제3장

수치를 이용한 기술통계학 기법

NUMERICAL DESCRIPTIVE TECHNIQUES

수치를 이용한 기술통계학 기법…

중심위치의 척도

- 평균(Mean), 중앙값(Median), 최빈값(Mode)

변동성의 척도

- 범위(Range), 표준편차(Standard Deviation), 분산(Variance), 변동계수(Coefficient of Variation)

상대위치의 척도

- 백분위수(Percentiles), (사분위수)Quartiles

선형관계의 척도

- 공분산(Covariance), 상관계수(Correlation Coefficient), 결정계수(Coefficient of Determination), 최소자승선(Least Squares Line)

중심위치의 척도

-산술평균(arithmetic mean) 또는 평균(mean) 은 가장 널리 사용되는 유용한 중심위치의 척도이다.

-산술평균은 모든 관측치들을 합하고 관측치의 수로 나누어서 계산된다.

$$\text{평균} = \frac{\text{모든 관측치의 합}}{\text{관측치의 수}}$$

기호

N = 모집단에 속한 관측치의 수

n = 표본에 속한 관측치의 수

μ = 모평균(모집단의 산술평균) “mu”

$\bar{x}$ = 표본평균(표본의 산술평균) “x-bar”

산술평균(Arithmetic Mean)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

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

모평균(population mean)

표본평균(Sample Mean)

산술평균

-분포의 집중도를 나타내는 중심개념을 말하는데 간단히 말해 평균이라고 한다.

-산술평균은 측정데이터 (예: 키, 점수, 등)의 중심위치를 나타내는데 적정한 척도이다.

-산술평균은 “이상치(outliers)”라고 부르는 극단값들에 의해 크게 영향을 받는다.

예: 억만장자가 이웃으로 이사오면 평균가계소득이 크게 증가한다…

산술평균

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_n}{N} = \frac{\sum_{i=1}^{N} X_1}{N}$$ 모평균

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$ 표본평균

산술평균의 예

1. 경영통계학과 학생의 성적이 다음과 같다면 평균성적은 얼마인가.

89,55,45,62,75,90,67,82,72,81

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$= \frac{89+55+45+62+75+90+67+82+72+81}{10}$$

$$71.8$$

2. 어느 지역 가구들의 월 평균 소비전력을 측정하기 위하여 8가구를 표본으로 뽑아 조사했더니 다음과 같았다. 표본의 평균값은 얼마인가

120,184,220,85,68,146,162,95

도수분포표에서 평균값을 구하라.

계급	중간점	도수(fi)
10-14	12	2
15-19	17	4
20-24	22	7
25-29	27	13
30-34	32	3
		$n = 29 \left(= \sum f_i \right)$

도수분포표에서 산술평균 식

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{n} = \frac{\sum f_i x_i}{n}$$

$$\bar{X} = \frac{12 \times 2 + 17 \times 4 + 22 \times 7 + 27 \times 13 + 32 \times 3}{29}$$

$$= \frac{693}{29}$$

$$= 23.9$$

중앙값

-중앙값(median)은 모든 관측치를 순서대로 정렬할 때 중심에 있는 관측치를 의미한다.

-중앙값은 숫자로 표시되는 양적자료에만 사용.

-중앙값은 수치로 된 자료를 크기 순서로 나열할 때 가장 가운데에 위치하는 관찰값을 의미.

-전체자료를 크기 순서로 나열할 때 중앙에 위치하는 값

중앙값의 계산방법

$$\frac{N+1}{2}$$ N은 관찰수

조선대학교 생협에서 판매하는 자동차 판매원 9명의 월간판매량을 크기 순서에 따라 정리한 결과이다.

22,24,24,25,27,30,31,35,40 일 때 중앙의 위치는

$\frac{9+1}{2}$ =5 , 다섯번째 위치한 자료 27 이 중앙값

만약 판매원 10명 22,24,24,25,27,30,31,35,40,42 이면

$\frac{10+1}{2}=5.5$ 이므로 중앙값은 5번째와 6번째 사이.
따라서 28.5이 중앙값이다.

최빈값

-관측치들의 최빈값(mode)은 발생되는 빈도수가 가장 많은 관측치이다.

-한 세트의 데이터에는 최빈값이 하나 또는 둘 이상이 존재할 수 있다.

-최빈값은 주로 명목데이터의 경우에 사용되지만 모든 데이터 유형에 대하여 유용한 중심위치의 척도이다.

-대규모 데이터 세트의 경우 최빈계급구간(modal class)가 단일 값을 가지는 최빈값보다 더 유용하다.

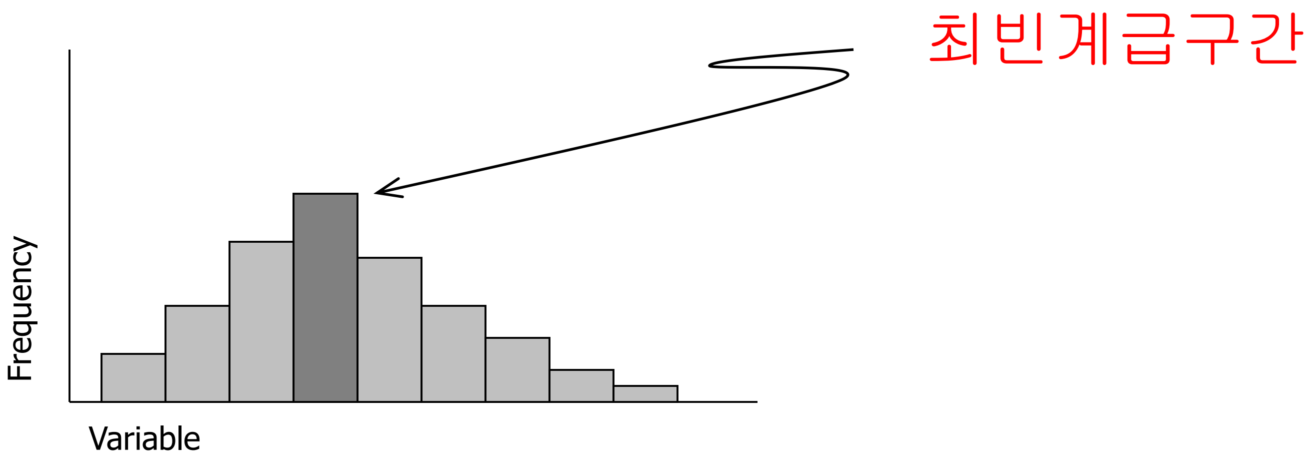
최빈값

예: 데이터: {0, 7, 12, 5, 14, 8, 0, 9, 22, 33} N=10

-어느 관측치가 가장 많이 나타나는가?

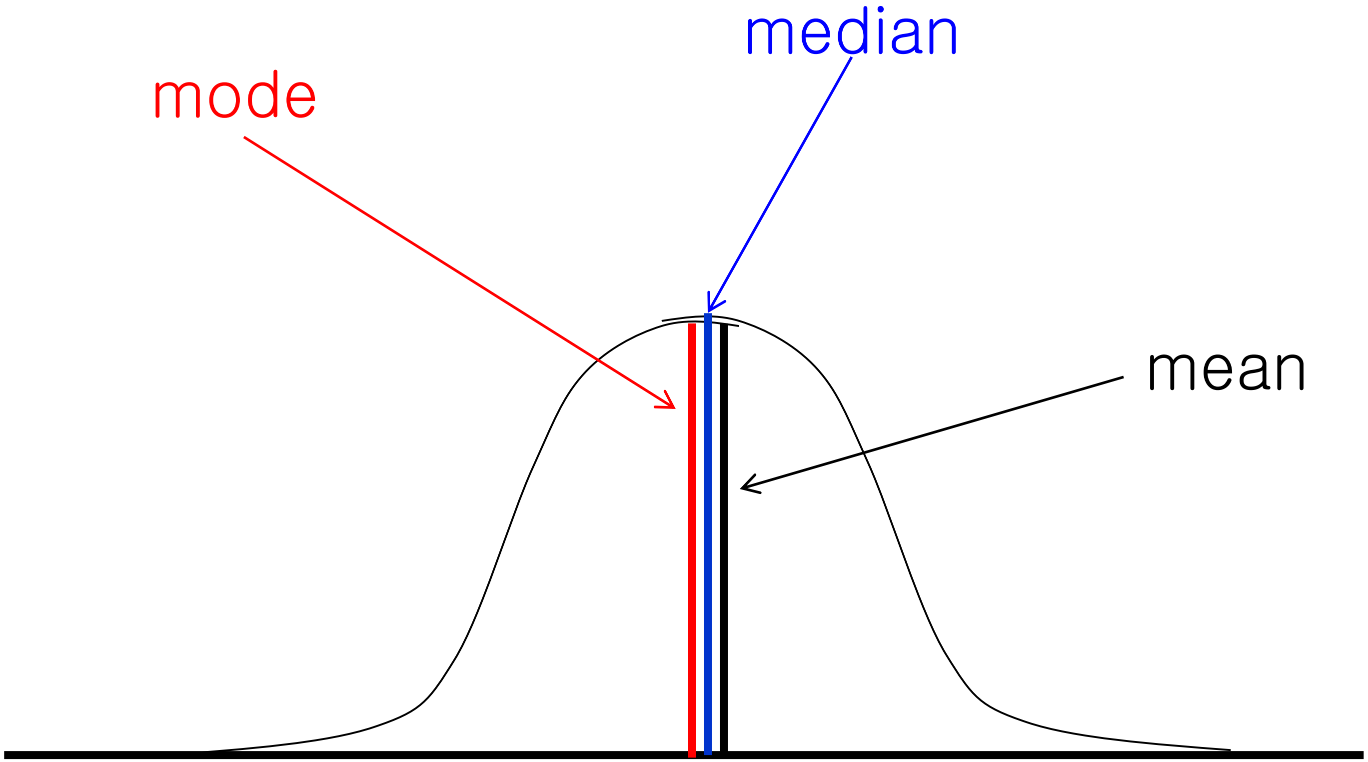
-이 데이터 세트의 최빈값은 0 이다.

-이와 같은 최빈값은 어떻게 중심위치의 척도가 되는가?



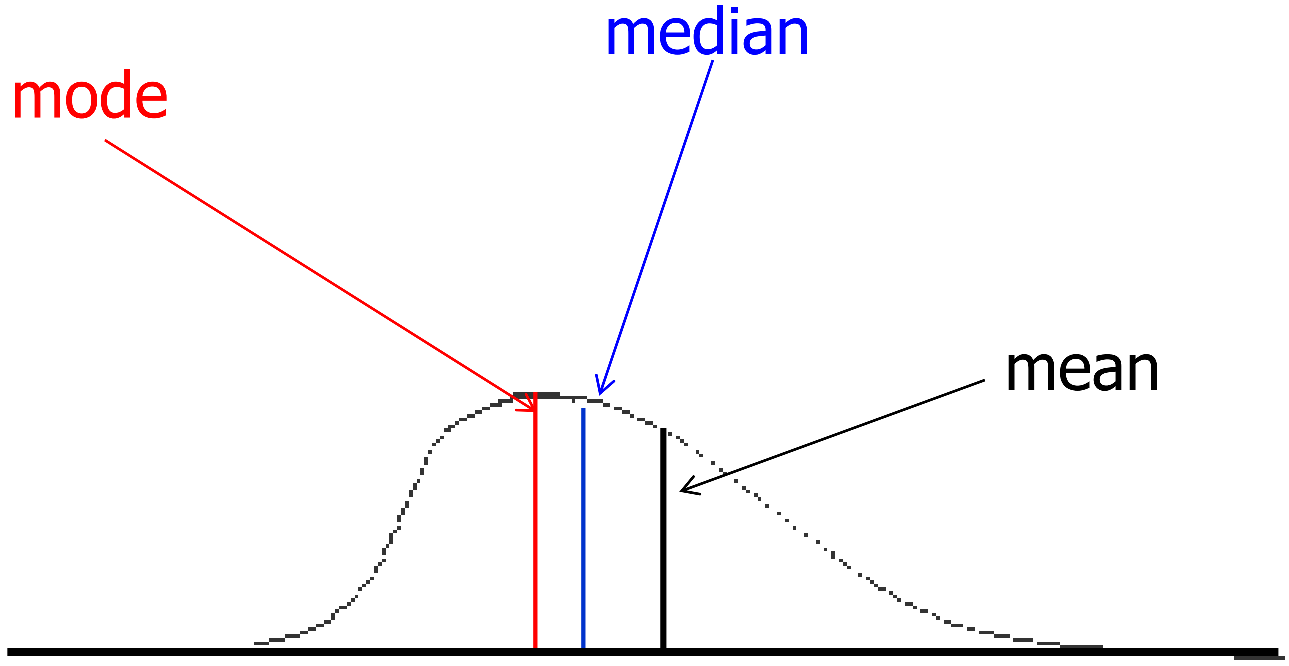
평균(Mean),중앙값(Median),최빈값(Mode)

-만일 변수의 분포가 대칭이면, 평균, 중앙값, 최빈값은 모두 동일할 수 있다…



평균(Mean),중앙값(Median),최빈값(Mode)

- 만일 변수의 분포가 비대칭이면, 즉 왼쪽으로 기울져 있거나 또는 오른쪽으로 기울어져 있으면, 평균, 중앙값, 최빈값은 서로 다를 수 있다.



평균, 중앙값, 최빈값 중에서 어느 것이 가장 좋은 중심위치의 척도인가?

- 평균은 일반적으로 가장 널리 사용되는 유용한 중심위치의 척도이다. 그러나 중앙값이 더 좋은 중심위치의 척도인 상황들이 존재한다.

- 최빈값은 결코 가장 좋은 중심위치의 척도는 아닌다.

- 중앙값이 가지고 있는 한가지 장점은 평균과는 달리 극단값들에 대하여 민감하지 않다는 점이다.

평균, 중앙값, 최빈값 중에서 어느 것이 가장 좋은 중심위치의 척도인가?

-예제 3.1 인터넷의 평균사용시간을 살펴보자.

-평균은 11.0이고 중앙값은 8.5이다.

-이제 33시간을 보고한 응답자가 실제로 133시간을 보고하였다고 하자. 이 경우 평균은

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+7+12+5+133+14+8+0+22}{10} = \frac{210}{10} = 21.0$$

평균, 중앙값, 최빈값 중에서 어느 것이 가장 좋은 중심위치의 척도인가?

-표본에는 평균(21)보다 큰 관측치들은 두개 존재한다. 이와 같이 극단값의 존재는 평균이 중심위치의 척도가 되지 못하게 만든다.

-그러나 중앙값은 극단값에 관계없이 동일하다. 상대적으로 적은 수의 극단값들이 존재할 때 중앙값은 일반적으로 데이터의 중심을 나타내는 더 양호한 척도가 된다.

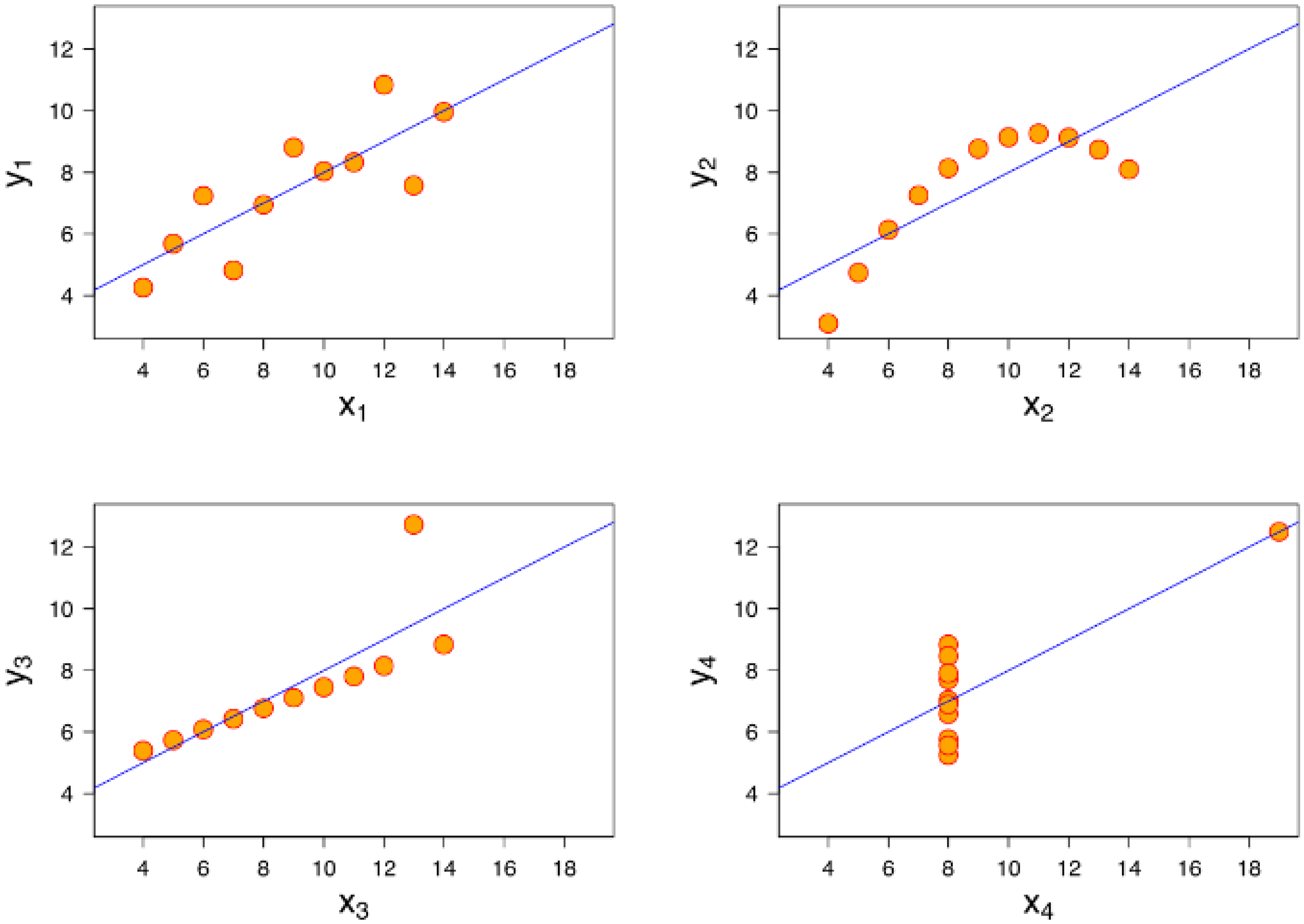
서열 및 범주데이터의 평균,중앙값, 최빈값

-서열데이터와 범주데이터의 경우 평균의 계산은 의미가 없다.

-서열데이터의 경우 중앙값은 중심위치의 척도가 된다.

-범주데이터의 경우 최빈값은 유용한 빈도 척도이나 “중심위치”의 척도는 아니다.

ANSCOMBE'S QUARTET



기하평균(Geometric Mean)

-산술평균은 가장 널리 사용되는 유용한 중심위치의 척도이다.

-그러나, 산술평균이나 중앙값이 최선의 중심척도가 아닌 상황이 존재한다.

-변수가 성장률 또는 변화율일 때, 기하평균이 유용한 중심위치 척도가 될 수 있다.

기하평균(Geometric Mean)

-<예시>당신이 $1,000를 2년간 투자한다고 하자.
첫째 해에 투자가치가 100% 증가하여 $2,000가 되고 두번째 해에 투가가치가 -50% 감소하여(손실발생) 다시 $1,000가 된다고 하자.

-연도 1과 연도 2의 수익률은 각각 $R_1 = 100\%$과 $R_2 = -50\%$이다. 두 연도 수익률의 산술평균(과 중앙값)은 다음과 같이 계산된다.

$$\overline{R} = \frac{R_1 + R_2}{2} = \frac{100 + (-50)}{2} = 25\%$$

기하평균(Geometric Mean)

-그러나 이 수치는 오도적이다. 투자가 이루어지는 2년 동안 투자가치는 변화가 없기 때문에, “평균”복리수익률은 0%이다.

-이와 같은 “평균”복리수익률은 기하평균의 값이다.

기하평균(Geometric Mean)

- R_i 는 기간 i의 수익률 (소수점으로 표시한 수익률)이라고 하자 (i = 1, 2, ⋯, n). 수익률들의 기하평균(geometric mean)R_g 는 다음과 같이 정의된다.

$$(1+R_g)^n = (1+R_1)(1+R_2)...(1+R_n)$$

- R_g 에 대하여 풀면,

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)...(1+R_n)} - 1$$

기하평균(Geometric Mean)

-따라서 주어진 예에서 투자수익률의 기하평균은

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)\dots(1+R_n)} - 1$$

$$= \sqrt[2]{(1+1)(1+[-.50])} - 1 = 1 - 1 = 0$$

-따라서 투자수익률의 기하평균은 0%이다. 따라서 0%의 복리이자율 공식을 사용하면

투자기간 말의 투자가치 = $1{,}000(1 + R_g)^2$

$= 1{,}000(1 + 0)^2 = 1{,}000$

변동성의 척도

-관측치들이 평균 주위에서 얼마나 흩어져 있는가를 측정하는 척도가 변동성의 척도이다.

예를 들면, 두 과목의 점수들이 주어져 있다고 하자. 평균은 두 과목 모두 50으로 같다…

그러나 붉은색으로 나타낸 과목의 점수가 파란색으로 나타낸 과목의 점수보다 변동성이 더 크다 (평균 주위에서 더 많이 흩어져 있다).