2장

데이터의 탐색(Exploratory)

▶ 어떠한 종류의 데이터인지 분류할 필요가 있음
▶ 그림, 표, 측도 등을 이용하여 데이터 요약
▶ 데이터 품질 탐색 (정확한지, 결측값은 없는지, 특이값은 없는지)

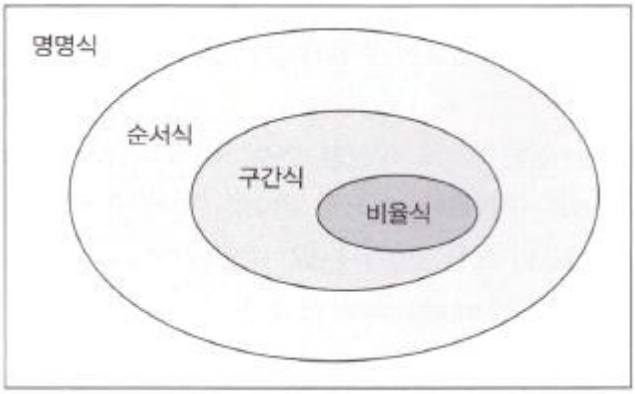
2.1.1 데이터의 분류

▶ 데이터: 관심의 대상이 되는 사람이나, 사물, 사건의 속성 (변수(variable), 특징(feature), 항목(item)이라 함)을 일정한 규칙에 의해 관찰한 결과이거나 측정한 값들
예) 백화점에서 옷을 구입하는 고객의 성별과 신장
성별 측정값: 여자, 남자 (변수의 측정값: 문자나 기호)
신장 측정값: 180cm, 175cm ... (변수의 측정값: 숫자)

▶ 변수
(1) 이산형 변수(discrete variable): 유한개이거나 셀 수 있는 변수, 데이터 값이 유한개의 범주로 표시
예) 성별, 불량품(inferior)의 수
→ 이러한 변수들의 데이터를 이산형 데이터라고 함
(2) 연속형 변수(continuous variable): 셀 수 없는 변수
예) 신장, 체중
→ 이러한 변수들의 데이터를 연속형 데이터라고 함

▶ 데이터를 측정하는 방법에 따라 명명식, 순서식, 구간식, 비율식 데이터로 나눔 (그림 2-1 참조)
(1) 명명식(nominal) 데이터: 예) 성별에 대한 조사
남자=1, 여자=2 : '1'이 '2'보다 좋다거나 우열을 나타내는 것이 아님
(2) 순서식(ordinal) 데이터: 관심의 대상이 되는 사물이나 사건을 순서에 의해 측정한 데이터
예) 개인의 월수입을 고소득, 중간소득, 저소득 구분하여 측정
(3) 구간식(interval) 데이터: 순서식 데이터처럼 한 대상이 다른 것보다 크고 작은 것을 구별할 뿐만 아니라, 온도와 같이 얼마나 크고 작은지를 측정하는 데이터
한 측정값을 다른 측정값으로 나눈 값(비율)에 의미가 없음
(4) 비율식(ratio) 데이터: 구간식 데이터처럼 한 대상이 다른 것보다 얼마나 크고 작은지 구체적으로 측정할 뿐 아니라, 무게, 길이와 같이 두 측정된 값의 비율이 의미가 있음
예) 체중, 키

비율식 데이터와 구간식 데이터의 차이는 절대 영점이 있느냐 vs. 없느냐의 차이(예: 무게 vs. 온도)



예제 2.1.1

[표 2.1] 한 신용카드 회사의 고객 데이터

카드번호	성명	성별	나이	지역	직업	월수입 (단위: 만원)	구매액 (단위: 만원)
2132001	김하나	여	25	서울	회사원	150	20
2132002	이우리	남	32	부산	공무원	200	30
2132003	정민국	남	41	광주	자영업	500	150
2132004	박대한	남	57	광주	회사원	400	120
2132005	김현지	여	21	서울	학생	70	30
2132006	이수창	남	33	서울	회사원	300	200
2132007	김대현	남	47	부산	자영업	500	150
2132008	장대호	남	35	서울	회사원	330	150
2132009	김영주	여	53	부산	공무원	350	100
2132010	김선규	남	20	부산	학생	70	20
2132011	김기문	남	50	광주	자영업	600	300
2132012	이수진	여	55	서울	자영업	700	200
2132013	박준일	남	34	서울	회사원	270	70
2132014	김성경	여	40	서울	공무원	450	50
2132015	홍현일	남	38	서울	회사원	400	100
2132016	이경은	여	35	광주	공무원	300	110
2132017	김승현	남	28	서울	회사원	250	120
2132018	이재군	남	25	서울	학생	50	15
2132019	김종선	남	42	부산	회사원	500	150

2.1.2 데이터의 품질

: 데이터의 품질을 높이기 위한 정제 작업 필요

가) 정확도(Accuracy), 오차(Error) 및 정밀도(Precision)

▷ 정확도(Accuracy): 실제값과 측정된 값의 가까운 정도
▷ 정확도에 대한 측도
측정오차(measurement error): 기구 등을 통하여 데이터 값을 측정하는 과정에서 발생하는 오차
측정오차를 줄이기 위해 기구의 교체, 실험환경의 개선 등이 필요
정밀도(precision): 한 속성변수 값을 반복적으로 측정하였을 때 이 값들이 서로 얼마나 가까운지를 나타내는 측도. 데이터의 표준편차 등이 이용됨.
편향(bias): 측정된 값들의 조직적인 변이도(variation)에 대한 측도.
측정값과 측정된 값의 평균과의 차이 이용
수집오차(collection error): 인위적인 실수로 데이터의 수집과정에서 나타나는 오차
데이터를 입력할 때 실수로 레코드를 생략한다든가, 중복해서 입력하는 경우 발생
주의를 기울이고 재검정하면서 개선할 수 있음.

나) 결측값(missing value)

▷ 정보가 수집되지 않은 경우, 응답자가 거부하여 발생
▷ 데이터의 수가 충분한 경우 간단한 방법은 결측값을 포함하는 데이터를 제거
← 해당변수의 결측 데이터가 많지 않은 경우 주로 사용
▷ 다른 방법
결측값 예측(평균이나 중앙값 이용) 또는 결측 데이터의 나머지 변수값과 같은 데이터들만 찾은 후, 이 데이터들의 변수 값의 평균이나 중앙값 이용

다) 특이값(outlier)

▷ 한 데이터가 나머지 대부분의 데이터의 변수와 속성이 매우 다르거나, 변수 값이 통상 기대되는 값의 범위를 벗어나는 경우
▷ 적법한 데이터로서 변수값이 다른 데이터와 많이 다른 것을 의미
▷ 잡음(noise)와 혼동하지 말 것
잡음: 측정기구의 잘못, 인위적인 실수에 의해 나타난 이상한 변수 값
▷ 특이 데이터를 찾아내는 분석기법은
예) 신용카드 사용자 데이터에서 최근 사용액이나 사용 횟수가 갑자기 증가한 사용자를 찾아내 혹시 도난카드인지 살펴보는 사기 탐지(fraud detection)를 들 수 있음

2.2 그림을 이용한 데이터 탐색

2.2.1 한 변수에 대한 그림

▶ 이산형 변수 - 막대그림, 원그림
▶ 연속형 변수 - 히스토그램

가) 막대그림 (bar chart)

- 빈도수의 분포를 한 눈에 비교할 수 있다
- 막대 사이의 간격을 띄움 (이산형 변수이기 때문)

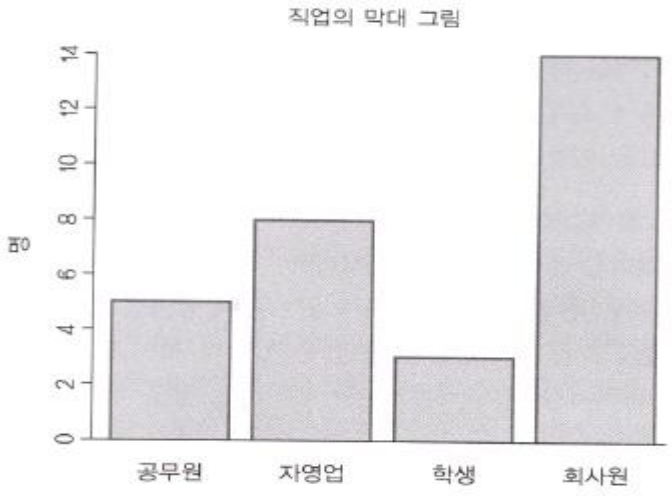


그림 2-2 [표 2.1] 데이터에서 직업에 대한 막대그림

나) 원그림(pie chart)

- 한 원을 변수값의 도수에 비례하게 몇 개의 조각으로 나누어 그린 그림
- 각 조각에 해당 변수 값과 상대도수 기입

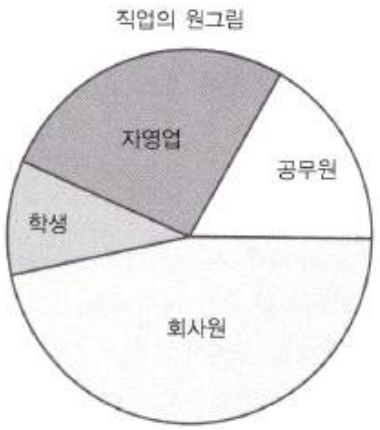


그림 2-3 [표 2.1] 데이터에서 직업에 대한 원그림

다) 히스토그램(histogram)

- 여러 개의 계급구간을 나누어 도수분포표를 작성한 후 이에 대한 막대그림(막대 사이에 간격이 없는)을 그림

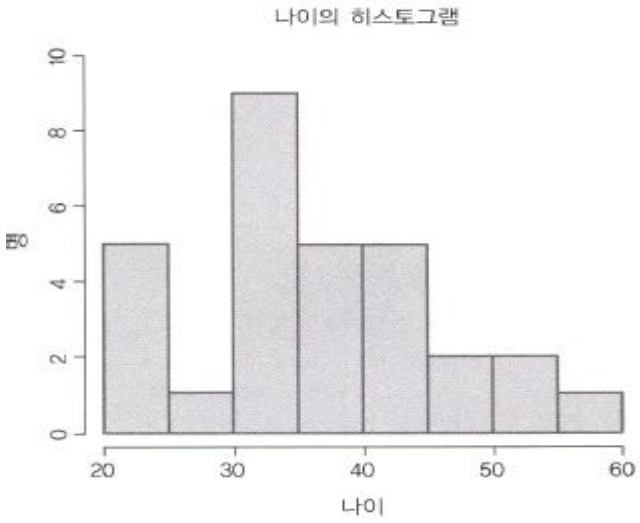


그림 2-4 [표 2.1] 데이터의 나이에 대한 히스토그램

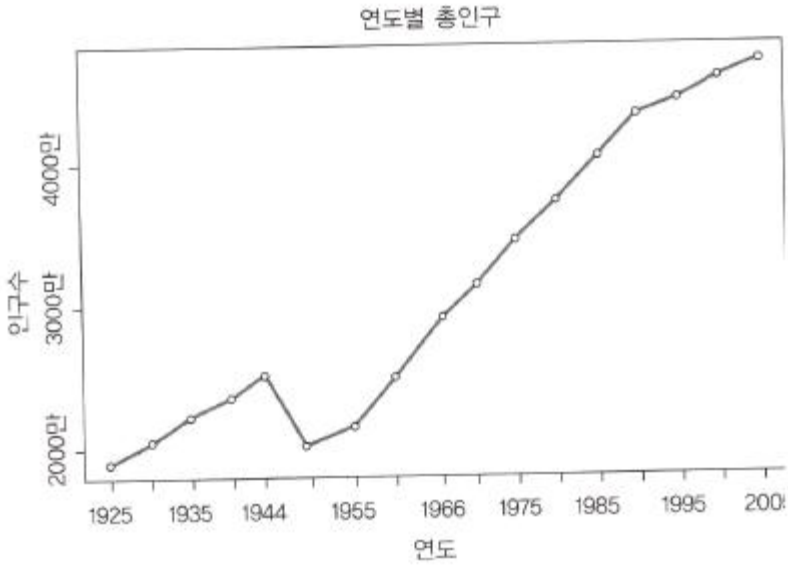
2.2.2 두 변수의 그림

▶ 두 변수가 모두 연속형일 경우 선그림이나 산점도를 이용하여 두 변수의 관련성을 조사

가) 선그림 및 시계열 그림

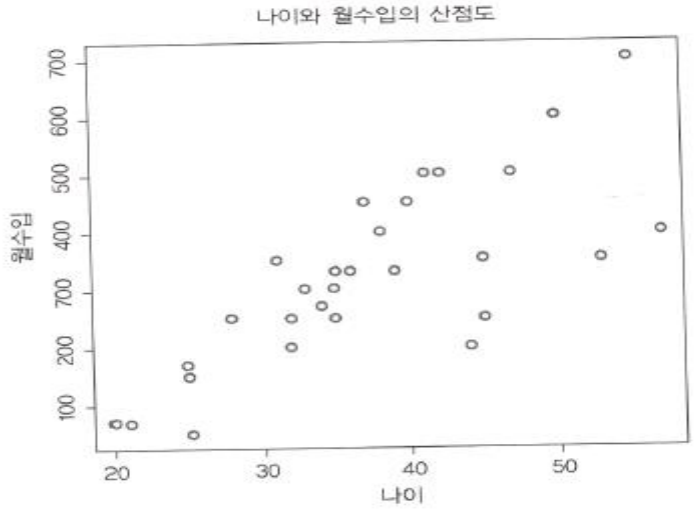
▷ 선그림(line graph) 및 시계열 그림(time series plot)

- 선그림: X축 변수에 대한 대소를 구별하지 않고 데이터 순서대로 그림
- 시계열 그림: X축 시간 변수의 크기에 따라 점들을 연결
- 변수가 시간에 따라 어떠한 추세를 보이는지 또는 주기성을 가지고 있는지 알 수 있음



나) 산점도

- 각각의 관찰값을 X-Y평면상의 좌표값으로 하여 점으로 나타냄
- 두 변수가 관련이 있다면 점들은 한 직선이나 곡선 주위에 위치할 수 있음.



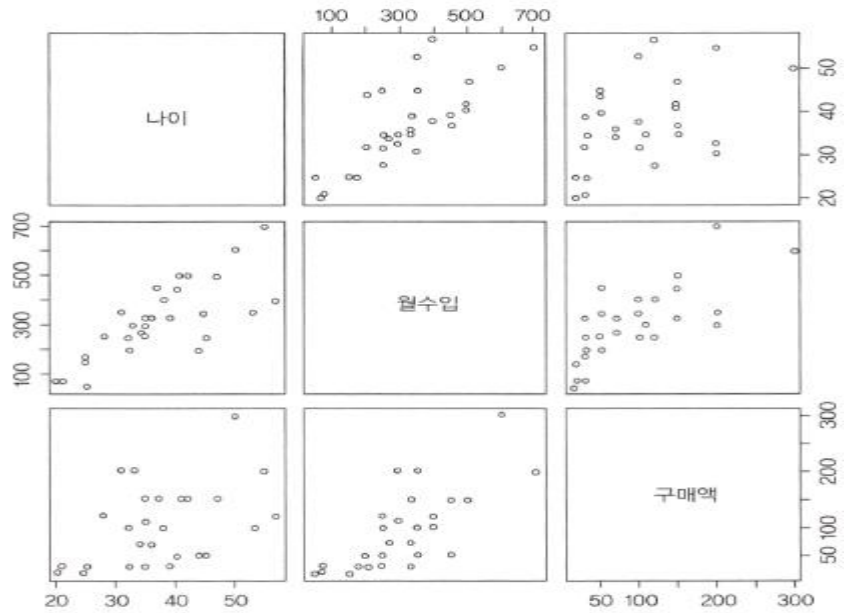
2.2.3 다중 변수의 그림

▶ 연속형 변수의 수가 여러개일 경우 산점도 행렬, 레이더차트, 평행좌표그림 이용
▶ 이산형 변수일 경우는 모자이크 그림 이용

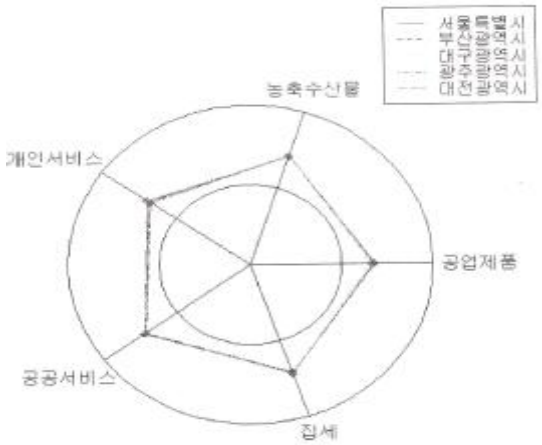
가) 산점도 행렬(scatter plot matrix)

- 여러 개의 X축 변수와 여러 개의 Y축 변수에 대한 산점도를 한 화면에 보여준다

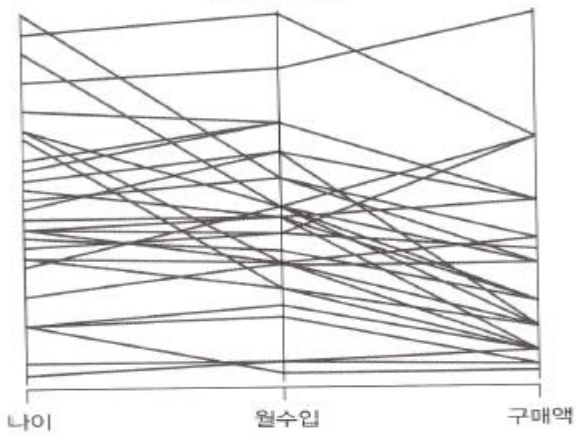
그림 2-7. 나이와 월수입 양의 상관
월수입과 구매액 상당한 상관을 보임
나이와 구매액의 상관은 높지 않음

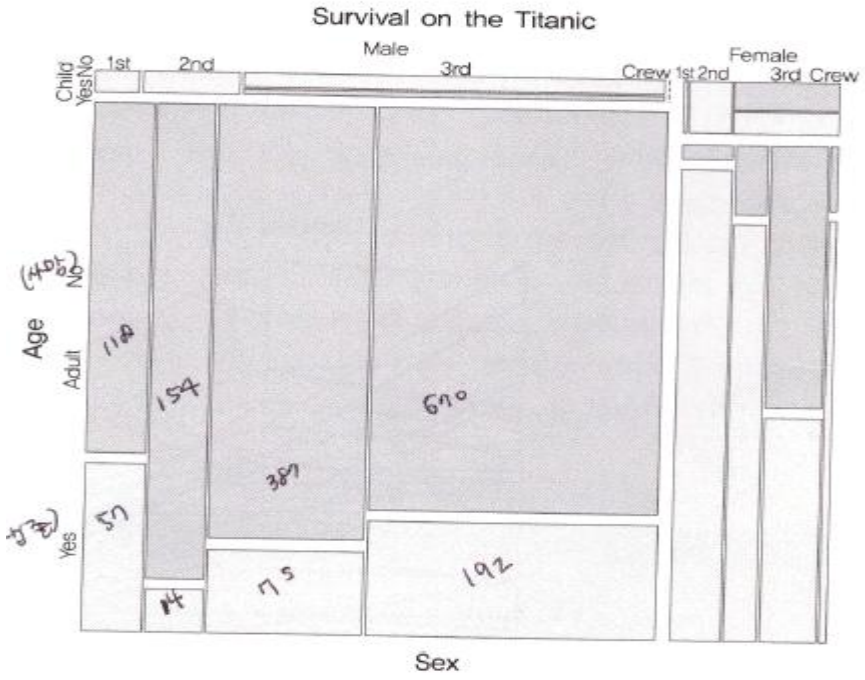


상품성질별 소비자 물가지수



평행좌표 그림





체르노프 안형 그림

2.3 통계표/통계량을 이용한 데이터 탐색

- 구체적인 데이터의 속성을 알기 위해 통계표나 통계량을 이용

2.3.1 한 변수의 통계표/통계량

▶ 이산형 데이터 요약 - 도수분포표 이용

▶ 연속형 데이터 요약 - 데이터의 중심위치, 데이터의 산포, 데이터의 왜도, 첨도

가) 도수분포표 (frequency table)

- 한 변수의 측정값이 나타나는 빈도수(frequency)를 정리한 것

▷ 상대도수(relative frequency or 백분율(%))= $\frac{\text{도수}}{\text{전체데이터의개수(결측값포함)}}$

▷ 누적상대도수(cumulative relative frequency): 상대도수를 누적함

▷ 유효백분율, 누적유효백분율= $\frac{\text{도수}}{\text{전체데이터의개수(결측값제외)}}$

나) 중심위치의 측도(measure of central tendency)

▷ 평균(mean): 가장 많이 사용

$$= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

모집단의 평균 = 모평균(population mean) = μ

표본의 평균 = 표본평균(sample mean) = $\overline{x}$

평균은 극단값에 영향을 받음

▷ 중앙값(median): 데이터를 크기 순서로 나열할 때 중앙에 놓이는 값

n(데이터 수) n: 홀수 median = $\left(\frac{n+1}{2}\right)$번째 값

n: 짝수 median= $\frac{\frac{n}{2}\text{번째 값} + \left(\frac{n}{2}+1\right)\text{번째 값}}{2}$

중앙값은 극단값에 민감하지 않아 평균보다 중심위치의 측도로 더 자주 쓰임

▷ 최빈값(mode): 데이터 중 가장 빈도가 높은 값

이산형 데이터 - 도수분포표를 보면 쉽게 구할 수 있음

연속형 데이터 - 데이터를 몇 개의 계급구간으로 나누어 가장 도수가 높은 계급구간의 중앙값을 최빈값으로 정함

다) 산포의 측도(measure of dispersion)

- 분산, 표준편차, 변이계수, 범위, 사분위수범위

▷ 분산(variance) - 각 데이터 값과 평균과의 거리를 제곱하여 합을 구한 후 이를 데이터 수로 나눈 것

- 데이터가 평균을 중심으로 많이 흩어져(scattered) 있으면 분산 ↑
  데이터가 평균을 중심으로 많이 몰려(gathered) 있으면 분산 ↓
- 모집단의 분산 = 모분산(population variance)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

표본의 분산 = 표본분산(sample variance)

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

▷ 표준편차(standard deviation) - 분산의 제곱근

모표준편차 $\sigma = \sqrt{\sigma^2}$

표본표준편차 $s = \sqrt{s^2}$

▷ 변이계수(coefficient of variation)

- 데이터 수나 측정단위가 다른 두 개 이상의 데이터를 비교할 때 사용

$$C.V = \frac{\sigma}{\mu}\times 100\left(\text{or } \frac{s}{\bar{x}}\times 100\right) = \frac{\text{표준편차}}{\text{평균}}$$

변이계수(모집단) $C = \frac{\sigma}{\mu}\times 100$ (단위 %)

변이계수(표본) $C = \frac{s}{\bar{x}}\times 100$ (단위 %)

▷ 범위(range) = 최대값 - 최소값

- 극단점이 있을 경우 올바른 산포의 측도가 되지 못함

▷ 사분위수범위(interquartile range, IQR) - 범위의 단점 보완

$IQR = Q_3 - Q_1$

▷ 백분위수(percentile) - 데이터를 작은 것부터 큰 것까지 순서대로 늘어놓았을 때 P%번째 데이터

P 백분위수 - 자기값 이하(값포함)로 적어도 P%의 관측값이 있고,
자기값 이상(값포함)으로 적어도 (100-P)% 관측값이 있는 수

일사분위수(1st quartile, Q_1 표시) = 25% 백분위수

이사분위수(2nd quartile, Q_2 표시) = 50% 백분위수 = m(중앙값)으로 표시

삼사분위수(3rd quartile, Q_3 표시) = 75% 백분위수

사분위수범위(IQR) = $Q_3 - Q_1$

▷ 상자그림(box-whisker plot) - 사분위수를 이용하여 데이터에 대한 요약그림을 그림

- 데이터 분포의 대칭성, 데이터의 중심위치, 산포의 정도, 극단점 등을 알 수 있음
- 여러 집단의 비교에 많이 이용

라) 왜도 및 첨도(skewness and kurtosis)

데이터의 형태 - 좌우대칭(symmetric) 형태

- 우측으로 꼬리가 긴 형태(skewed to the right)
- 좌측으로 꼬리가 긴 형태(skewed to the left)

▷ 왜도(skewness): 표본 데이터의 대칭성을 측정하기 위한 측도

$$\beta_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3$$

$\beta_3 = 0$ 대칭

$\beta_3 > 0$ 평균에서 오른쪽으로 꼬리가 긴 경우

$\beta_3 < 0$ 평균에서 왼쪽으로 꼬리가 긴 경우

s^3으로 나누는 이유: 두 개 이상의 데이터에 대한 왜도를 비교할 때 데이터 개수나 측정 단위가 서로 다를 수 있기 때문에 표준화 시켜주는 것

▷ 첨도(kurtosis): 표본 데이터가 정규분포와 비교하여 봉오리(bud)가 얼마나 높거나 낮은 지 알아보는 측도

$$\beta_4 = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

정규분포의 첨도 $\beta_4 = 0$

정규분포보다 높은 봉오리를 갖는 분포 $\beta_4 > 0$

정규분포보다 낮은 봉오리를 갖는 분포 $\beta_4 < 0$

2.3.2 두 변수의 통계표/통계량

▶ 두 변수의 요약을 위해, 이산형 데이터의 경우: 교차표(cross table) 이용

연속형 데이터의 경우: 중심위치, 평균벡터, 산포도, 공분산(covariance) 및 상관계수 측도(correlation coefficient) 이용

가) 교차표(cross table)

- 두 개의 이산형 변수를 정리하여 그 연관된 특성을 연구하는 데 효과적
- 한 변수는 행에, 한 변수는 열에 표시
- 행 변수의 속성과 열 변수의 속성이 교차하는 부분에 칸을 만든 후, 칸에 속하는 빈도수를 적은 것

나) 공분산, 상관계수

▷ 공분산(covariance): 두 변수 X와 Y의 상호관계를 나타내는 측도

- 모집단 공분산(population covariance)= $Cov(X, Y)$
- 표본: $(x_1, y_1), \cdots, (x_n, y_n)$

- 표본 공분산(sample covariance): $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- $s_{XY} > 0$: X, Y 양의 상관관계
- $s_{XY} < 0$: X, Y 음의 상관관계

▷ 상관계수(correlation coefficient): 두 변수 X와 Y의 상관관계를 나타내는 측도

- 모집단 상관계수(population correlation)=ρ

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

(i) $-1 \le \rho \le 1$: ρ의 값이 +1에 가까울수록 양의 상관관계
ρ의 값이 -1에 가까울수록 음의 상관관계
ρ의 값이 0에 가까울수록 선형관계는 약해짐

(ii) X와 Y에 대응되는 모든 값들이 한 직선상에 위치하면 ρ의 값은 -1 (직선기울기가 음인 경우) 또는 $+1$ (직선기울기가 양인 경우)의 값을 가짐

(iii) ρ는 두 변수의 선형관계만을 나타내는 측도
ρ=0인 경우 두 변수의 선형관계는 없지만 다른 관계를 가질 수 있음
표본상관계수 (sample correlation coefficient)

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

2.3.3 다중 변수의 통계표 및 통계량

가) **다차원 도수분포표(multi-dimensional table)**: 두 개 이상의 이산형 변수에 대한 교차표
예) 표 2.8: 변수 4개에 대한 4차원 도수분포표

나) 공분산행렬(covariance matrix), 상관계수행렬(correlation matrix)

- 변수 $X_1, X_2, \cdots, X_m$가 있을 때, 모든 두 변수 사이의 모집단 공분산행렬(covariance matrix)

$$\Sigma = \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_m) \\ Cov(X_2,X_1) & \ddots & & Cov(X_2,X_m) \\ \vdots & & \ddots & \vdots \\ Cov(X_m,X_1) & & \cdots & Cov(X_m,X_m) \end{bmatrix}$$

- 표본 평균벡터(sample mean vector): 확률벡터 $X = (X_1, \cdots, X_m)$의 n개의 표본 $x_1, \cdots, x_n$이 다음과 같다

$$x_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{m1} \end{bmatrix},\ x_2 = \begin{bmatrix} x_{12} \\ \vdots \\ x_{m2} \end{bmatrix},\ \cdots, x_n = \begin{bmatrix} x_{1n} \\ \vdots \\ x_{mn} \end{bmatrix}$$

$$\bar{x} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_m} \end{bmatrix},\ \bar{x}_i = \frac{\sum_{k=1}^{n} x_{ik}}{n}$$

$(m \times m)$ 표본 공분산행렬(sample covariance matrix)

$$S = \begin{bmatrix} s_1^2 & s_{12} \cdots s_{1m} \\ s_{21} & s_2^2 \cdots s_{2m} \\ \vdots & \ddots \\ s_{m1} & s_{m2} \cdots s_m^2 \end{bmatrix} \quad s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

$s_{ij} = s_{ji}$: S는 대칭행렬

- 표본 상관계수행렬(sample correlation matrix)

$$R = \begin{bmatrix} 1 & r_{12} \cdots r_{1m} \\ r_{21} & 1 \cdots r_{2m} \\ \vdots & \ddots \\ r_{m1} & r_{m2} \cdots 1 \end{bmatrix},\ r_{ij} = \frac{s_{ij}}{s_i s_j}$$

2.4 통계분포를 이용한 데이터 모형

▶ 표본공간(sample space, S): 통계적 실험의 모든 가능한 결과의 집합
▶ 사건(event): 표본공간의 부분집합
▶ 확률(probability): 사건 A가 발생할 가능성을 0과 1사이의 실수로 표시
▶ 확률변수(random variable, X): 표본공간의 각 원소에 하나의 실수값을 대응시켜 주는 함수
▶ 확률분포함수(probability distribution function): 확률 변수 X가 가질 수 있는 값 x에 대한 확률을 함수형태로 정리하여 놓은 것: $P(X=x)$, $f_X(x)$, $f(x)$로 표시

이산형 평균 $E(X) = \mu = \sum_{x \in S} x f(x)$ 연속형 $E(X) = \int x f(x) dx$

분산 $V(X) = \sigma^2 = \sum_{x \in S} (x-\mu)^2 f(x)$ $V(X) = \int (x-\mu)^2 f(x) dx$

2.4.1 한 변수의 통계분포

가) 이항분포 (Binomial distribution)

베르누이 시행: 표본공간의 원소가 2개인 통계적 실험.
예) 동전을 하나 던지는 실험의 경우: 결과는 '앞' '뒤'
공장에서 제품을 하나 만들 때: 결과는 '우량', '불량'
p =P(성공) 확률변수(X): $X=1$ or 0으로 표시

▷ 베르누이 분포 (Bernoulli distribution)

$f(x) = p^x (1-p)^{1-x},\ x = 0, 1$

$E(X) = p,\ V(X) = p(1-p)$

▷ 이항분포: 베르누이 시행을 n번 반복하여 '성공'이 나타나는 횟수에 대한 분포 함수

성공확률=p, 총 시행횟수=n, 성공 횟수=X

$f(x) = {}_nC_x p^x (1-p)^{n-x},\quad x = 0, 1, \cdots, n$

$E(X) = np,\quad V(X) = np(1-p)$

나) 정규분포 (Normal distribution)

- 연속형 데이터에 대해 데이터들이 평균 근처에 모여 있고, 평균에서 멀어질수록 데이터들의 수가 적어지며, 평균 중심으로 좌우 대칭

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ -\infty < x < \infty$$

$X \sim N(\mu, \sigma^2),\qquad E(X) = \mu,\quad V(X) = \sigma^2$

- 정규분포의 특징

(1) 종모양

(2) 평균 μ에 대해 대칭. $P(X \le \mu) = P(X \ge \mu) = 0.5$

(3) μ나 σ의 값에 따라 정규분포는 무한이 많이 있을 수 있다.

(4) $P(X \in [\mu-\sigma, \mu+\sigma]) = 0.683$

$P(X \in [\mu-2\sigma, \mu+2\sigma]) = 0.954$

$P(X \in [\mu-3\sigma, \mu+3\sigma]) = 0.997$

2.4.2 다중 변수의 통계분포

▶ 다변량 정규분포 (Multivariate normal distribution)

- m개의 변수에 대한 확률변수 $X = (X_1, X_2, \cdots, X_m)$

평균벡터 $\mu = (\mu_1, \mu_2, \cdots, \mu_m)$

$m \times m$ 공분산행렬 Σ

$$f(x) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)},\ -\infty < x_j < \infty,\ j = 1, 2, \cdots, m$$

$X \sim N(\mu, \Sigma)$

평균 $E(X) = \mu$, 공분산행렬 $V(X) = \Sigma$

2.4.3 분포함수(distribution function)의 추정

표본 $x_1, x_2, \cdots, x_n$

미지의 모집단 분포를 추정하고자 함

▶ 추정방법 1) 모수적 방법 (parametric estimation)

2) 비모수적 방법 (nonparametric estimation)

▶ 모수적 추정 방법

▷ 모집단의 분포함수 형태 $f(x)$를 가정

▷ 이 분포함수의 모수(parameter) θ를 표본 데이터를 이용하여 추정

▷ 대표적인 모수적 추정방법은 최대우도추정법 (maximum likelihood estimation: MLE)

▷ 확률변수들 $X_1, X_2, \cdots, X_n$이 서로 독립이고, 동일한 모수 $\theta = (\theta_1, \cdots, \theta_k)$를 갖는 확률분포 $f(x;\theta)$를 따른다고 가정하자.

$f(x_1, \cdots, x_n;\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta)$: 우도함수 (likelihood function)

- 우도함수는 모수 $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$에 대한 함수
- 우도함수가 최대가 되는 모수 $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$를 찾아서 분포함수를 추정하는 방법을 최대우도추정법이라 함
- 우도함수의 최대값은 $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$에 관해 미분하여 0으로 놓고 방정식을 품

$$\frac{\partial f(x_1, x_2, \cdots, x_n;\theta)}{\partial \theta} = 0 \quad \text{또는} \quad \frac{\partial \ln f(x_1, x_2, \cdots, x_n;\theta)}{\partial \theta} = 0$$

- 우도함수의 최대값을 $\hat{\theta}$라 할 때, 추정된 분포함수는 $f(x;\hat{\theta})$이다.

예제 2.4.1) 세 표본의 검사 결과: '우량', '불량', '우량'

모집단 분포: 베르누이 분포를 가정하고 최대우도추정법으로 분포함수 추정하여라.

=> 우량이면 1, 불량이면 0 => $P(\text{우량}) = p$, $P(\text{불량}) = 1-p$

=> 세 표본 $x_1 = 1,\ x_2 = 0,\ x_3 = 1$

우도함수:

$$f(x_1, x_2, x_3;p) = f(x_1;p)f(x_2;p)f(x_3;p) = p^{x_1}(1-p)^{(1-x_1)}p^{x_2}(1-p)^{(1-x_2)}p^{x_3}(1-p)^{(1-x_3)}$$

$$f(1,0,1;p) = p(1-p)p = p^2(1-p)$$

$$f'(1,0,1;p) = 2p(1-p) + p^2(-1) = 2p - 3p^2 = p(2-3p) = 0 \ \Rightarrow\ p = 0, p = \frac{2}{3}$$

$\rightarrow p = \frac{2}{3}$일 때 가능도함수가 최대가 됨

$$f(x) = \left(\frac{2}{3}\right)^x \left(1 - \frac{2}{3}\right)^{1-x},\ x = 0,1$$

일변수 표본

$x_1, x_2, \cdots, x_n \sim N(\mu, \sigma^2)$에 대해 최대우도추정법으로 구한 모수 추정량

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

다변량 표본데이터

$x_1, x_2, \cdots, x_n \sim N(\mu, \Sigma)$에 대해 최대우도추정법으로 구한 모수 추정량

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \hat{\Sigma}=\frac{(n-1)}{n}S \quad (\text{여기서 } S\text{는 표본의 공분산행렬})$$

2.5 데이터의 변환

수집된 데이터를 탐색한 후에는 데이터마이닝 기법을 적용하기에 적합한 데이터로 정제할 필요가 있음

: 전처리(preprocessing) 1) 데이터의 변환(transformation)

2) 차원의 축소(dimension reduction)

2.5.1 데이터의 변환 (transformation)

가) 수학적 변환

- 연속형 데이터의 경우 데이터마이닝 모형을 사용하기 위해 변환 이용

예) 여러 변수들의 단위가 다른 경우 같은 단위로 만들기 위해 [0-1] 변환이나 표준화변환 사용

▷ [0-1] 변환: 변수의 값 x를 0과 1사이의 값으로 변환하는 것.

▷ $y=\frac{x-\min}{\max-\min}$: 데이터의 최소값과 최대값 사이의 상대적인 위치를 의미

▷ 표준화변환(standardization)

: 원 데이터(raw data)를 평균 0이고 분산이 1인 데이터로 변환하여 주는 것

Z 변환: $Z=\frac{x-\bar{x}}{s}$, 여기서 $\bar{x}$: 평균, s : 표준편차

▷ 지수변환($\exp(x)$), 로그변환($\log(x)$), 역함수변환($1/x$), 제곱근변환($\sqrt{x}$)

박스-콕스 변환(Box-Cox transformation)

$$y=\begin{cases}\frac{x^p-1}{p}, & p\neq 0\\ \log(x), & p=0\end{cases} \quad (\text{모수 } p\text{는 주어진 값})$$

나) 집계 (aggregation)

- 대량의 데이터를 분석하는 경우, 수집된 데이터를 이용하는 것보다 어느 집단으로 집계하여 분석하는 것이 나음

예) 대형 슈퍼마켓의 경우, 모든 고객의 구매상품에 대한 데이터가 만들어짐

→ 그러나 일별 매출액(sales)의 합계, 상품별 일별 판매개수(sale)를 합계할 수 있음

→ 집계를 하면 데이터의 수를 줄이면서 필요한 데이터마이닝 분석을 용이하게 할 수 있음

→ 개별 데이터에서 나타내는 실제변화를 감지 못할 위험성

다) 이산형화 (discretization)

- 연관분석이나 분류분석모형 중에는 범주형 변수인 경우 적용할 수 있는 모형이 있음
- 연속형 변수를 범주형 변수로 변환하는 것

예) 나이변수를 10대(teenager), 20대(twenties), 30대(thirties)

라) 이항변수화 (binarization)

- 연관분석 모형에는 이항변수인 경우에만 적용할 수 있는 모형이 있음
- 연속형이나 두 개 이상의 값이 있는 범주형 변수를 여러 개의 이항변수 변환

예) 나이 '20대', '30대', '40대', '50대 이상' 네 범주

→ 네 개의 이항변수 생성 (표 2.11)

2.5.2 데이터 차원의 축소(Dimension Reduction)

▶ 데이터가 워낙 대량이어서 데이터 처리 시간이 오래 걸릴 수 있음

▶ 변수의 수가 너무 많아 모형을 적용할 수 없는 경우

→ 이러한 경우 적절한 데이터 수의 표본추출, 불필요한 변수 버리고 주성분 분석을 이용하여 변수 수를 축소하는 방법 고려

가) 데이터 수의 축소 - 표본추출(sampling)

▷ 데이터의 크기가 워낙 커서 데이터 처리의 시간과 경비를 줄이기 위해 전체 데이터의 일부 데이터를 표본추출(sampling)하여 분석함

▷ 추출한 일부를 이용하여 전체 데이터의 경향을 파악

- 좋은 표본추출 방법이 필요
- 전체 데이터를 충분히 잘 대표하고 비슷한 특성을 지닌 표본을 추출

▷ 단순 확률 추출법(simple random sampling)

- 모집단의 모든 데이터가 표본으로 뽑힐 확률이 같도록 표본을 추출하는 방법

  i) 복원추출(with replacement): 한번 추출된 데이터를 다시 모집단에 포함시킴

  ii) 비복원추출(without replacement): 한번 추출된 데이터를 다시 모집단에 넣지 않음

- 모집단의 데이터가 표본으로 뽑힐 확률이 같도록 하기 위해 난수표(random number table) 이용

▷ 층화추출법(stratified sampling)

- 모집단을 적당한 개수의 동질적인 층(strata)으로 분할하여 각 층에서 정해진 크기의 표본을 추출하는 방법

  예) 남자와 여자로 구성된 모집단에서 평균 임금에 관심이 있음

  모집단을 남녀 층으로 나눔: 남 → 단순 확률 추출로 표본을 뽑음

  여 → 단순 확률 추출로 표본을 뽑음

  → 전체 평균 임금 추정

- 각 층 내의 데이터들이 동질적이 되도록 함
- 층끼리의 데이터들이 이질적이 되도록 함
- 장점: 추정값의 분산, 추출비용도 감소

나) 변수 수의 축소 - 주성분분석(principle component analysis)

▷ 변수들의 개수가 많은 경우에 변수들의 상관관계를 이용하여 적은 수의 새로운 변수를 만들어 이용하면 분석이 간단해짐

▷ 주성분분석: 변수들의 공분산행렬이나 상관행렬을 이용하여 변수들의 선형결합으로 적은 차원의 새로운 변수를 찾아냄

확률벡터 $X=(X_1, X_2, \cdots, X_m)$가 평균벡터 μ, 공분산행렬 Σ 가진다고 가정하자.

$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{bmatrix}$: m개의 고유값(eigen value)을 내림차순으로 $\lambda_1, \lambda_2, \cdots, \lambda_m$라 하고, 각 고유값에 대응되는 고유벡터(eigen vector)를 $e_1, \cdots, e_m$이라 하자.

$E=[e_1, e_2, \cdots, e_m]$ (m개의 고유벡터를 열로 가지는 $m \times m$ 행렬)

선형결합 $Y=EX$는 새로운 m개의 변수 $Y_1, Y_2, \cdots, Y_m$ 생성함 (이를 주성분이라 함)

$=[e_1, e_2, \cdots, e_m](X_1, X_2, \cdots, X_m)$

주성분변수 $Y=(Y_1, Y_2, \cdots, Y_m)$의 공분산행렬은 다음과 같다

$$Cov(Y) = Cov(EX) = E' Cov(X) E = E' \Sigma E = \Lambda$$

Σ의 고유값 $\lambda_1, \cdots, \lambda_m$를 대각원소로 하는 대각행렬 $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_m \end{bmatrix}$

위의 식에서

- 주성분 Y_1(첫번째 주성분), $Y_2, \cdots, Y_m$(마지막 주성분)은 서로 독립임을 알 수 있음

  $Var(Y_j)=\lambda_j$임을 알 수 있음

  고유값들이 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$로 가정했으므로

  $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_m)$임을 알 수 있음
- Σ의 대각원소들의 합 = 고유값들의 합과 같음

  $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_m^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_m$
- 고유값은 크기순으로 정렬되어 있기 때문에 주성분의 분산 몇 개가 전체 변수들의 분산합의 많은 부분을 설명할 수도 있음을 의미함

  $\therefore$ 소수 k개의 주성분 $Y_1, Y_2, \cdots, Y_k$의 분산만으로 전체 m개의 X_i의 분산을 잘 설명할 수 있다는 뜻
- 상관행렬을 이용하는 경우가 많다

  X_i 변수들을 표준화하였다면 $X=(X_1, X_2, \cdots, X_m)$의 공분산행렬은 상관행렬이 되고,

  주성분 $Y_1, Y_2, \cdots, Y_m$의 분산의 합은 m이 됨

  이 경우 주성분의 분산이 1보다 크면 이 주성분의 공헌도가 평균 1보다 크다는 것을 알 수 있음

  전체 $X_1, X_2, \cdots, X_m$개의 변수 대신 공헌도가 1보다 큰 적은 수의 주성분 $Y_1, Y_2, \cdots, Y_k$로 차원을 축소하여 분석함

다) 데이터의 압축 - 웨이블렛 변환(wavelet transformation)

▶ 대량의 원시데이터를 적절한 압축변환을 하여 적은 공간에 저장하였다가 필요하면 다시 원 데이터를 재생하는 것을 의미

▶ 많이 이용되는 데이터 압축방법은 푸리에 변환(Fourier transformation)과 웨이블렛 변환(wavelet transformation)

웨이블렛 변환: 압축시 정보 손실이 적어 더 많이 이용됨.

2.6 데이터의 유사성 측도

- ▶ 데이터와 데이터 사이의 유사성(similarity) 또는 비유사성(dissimilarity)을 조사하여 데이터들이 서로 유사한지 살펴봄
  비유사성: 거리개념(distance)이다
- ▶ 군집분석은 데이터의 유사성/비유사성에 근거하여 비슷한 데이터를 군집화하는 방법