경영통계

원광대학교 경영학부

담당교수: 정호일

제3장 기술통계학: 분포의특성

- 중심경향치
- 산포도
- Chebyshev의 정리
- 변동계수
- 상대위치의 측정치
- 형태의 측정치

1. 중심경향치

중심경향치: 자료분포의 중심이 되어 전체자료를 대표하는 값

◆ 산술평균(arithmetic mean)

모든 자료의 합을 자료의 수로 나눈 값

모평균 : $$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

표본평균 : $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

x_i : i번째 관측치, N : 모집단 크기, n : 표본크기

◆중앙치(Median:Md)

- 자료를 크기 순으로 나열하여 중간에 위치한 관측치.

중앙치를 구하는 절차

- 자료를 크기 순으로 배열
- 자료의 수(n)가 홀수이면 $\frac{(n+1)}{2}$ 번째 자료값이 중앙치
- 자료의 수(n)가 짝수이면 $\frac{n}{2}$ 번째와 $(\frac{n}{2}+1)$ 번째 자료값을 평균한 값

◆최빈치(Mode: Mo)

- 자료 중에서 빈도가 가장 많은 관측치
- 질적 자료와 양적 자료에 모두 사용가능하며 주로 질적 자료의 대표값을 구하는데 사용
- 도수가 모두 같은 자료는 최빈치를 갖지 않음
- 동시에 두 개 이상의 최빈치가 있을 수 있음.
- 최빈치가 두 개이면 쌍봉(bimodal), 세 개 이상이면 다봉(multimodal)이라 한다.

◆ 통계분석의 대표치로서 평균이 자주 사용되는 이유

- 수학적 연산이 가능하다.
- 평균은 모든 자료의 영향을 고려한다.
- 가중평균을 계산할 수 있다.
- 분산의 계산과 모수의 추정 외에 가설검증 등 통계분석의 대표치로서 가장 널리 사용

• 평균,중앙치,최빈치의 비교

1.평균이 신뢰성 있는 대표치로 선호되고 특히 모집단의 모수를 추정하는데 이용된다.

2.평균은 수학적 연산이 가능하지만 중앙치와 최빈치는 불가능

3.자료 속에 극단적인 이상치(outlier)가 있는 경우 극단적인 관측치에 덜 민감한 중앙치가 대표치로 사용될 수 있다.

4.개방구간을 갖는 도수분포표의 경우 중앙치 또는 최빈치가 사용된다.

5. 특히 명목자료와 서열자료에 대해서는 평균과 중앙치를 계산할 수 없으므로 최빈치를 대표치로 사용된다.

2. 산포도

자료분포의 특성을 분석할 때에는 집중경향치와 동시에 산포도를 고려할 필요가 있다.

- 산포도 또는 분산도(measure of dispersion)는 자료들의 흩어진 정도를 측정한다.

- 산포도는 두 분포에서 자료의 흩어짐을 비교하는데 이용된다.

- 분산도를 측정하는 요약특성치로는 범위, 중간범위, 평균절대편차, 분산,
  표준편차, 변동계수 등이 있다.

범위(Range)

자료에서 최대치와 최소치의 차이를 말한다.

– 간단히 구할 수 있지만 두 개의 극단적인 값만을 고려하기 때문에 다른 값들에 대해서는 아무 것도 말해주지 않는다.

중간범위

범위의 문제점을 극복하기 위해서 고안된 것으로서 자료의 중간부분만 고려하여 구한다. 자료의 중간 50%인 1사분위수와 3사분위수의 차이 혹은 중간 80%인 10번째 백분위수와 90번째 백분위수의 차이로 구한다.

평균절대편차 (mean absolute deviation: MAD)

편차의 절대값을 모든 자료에 대해 구한 총절대편차를 평균하여 구한다.

*편차 (Deviation): 관측값과 평균과의 차이

모집단 : $$MAD = \frac{\sum_{i=1}^{N} |X_i - \mu|}{N}$$

표 본 : $$MAD = \frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n}$$

•분산과 표준편차

분산(Variance)

모든 편차 제곱의 합을 자료의 총수로 나눈 것

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

표본분산 $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

- 분산의 특징

1. 분산은 주어진 자료가 평균 주위로 얼마나 집중되어 있는가를 측정한다. 분산의 값이 작으면 자료의 변동이 심하지 않으며 대체로 평균 가까이에 분포하고 있음을 의미한다.(변동성, 이질성, 위험의 척도)

2. 표본분산은 모분산을 구하고자 할 때 추정치로서 사용된다.

3. 표본분산의 공식에서 (n-1) 대신에 n을 사용하여 편차제곱승의 평균을 구하면 모분산을 과소평가하게 되어 편의추정치(biased estimate)를 제공하게 된다. 즉 (n-1)을 사용하여 구한 표본분산은 모분산의 불평추정치 (unbiased estimate) 가 된다.

4. 분산은 각 자료에 대한 편차제곱으로 구하기 때문에 원자료의 단위가 왜곡되며, 편차정보가 부풀려 진다.

표준편차 (standard deviation)

분산의 제곱근이다.

분산의 제곱근인 표준편차를 구하면 원래 자료의 단위로 환원되어 같은 단위로 측정된 평균이나 다른 통계량과 쉽게 비교할 수 있는 이점을 갖는다. 따라서 산포도는 보통 표준편차로 측정하게 된다.

모집단 $\sigma = \sqrt{\sigma^2}$

표 본 $S = \sqrt{S^2}$

3.3 Chebyshev의 정리

자료의 분포가 정규분포가 아니거나 또는 이를 모르는 경우 체비셰프의 정리(Chebyshev's theorem)를 이용하여 일정구간에서 자료가 속할 가능성의 크기(비율)를 알아내는데 활용될 수 있다.

체비셰프의 정리

어떤 자료에 있어서 평균+k 표준편차 내에 존재할 자료의 비율은 적어도

$$P(\text{평균} - k\text{표준편차} \le x \le \text{평균} + k\text{표준편차}) = 1 - \frac{1}{k^2}$$

3.4 변동계수

두 자료군의 상대적 변동을 측정하는데 이용되는 기법이 변동계수(coefficient of variaton : CV)이다. 상대적 표준편차라고도 한다.

변동계수

모집단 : $CV = \frac{\text{표준편차}(\sigma)}{\text{평균}(\mu)} \times 100\%$

표 본 : $CV = \frac{\text{표준편차}(S)}{\text{평균}(\overline{X})} \times 100\%$

두 자료군 사이의 측정단위가 다르거나 평균에 있어 큰 차이가 있을 경우 표준편차를 비교하는데 무리가 따르는데 이러한 경우 변동계수를 이용하여 비교하여야 한다.

3.5 상대위치의 측정치

- 백분위수

백분위수란 자료를 크기 순으로 정리하여 백등분 하였을 때 각 등분점에 위치하는 자료를 말한다.

P번째 백분위수를 계산하는 절차
- 자료를 작은 것부터 큰 순서로 정렬한다.
- 지수 i를 다음과 같이 계산한다.

$$i = n\left(\frac{P}{100}\right)$$

단, P : 관심있는 백분위수, n : 자료의 수

-만일 i가 정수이면 l와 $(l+1)$의 위치에 있는 자료를 평균한 값이 P번째 백분위수이다. 만일 i가 정수가 아니면 i보다 큰 가장 가까운 정수가 P번째 백분위수이다.

•사분위수

100분위수 중 25번째 백분위수를 1사분위수 Q_1, 50번째 백분위수를 2사분위수 또는 중앙치 Q_2, 75번째 백분위수를 3사분위수 Q_3 이라고 한다.

– 사분위수범위 : 1사분위수와 3사분위수의 차이

•Z값

Z값(Z score, Z value)이란 백분위수처럼 특정 관측치가 평균의 위 아래로부터 몇 개의 표준편차만큼 떨어져 있는가를 상대적으로 나타내는 상대적 위치를 결정한다.

Z값

모집단 : $Z = \frac{x_i - \mu}{\sigma}$

표 본 : $Z = \frac{x_i - \bar{x}}{s}$

3.6 형태의 측정치

비대칭도(왜도)

자료분포의 모양을 측정하는 비대칭도 (skewness)는 자료분포의 좌우대칭정도를 측정한다.

- 좌우대칭
- 오른쪽 꼬리분포
- 왼쪽꼬리분포

비대칭도계수

비대칭도를 결정하는 한 방법은 Pearson의 비대칭도 계수(Pearson's cofficient of skewness)가 있다.

$$Sk = \frac{3(\overline{X} - Md)}{S} \qquad (\text{단}, -3 \le Sk \le 3)$$

- $Sk = 0$인 경우 $\overline{X} = Md = Mo$: 분포는 좌우대칭
- $Sk > 0$인 경우 $\overline{X} \ge Md \ge Mo$: 분포는 오른쪽으로 긴꼬리
- $Sk < 0$인 경우 $\overline{X} \le Md \le Mo$: 분포는 왼쪽으로 긴 꼬리

첨도

자료분포의 뾰족함(peakedness)의 정도를 측정하는 것이 첨도(kurtosis)이다.

첨도가 크면 뾰족한 봉우리를 가지고 작으면 평평한 봉우리를 갖는다.