통계기초 2

차이를 설명하는 실마리, 관계

1. 데이터 공간의 개념

2. 두 범주형 변수의 관계

3. 두 수치형 변수의 관계

4. 한 범주형 변수와 한 수치형 변수의 관계

데이터 공간의 형성

데이터 공간이란?

– 데이터마다 다른, 변수와 관측치 구성에 따라 만들어진 공간

데이터 공간의 구성과 특성

- 변수 수 만큼의 차원이 생성
  - 예제)키 : 1차원, 연령대 : 1차원, 연령대와 키 : 2차원
- 관측치 수 만큼의 점이 공간에 표현

데이터 공간과 분석의 재정의

① 키, 몸무게 같은 변수가 만들어 내는 공간에서
② 관측치들이 만들어 내는 차이를 숫자와 그래프로 확인하고
③ 더 자세히 **상대적인 차이**를 확인한 다음
④ 가능하다면 차이를 설명하는 과정

공간에서 범주형 변수 요약 다시 살펴보기

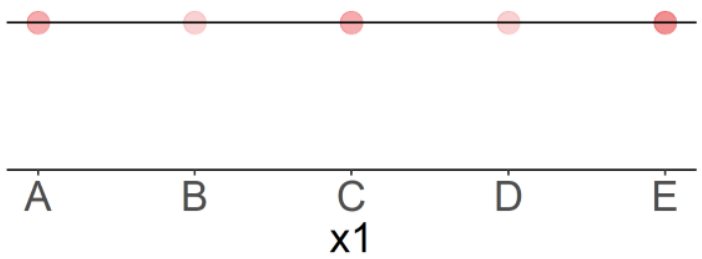
한 범주형 변수 = 1차원

: 관측치들이 1차원에서 정해진 k개 수준 중에 하나의 값을 가짐

① 빈도표 : 각 수준에 관측치들이 몇 개씩 있는지를 표로 요약

② 막대그래프 : 요약된 표의 빈도만큼 높이로 표현

- 예제)한 범주형 변수의 공간과 요약



A	B	C	D	E	합계
2	1	2	1	3	9

공간에서 수치형 변수 요약 다시 살펴보기

한 수치형 변수 = 1차원

: 관측치들이 1차원 수직선에서 다양한 값을 가짐

① 다섯 숫자 요약/상자그림 : 최솟값, 최댓값 등의 위치를 확인

② 히스토그램 : 1차원을 구간화하고 구간별 관측치 수를 확인

③ 평균 : 1차원 공간의 무게 중심

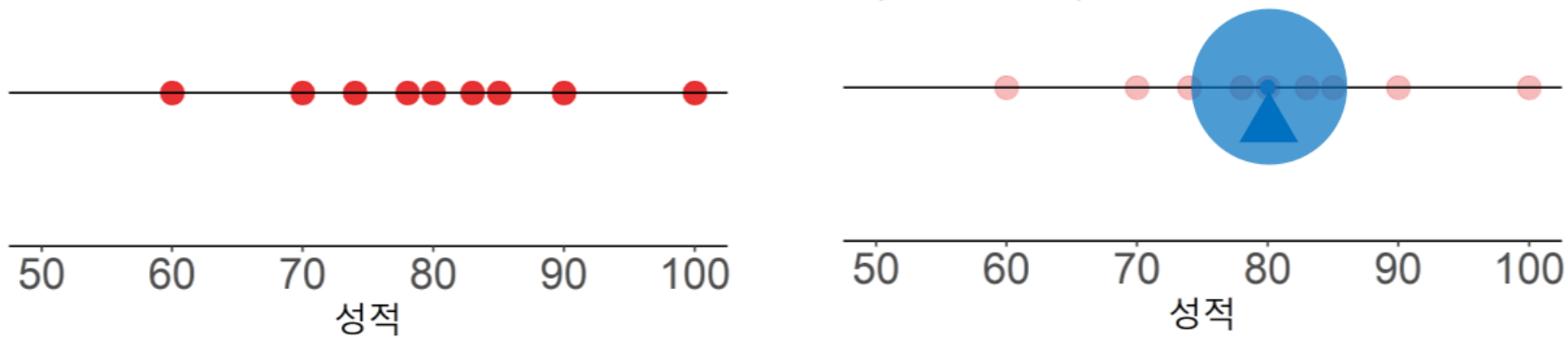
④ 분산 : 평균을 중심으로 관측치들이 흩어진 정도

공간에서 수치형 변수 요약 다시 살펴보기

- 예제) 9명 학생의 시험 점수

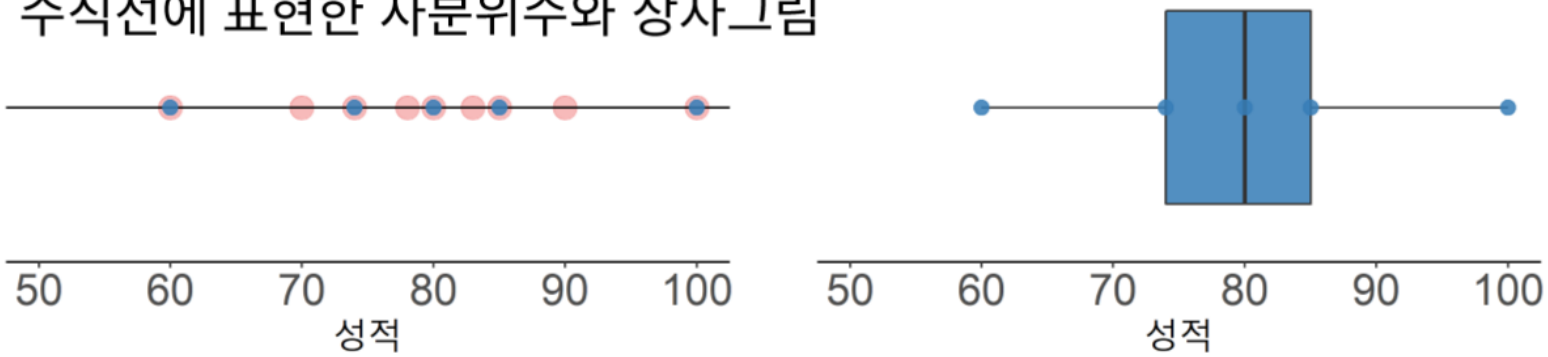
점수	60	78	83	74	100	80	90	85	70

- 수직선에 표현한 1차원 수치형 변수와 평균(무게중심)

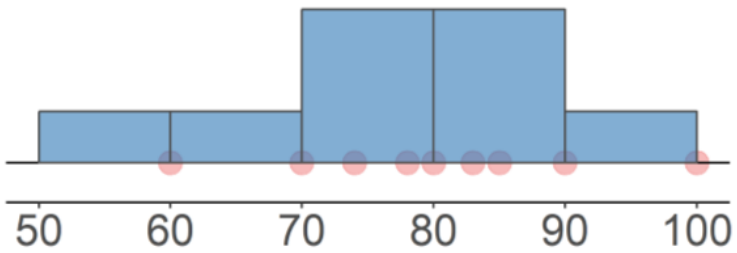


공간에서 수치형 변수 요약 다시 살펴보기

- 수직선에 표현한 사분위수와 상자그림



- 수직선의 구간화와 히스토그램



1차원에서 2차원으로

① 한 변수의 분석
- 1차원 공간에서 관측치들의 흩어진 패턴을 파악
- 주로 변수의 특성을 확인하는데 초점

② 두 변수의 분석
- 2차원 공간에서 관측치들의 흩어진 패턴을 파악
- 두 변수의 관계를 설명하는 데 초점

교차표를 활용한 두 범주형 변수의 요약

두 범주형 변수의 관계

- 두 범주형 변수의 수준들 간의 관계로 확인가능

교차표(contingency table, 분할표)

- 두 범주형 변수의 요약을 위한 2차원 표
- 두 범주형 변수의 수준 조합에 대한 빈도표
- 수준 조합의 절대적인 차이를 확인

교차표를 활용한 두 범주형 변수의 요약

- 예제) 회원 9명의 성별, 연령대 데이터와 교차표

성별	연령대
남	20대
여	30대
여	20대
남	30대
여	30대
여	20대
여	30대
남	30대
남	30대

	20대	30대
남	1	3
여	2	3

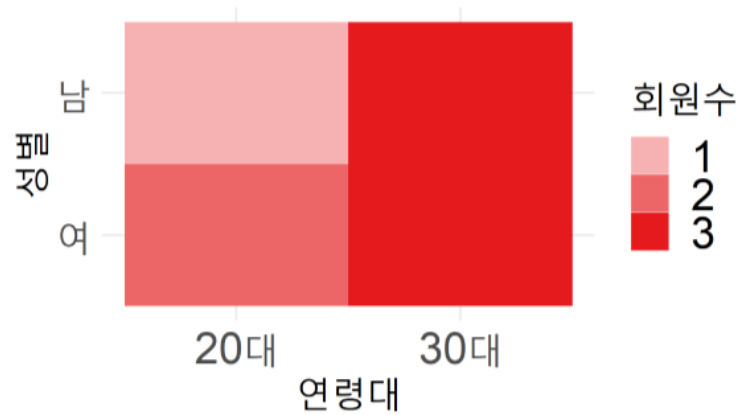
열지도를 활용한 교차표의 시각화

열지도(heatmap)

- 2차원 교차표를 숫자 대신 색으로 표현한 그림
- 숫자 대신 색의 진하기로 크기를 표현

  - 예제) 성별/연령대 교차표의 열지도

	20대	30대
남	1	3
여	2	3

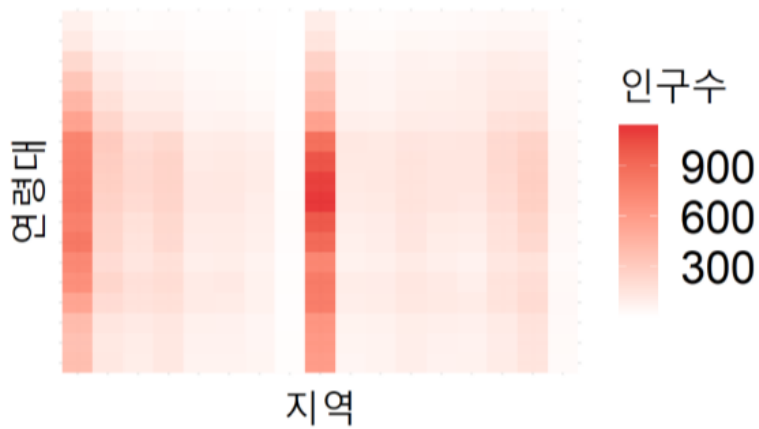


큰 교차표에서 더 효율적인 열지도

숫자보다 더 직관적인 그림이 크기를 비교하는데 편리

- 예제) 17개 광역시도별 5세 단위 인구분포 교차표와 열지도

	서울특별시	부산광역시	대구광역시	인천광역시	광주광역시	대전광역시	울산광역시	세종특별자치시	경기도	강원도	충청북도	충청남도	전라북도	전라남도	경상북도	경상남도	제주특별자치도
0~4세	384	133	100	133	68	71	59	14	600	58	69	96	76	77	112	155	30
5~9세	368	128	104	133	74	73	56	14	620	63	71	98	79	75	107	158	32
10~14세	401	142	121	138	84	79	58	12	643	72	77	100	91	85	117	165	34
15~19세	543	204	165	177	113	111	77	12	787	97	106	135	123	111	163	207	39
20~24세	681	241	176	196	114	125	77	14	805	105	108	136	121	94	163	194	36
25~29세	722	204	141	180	90	100	70	11	731	77	87	114	92	76	137	169	29
30~34세	824	233	160	215	107	110	86	17	904	86	102	141	105	98	164	222	39
35~39세	768	241	175	226	116	115	91	20	994	96	108	148	118	110	172	243	43
40~44세	809	264	208	246	130	130	99	19	1116	116	122	160	141	130	198	275	52
45~49세	790	279	221	252	130	131	109	14	1092	123	131	159	144	142	213	286	52
50~54세	767	293	216	254	116	123	105	13	1014	127	129	159	143	144	221	275	48
55~59세	749	307	195	224	104	112	91	11	868	131	122	148	140	142	218	258	43
60~64세	558	238	143	143	70	76	59	8	567	94	88	111	108	110	170	186	30
65~69세	436	179	105	105	57	55	38	6	420	70	64	92	95	101	132	138	25
70~74세	338	139	85	81	45	44	27	5	345	71	62	83	82	97	121	117	21
75~79세	226	97	63	60	32	32	18	4	261	58	52	75	70	85	106	98	17
80~84세	122	53	37	36	19	20	11	3	155	32	32	50	47	54	67	62	12
85세 이상	81	32	21	25	13	13	7	2	104	23	20	31	30	35	42	38	8



행 백분율과 열 백분율의 의미

교차표의 상대적인 차이 확인

- 절대적인 차이 : “이 칸에 관측치가 많다”
- 상대적인 차이 : “이 칸이 상대적으로 비율이 높다“

행(열) 백분율

- 전체가 아닌 각 행(열)에서 상대빈도를 계산
- 각 수준의 전체 상대빈도와 비교

행 백분율과 열 백분율 예제

예제) 남녀의 찬반 교차표

- 전체 상대빈도의 계산
  - 찬성 30%, 반대 70%
  - 남 60%, 여 40%

	남	여	합계
찬	15	15	30
반	45	25	70
합계	60	40	100

- 행 백분율(왼쪽)과 열 백분율의 계산과 비교

	남	여	합
찬	50%	50%	100%
반	64%	36%	100%
합	60%	40%	100%

	남	여	합
찬	25%	38%	30%
반	75%	62%	70%
합	100%	100%	100%

산점도를 활용한 2차원 공간의 시각화

산점도(scatter plot)

- 두 수치형 변수를 가로축, 세로축으로 활용하여 그린 그래프
- 2차원 공간에 관측치의 수만큼 찍힌 점의 패턴을 파악

보조선의 중요성

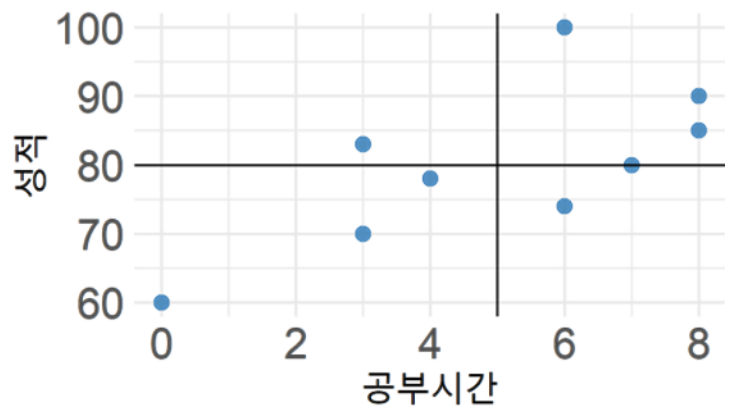
- 두 변수의 평균을 활용해서 수직/수평선 추가
- 두 직선이 만나는 지점이 2차원 공간의 무게 중심

2차원 데이터와 산점도 예제

예제)공부시간과 점수의 산점도

공부시간	점수
0	60
4	78
3	83
6	74
6	100
7	80
8	90
8	85
3	70

- 평균 공부시간 = 5
- 평균 점수 =80



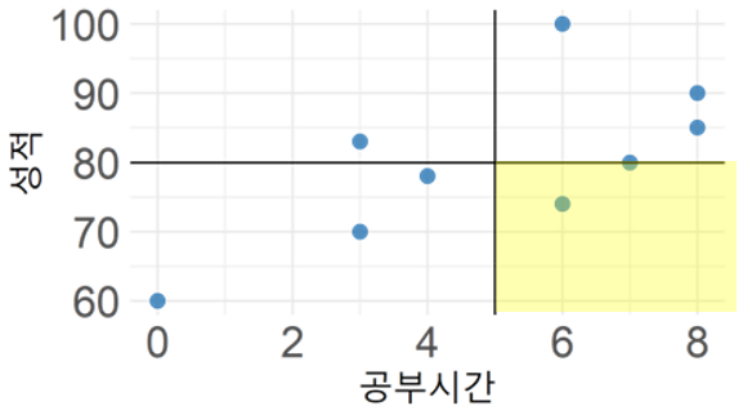
사분면과 관측치들의 분류

사분면(quadrant)

- 2차원 공간에서 무게 중심 기준으로 나눠진 4개 면
- 오른쪽 위 제 1사분면부터 반시계 반향으로 순서를 지정

각 사분면과 관측치들의 특성

사분면	첫번째 변수	두번째 변수
1	평균 이상	평균 이상
2	평균 이하	평균 이상
3	평균 이하	평균 이하
4	평균 이상	평균 이하

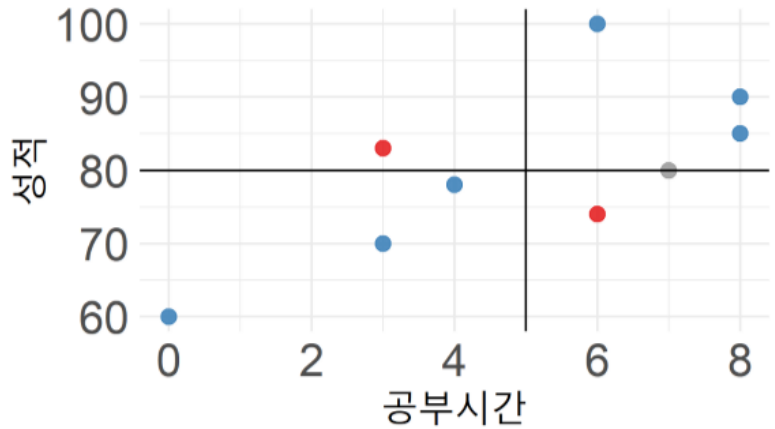


사분면과 두 변수의 상관 관계

① 두 수치형 변수의 관계

–양의 상관 = “같이 간다”

–음의 상관 = “반대로 간다”



② 산점도의 패턴과 두 변수의 관계

–제 1, 3사분면의 관측치 수 ↑ : 두 변수의 양의 상관을 의미

–제 2, 4사분면의 관측치 수 ↑ : 두 변수의 음의 상관을 의미

공분산을 활용하여 변수 관계 확인하기

공분산(covariance)

$$q_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

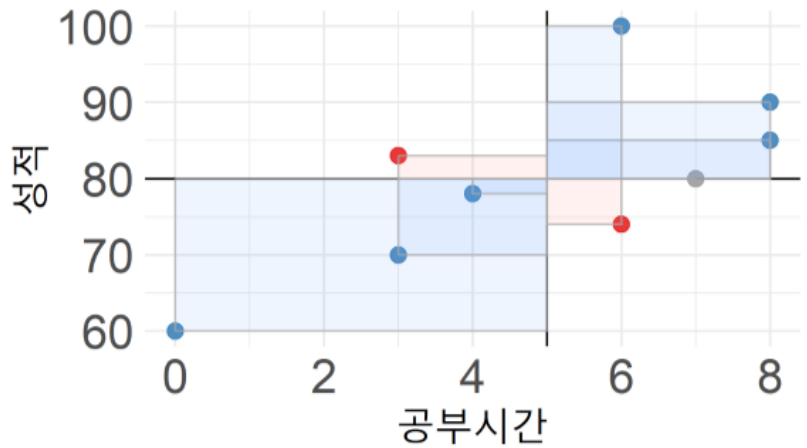
: 두 수치형 변수의 관계를 계산한 기술 통계량

- 0에 가까울수록 관련이 없음
- 큰 양수가 나올 수록 두 변수가 양의 상관을 가짐
- 큰 음수가 나올 수록 두 변수가 음의 상관을 가짐

공간에서 공분산의 의미 확인하기

관측치별 $(x_i - \bar{x})(y_i - \bar{y})$ 값 계산

- 각 관측치가 중심 $(\bar{x},\ \bar{y})$로 부터 떨어진 면적
- 제 1, 3사분면의 관측치는 양의 면적이 계산됨
- 제 2, 4사분면의 관측치는 음의 면적이 계산됨



공분산의 계산 예제

예제)공부시간(x)과 점수(y)의 공분산

공부시간	점수	$x_i - \bar{x}$	$y_i - \bar{y}$	×(곱)
0	60	-5	-20	100
4	78	-1	-2	2
3	83	-2	3	-6
6	74	1	-6	-6
6	100	1	20	20
7	80	2	0	0
8	90	3	10	30
8	85	3	5	15
3	70	-2	-10	20

$$q_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$= \frac{175}{9-1}$$

$$= 21.875$$

공분산의 한계와 대체방안

공분산의 특성과 한계

- 두 변수의 관계에 대한 절대적인 요약 값
- 단위 문제 발생 :
  - scale : 공부시간과 점수의 분포에 비해 단위가 큰 숫자로 계산된 공분산
  - unit : 공부시간과 점수의 공분산의 단위는 “시간×점”

상관계수의 도입

- 두 변수를 표준화해서 공분산의 단위의 문제를 해결
- 두 변수의 관계에 대한 상대적인 요약 값

상관계수의 활용

피어슨 상관계수(Pearson's correlation coefficient)

$$r_{xy} = \frac{q_{xy}}{s_x\, s_y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}\,, \quad -1 \le r_{xy} \le 1$$

- 표준화된 두 수치형 변수로 계산된 공분산
- 공분산의 단위(scale/unit)의 문제를 해결
  - $r_{xy} = 0$: 두 변수가 상관이 없음
  - $r_{xy} > 0$: 두 변수가 함께 증가하거나 감소하는 양의 상관을 가짐
  - $r_{xy} < 0$: 한 변수가 증가하면 나머지 한 변수는 감소하는 음의 상관을 가짐

상관계수 계산의 예제

예제)공부시간(x)과 점수(y)의 상관계수

- 상관계수와 공분산, 두 변수의 표준편차의 관계를 활용

  - $q_{xy} = 21.875,\ \ s_x = 2.693,\ s_y = 11.587$

$$r_{xy} = \frac{q_{xy}}{s_x\, s_y} = \frac{21.875}{2.693 \times 11.587} = \mathbf{0.70}$$

- "공부시간과 점수는 양의 상관 관계를 가지고 있다."

변수 형식에 따른 두 변수의 관계와 요약

변수의 형식에 맞는 2차원 분석 방법을 활용

①두 범주형 변수의 관계 : **교차표**를 활용한 수준간 관계 확인

②두 수치형 변수의 관계 : **산점도**를 활용한 관측치 패턴 확인

③한 범주형 변수와 한 수치형 변수의 관계

- 범주형 변수를 그룹으로 활용한 수치형 변수의 그룹별 평균 계산
- 수치형 변수를 조건으로 활용한 범주형 변수의 조건부 비율 계산

조건부 평균의 계산과 활용

조건부 평균(conditional mean)

- 범주형 변수의 수준별로 관측치를 나누기
- 각 수준별로 수치형 변수의 평균을 계산
- 그룹(수준)에 따른 절대적인 차이를 확인

상대적인 차이의 확인

- 전체 평균 대비 그룹별 평균을 비교

조건부 평균의 계산과 활용

예제) 공부방법과 점수를 활용한 그룹별 평균 계산

공부방법	점수
B	60
B	78
B	83
B	74
A	100
A	80
A	90
B	85
A	70

- 전체 평균 = 80점
- A 평균 = 85점, B 평균 = 76점

공부방법 A				평균
100	80	90	70	85

공부방법 B					평균
60	78	83	74	85	76

그룹별 상자그림을 활용한 시각화

히스토그램과 상자그림

: 한 수치형 변수의 분포를 확인하는데 활용

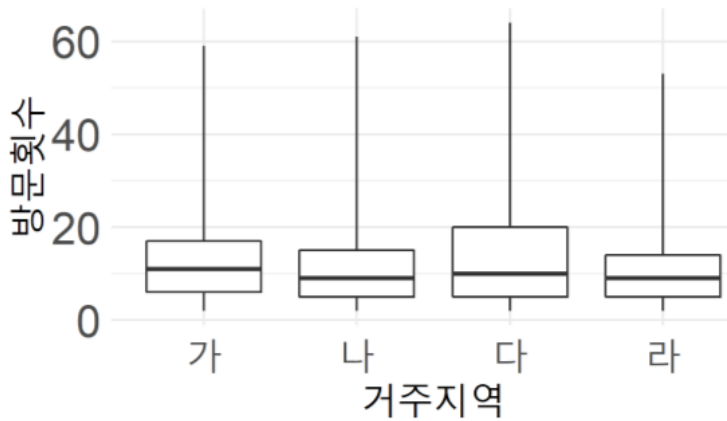
그룹별 상자그림

- 각 수준별로 수치형 변수의 상자그림을 작성
- 동일한 축을 활용하여 나란하게 표현
- 그룹 간의 분포 비교에 활용

그룹별 상자그림을 활용한 시각화

예제) 고객 1,000명의 거주지역과 방문횟수의 요약

고객번호	거주지역	방문횟수
1	가	6
2	다	30
⋮	⋮	⋮
1,000	라	13



- 평균 방문횟수 요약

거주지역	가	나	다	라	전체
평균 방문횟수	13.9	12.9	15.2	12.8	13.8

수치형 변수의 구간화 활용

수치형 변수의 구간화

- 히스토그램을 그리기 위해 수치형 변수를 구간화
- 구간 값을 활용하여 범주형 변수로 변환 후 요약 가능

범주형 변수와 구간화 된 수치형 변수의 관계

- 사실상 두 범주형 변수의 관계
- 교차표 등을 활용해서 분석

수치형 변수의 구간화 활용

예제) 공부방법과 구간화된 점수의 요약

공부방법	점수	점수구간
B	60	0~75
B	78	76~85
B	83	76~85
B	74	0~75
A	100	86~100
A	80	76~85
A	90	86~100
B	85	76~85
A	70	0~75

- 교차표의 만들기

점수구간	공부방법 A	공부방법 B
0~75	1	2
76~85	1	3
86~100	2	0

두 변수의 관계 설명

첫번째 변수	두번째 변수	기술 통계량	시각화
범주형	범주형	교차표	열지도
수치형	수치형	상관계수	산점도
범주형	수치형	그룹별 평균	그룹별 상자그림