통계기초

데이터 분석

데이터 분석의 목적

- 변수 속에서 관측치 간의 차이를 확인
- 변수 간의 관계를 확인
- 차이와 관계를 확인하고 설명

데이터 분석의 과정

- 숫자와 그래프로 차이를 확인
- 모형으로 차이를 설명

두 가지 차이

① 절대적인 차이

: 관측치의 실제 값이나 데이터를 요약해서 얻은 숫자의 차이

② 상대적인 차이

: 절대적인 차이를 상대적인 값으로 바꾼 숫자의 차이

- 예)시험 점수와 시험 등수

절대적인 점수	상대적인 등수
90점	30명 중 3등

기술 통계량의 활용

①통계량(statistic)

–데이터로부터 계산된 모든 숫자

②기술 통계량(descriptive statistic)

–변수나 변수의 관계 등 데이터의 특성을 설명하는 통계량

- 예제) 표, 평균, 최댓값, 분위수 등

범주형 변수와 수준

범주형 변수

– 관측치들이 몇 개의 정해진 값만 가질 수 있음

범주형 변수의 수준(levels)

– 어떤 범주형 변수의 관측치들이 가질 수 있는 값들의 묶음

- 예) 변수 "성별"의 수준 : (남, 여)

  변수 "연령대"의 수준 : (10대, 20대, 30대, 40대, 50대, 60대 이상)

– 처리(treatment), 그룹(group)이라고도 표현

범주형 변수의 요약

예제) 고객 9명의 성별 요약하기

- 범주형 변수 “성별” 확인하기

고객	1	2	3	4	5	6	7	8	9
성별	남	여	여	남	여	여	여	남	남

- 수준별로 관측치 나누기

남	남			남				남	남
여		여	여		여	여	여		

- 수준별로 관측치 개수 세고 표로 정리하기

남	4
여	5

혹은

남	여
4	5

표와 차이

① 빈도표(frequency table)

–범주형 변수의 수준별 관측치 수를 정리한 표

–수준 간 절대적인 차이를 확인

② 상대빈도(relative frequency)

–빈도표에서 각 수준의 비율(proportion)을 계산

–수준 간 상대적인 차이를 확인

상대빈도의 계산

예제) 고객 9명의 성비 계산

- 범주형 변수 “성별”의 빈도표 확인하기

남	여
4	5

- 전체 합계 계산하기

남	여	합계
4	5	9

- 각 수준의 숫자를 전체 합계로 나눠 비율 계산하기

남	여	합계
0.44	0.56	1.00

$$\frac{5}{9} = 0.56$$

범주형 변수의 시각화

① 막대 그래프(bar chart)

–계산된 빈도표를 활용하여 각 수준의 값을 높이로 표현

–절대적인 차이를 확인

② 원 그래프(pie chart)

–계산된 상대빈도를 활용하여 원을 부채꼴로 분할

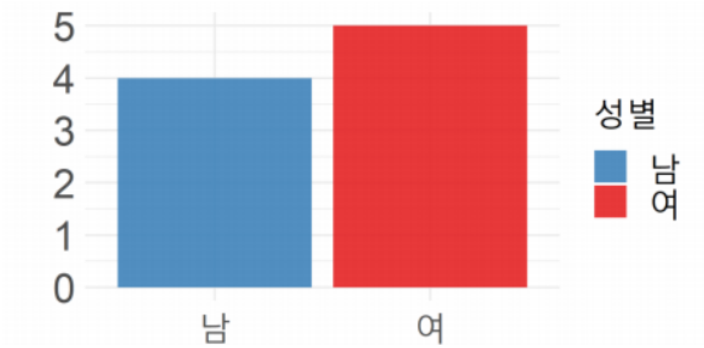
–상대적인 차이를 확인

범주형 변수의 시각화

예제) 성별 변수의 시각화

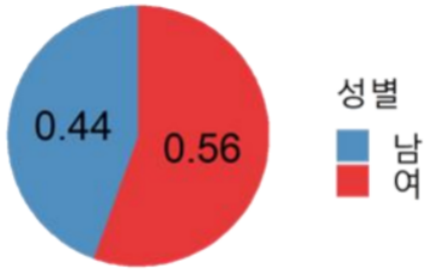
· 빈도표와 막대그래프

남	여	합계
4	5	9



· 상대빈도와 원 그래프

남	여	합계
0.44	0.56	1.00



수치형 변수의 처리

수치형 변수
–관측치들이 다양한 숫자 값을 가질 수 있음

①정렬을 활용한 수치형 변수의 요약
–최솟값, 최댓값, 중앙값 등 관측치들의 전반적인 위치를 확인

②합계를 활용한 수치형 변수의 요약
–평균, 분산 등 관측치들의 전반적인 특성을 확인

수치형 변수의 정렬

예제) 학생 9명의 시험 점수와 정렬

- 수치형 변수 “점수” 살펴보기

학생	1	2	3	4	5	6	7	8	9
점수	60	78	83	74	100	80	90	85	70

- “점수”의 오름차순으로 정렬하기

순서	1	2	3	4	5	6	7	8	9
점수	60	70	74	78	80	83	85	90	100

분위수의 활용

분위수(quantile)의 의미와 활용

–오름차순을 기준으로 관측치를 동일한 비율로 나누는 경계값

–수치형 변수 속 관측치들의 전반적인 분포를 확인

①백분위수(percentile)

–관측치를 1%씩 나누는 101개 숫자(0%, 1%, …, 99%, 100%)

②사분위수(quartile)

–관측치를 25%씩 나누는 5개 숫자(0%, 25%, 50%, 75%, 100%)

사분위수와 다섯숫자요약

다섯숫자요약(5 number summary)

–사분위수(5개 숫자) 계산하는 요약

–수치형 변수를 정렬하고 순서를 활용해서 값을 계산

① 0% : 최솟값(minimum)

② 25% : Q1(1st Quartile)

③ 50% : 중앙값(median)

④ 75% : Q3(3rd Quartile)

⑤ 100%: 최댓값(maximum)

사분위수의 의미와 계산

예제) 학생 9명의 시험 점수와 사분위수

- 오름차순으로 정렬된 점수의 확인

순서	1	2	3	4	5	6	7	8	9
점수	60	70	74	78	80	83	85	90	100

0% 25% 50% 75% 100%

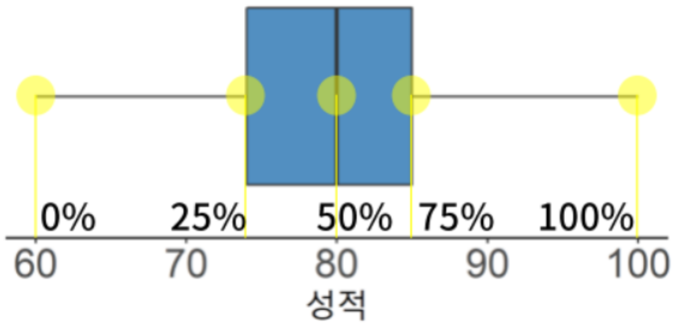
- 최솟값(0%) = 60 : 가장 먼저 나온 가장 작은 값
- 중앙값(50%) = 80 : 사람수를 봤을 때 중간에 있는 값
- 최댓값(100%) = 100 : 가장 나중에 나온 가장 큰 값
- Q1 = 73, Q3 = 86.25

상자그림을 활용한 시각화

예제) 학생 9명의 시험 점수의 다섯숫자요약과 상자그림

- 다섯숫자요약과 상자그림

분위수	0%	25%	50%	75%	100%
점수	60	73	80	86.25	100



도수분포표와 히스토그램

① 도수분포표(frequency table)

–수치형 변수를 적절한 구간 값을 활용하여 구간화

–각 구간의 관측치 수를 정리한 표

–수치형 변수에서 관측치 분포를 확인

② 히스토그램(histogram)

–도수분포표를 높이로 표현한 그림

–각 구간의 비중을 확인

도수분포표와 히스토그램

예제) 학생 9명의 시험 점수의 구간화와 히스토그램

- 오름차순으로 정렬된 시험점수

순서	1	2	3	4	5	6	7	8	9
점수	60	70	74	78	80	83	85	90	100

- 10점 간격 구간을 활용한 도수분포표 작성

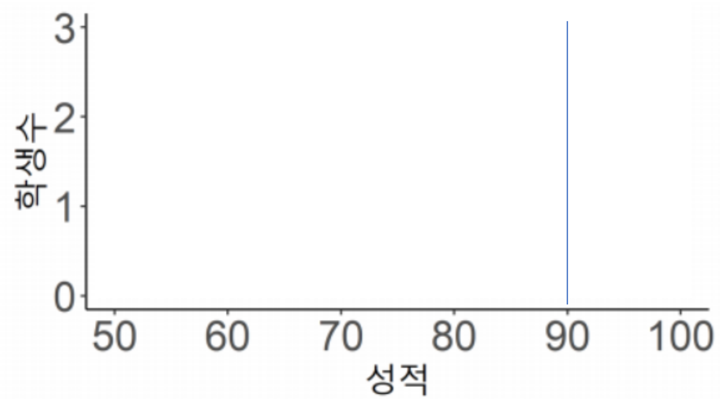
구간	51~60	61~70	71~80	81~90	91~100
학생수					

도수분포표와 히스토그램

예제) 학생 9명의 시험 점수의 구간화와 히스토그램

- 10점 간격 구간을 활용한 도수분포표와 히스토그램

구간	51~60	61~70	71~80	81~90	91~100
학생수	1	1	3	3	1



수치형 변수의 합계

수치형 변수와 범주형 변수의 차이

–수치형 변수는 사칙연산(+ − × ÷) 가능

수치형 변수의 합계를 활용한 요약

–평균 : 관측치들의 전반적으로 큰 정도

–분산/표준편차 : 관측치들 사이의 전반적인 차이의 정도

수치형 변수와 표기법

표기법(notation)

- 복잡한 계산을 표현하기 위해 미리 의미를 약속해둔 기호

- n : 관측치 개수

평균의 의미와 계산

평균(mean)의 계산

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 수치형 변수의 모든 관측치를 더하고 관측치 개수로 나눈 값
- 고정된 합계를 관측치 간 차이가 없게 나눠 가진 값
- 관측치들의 전반적인 크기를 의미

평균의 의미와 계산

예제) 9명 학생의 평균 시험 점수

- 오름차순으로 정렬된 시험점수

순서	1	2	3	4	5	6	7	8	9
점수	60	70	74	78	80	83	85	90	100

- 관측치 개수와 합계를 활용한 평균 계산

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{1}{9} \times 720 = 80$$

평균과 중앙값의 비교

대푯값

–관측치들의 전반적인 크기를 설명하는 값

–평균과 중앙값이 대표적

①평균

: 전반적인 크기를 잘 설명하지만 특이값에 따라 영향을 많이 받음

②중앙값

: 관측치 개수를 활용하기 때문에 특이값의 영향이 제한적

평균과 중앙값의 비교

예제) 9명 학생의 시험 점수의 평균과 중앙값

- 평균 80점, 중앙값 80점
- 0점 처리된 한 학생이 추가된 10명의 시험 점수의 대푯값 계산

순서	1	2	3	4	5	6	7	8	9	10
점수	0	60	70	74	78	80	83	85	90	100

- 10명의 평균 72점 : $\frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10} \times 720 = 72$
- 10명의 중앙값 79점 : $\frac{(78+80)}{2} = \boxed{79}$

분산의 의미와 계산

분산(variance)의 계산

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

- 관측치에서 평균을 뺀 것의 제곱의 평균
- 관측치들이 평균을 중심으로 흩어져 있는 정도(면적)

표준편차의 의미와 계산

표준편차(standard deviation)의 계산

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

- 계산된 분산의 제곱근
- 분산이 가지는 단위(scale/unit)의 문제를 해결
- 관측치들이 평균을 중심으로 흩어져 있는 정도(길이)

분산과 표준편차의 계산

예제) 9명 학생의 시험 점수의 분산과 표준편차

- 오름차순으로 정렬된 시험점수

점수	60	70	74	78	80	83	85	90	100

- 수식과 평균 80을 활용한 분산의 계산

$$s_x{}^2 = \frac{1}{9-1}\{ (60-80)^2 + (70-80)^2 + \cdots + (100-80)^2\} = \frac{1074}{8} = 134.25$$

- 표준편차의 계산

$$s_x = \sqrt{s_x{}^2} = \sqrt{134.25} = 11.59$$

평균과 분산, 표준편차의 계산

예제) 5개 관측치를 가진 3개 수치형 변수의 비교

① 첫번째 변수 :

1	2	3	4	5

- 평균 : $\frac{1}{5} \times 15 = 3$, 분산 : $\frac{10}{4} = 2.5$, 표준편차 : $\sqrt{2.5} = 1.58$

② 두번째 변수 :

4	5	6	7	8

- 평균 : $\frac{1}{5} \times 30 = 6$, 분산 : $\frac{10}{4} = 2.5$, 표준편차 : $\sqrt{2.5} = 1.58$

③ 세번째 변수:

2	4	6	8	10

×4 ×2

- 평균 : $\frac{1}{5} \times 30 = 6$, 분산 : $\frac{40}{4} = 10$, 표준편차 : $\sqrt{10} = 3.16$

다양한 상대적인 위치

① 백분율(percentage)

–전체 관측치 중 특정 값보다 작은 관측치 개수의 비율을 0~1로 계산

② 최소-최대 정규화(min-max normalization)

–최솟값과 최댓값의 구간에서 특정 값의 상대적 위치를 0~1로 계산

③ 표준화(standardization)

–특정 값이 평균으로부터 떨어진 정도를 표준편차의 단위로 표현

–평균과 표준편차에 따라 부호에 상관없이 다양한 값을 가질 수 있음

다양한 상대적인 위치

예제) 어떤 시험 점수의 요약

- 응시자수 50명, 60점 이하 득점자 5명, 90점 초과 득점자 5명
- 최솟값 50, 최댓값 100, 평균 70, 표준편차 10,

60점과 90점의 상대점수 계산

시험 점수	백분위	최소-최대	표준화
60점	$\frac{5}{50}$=0.1, 10%	$\frac{60-50}{100-50}$=0.2	$\frac{60-70}{10}=-1$
90점	$\frac{45}{50}$=0.9, 90%	$\frac{90-50}{100-50}$=0.8	$\frac{90-70}{10}=2$