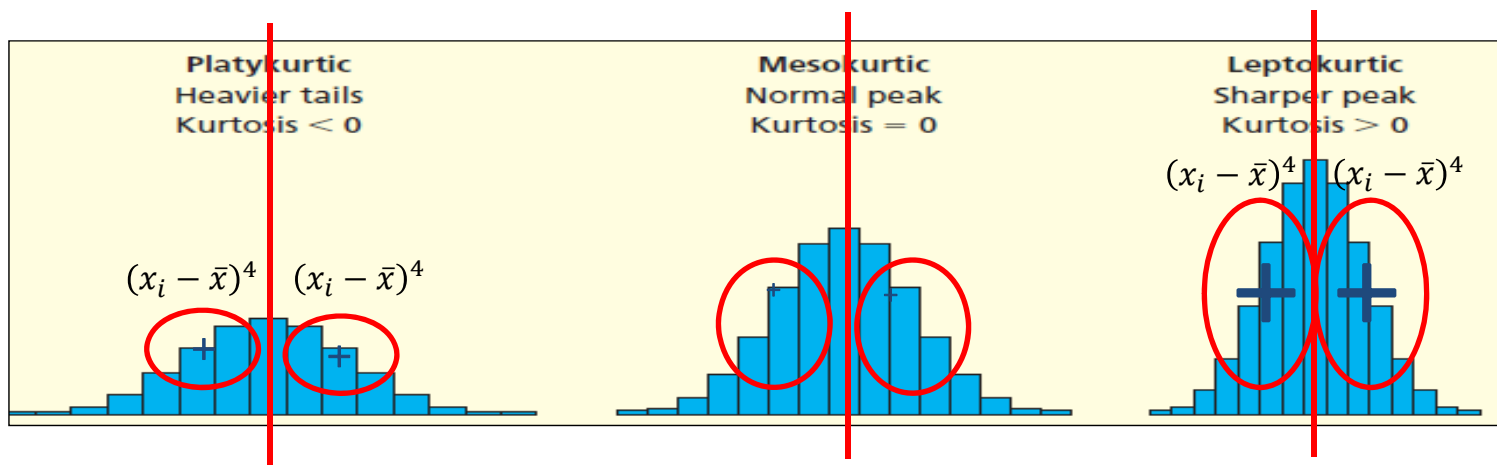
$$\sqrt{b_1} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$



❖ 자료의 분포 형태

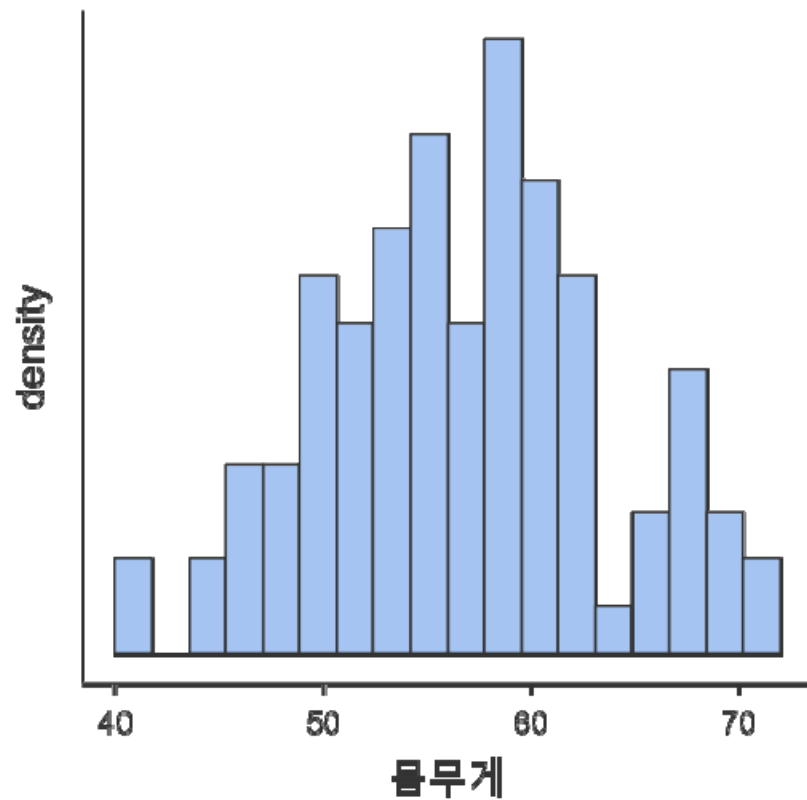
- 첨도(kurtosis): 양쪽 꼬리가 얼마나 두터운지
- 자료에 이상점이 있는지 점검

$$\sqrt{b_2} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3$$

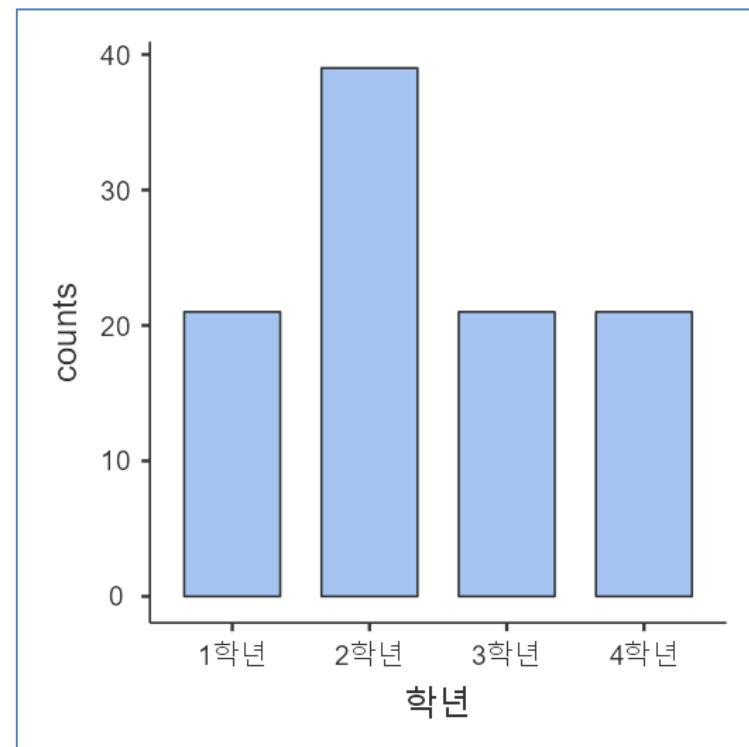


수치형 자료의 그래프

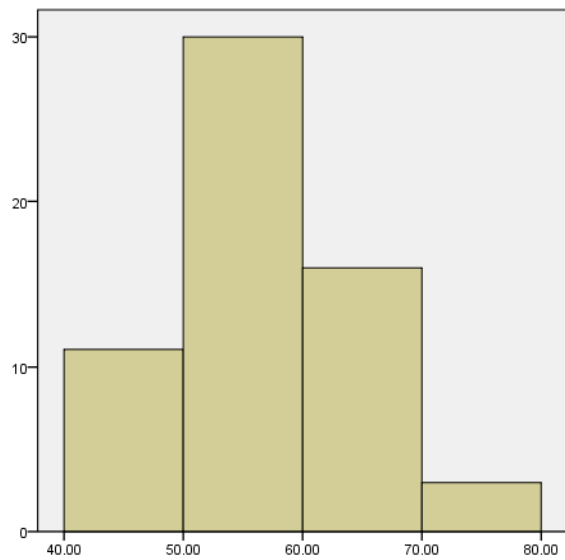
- ❖ 수치형 자료의 그래프
 - 히스토그램



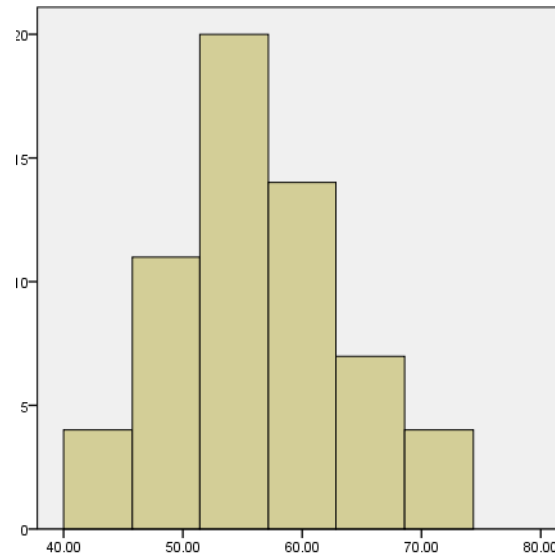
막대그래프



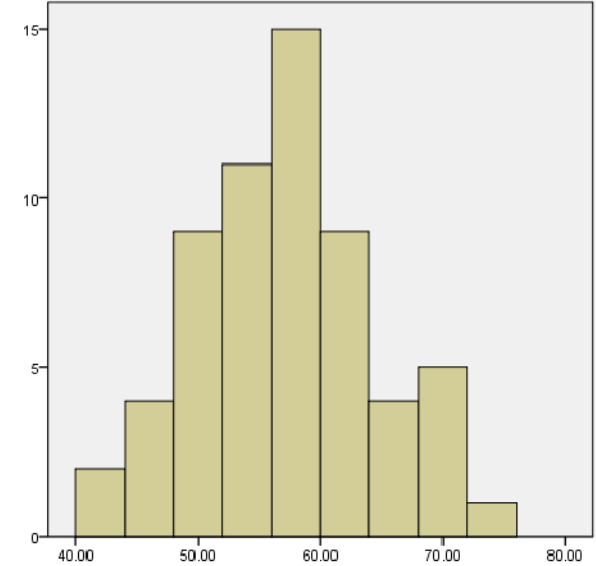
❖ 계급에 따른 그래프의 변화



계급:4



계급:7

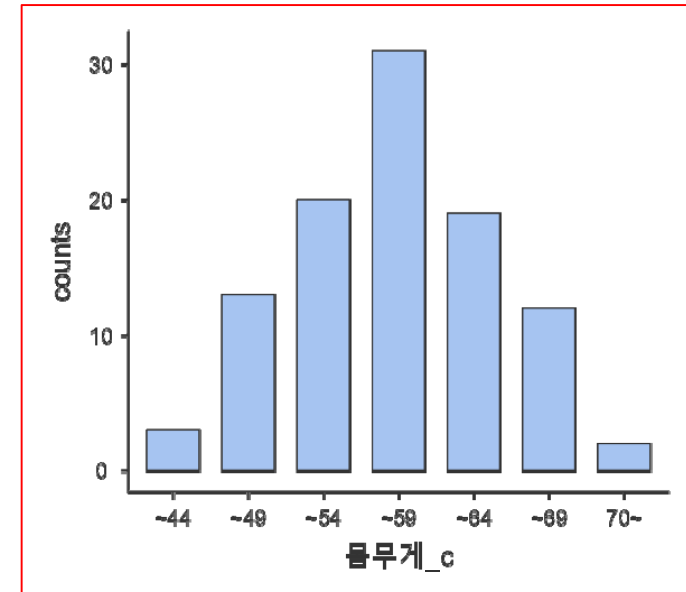


계급:10

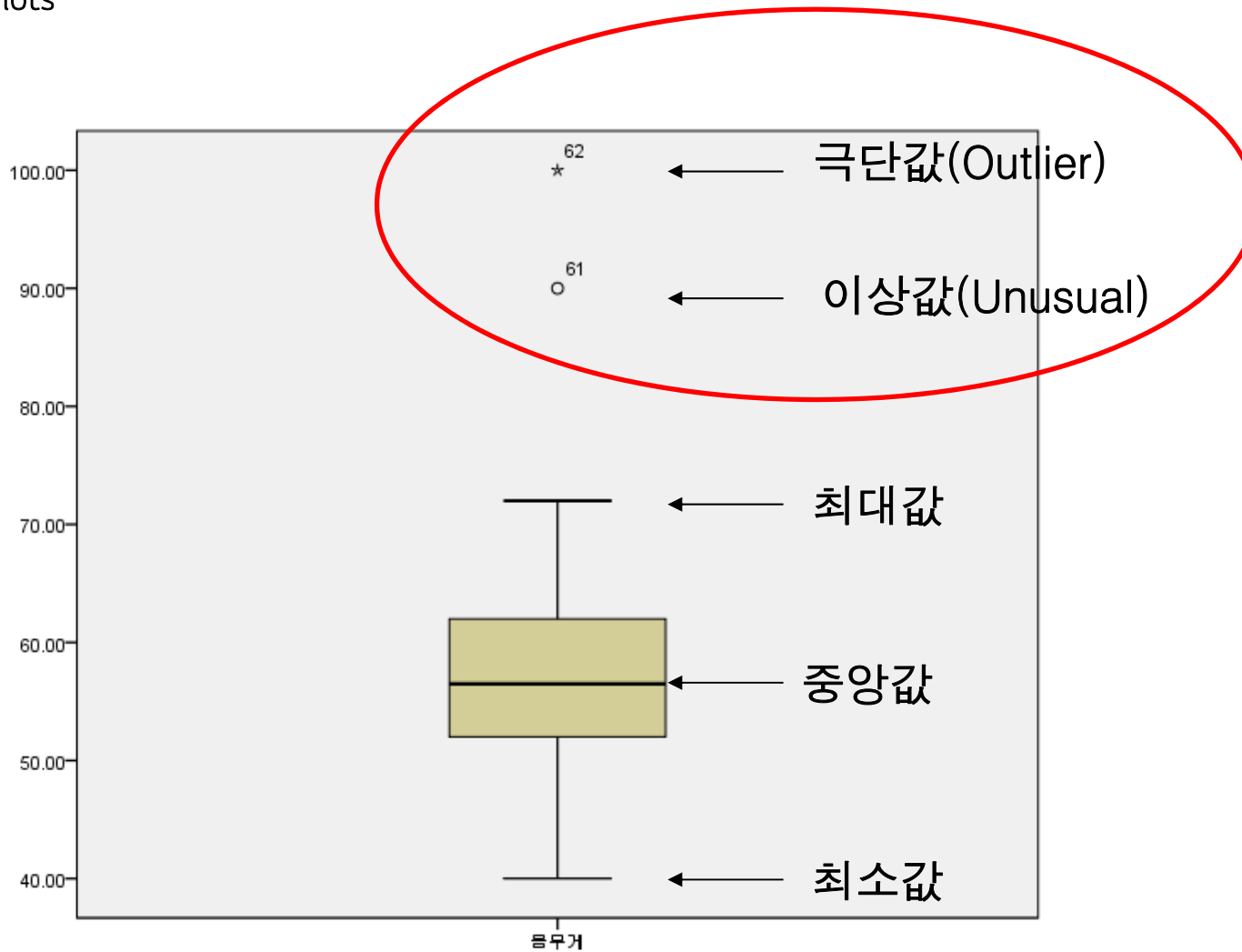
수치형 자료의 범주화

❖ 범주형 자료로 변환 후 정리

몸무게_범주	dot수	%	누적%
~44	3	3	3
~49	13	13	16
~54	20	20	36
~59	31	31	67
~64	19	19	86
~69	12	12	98
70~	2	2	100



❖ Box Plots

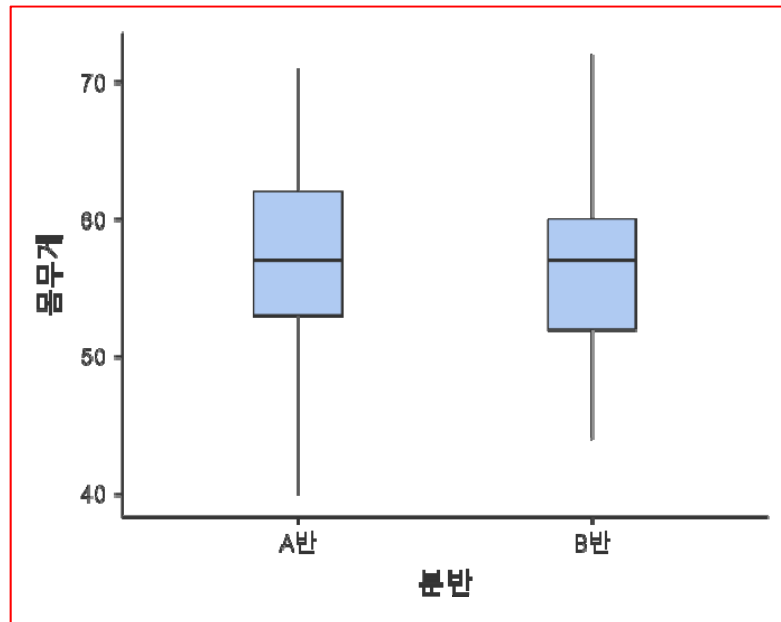


그룹별 수치자료 비교

❖ 그룹간 수치자료 비교

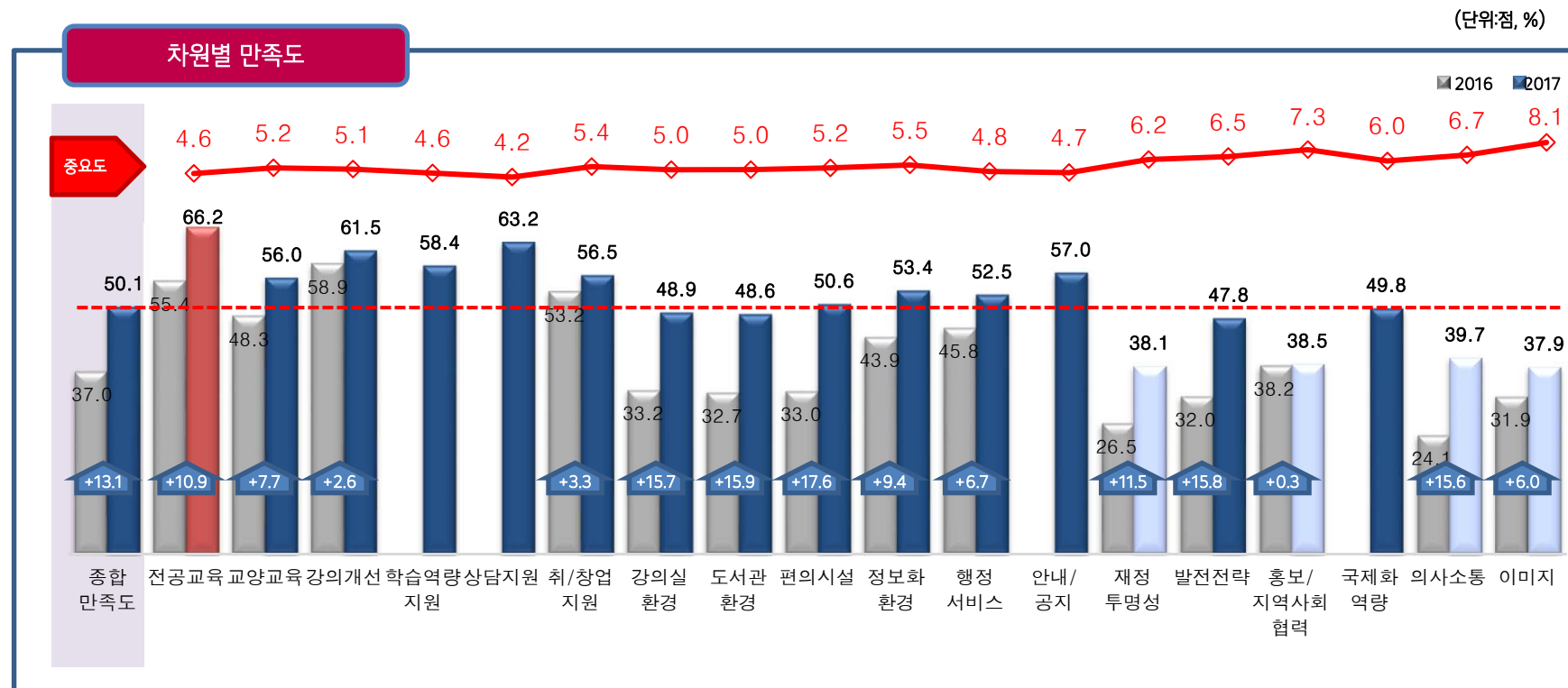
- 범주형 자료 + 수치형 자료
- 통계값: 표본크기, 평균, 표준편차

분반	N	Mean	Median	SD	Skewness	Kurtosis
A반	51	56.96	57	7.21	-0.17	-0.29
B반	49	56.59	57	6.41	0.22	-0.2



다변량 수치형 자료의 정리

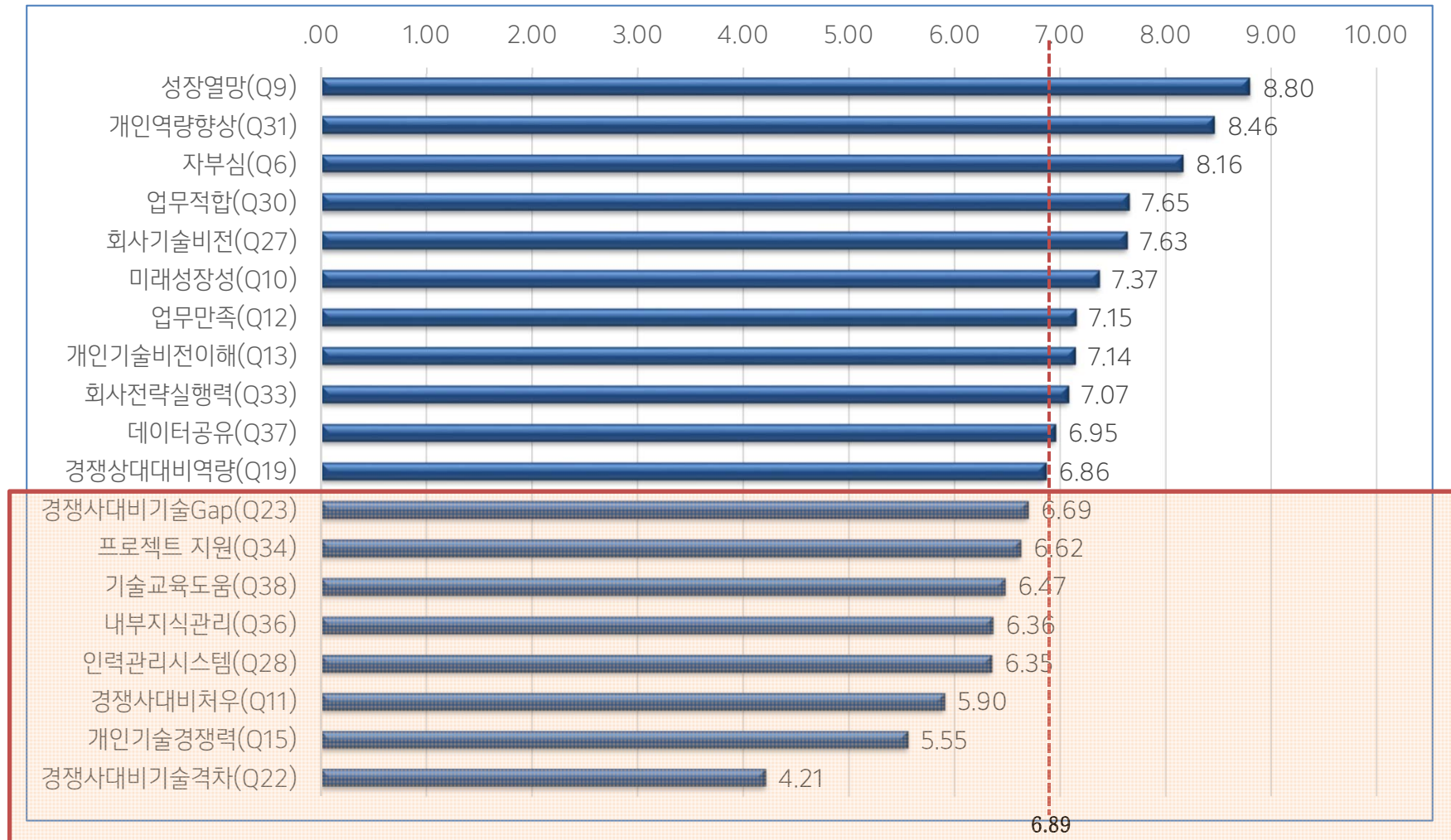
LGE Internal Use Only



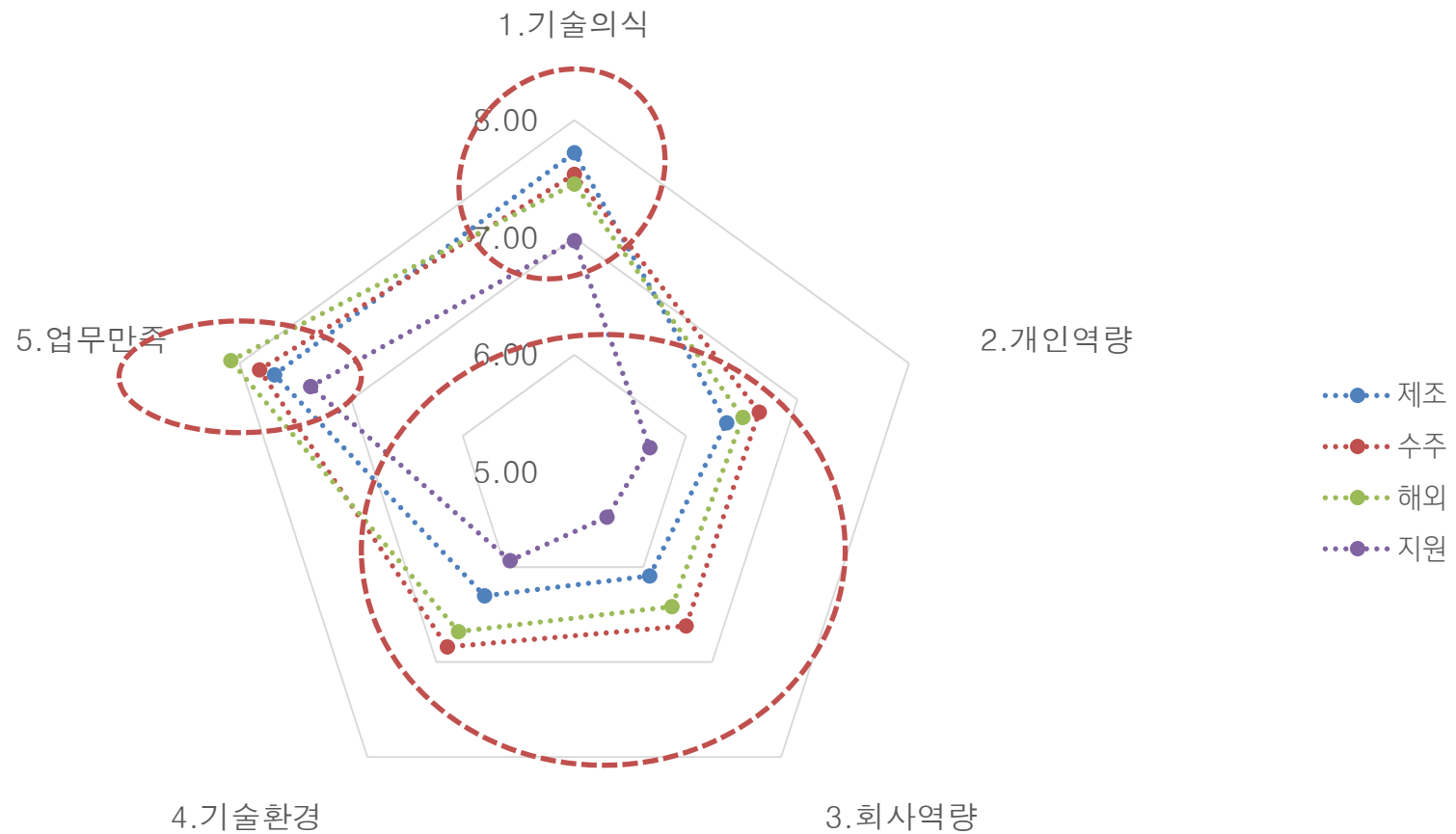
강서대학 2017 교육수요자 만족도조사

다변량 수치형 자료의 정리

LGE Internal Use Only



다변량 수치형 자료의 정리



✓ 1.package 설치

```
[1] # 1.기본
import numpy as np # numpy 패키지 가져오기
import matplotlib.pyplot as plt # 시각화 패키지 가져오기
import seaborn as sns # 시각화

# 2.데이터 가져오기
import pandas as pd # csv -> dataframe으로 전환
```

2. 데이터 불러오기

2.1 데이터 프레임으로 저장

인버테이드(csv)를 dataframe 형태로 가져오기(pandas) ✓ 0초 오후 6:18에 완료됨

2.데이터 불러오기

2.데이터 불러오기

2.1 데이터 프레임으로 저장

- 원본데이터(csv)를 dataframe 형태로 가져오기(pandas)

0초

▶

```
url = "https://raw.githubusercontent.com/leecho-bigdata/statistics-python/main/01_1.EDA.csv"
eda_df = pd.read_csv(url, encoding="cp949")
eda_df.head(10)
```

id

성별

분반

학년

몸무게

출석

중간

기말

0	1	남자	1	1	40	100	87	80
1	2	여자	2	2	50	100	83	60
2	3	남자	1	3	56	100	84	60
3	4	여자	2	4	51	100	73	60
4	5	남자	1	1	55	100	68	60
5	6	남자	2	2	61	100	77	50
6	7	여자	1	3	69	100	40	80
7	8	여자	2	2	44	100	73	30
8	9	여자	1	2	66	80	64	40
9	10	남자	2	2	60	100	66	40

Next steps:

View recommended plots

2.2 자료구조 살펴보기

0초

[4]


```
eda_df.info()
```

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 100 entries, 0 to 99

0초

오후 6:18에 완료됨


 LG
Life's Good

135/862

3. 수치형 변수(1개) (one numerical)

3. 수치형 변수(1개) (one numerical)

3.1 기술통계

```
[10] # 기술통계 사용 함수
print(eda_df.몸무게.count())
print(eda_df.몸무게.mean())
print(eda_df.몸무게.std())
print(eda_df.몸무게.min())
print(eda_df.몸무게.quantile(0.25))
print(eda_df.몸무게.quantile(0.50))
print(eda_df.몸무게.quantile(0.75))
print(eda_df.몸무게.max())
print(eda_df.몸무게.median())
print(eda_df.몸무게.mode().values[0])
print(eda_df.몸무게.skew())
print(eda_df.몸무게.kurtosis())
```

```
102
57.72549019607843
9.74339772678295
40
52.0
57.0
61.0
120
57.0
56
2.859762999904391
16.388423380051602
```

```
[11] # 필요한 변수 1개 선택
# describe() 사용
eda_df.몸무게.describe()
```

```
count    102.000000
mean      57.725490
std        9.743398
min       40.000000
25%       52.000000
```

✓ 0초 오후 6:18에 완료됨

3. 수치형 변수(1개) (one numerical)

```
[11] # 필요한 변수 1개 선택
# describe() 사용
eda_df.몸무게.describe()
```

```
count    102.000000
mean      57.725490
std        9.743398
min       40.000000
25%       52.000000
50%       57.000000
75%       61.000000
max       120.000000
Name: 몸무게, dtype: float64
```

```
[12] # table로 저장
몸무게_df = pd.DataFrame(eda_df.몸무게.describe()).T
몸무게_df
```

	count	mean	std	min	25%	50%	75%	max
몸무게	102.0	57.72549	9.743398	40.0	52.0	57.0	61.0	120.0

```
[13] # 필요한 통계수치 추가
몸무게_df["skew"] = eda_df.몸무게.skew()
몸무게_df["kurtosis"] = eda_df.몸무게.kurtosis()
몸무게_df
```

	count	mean	std	min	25%	50%	75%	max	skew	kurtosis
몸무게	102.0	57.72549	9.743398	40.0	52.0	57.0	61.0	120.0	2.859763	16.388423

```
[14] # agg 이용해서 필요한 항목만 추출
eda_df.agg({"몸무게": ["count", "mean", "std", "min", "max", "median", "skew", "kurtosis"]}).T
```

	count	mean	std	min	max	median	skew	kurtosis
몸무게	102.0	57.72549	9.743398	40.0	120.0	57.0	2.859763	16.388423

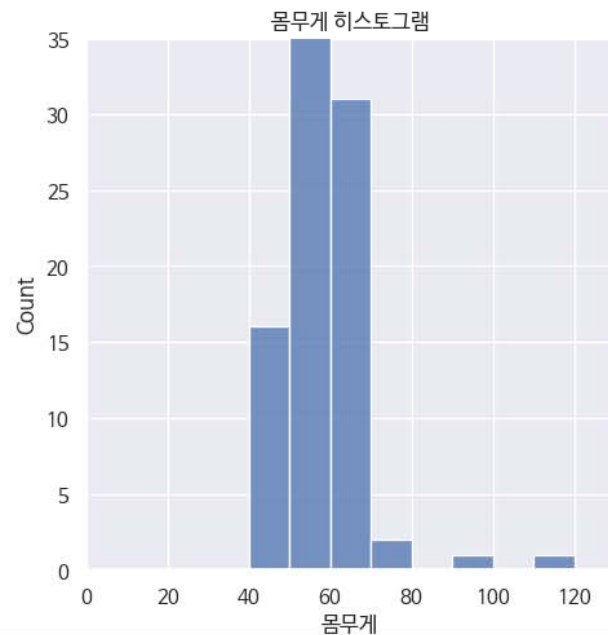
✓ 0초 오후 6:18에 완료됨

3. 수치형 변수(1개) (one numerical)

3.2 그래프 그리기

- histogram, boxplot

```
[15] # histplot
g = sns.displot(data = eda_df,
                 x = "몸무게",
                 binwidth = 10,
                 kind = "hist")
g.set(title = "몸무게 히스토그램",
      xlim = (0, 130),
      ylim = (0, 35))
plt.show()
```



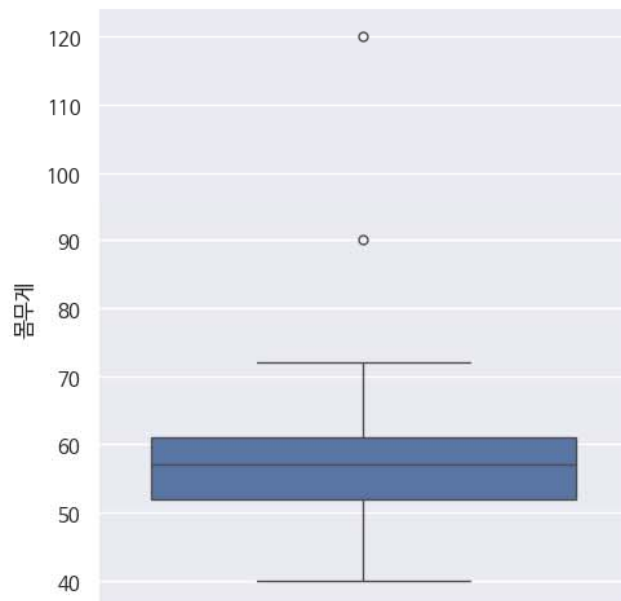
✓ 0초 오후 6:18에 완료됨

4.이상치 제거

4.이상치 제거

4.1 이상치 확인

```
[16] sns.catplot(y = "몸무게",  
             kind = "box",  
             data = eda_df)  
  
plt.show()
```



```
[17] filter = (eda_df["몸무게"] >= 80)  
eda_df.loc[filter]
```

✓ 0초 오후 6:18에 완료됨

4.이상치 제거

```
[17] filter = (eda_df["몸무게"] >= 80)
      eda_df.loc[filter]
```

	id	성별	분반	학년	몸무게	출석	중간	기말
100	101	남자	B반	2학년	90	90	54	10
101	102	여자	A반	2학년	120	100	49	40

4.2 이상치 제거

```
[18] eda_df.drop(eda_df[filter].index, inplace = True)
```

```
[19] sns.catplot(y = "몸무게",
              kind = "box",
              data = eda_df)
      plt.show()
```



✓ 0초 오후 6:18에 완료됨

5. 수치형 변수를 범주형으로 변환

5. 수치형 변수를 범주형으로 변환

- 예제: 50세
- `[: <= , >=`
- `(: < , >`
- `right = False`: `45 <= x < 50` (~미만) [50, 55)
- `right = True(default)`: `45 < x <= 50` (~이하) (45, 50]

5.1 범위 확인

```
[20] pd.cut(x = eda_df['몸무게'], bins = 7) #
      .value_counts()
```

```
(53.714, 58.286]    28
(58.286, 62.857]    23
(49.143, 53.714]    16
(44.571, 49.143]    13
(62.857, 67.429]     9
(67.429, 72.0]      8
(39.968, 44.571]     3
Name: 몸무게, dtype: int64
```

5.2 범주형 변환후 저장

```
[21] bins = [0, 45, 50, 55, 60, 65, 70, 100]
      eda_df['몸무게_bin'] = pd.cut(x = eda_df['몸무게'],
                                   bins = bins,
                                   right = False)

      eda_df.head(10)
```

	id	성별	분반	학년	몸무게	출석	중간	기말	몸무게_bin
0	1	남자	A반	1학년	40	100	87	80	[0, 45)
1	2	여자	B반	2학년	50	100	83	60	[50, 55)

✓ 0초 오후 6:18에 완료됨

5. 수치형 변수를 범주형으로 변환

Next steps: [View recommended plots](#)

```
[22] bins = [0, 45, 50, 55, 60, 65, 70, 100]
      label = ["~45미만", "45~50미만", "50~55미만", "55~60미만",
              "60~65미만", "65~70미만", "70이상~"]
      eda_df['몸무게_c'] = pd.cut(x = eda_df['몸무게'],
                                bins = bins,
                                labels = label,
                                right = False)

      eda_df.head(10)
```

	id	성별	분반	학년	몸무게	출석	중간	기말	몸무게_bin	몸무게_c
0	1	남자	A반	1학년	40	100	87	80	[0, 45)	~45미만
1	2	여자	B반	2학년	50	100	83	60	[50, 55)	50~55미만
2	3	남자	A반	3학년	56	100	84	60	[55, 60)	55~60미만
3	4	여자	B반	4학년	51	100	73	60	[50, 55)	50~55미만
4	5	남자	A반	1학년	55	100	68	60	[55, 60)	55~60미만
5	6	남자	B반	2학년	61	100	77	50	[60, 65)	60~65미만
6	7	여자	A반	3학년	69	100	40	80	[65, 70)	65~70미만
7	8	여자	B반	2학년	44	100	73	30	[0, 45)	~45미만
8	9	여자	A반	2학년	66	80	64	40	[65, 70)	65~70미만
9	10	남자	B반	2학년	60	100	66	40	[60, 65)	60~65미만

Next steps: [View recommended plots](#)

5.3 그래프 그리기

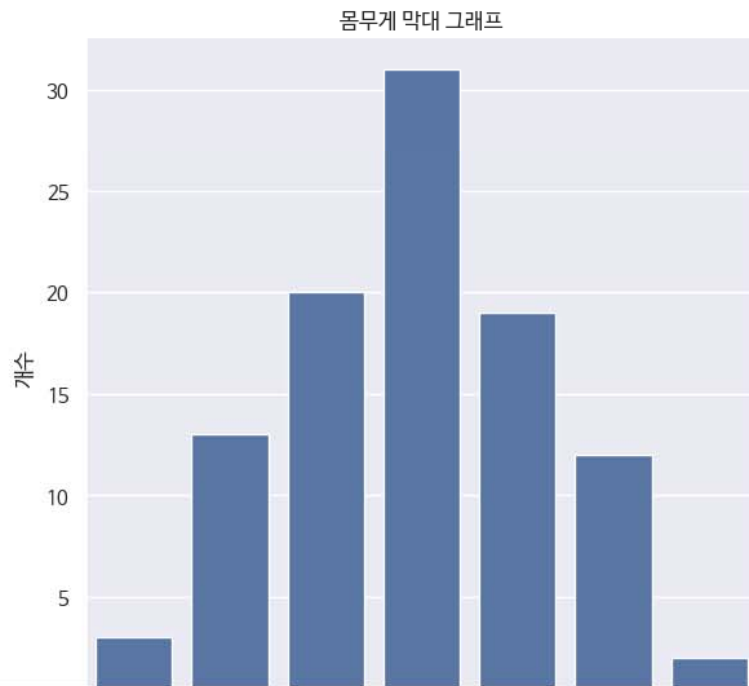
```
[23] # 범주형 변수(막대그래프)
      g = sns.catplot(data = eda_df,
                      height = 6,
                      x = "몸무게_c",
```

✓ 0초 오후 6:18에 완료됨

5. 수치형 변수를 범주형으로 변환

5.3 그래프 그리기

```
[23] # 범주형 변수(막대그래프)
g = sns.catplot(data = eda_df,
                height = 6,
                x = "몸무게_c",
                kind = "count")
g.set(title = "몸무게 막대 그래프",
      xlabel = "몸무게",
      ylabel = "개수")
plt.show()
```



✓ 0초 오후 6:18에 완료됨

6.수치형 1개 + 범주형 1개

6. 수치형 1개 + 범주형 1개

6.1 기술통계(그룹별)

✓ 0초 [24] # 그룹별 분석
eda_df.groupby('성별')['몸무게'].mean().T.round(2)

성별
남자 57.04
여자 56.49
Name: 몸무게, dtype: float64

✓ 0초 [25] eda_df.groupby('성별')['몸무게'].describe().round(2)

	count	mean	std	min	25%	50%	75%	max
성별								
남자	53.0	57.04	6.58	40.0	53.0	57.0	61.0	71.0
여자	47.0	56.49	7.10	44.0	51.0	57.0	60.5	72.0

✓ 0초 [26] eda_df.groupby('성별') #
.agg({"몸무게": ["count", "mean", "std", "min", "max", "median", "skew"]}) #
.round(2) # groupby에서는 kurtosis 지원x

	count	mean	std	min	max	median	skew
몸무게							
성별							
남자	53	57.04	6.58	40	71	57.0	-0.31
여자	47	56.49	7.10	44	72	57.0	0.29

6.2 그래프 그리기

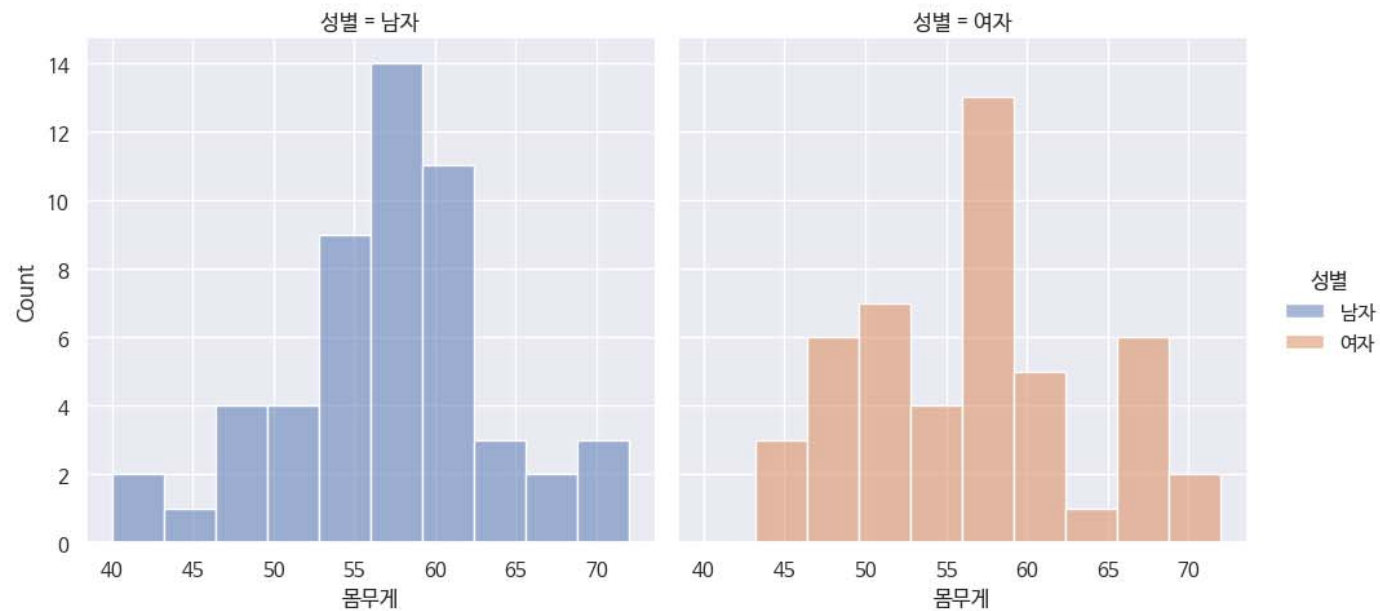
✓ 0초 오후 6:18에 완료됨

6.수치형 1개 + 범주형 1개

6.2 그래프 그리기

✓ 2초
[27] # 범주형 변수로 구분
sns.displot(data = eda_df,
 x = "몸무게",
 bins = 10,
 hue = "성별",
 col = "성별",
 kind = "hist")

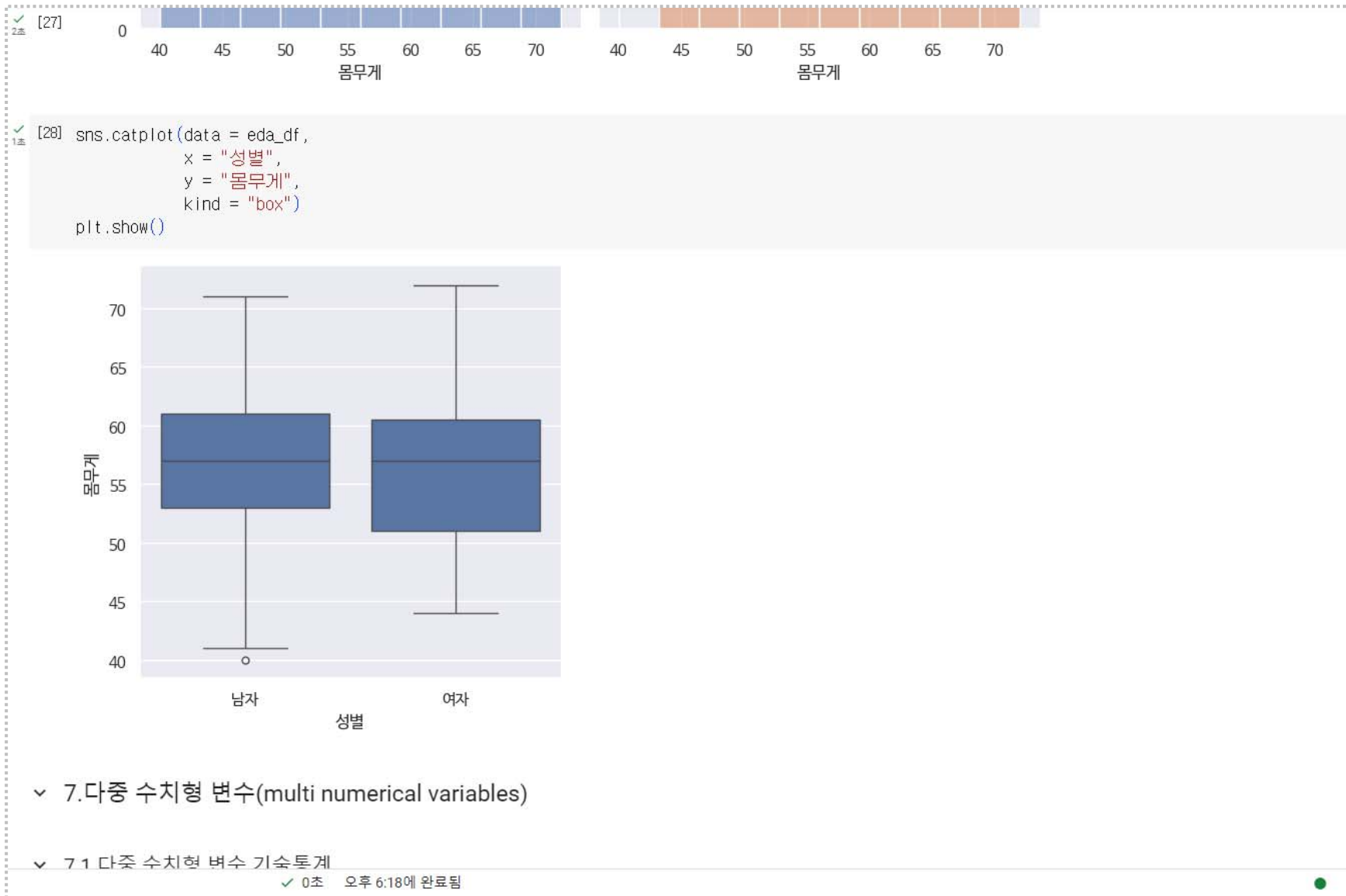
plt.show()



✓ 1초
[28] sns.catplot(data = eda_df,
 x = "성별",

✓ 0초 오후 6:18에 완료됨

6.수치형 1개 + 범주형 1개



7.다중 수치형 변수(multi numerical variables)

7.다중 수치형 변수(multi numerical variables)

7.1 다중 수치형 변수 기술통계

0초 # pd.describe 이용: 수치형 자료만 분석
eda_df.describe()

	id	몸무게	출석	중간	기말
count	100.000000	100.000000	100.000000	100.000000	100.000000
mean	50.500000	56.780000	95.400000	55.710000	36.500000
std	29.011492	6.80104	10.290723	19.692738	26.982037
min	1.000000	40.000000	60.000000	5.000000	10.000000
25%	25.750000	52.000000	100.000000	44.000000	10.000000
50%	50.500000	57.000000	100.000000	58.000000	30.000000
75%	75.250000	61.000000	100.000000	71.000000	56.250000
max	100.000000	72.000000	100.000000	92.000000	100.000000

0초 [30] # 범주형 변수까지 분석
eda_df.describe(include = 'all').T

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
id	100.0	NaN	NaN	NaN	50.5	29.011492	1.0	25.75	50.5	75.25	100.0
성별	100	2	남자	53	NaN	NaN	NaN	NaN	NaN	NaN	NaN
분반	100	2	A반	51	NaN	NaN	NaN	NaN	NaN	NaN	NaN
학년	100	4	2학년	37	NaN	NaN	NaN	NaN	NaN	NaN	NaN
몸무게	100.0	NaN	NaN	NaN	56.78	6.80104	40.0	52.0	57.0	61.0	72.0
출석	100.0	NaN	NaN	NaN	95.4	10.290723	60.0	100.0	100.0	100.0	100.0
중간	100.0	NaN	NaN	NaN	55.71	19.692738	5.0	44.0	58.0	71.0	92.0

0초 오후 6:18에 완료됨

7.다중 수치형 변수(multi numerical variables)

```
[30] 기말 100.0 NaN NaN NaN 36.5 26.982037 10.0 10.0 30.0 56.25 100.0
      몸무게_bin 100 7 [55, 60) 31 NaN NaN NaN NaN NaN NaN NaN
      몸무게_c 100 7 55~60미만 31 NaN NaN NaN NaN NaN NaN NaN
```

```
[31] # 필요한 변수만 선택
eda_df.columns

Index(['id', '성별', '분반', '학년', '몸무게', '출석', '중간', '기말', '몸무게_bin', '몸무게_c'], dtype='object')
```

```
[32] num_feature = ['몸무게', '출석', '중간', '기말']
eda_df[num_feature].describe().T.round(2)
```

	count	mean	std	min	25%	50%	75%	max
몸무게	100.0	56.78	6.80	40.0	52.0	57.0	61.00	72.0
출석	100.0	95.40	10.29	60.0	100.0	100.0	100.00	100.0
중간	100.0	55.71	19.69	5.0	44.0	58.0	71.00	92.0
기말	100.0	36.50	26.98	10.0	10.0	30.0	56.25	100.0

```
[33] # agg 이용해서 필요한 항목만 추출
eda_df[num_feature].agg(["count", "mean", "std", "min", "max", "median", "skew", "kurtosis"]).T.round(3)
```

	count	mean	std	min	max	median	skew	kurtosis
몸무게	100.0	56.78	6.801	40.0	72.0	57.0	-0.004	-0.290
출석	100.0	95.40	10.291	60.0	100.0	100.0	-2.209	4.052
중간	100.0	55.71	19.693	5.0	92.0	58.0	-0.392	-0.238
기말	100.0	36.50	26.982	10.0	100.0	30.0	0.817	-0.328

7.2 그룹별 기술통계

```
[34] # 그룹별 분석
eda_df.groupby('성별')[num_feature].mean().T.round(2)
```

✓ 0초 오후 6:18에 완료됨

7.다중 수치형 변수(multi numerical variables)

7.2 그룹별 기술통계

[34] # 그룹별 분석
eda_df.groupby('성별')[num_feature].mean().T.round(2)

성별	남자	여자
몸무게	57.04	56.49
출석	96.23	94.47
중간	53.17	58.57
기말	35.94	37.13

[35] eda_df.groupby('성별')[num_feature].describe().round(2)

	몸무게								출석								중간								기말							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
성별																																
남자	53.0	57.04	6.58	40.0	53.0	57.0	61.0	71.0	53.0	96.23	...	67.0	87.0	53.0	35.94	26.80	10.0	10.0	30.0	50.0	100.0											
여자	47.0	56.49	7.10	44.0	51.0	57.0	60.5	72.0	47.0	94.47	...	73.0	92.0	47.0	37.13	27.46	10.0	10.0	30.0	60.0	100.0											

2 rows x 32 columns

[36] eda_df.groupby('성별')[num_feature].agg(["count", "mean", "std", "min", "max", "median", "skew"]).round(3)

	몸무게								출석								중간								기말							
	count	mean	std	min	max	median	skew	count	mean	std	min	max	median	skew	count	mean	std	min	max	median	skew	count	mean	std	min	max	median	skew				
성별																																
남자	53	57.038	6.581	40	71	57.0	-0.31	53	96.226	8.82	...	87	53.0	-0.465	53	35.943	26.802	10	100	30.0	0.837											
여자	47	56.489	7.101	44	72	57.0	0.29	47	94.468	11.76	...	92	62.0	-0.381	47	37.128	27.460	10	100	30.0	0.820											

2 rows x 28 columns

0초 오후 6:18에 완료됨

7.다중 수치형 변수(multi numerical variables)

✓ 0초

[36]

성별

	count	mean	std	min	max	median	skew
남자	53	57.038	6.581	40	71	57.0	-0.31
여자	47	56.489	7.101	44	72	57.0	0.29

2 rows × 8 columns

✓ 0초

[37]

```
for num in num_feature:
    print("----", num, "----")
    results = eda_df.groupby('성별')[num].describe().round(2)
    print(results, "\n\n")
```

---- 몸무게 ----

	count	mean	std	min	25%	50%	75%	max
성별								
남자	53.0	57.04	6.58	40.0	53.0	57.0	61.0	71.0
여자	47.0	56.49	7.10	44.0	51.0	57.0	60.5	72.0

---- 흡석 ----

	count	mean	std	min	25%	50%	75%	max
성별								
남자	53.0	96.23	8.82	60.0	100.0	100.0	100.0	100.0
여자	47.0	94.47	11.76	60.0	100.0	100.0	100.0	100.0

---- 중간 ----

	count	mean	std	min	25%	50%	75%	max
성별								
남자	53.0	53.17	19.31	5.0	45.0	53.0	67.0	87.0
여자	47.0	58.57	19.93	10.0	43.5	62.0	73.0	92.0

---- 기말 ----

	count	mean	std	min	25%	50%	75%	max
성별								
남자	53.0	35.94	26.80	10.0	10.0	30.0	50.0	100.0
여자	47.0	37.13	27.46	10.0	10.0	30.0	60.0	100.0

✓ 1초

[38]

```
# 그룹별
sns.relplot(data = eda_df,
            y = "주거")
```

✓ 0초

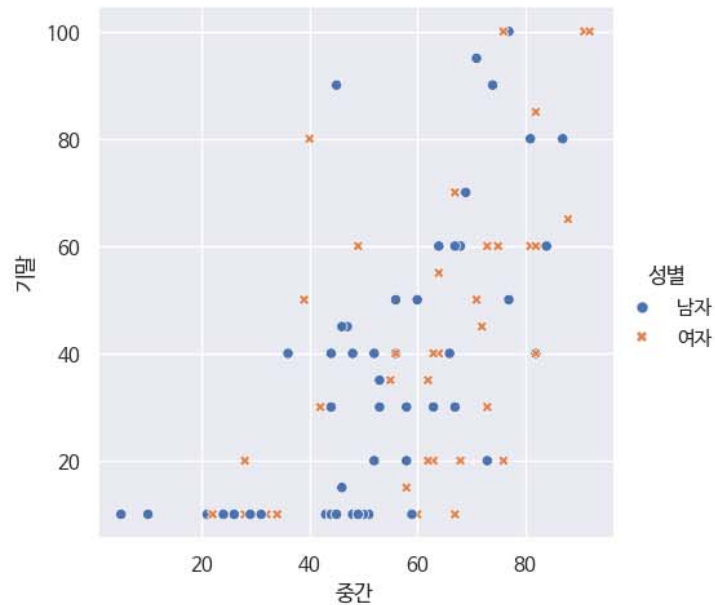
오후 6:18에 완료됨

7.다중 수치형 변수(multi numerical variables)

7.3 그래프 그리기

```
[38] # 그룹별
sns.relplot(data = eda_df,
            x = "중간",
            y = "기말",
            hue = "성별",
            style = "성별")

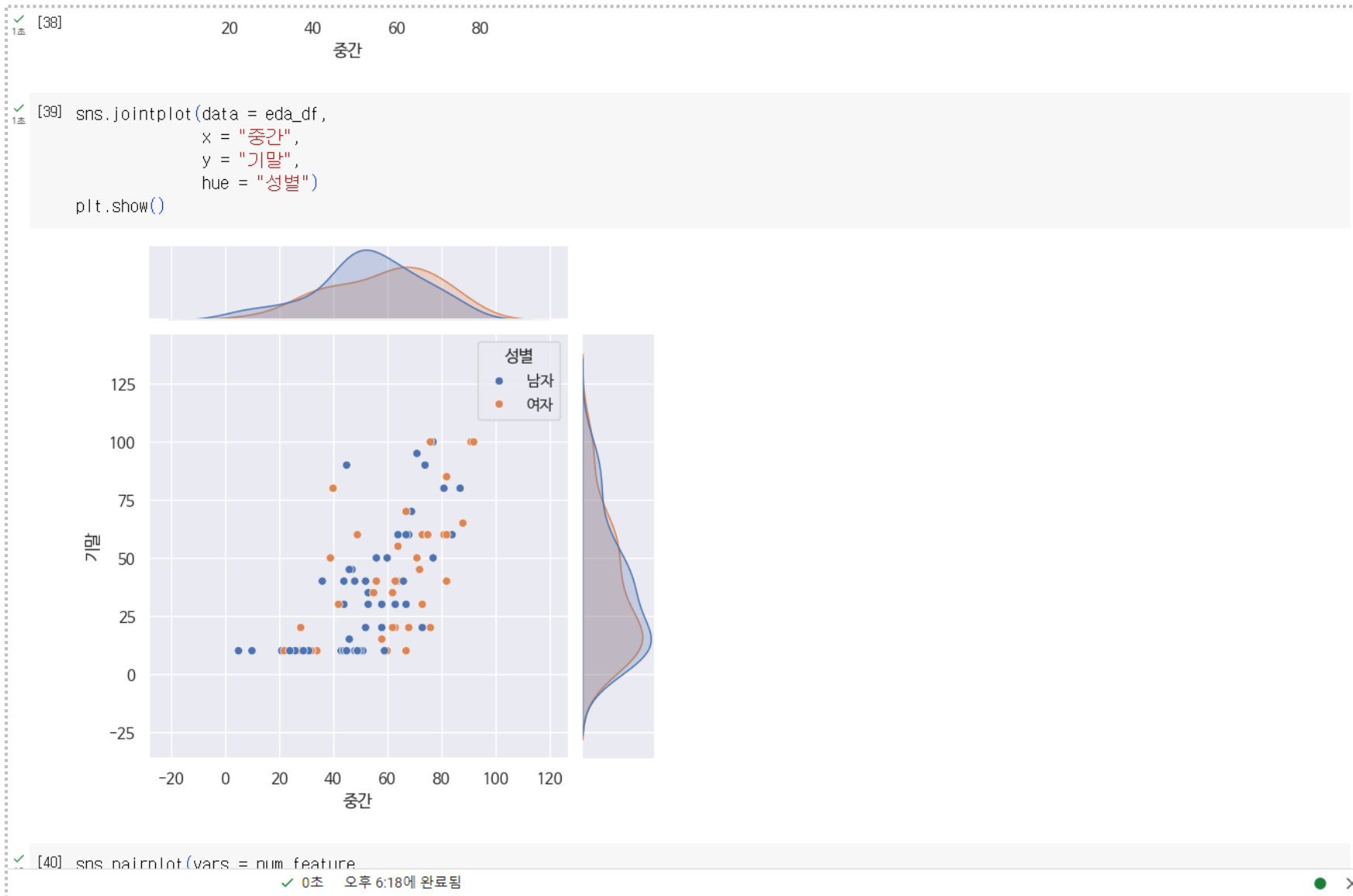
plt.show()
```



```
[39] sns.jointplot(data = eda_df,
                x = "중간",
                y = "기말",
```

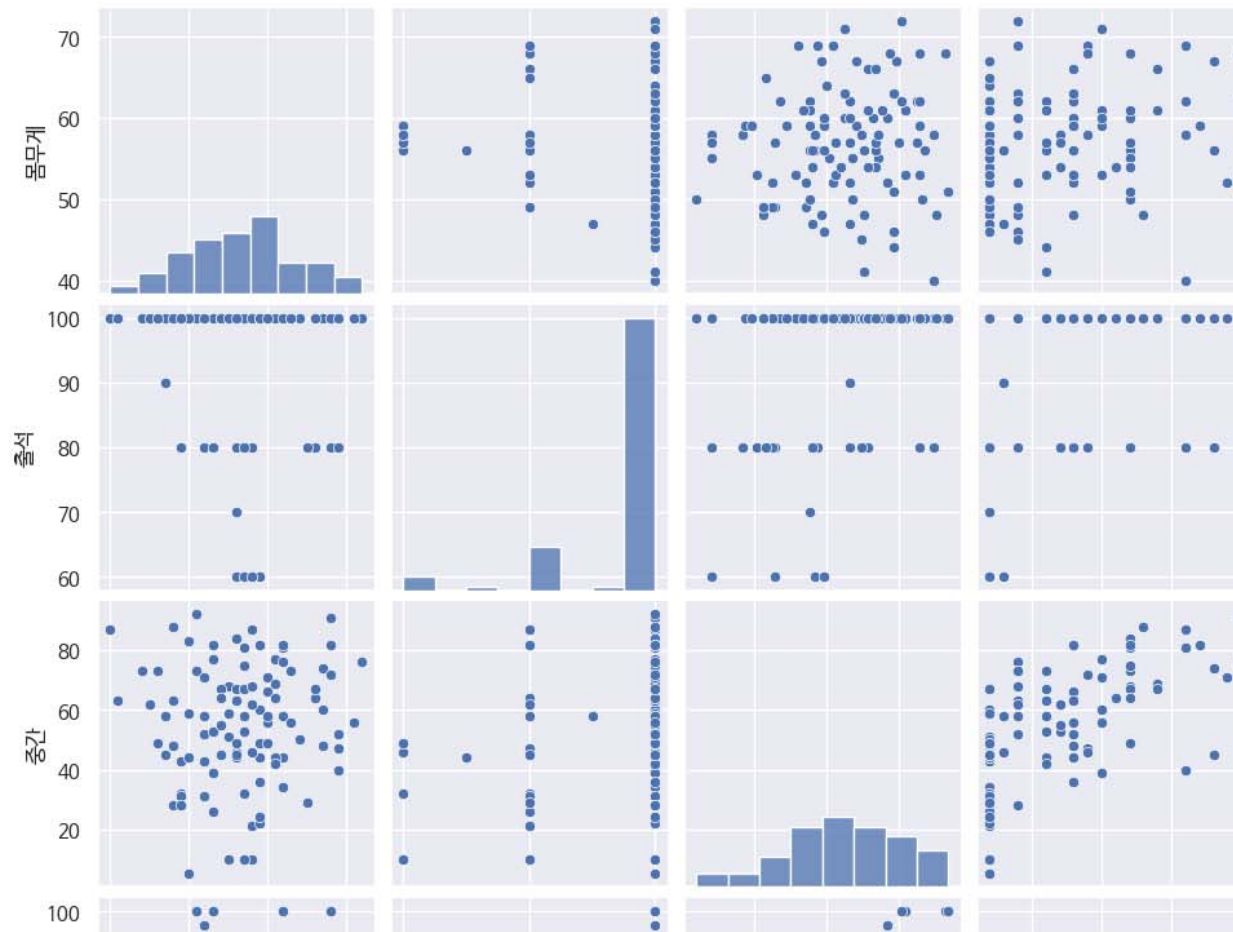
0초 오후 6:18에 완료됨

7.다중 수치형 변수(multi numerical variables)



7.다중 수치형 변수(multi numerical variables)

```
[40] sns.pairplot(vars = num_feature,
               data = eda_df)
      plt.show()
```



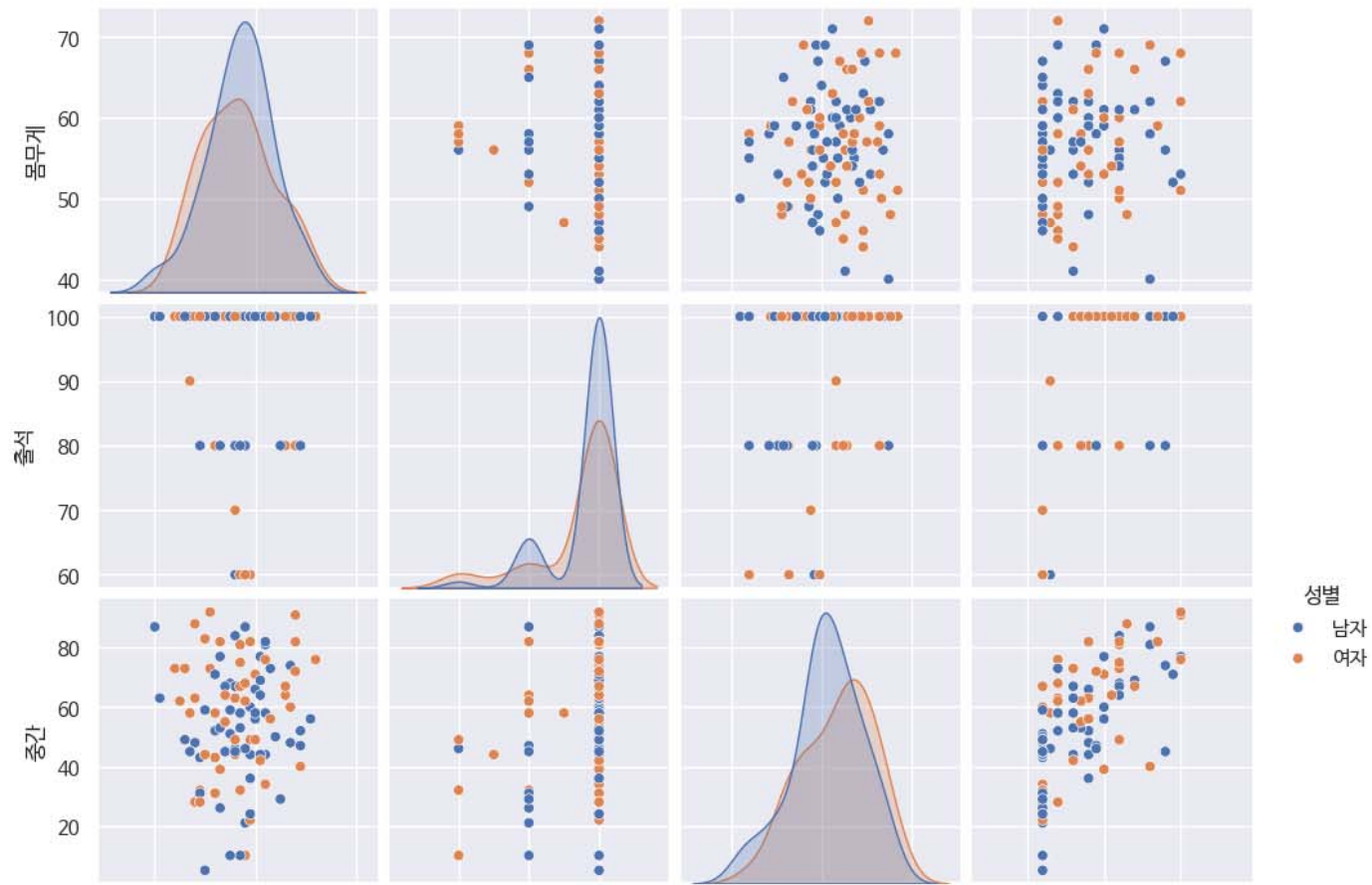
✓ 0초 오후 6:18에 완료됨

● X

7.다중 수치형 변수(multi numerical variables)

```
[41] sns.pairplot(vars = num_feature,
               hue = "성별",
               data = eda_df)

plt.show()
```



✓ 0초 오후 6:18에 완료됨

8. 평균

8. 평균

8.1 가중평균

```
[42] 가중평균 = [(10,60), (50,70),(40,80)]
sum = 0
n = 0
for i,j in 가중평균:
    sum = sum+(i*j)
    n = i+n
    weighted_mean = sum/n

weighted_mean
```

73.0

```
[43] eda_df["score"] = 0.2*eda_df["출석"]+0.3*eda_df["중간"]+0.5*eda_df["기말"]
eda_df.head()
```

	id	성별	분반	학년	몸무게	출석	중간	기말	몸무게_bin	몸무게_c	score
0	1	남자	A반	1학년	40	100	87	80	[0, 45)	~45미만	86.1
1	2	여자	B반	2학년	50	100	83	60	[50, 55)	50~55미만	74.9
2	3	남자	A반	3학년	56	100	84	60	[55, 60)	55~60미만	75.2
3	4	여자	B반	4학년	51	100	73	60	[50, 55)	50~55미만	71.9
4	5	남자	A반	1학년	55	100	68	60	[55, 60)	55~60미만	70.4

Next steps: [View recommended plots](#)

```
[44] eda_df["score"].mean()

54.043
```

✓ 0초 오후 6:18에 완료됨

8.2 기하평균

```
[45] import statistics as st

cagr = [998/635, 1265/998, 1701/1265, 2363/1701]
st.geometric_mean(cagr)-1

0.3889048648162128
```

```
[46] cagr = [0.572, 0.268, 0.345, 0.389]
      st.geometric_mean(cagr)

0.37872578967680914
```

```
[47] cagr = [635, 998, 1265, 1701, 2363]
np.power(2363/635, 1/4)-1

0.3089040640162120
```

8.3 조화평균

```
[48] import statistics as st

      harmonic = [400, 100]

      st.harmonic_mean(harmonic)

      160.0
```

```
[49] harmonic = 1000/((400/400)+(300/100)+(300/100))
harmonic

142.85714285714286
```

8. 평균

```
0.3889048648162128
```

8.3 조화평균

✓ 0초

[48] import statistics as st

harmonic = [400, 100]

st.harmonic_mean(harmonic)

160.0

✓ 0초

[49] harmonic = 1000/((400/400)+(300/100)+(300/100))
harmonic

142.85714285714286

Colab 유료 제품 - [여기에서 계약 취소](#)

✓ 0초 오후 6:18에 완료됨

연습문제

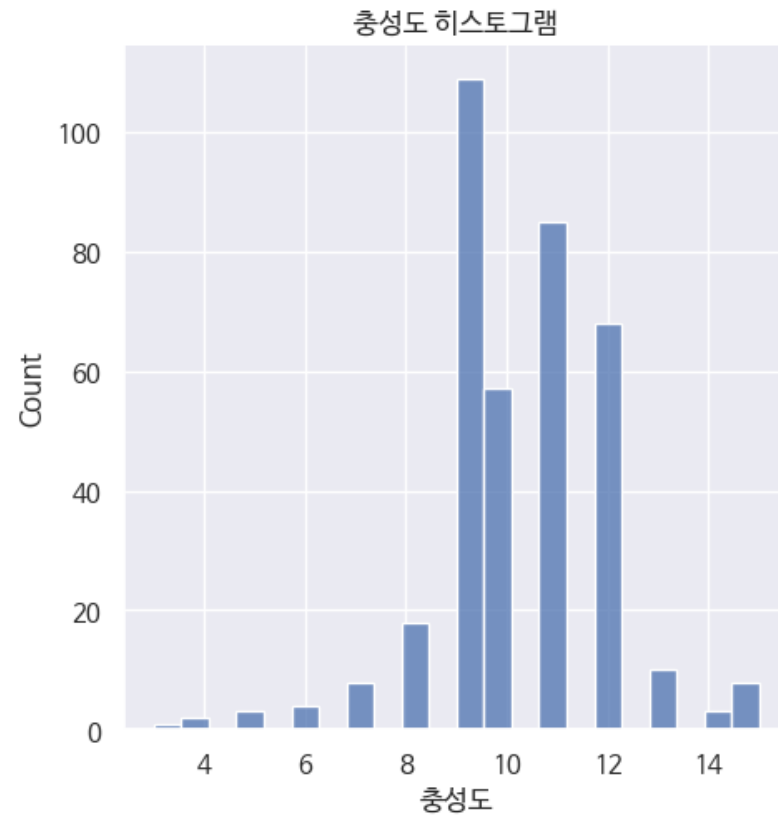
연습문제1

- ❖ 01_2.OnlineGame.csv를 이용하여 수치형 자료를 분석하세요.
- ❖ 성별: 1=남자, 2=여자
- ❖ 충성도의 기초통계분석을 하고, 히스토그램을 그리세요.
- ❖ 충성도의 box plot을 그리고, 이상치를 제거하세요(5.5이상만 사용)
- ❖ 연령를 범주형으로 변환(연령대)(~19, ~29, ~39, 40~으로 구분)하고 막대그래프를 그리세요.
- ❖ 성별에 따른 충성도 점수를 구하고, box plot 그리세요.
- ❖ 도구~몰입의 수치형 변수의 기초통계분석을 하세요.
- ❖ 성별에 따른 수치형 변수의 그래프를 그리세요.

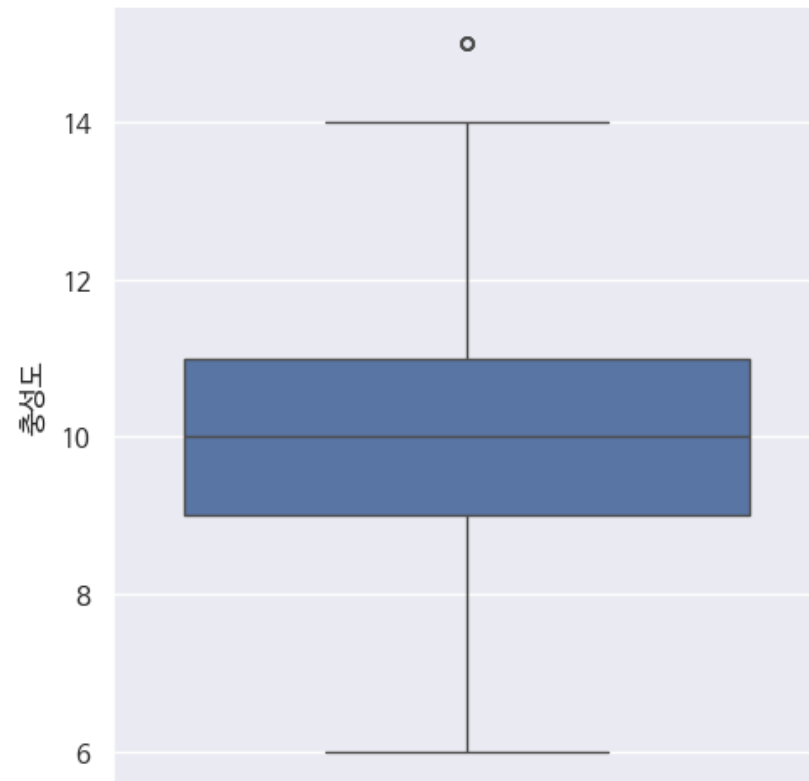
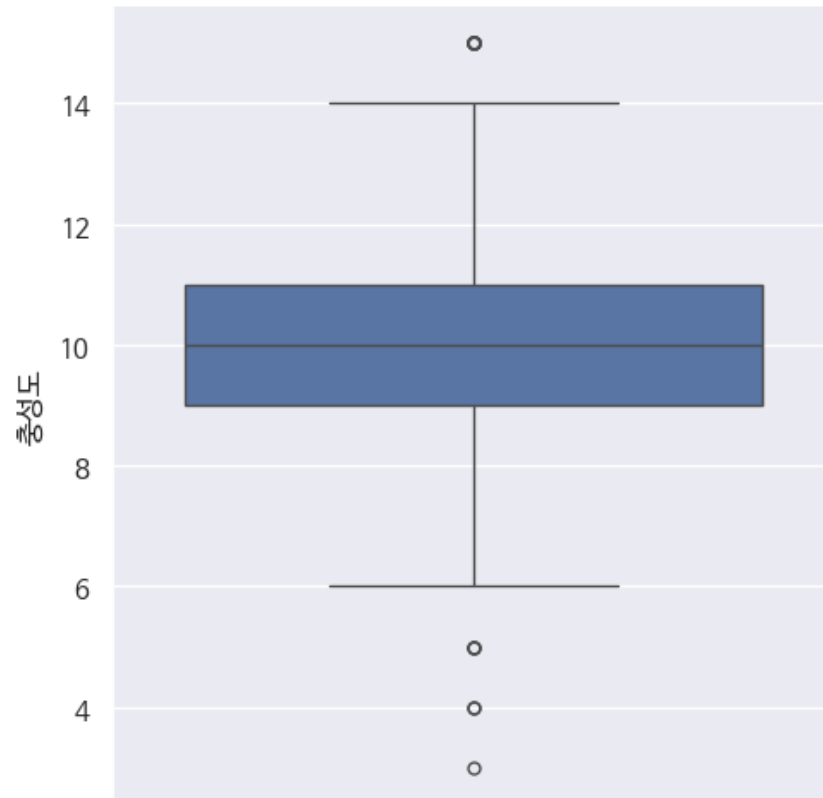
연습문제1

❖ 충성도의 기초통계분석을 하고, 히스토그램을 그리세요.

	count	mean	std	min	max	median	skew	kurtosis
충성도	376.0	10.223404	1.780812	3.0	15.0	10.0	-0.277817	1.516793

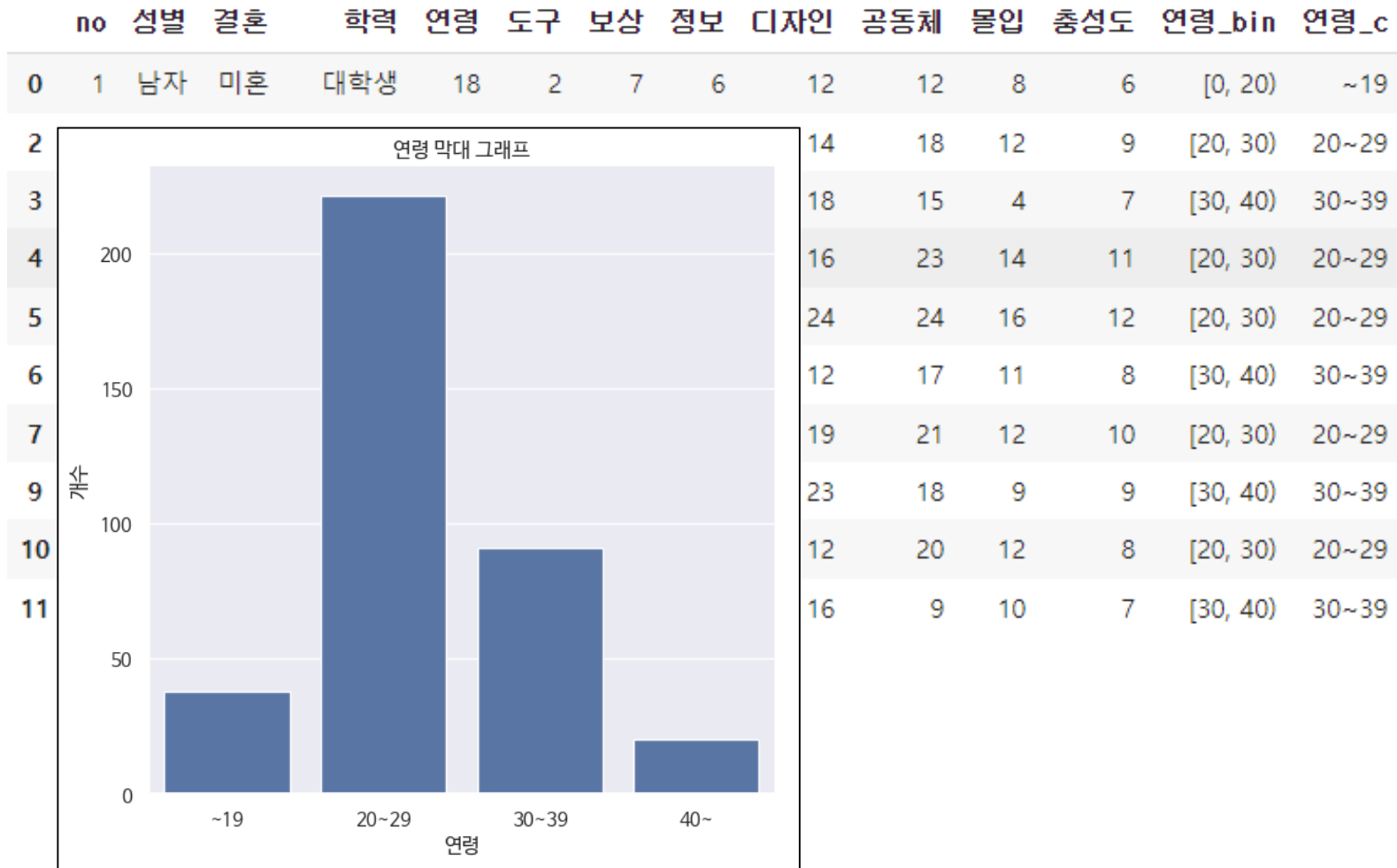


- ❖ 충성도의 box plot을 그리고, 이상치를 제거하세요(5.5이상만 사용)



연습문제1

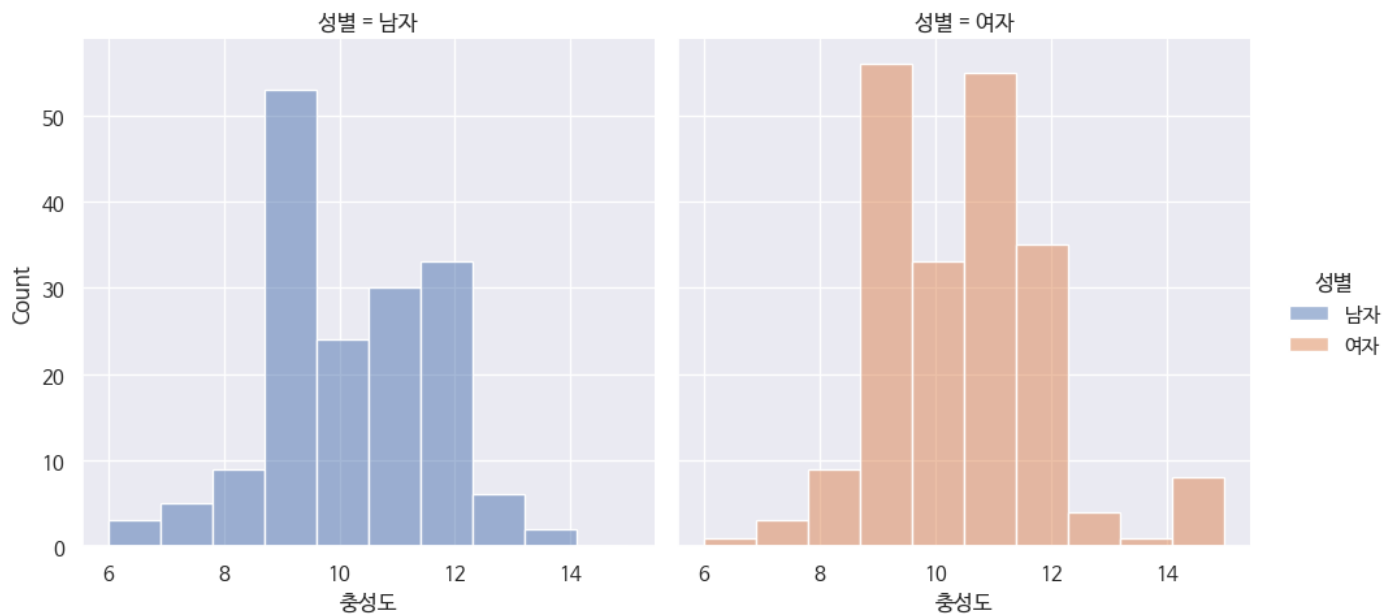
- ❖ 연령을 범주형으로 변환(연령대)(~19, ~29, ~39, 40~으로 구분)하고 막대그래프를 그리세요.



연습문제1

- ❖ 성별에 따른 충성도 점수를 구하고, box plot 그리세요.

충성도							
	count	mean	std	min	max	median	skew
성별							
남자	165	10.15	1.61	6	14	10.0	-0.08
여자	205	10.46	1.63	6	15	11.0	0.55

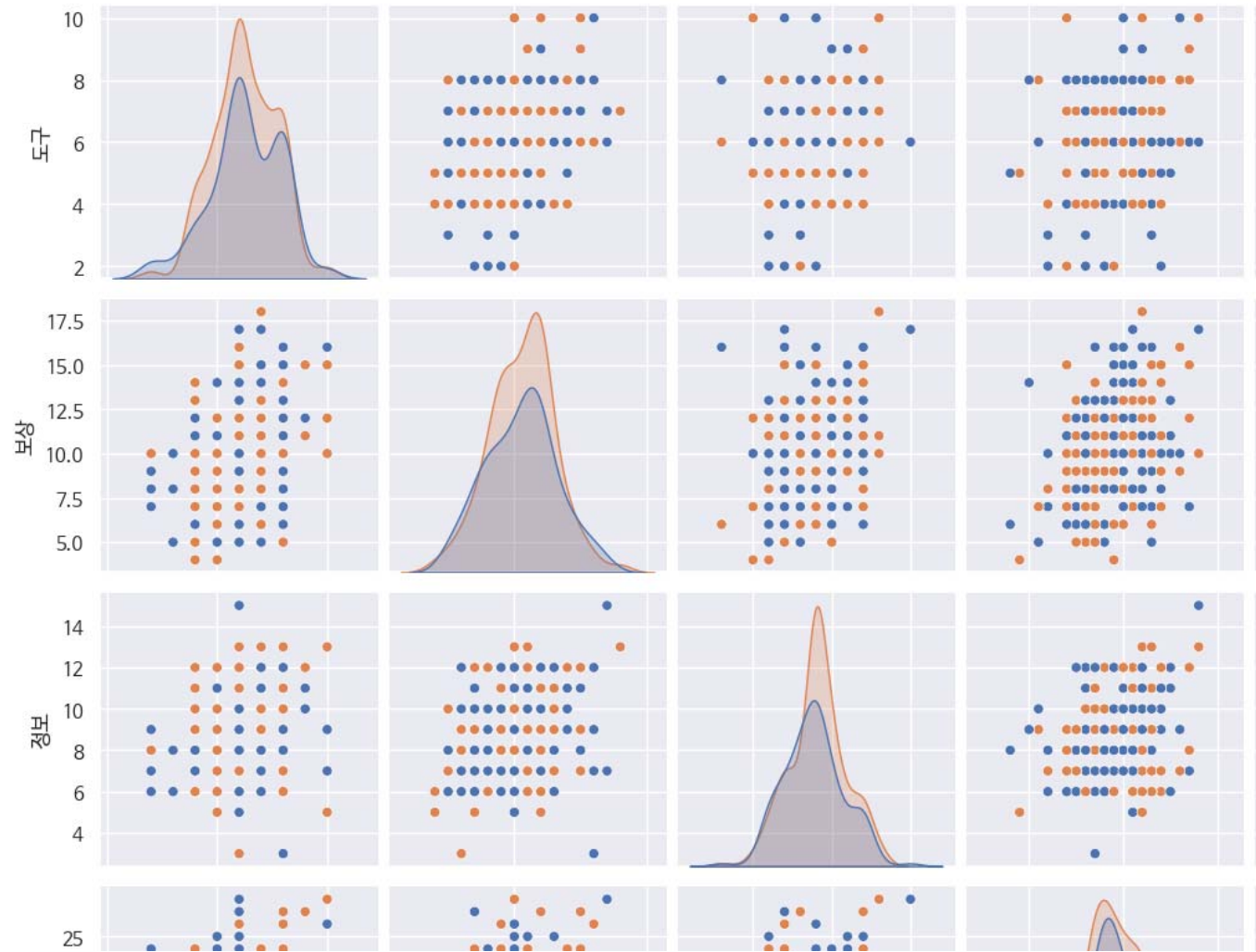


연습문제1

❖ 도구~몰입의 수치형 변수의 기초통계분석을 하세요.

	count	mean	std	min	max	median	skew	kurtosis
도구	370.0	6.273	1.503	2.0	10.0	6.0	-0.266	0.174
보상	370.0	10.543	2.561	4.0	18.0	11.0	0.017	-0.029
정보	370.0	8.968	1.822	3.0	15.0	9.0	0.019	0.184
디자인	370.0	19.051	3.161	8.0	28.0	19.0	-0.086	0.424
공동체	370.0	19.527	2.961	9.0	29.0	19.0	0.008	0.301
몰입	370.0	12.792	2.069	4.0	20.0	13.0	0.076	1.620

❖ 성별에 따른 수치형 변수의 그래프를 그리세요.



연습문제2

- ❖ G카페는 3개의 매장을 운영하고 있다. 만족도를 조사하기 위해 회원별 만족도를 조사하였다. 3개 매장별 만족도 점수는 아래의 표와 같다. G카페의 전체 만족도는 얼마인가?
- ❖ 답: 92.7

반	회원수	만족도
강남	30	90
강동	30	95
강서	40	93

연습문제3

- ❖ 다음은 이길동 교수가 투자한 회사의 4개년 수익율이다. 평균수익율은 얼마인가
- ❖ 답: 0.145(14.5%)

년도	수익
2020	100
2021	120
2022	110
2023	150

연습문제4

- ❖ 이길동 교수는 서울에서 광주로 (340km)를 출장을 갔다가 왔다. 갈 때는 비행기(500km/h)를 이용해서 가고, 올 때는 KTX(300km/h)타고 왔다. 왕복하는데 걸린 평균 시속은 얼마인가?(km/h)
- ❖ 답: 375km/h