Part 2

차이를 설명하는 통계 개념

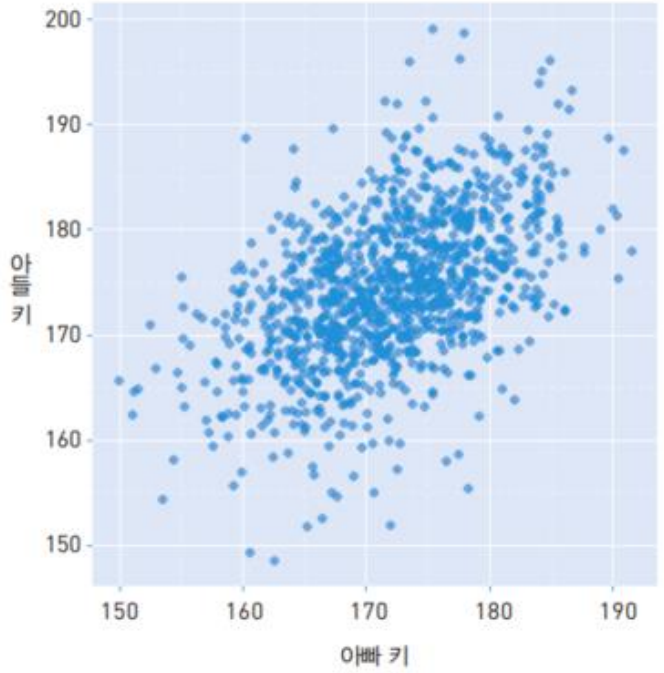
[2개 연속형 변수의 관계 탐색]

산점도(Scatter plot)

변수와 차원

- $n \times p$ 데이터는 p차원 공간에 n개 점이 찍힌 공간으로 설명 가능
- 두 연속형 변수가 있는 데이터는 2차원 공간으로 설명가능
- 산점도 : 관측치의 위치를 점으로 표현한 그림
- 예제) Pearson의 키 데이터
  - 아빠 키와 아들 키를 각각 x축, y축 좌표로 활용한 산점도

가족 번호	아빠 키(cm)	아들 키(cm)
1	162.2	151.8
2	160.7	160.6
...	...	...
1,077	179.7	176.0
1,078	178.6	170.2



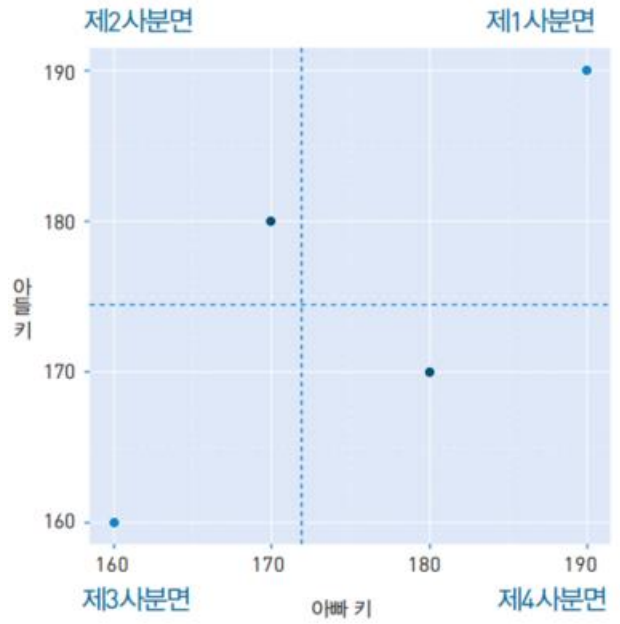
산점도의 사분면과 상관관계

사분면의 구성

- x축, y축의 평균을 기준으로 가로, 세로로 공간을 분할
- 오른쪽 위부터 반시계 방향으로 1, 2, 3, 4 사분면 생성

사분면과 상관관계

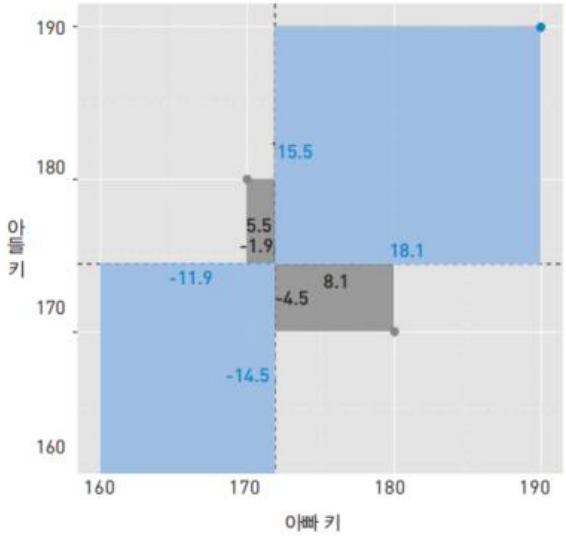
- 제 1, 3 사분면의 점은 “두 변수가 양의 상관관계를 가짐”에 영향
  - 아빠 키가 크면 아들 키도 크고, 아빠 키가 작으면 아들 키도 작음
- 제 2, 4 사분면의 점은 “두 변수가 음의 상관관계를 가짐”에 영향
- 아빠 키, 아들 키 데이터의 경우 제 1, 3사분면에 관측치가 더 많아 아빠 키와 아들 키는 양의 상관관계를 가진다고 볼 수 있음



공분산(Covariance)

공분산의 계산

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



- 산점도에서 두 변수의 상관관계를 숫자 하나로 계산한 값
- 계산과정
  - 각 점의 x변수가 평균으로 부터 떨어진 거리 $(x_i - \bar{x})$를 계산
  - 각 점의 y변수가 평균으로 부터 떨어진 거리 $(y_i - \bar{y})$를 계산
  - 두 거리를 곱해 직사각형의 면적을 계산
    - 제 1, 3 사분면의 점은 양(+)의 면적, 제 2, 4 사분면의 점은 음(-)의 면적이 계산됨
  - n개 각 관측치의 직사각형 면적을 모두 더해 $(n-1)$로 나눔

공분산의 활용과 한계

공분산의 활용

- 계산된 공분산으로 두 변수의 상관관계를 유추가능
  - 0에 가까울 수록 서로 관련이 없음
  - 큰 양수가 나오면 강한 양의 상관관계, 큰 음수가 나오면 강한 음의 상관관계가 있음

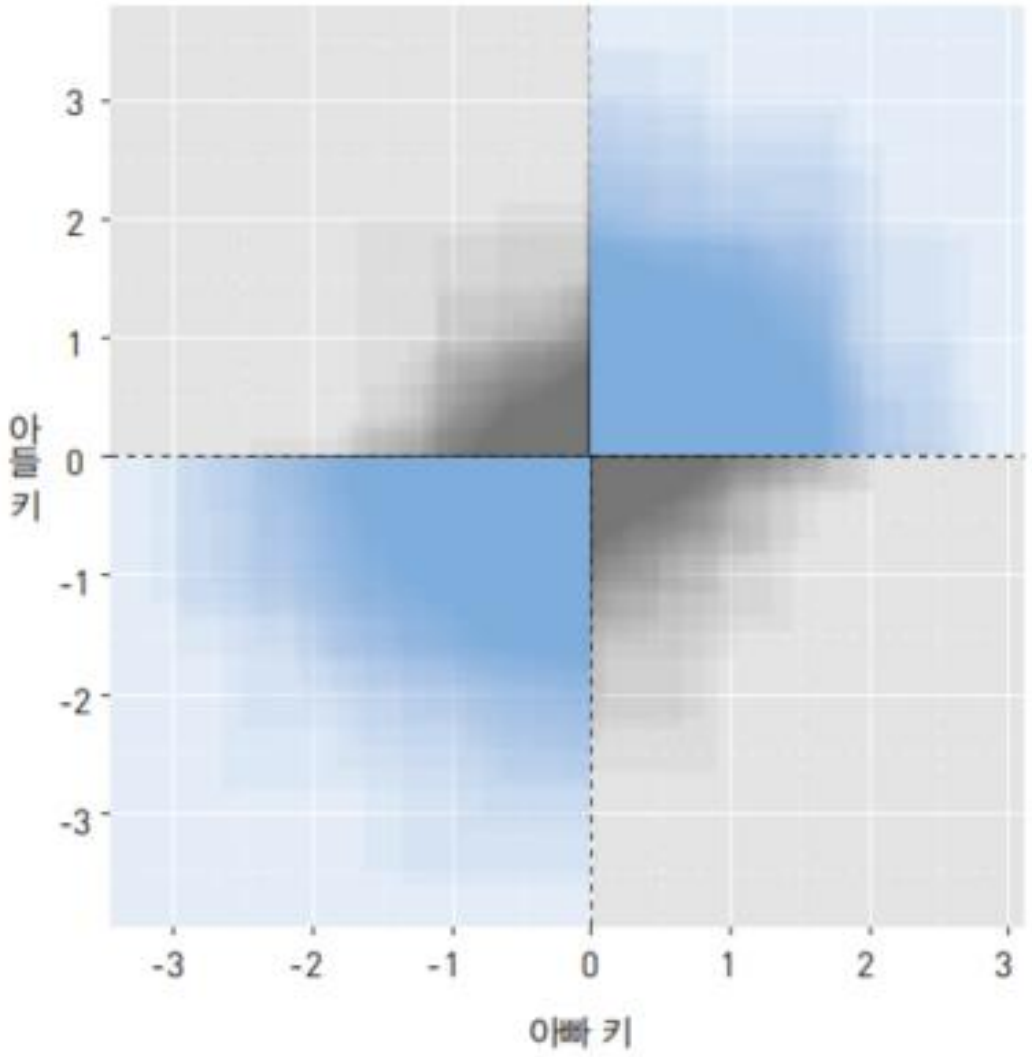
공분산의 한계

- 공분산의 단위(Unit)는 두 변수 x, y단위의 곱
  - 예제) x가 키(cm)이고 y가 몸무게(kg)일 때 공분산의 단위는 $cm \cdot kg$
  - 의미 파악과 활용이 어려움
- 공분산은 단위(Scale)에 영향을 받음
  - 예제) cm 단위인 x를 m 단위로 바꾸면, 동일한 데이터로 계산된 공분산이 100배 차이 남

상관계수(Correlation)

번호	아빠 키(cm)	아들 키(cm)	표준화된 아빠 키(cm)	표준화된 아들 키(cm)
1	162.2	151.8	-1	-3.2
2	160.7	160.6	-1.6	-1.9
…	…	…	…	…
1,077	179.7	176.0	1.1	0.2
1,078	178.6	170.2	1.0	-0.6

상관계수(Correlation)



상관계수(Correlation)

상관계수의 계산

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

- 표준화된 두 변수의 공분산을 계산
- x와 y가 완전히 똑같을 때$(x = y)$ 최대값 1, 완전히 반대로 갈 때$(x = -y)$ 최소값 1을 가짐
- 어떠한 두 연속형 변수에 대해서도 −1부터 1 사이의 값을 가짐
- 예제) 키 데이터의 상관계수
  - 아빠 키와 아들 키의 상관계수는 약 0.5

[2개 범주형 변수의 관계 탐색]

교차표와 백분율

교차표 (Contingency table)

- 두 범주형 변수의 수준 조합별 관측치 개수를 세어 표현한 표
- 연속형 변수와 달리 지정된 수준을 가지는 범주형 변수의 특성이 반영됨
- 연속형 변수의 산점도와 마찬가지로 두 변수의 관계를 확인 가능
- 예제) 최근 4회 하계 올림픽에서의 메달 성적 교차표

	금메달	은메달	동메달	행 합계
28회 아테네	9	12	9	30
29회 베이징	13	10	9	32
30회 런던	13	8	7	28
31회 리우	9	3	9	21
열 합계	44	33	34	111

교차표와 백분율

- 교차표로 표현된 두 변수의 관계를 확인하기 위해서는 백분율 계산이 필수적

행 백분율과 열 백분율

행 백분율

- 각 행의 합계를 1로 고정했을 때 각 열의 비중
- 전체 열 합계에 대한 행 백분율이 기준 값이 됨
- 예제)
  - 전체 열 합계에서 금메달의 비중은 40%
  - 29회 베이징은 전체와 비슷하지만,
  - 28회 아테네는 금메달의 비중이 30%로 상대적으로 낮고
  - 30회 런던은 금메달의 비중이 46%로 상대적으로 높음

	금메달	은메달	동메달	행 합계
28회 아테네	0.30	0.40	0.30	1.00
29회 베이징	0.41	0.31	0.28	1.00
30회 런던	0.46	0.29	0.25	1.00
31회 리우	0.43	0.14	0.43	1.00
열 합계	0.40	0.30	0.30	1.00

열 백분율

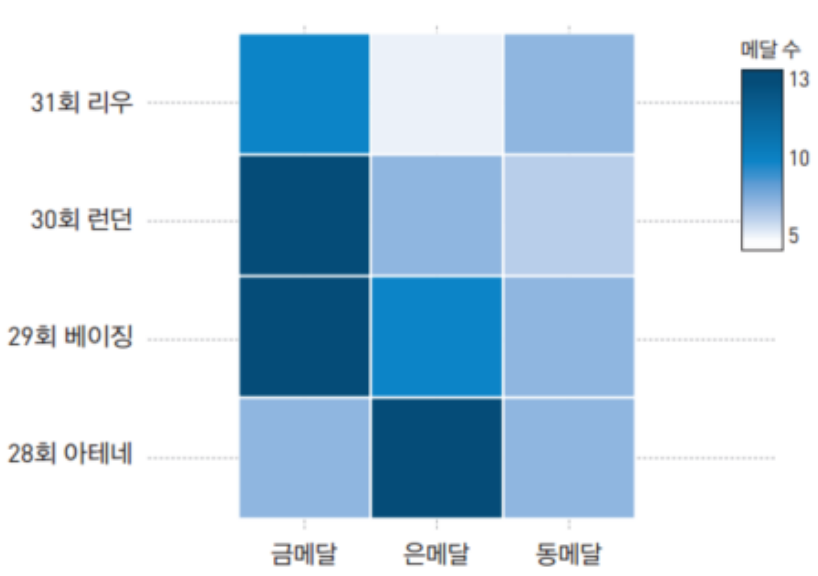
- 각 열의 합계를 1로 고정했을 때 각 행의 비중

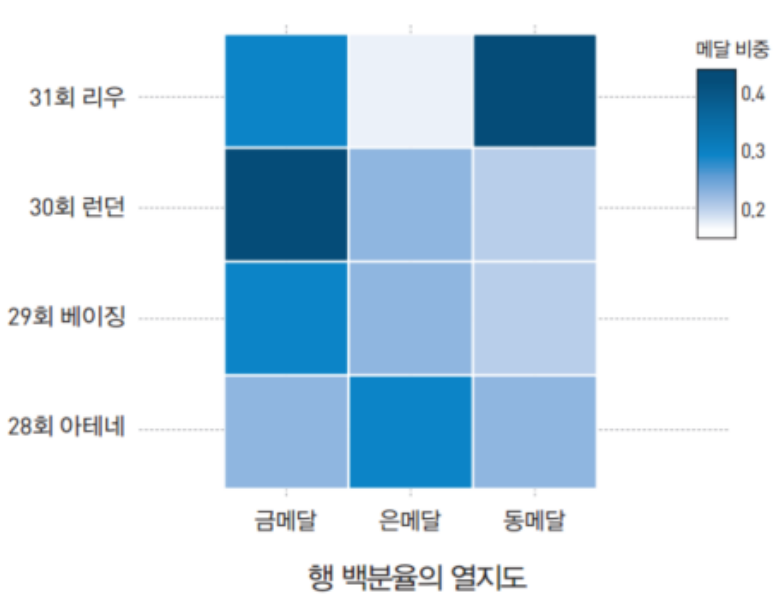
	금메달	은메달	동메달	행 합계
28회 아테네	0.20	0.36	0.26	0.27
29회 베이징	0.30	0.30	0.26	0.29
30회 런던	0.30	0.24	0.21	0.25
31회 리우	0.20	0.09	0.26	0.19
열 합계	1.00	1.00	1.00	1.00

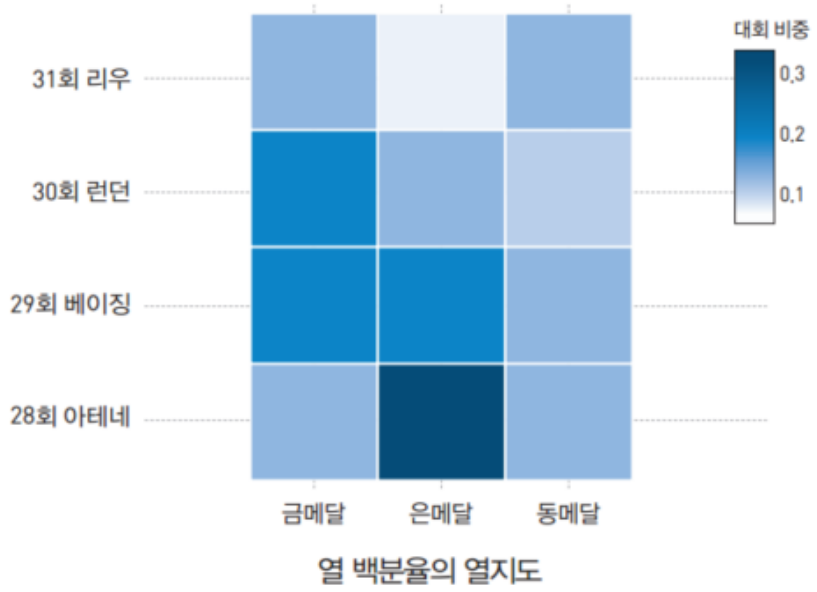
열지도(Heatmap)

열지도

- 교차표의 시각화 방법
- 수준 간의 차이를 숫자가 아닌 색깔로 확인
- 일반적으로 값이 크면 색을 진하게, 값이 작으면 색을 연하게 표현
- 예제) 올림픽 메달 데이터의 열지도와 행 백분율의 열지도







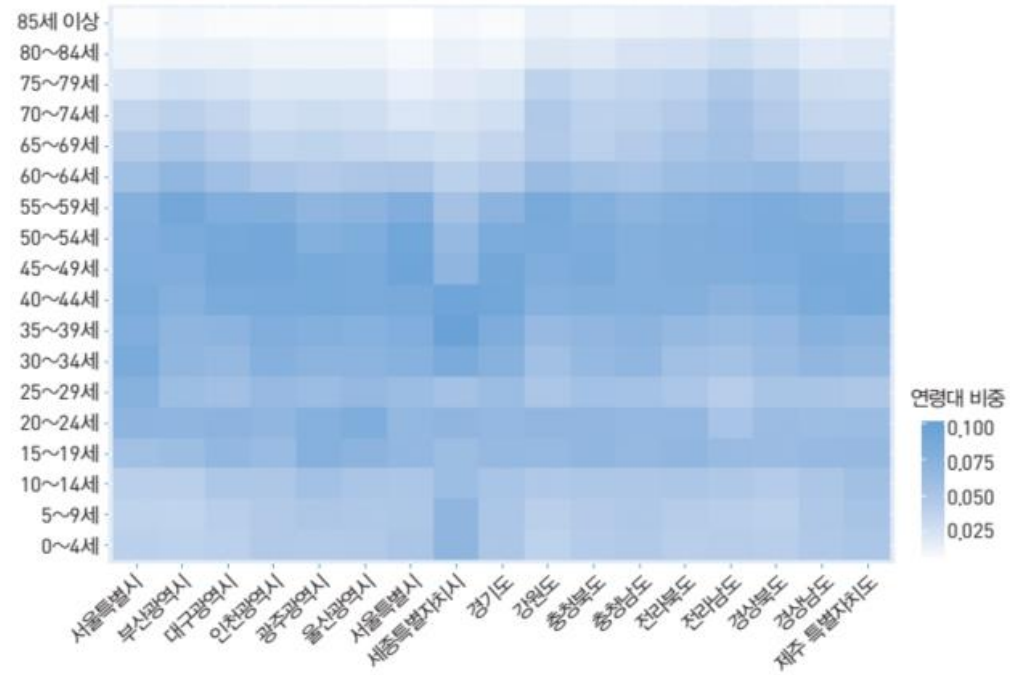
열지도 예제

2015년 인구 총 조사 기준 지역/연령대별 인구분포표

- 숫자로 채운 교차표보다 색깔로 표현한 열지도가 훨씬 더 효율적

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

2015년 인구총조사 기준 지역별 연령대 분포, 단위 : 천명



열 백분율 기준 지역별 연령대 분포의 열지도

[일반적인 두 변수의 관계 탐색]

독립(Independence)

독립

- 두 변수가 서로 관련이 없는 상태
- 연속형 변수의 독립
  - 상관계수가 0에 가까울 때
- 범주형 변수의 독립
  - 교차표에서 계산한 행/열 백분율과 전체 합계 행/열 백분율에 차이가 없을 때
- 예제) 독립을 가정한 올림픽 메달 교차표

	금메달	은메달	동메달	행 합계
28회 아테네	12	9	9	30
29회 베이징	13	10	10	32
30회 런던	11	8	9	28
31회 리우	8	6	6	21
열 합계	44	33	34	111

조건부 확률과 조건부 평균

조건(Condition)

- 데이터에 있는 변수에 기준 값을 지정한 것
- 전체 관측치가 아닌 부분 관측치를 선택할 때 활용
- 예제) 아빠 키가 $180cm$이상인 가족

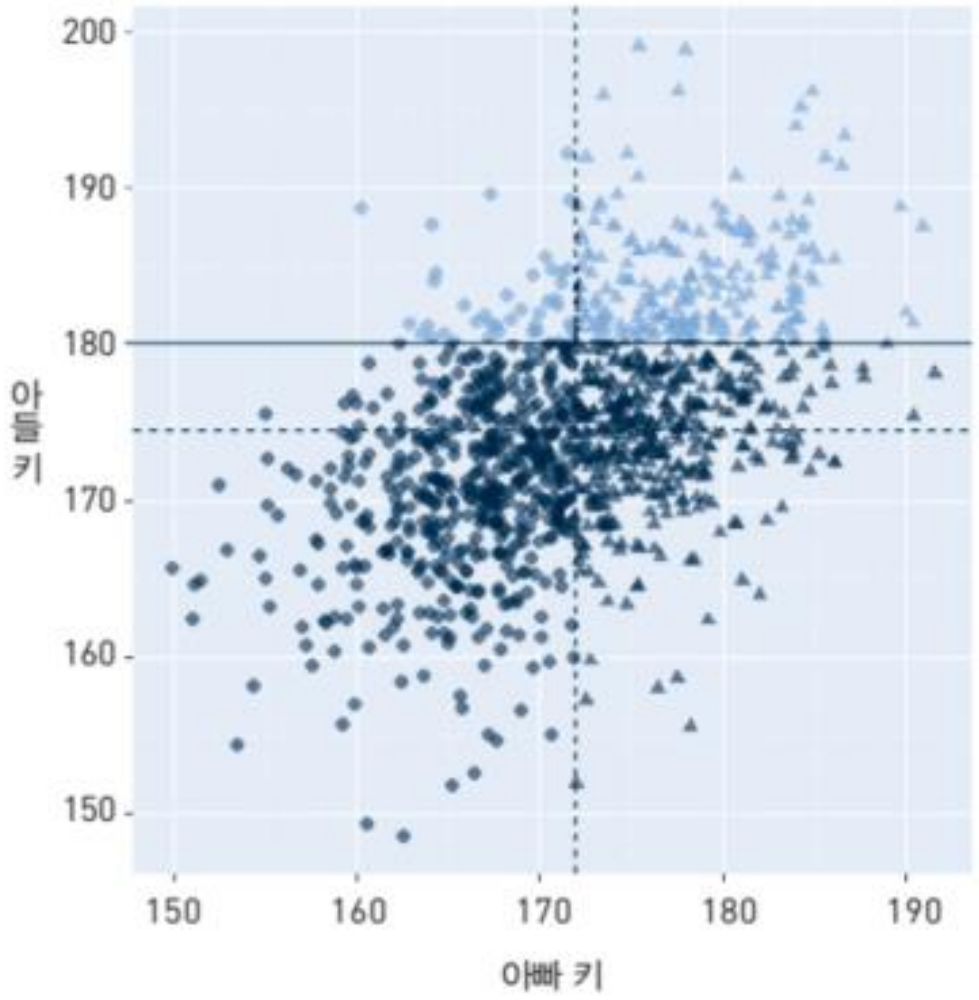
조건부 확률

- 전체 관측치에 대한 확률이 아닌 조건으로 선택된 부분 관측치에 대한 확률
- 예제)
  - 전체 가족에서 아들 키가 $180cm$ 이상일 확률 = 22%
  - 아빠 키가 $180cm$이상인 가족에서 아들 키가 $180cm$ 이상일 확률 = 35%

조건부 평균

- 전체 관측치에 대한 확률이 아닌 조건으로 선택된 부분 관측치에 대한 평균
- 예제) 아빠 키가 $180cm$이상인 가족에서 아들 키의 평균

조건부 확률과 조건부 평균



심슨의 역설(Simpson's paradox)

심슨의 역설

- 관측치의 구성에 따라 전체 확률과 조건부 확률이 모순처럼 보이는 경우
- 예제)
  - 성별 전체 합격률에서 남자의 합격율이 높은 것처럼 보이지만, 성/학과별 합격률을 계산해 보면 오히려 여자의 합격률이 높음

학과	성별	지원자	합격자	합격률
A학과	남자	80	64	80%
	여자	20	18	90%
B학과	남자	20	4	20%
	여자	80	24	30%

심슨의 역설(Simpson's paradox)

	A학과	B학과	합계
남자	80	20	100
여자	20	80	100
합계	100	100	200

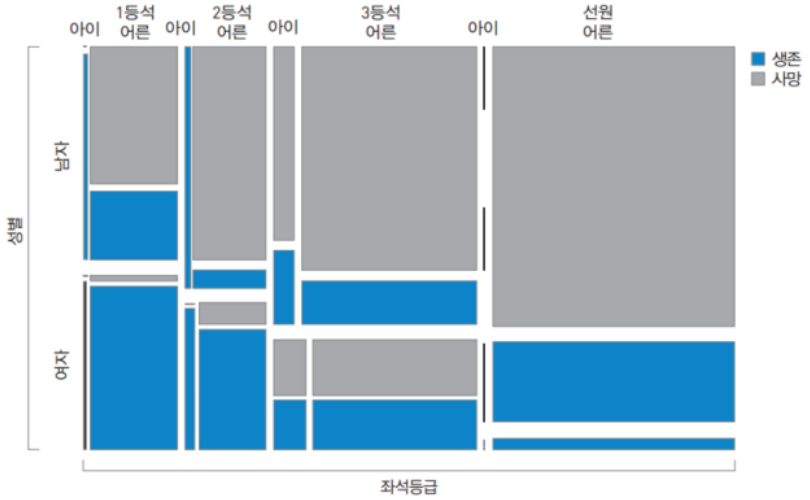
[더 복잡한 변수 간의 관계 탐색]

고차원 교차표와 모자이크 그림(Mosaic plot)

3차원 이상의 교차표

- 두 개의 변수로 계산한 (2차원) 교차표보다 훨씬 복잡한 정보를 포함
- 모자이크 그림 등으로 시각화 가능
- 예제) 타이타닉 호의 생존/사망 데이터

생존 여부		사망				생존			
합계		1,490				711			
연령 구분		아이		성인		아이		성인	
합계		52		1,438		57		654	
성별		남자	여자	남자	여자	남자	여자	남자	여자
합계		35	17	1,329	109	29	28	338	316
좌석등급	1등석	0	0	118	4	5	1	57	140
	2등석	0	0	154	13	11	13	14	80
	3등석	35	17	387	89	13	14	75	76
	선원	0	0	670	3	0	0	192	20

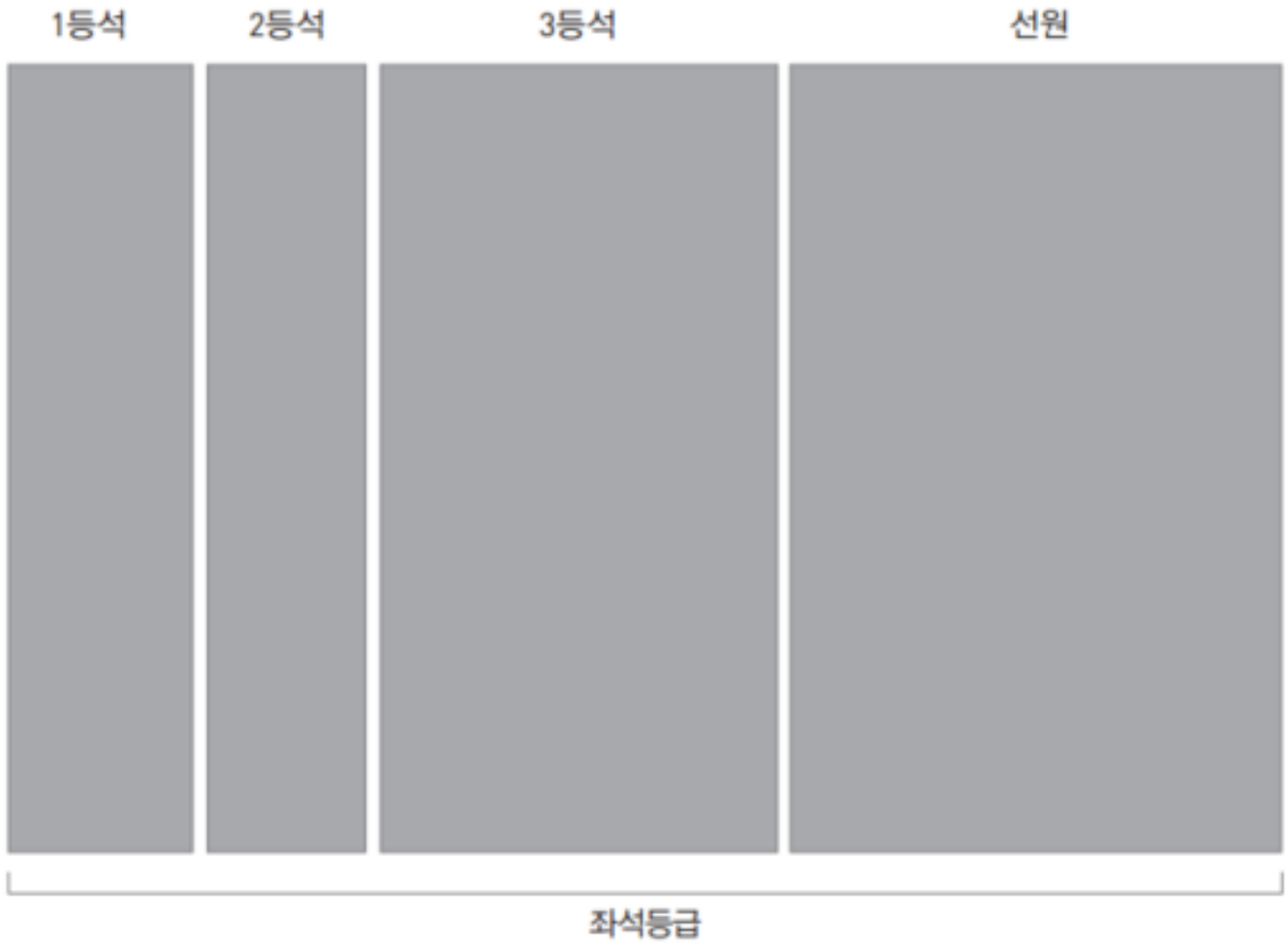


- “생존율에 가장 큰 영향을 미친 요인은 무엇일까?”에 대한 답을 주지는 못함

고차원 교차표와 모자이크 그림(Mosaic plot)

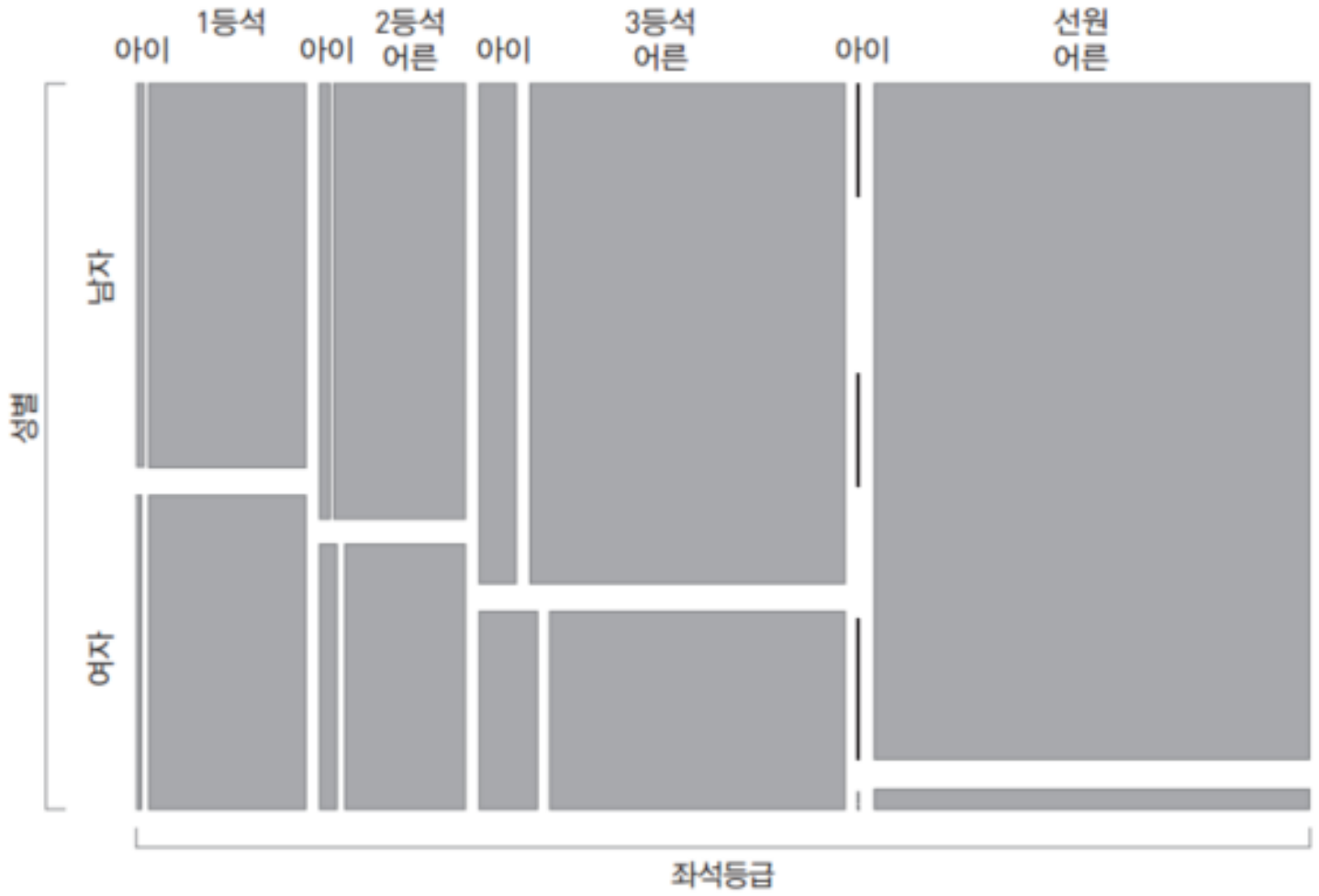
생존 여부		사망				생존			
합계		1,490				711			
연령 구분		아이		성인		아이		성인	
합계		52		1,438		57		654	
성별		남자	여자	남자	여자	남자	여자	남자	여자
합계		35	17	1,329	109	29	28	338	316
좌석 등급	1등석	0	0	118	4	5	1	57	140
	2등석	0	0	154	13	11	13	14	80
	3등석	35	17	387	89	13	14	75	76
	선원	0	0	670	3	0	0	192	20

고차원 교차표와 모자이크 그림(Mosaic plot)

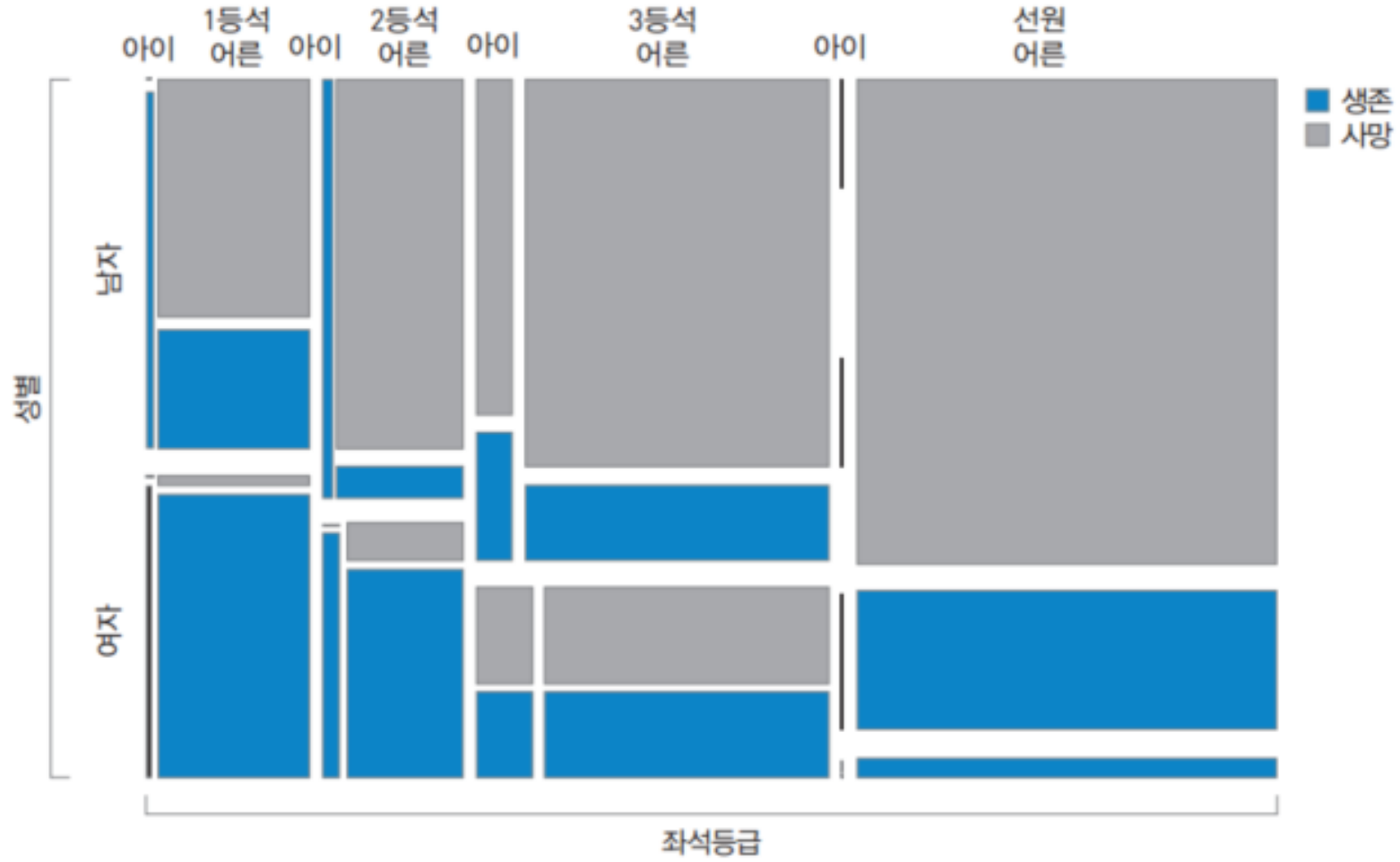


고차원 교차표와 모자이크 그림(Mosaic plot)

고차원 교차표와 모자이크 그림(Mosaic plot)



고차원 교차표와 모자이크 그림(Mosaic plot)



의사결정나무(Decision Tree) 모형

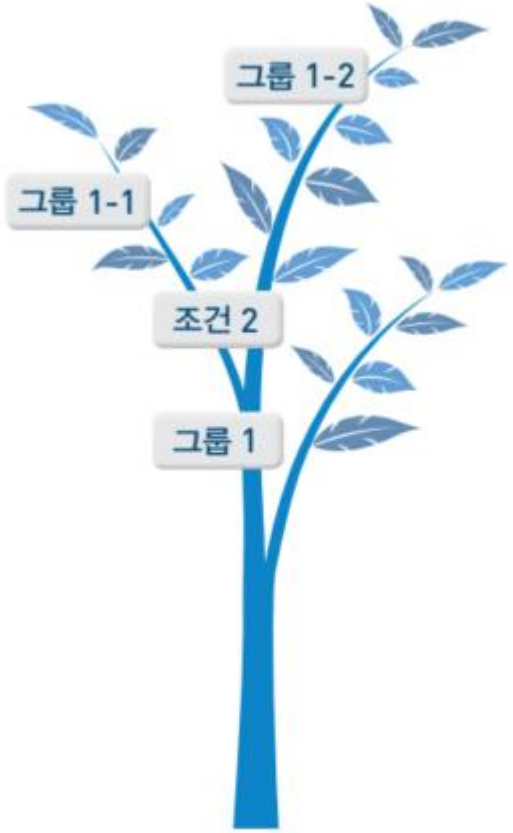
통계 모형(Statistical Model)

- 표현 방법은 다르나 “확률 모형”, “기계 학습” 등과 본질적으로 동일한 의미
- 관심있는 사건의 평균이나 확률을 설명하기 위해 다른 변수들을 활용
- 다른 변수를 조건으로 활용해서 조건부 확률, 조건부 평균을 계산

의사결정나무 모형

- 나무가지가 “Y” 형태로 갈라지듯 조건을 활용하여 관측치들을 두 개의 하위그룹으로 분할
- 나뉜 두 하위 그룹의 확률 혹은 평균의 차이가 가장 큰 조건을 선택
- 재귀분할(Recursive partitioning)로 모형 세분화
  - 하위 그룹을 계속해서 두 개의 더 작은 하위 그룹으로 분할
- 가지치기(Pruning)로 모형 단순화
  - 적정한 가지수를 포착하고 하위그룹을 단순화

의사결정나무(Decision Tree) 모형

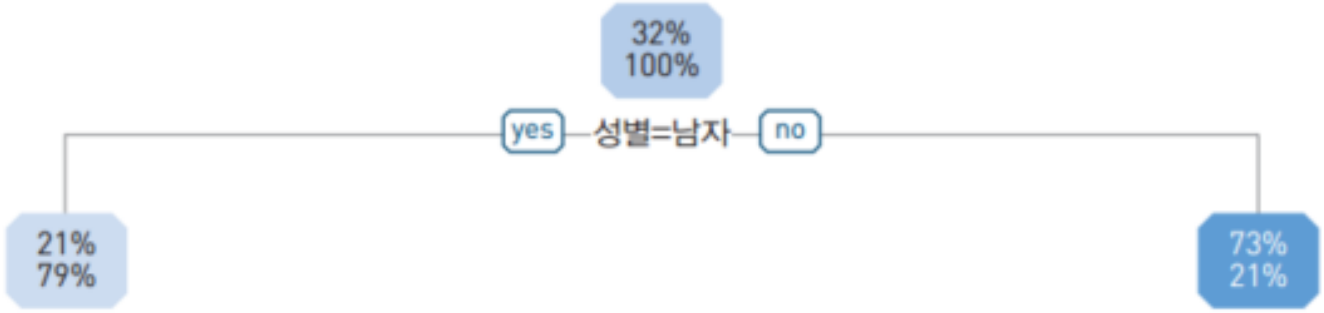


의사결정나무 모형의 예제

의사결정나무 모형을 활용한 타이타닉 생존 확률 확인

- 전체 생존율 : 전체 탑승객(100%)의 생존율은 32%

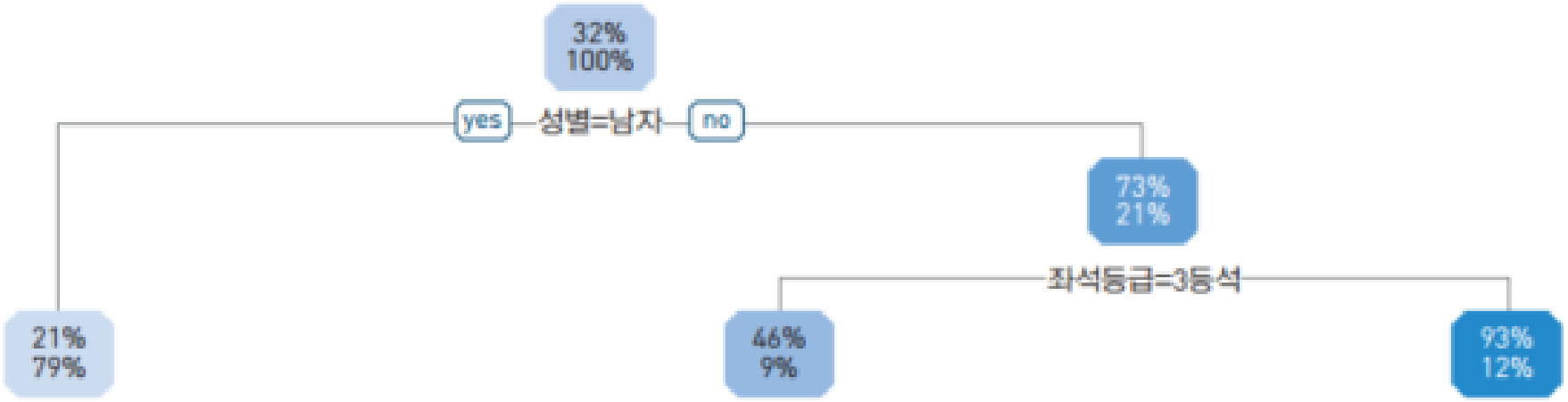
- 첫번째 분할 : 남자 탑승객(전체 중 79%)의 생존율은 21%지만,
  여자 탑승객(전체 중 21%)의 생존율은 73%
  - 다른 조건들보다 성별로 관측치를 나눴을 때 두 그룹의 생존율이 가장 큰 차이를 보임



의사결정나무 모형의 예제

의사결정나무 모형을 활용한 타이타닉 생존 확률 확인

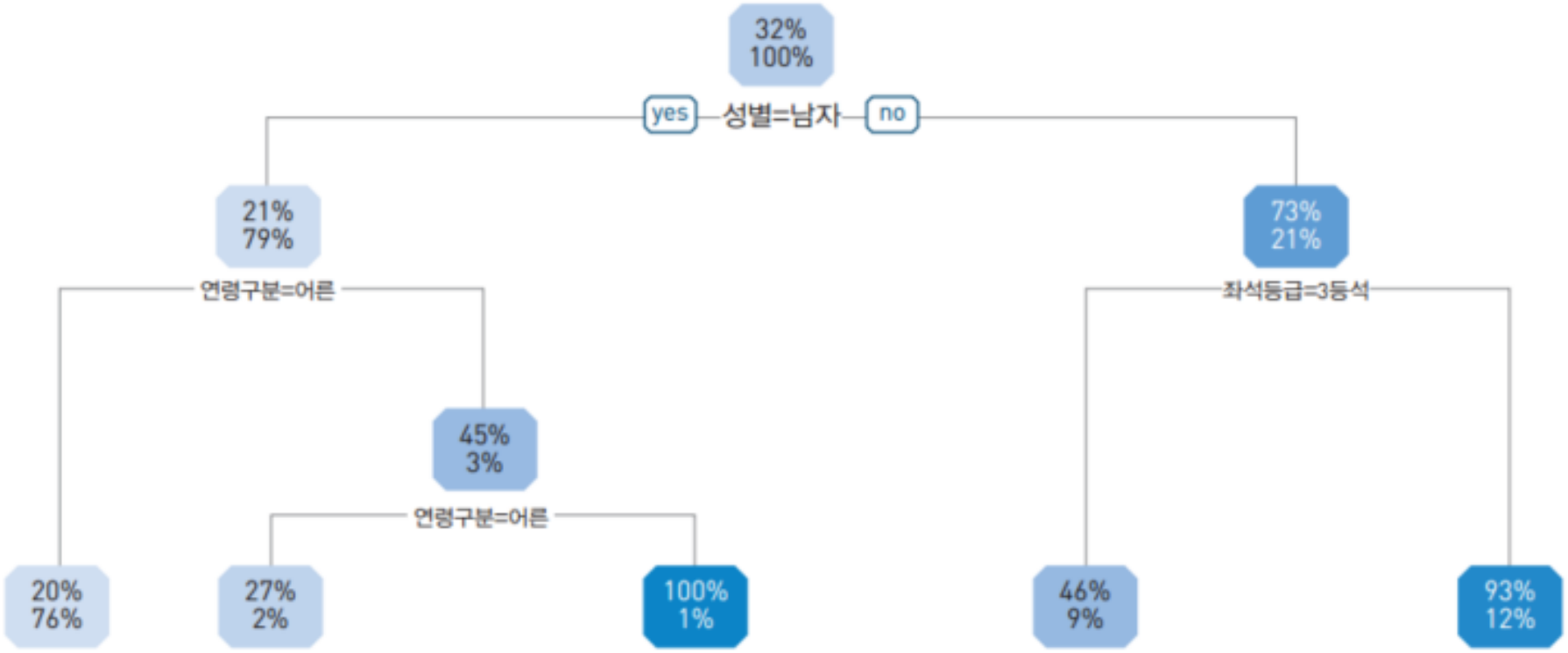
- 두번째 분할 : 여자 탑승객 중 3등석 탑승객(전체 중 9%)의 생존율은 46%지만,
  여자 탑승객 중 1, 2등석 및 선원(전체 중 12%)의 생존율은 93%



의사결정나무 모형의 예제

의사결정나무 모형을 활용한 타이타닉 생존 확률 확인

- 끊임없이 분할 가능



분산과 분산분석(Analysis of variance)

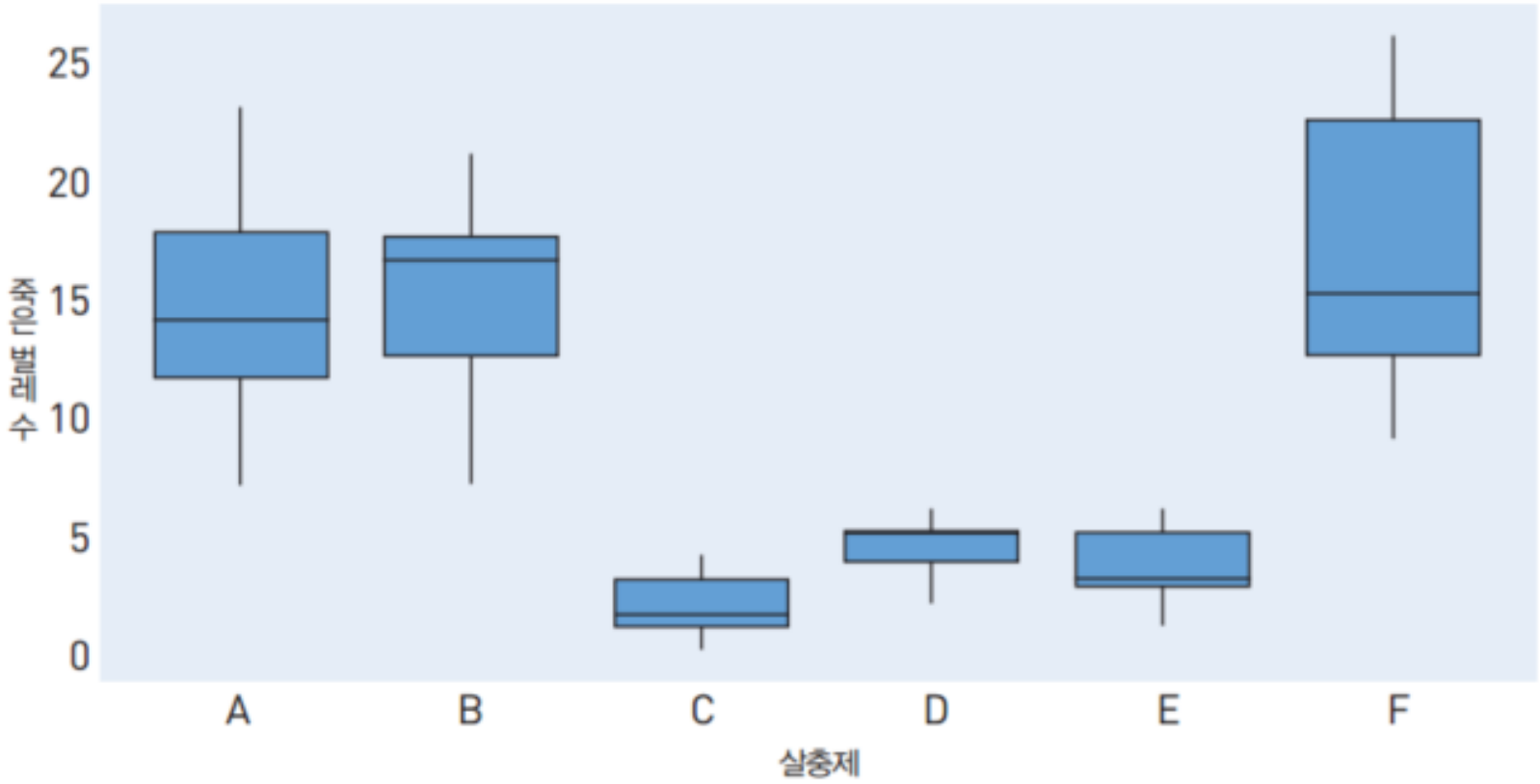
분산분석

- 하나의 연속형 변수와 범주형 변수 간의 관계에 대한 분석
- 영어 앞 글자를 따서 “ANOVA”로도 표현
- 최근 마케팅에서 “AB테스트” 등에 활용

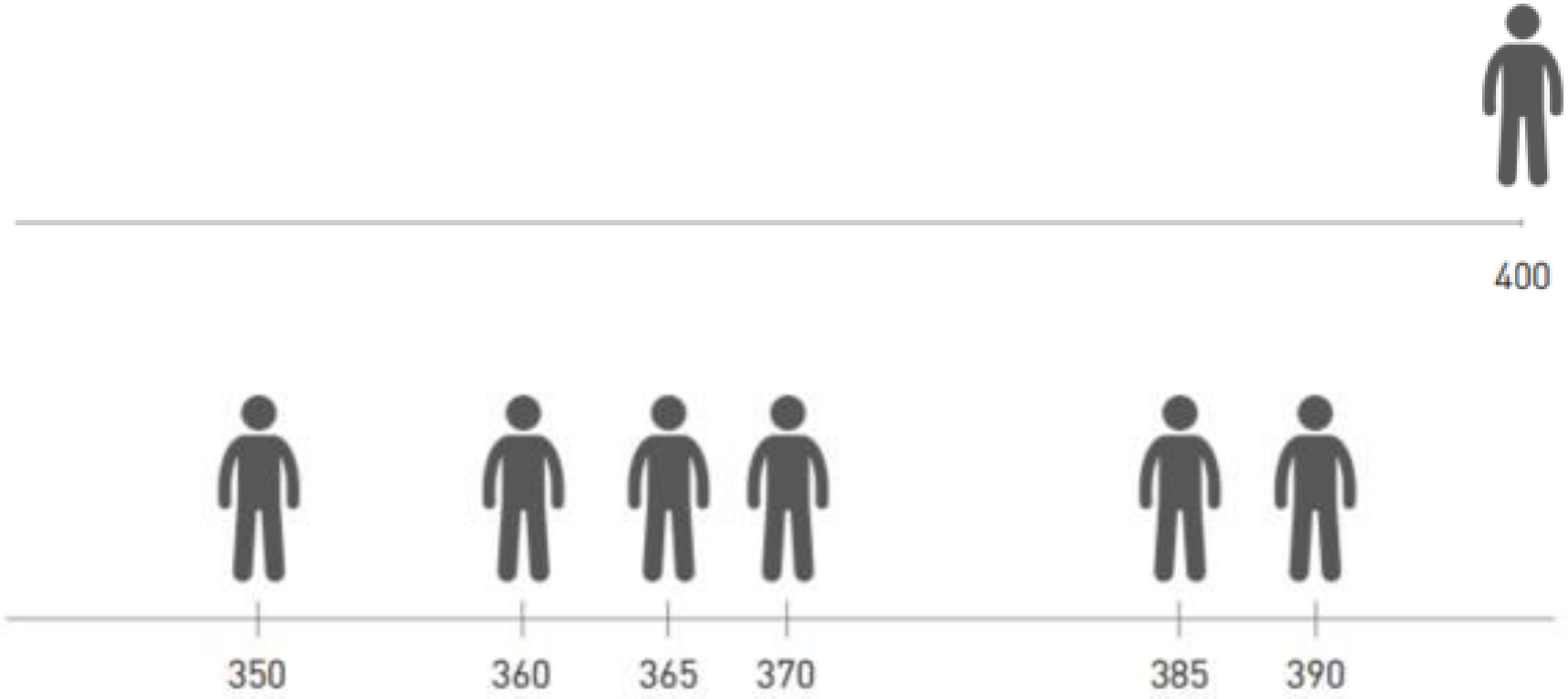
분산분석 예제

- “6개 살충제의 살충효과에 차이가 있을까?”
- 실제 살충제별 평균 죽은 벌레 수를 계산하거나 그룹별 상자그림을 통해 그룹간 차이를 확인 가능
- 이후 제곱합을 계산해서 통계적 유의성을 확인

분산과 분산분석(Analysis of variance)

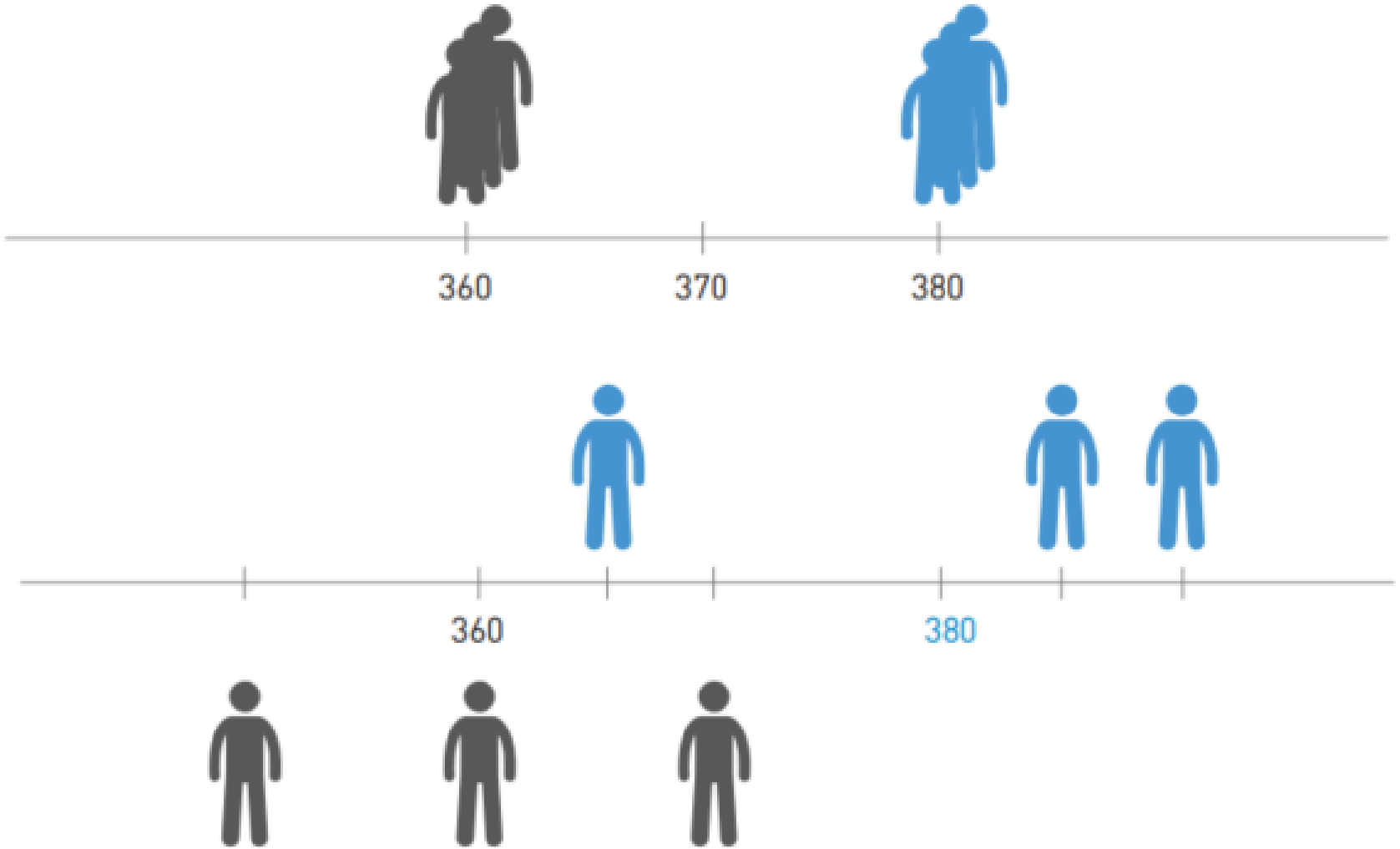


분산과 분산분석(Analysis of variance)



분산과 분산분석(Analysis of variance)

분산과 분산분석(Analysis of variance)



분산과 분산분석(Analysis of variance)

$$\sum_{i=1}^{72}(y_i - \bar{y})^2 = \sum_{i=1}^{72}(y_i - 9.5)^2 = 3{,}684$$

분산과 분산분석(Analysis of variance)

$$\sum_{i=1}^{72}(yg_i-\bar{y})^2=12\times(14.5-9.5)^2+12\times(15.3-9.5)^2+12\times(2.1-9.5)^2$$
$$+12\times(4.9-9.5)^2+12\times(3.5-9.5)^2+12\times(16.7-9.5)^2=2,669$$

분산과 분산분석(Analysis of variance)

$$\sum_{i=1}^{72}(y_i - yg_i)^2 = 1{,}015$$

분산과 분산분석(Analysis of variance)

$$\sum_{i=1}^{72}(y_i-\bar{y})^2=\sum_{i=1}^{72}(yg_i-\bar{y})^2+\sum_{i=1}^{72}(y_i-yg_i)^2$$