4. 표본 데이터의 요약 (2)

수치적 기술통계

- **중심 위치 척도**

  표본 자료 $x_1, \dots, x_n$ 이 주어졌을 때,

  - **표본 평균(Sample Mean) – 산술평균 (Arithmetic Mean)**

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

    - 가장 대표적인 중심위치 척도
    - 이상치(outlier)에 민감함.

  - **표본 중위수, 중앙값 (Sample Median)**
    - 표본 자료를 오름차순 정렬한 자료 $x_{(1)} \leq \cdots \leq x_{(n)}$ 에 대하여,

      $$x_{med} = x_{((n+1)/2)}$$

    - 이상치(outlier)에 민감하지 않음.

  - **표본 최빈값 (Sample Mode)**
    - 가장 빈도가 높은 값 또는 구간

수치적 기술통계

- 예제
  - 8 개의 금융기관 별로 어느 달에 기업고객 중 부도업체 수를 다음과 같이 정리하였다. 표본평균과 표본 중위수를 구하여라.

    2, 8, 3, 5, 6, 9, 4, 1

수치적 기술통계

- **중심 위치 척도**

  표본 자료 $x_1, \dots, x_n$ 이 주어졌을 때,

  - **기하평균(Geometric Mean)**

$$G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (x_1 \cdot x_2 \cdots x_n)^{1/n}$$

    - 증가율, 성장률, 변화율 등의 배수 자료에 대한 중심값으로 활용됨.
    - 모든 자료의 값이 양수여야 계산할 수 있음.

수치적 기술통계

- 예제
  - 다음은 어느 기업의 최근 5년간 주당순이익(EPS) 자료이다.

사업연도	제 45 기	제 46 기	제 47 기	제 48 기	제 49 기
EPS	1078	7369	4815	2179	1981

이 기업의 5년간 연평균 EPS 성장률은 얼마인지 구하여라.

수치적 기술통계

- **상대적 위치 척도**
  - **백분위수, 퍼센타일(Percentile)**

    $x_{(L_p)}$, 단, $L_p = \frac{p}{100}(n+1)$임.

    - 전체 자료의 $p\%$ 는 p 백분위수보다 작은 것으로 해석

  - **사분위수 (Quartile)** : Q1, Q2, Q3
    - **Q1**$\left(= x_{((n+1)/4)}\right)$: 25퍼센타일, 1사분위수
    - **Q2**$\left(= x_{((n+1)/2)}\right)$: 중위수, 50 퍼센타일, 2사분위수
    - **Q3**$\left(= x_{(3(n+1)/4)}\right)$: 75 퍼센타일, 3사분위수

수치적 기술통계

- 예제
  - 8 개의 금융기관 별로 어느 달에 기업고객 중 부도업체 수를 다음과 같이 정리하였다. `사분위수 Q1과 Q3를 구하여라.

    2, 8, 3, 5, 6, 9, 4, 1

수치적 기술통계

- **변동성 척도**

  표본 자료 $x_1, \ldots, x_n$ 이 주어졌을 때,

  - **범위 (Range)**

$$max(x_i) - min(x_i)$$

    - 이상치에 민감함.

  - **사분위간 범위 (IQR, Inter Quartile Range)**

$$Q3 - Q1$$

    - 가운데 50%에 해당하는 자료의 범위로 이상치에 민감하지 않음.

수치적 기술통계

- **표본 분산 (Sample Variance)**

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

  - 가장 대표적인 변동성 척도
  - 모든 자료값을 반영하여 변동성을 측정함.
  - 음의 값을 가질 수 없음.
  - 값이 클수록 자료값의 변동성이 크다는 것을 의미함.
  - 자료 값의 단위가 제곱됨.

수치적 기술통계

- **표본 표준편차 (Sample Standard Deviation)**

$$s = \sqrt{s^2}$$

  - 자료 값의 단위와 표본 표준편차의 단위는 동일하기 때문에 해석이 용이함.

- **변동계수 (Coefficient of Variation)**

$$cv = s/\bar{x}$$

  - 단위가 다르거나, 평균 차이가 큰 여러 변수의 변동성 비교에 활용됨.

수치적 기술통계

- 예제
  - 어느 금융기관 여신담당자로부터 9개 기업의 신뢰도를 설문조사한 결과 다음과 같은 점수를 얻었다. 이 자료에 대한 표본 표준편차를 구하고 이를 해석하여라.

    4, 9, 2, 5, 6, 9, 8, 6, 5

수치적 기술통계

- 예제
  - A 사와 B 사에 대하여 주당수익률 자료를 조사한 결과 A사의 주당수익률은 평균이 2000원, 표준편차는 300원이었고, B사는 평균이 5000원, 표준편차는 600원이었다고 한다. 어느 회사의 주당수익률이 더 안정적이라고 말할 수 있는가?

수치적 기술통계

- **통신요금 예제**
  - 통신요금 데이터(telephone_bills.csv)을 이용하여 월 청구액에 대한 다음의 기술통계량 값을 구하여라.
    - 표본 평균, 표본 중위수, 범위, 표본 분산, 표본 표준편차, 변동계수.
    - Q1, Q3, IQR
    - 30th 퍼센타일, 80th 퍼센타일

```
> longdist <- read.csv('telephone_bills.csv')
> head( longdist )
   Bills
1  42.19
2  38.45
3  29.23
4  89.35
5 118.04
6 110.46
```

수치적 기술통계

```
> bills <- longdist$Bills
> mean( bills )
[1] 43.5876
> median( bills )
[1] 26.905
> range( bills )
[1]   0.00 119.63
> diff( range( bills ) )
[1] 119.63
> var( bills )
[1] 1518.638
> sd( bills )
[1] 38.96971
> sd( bills )/mean( bills )
[1] 0.8940549
```

수치적 기술통계

```
> quantile( bills )
      0%       25%       50%       75%      100% 
  0.0000    9.3850   26.9050   84.8275  119.6300 
> quantile( bills , prob=c(0.3, 0.8))
   30%    80% 
11.529 90.428 
```

수치적 기술통계

- **선형적 연관성 척도**

  표본 자료 $(x_1, y_1), \dots, (x_n, y_n)$이 주어졌을 때,

  - **표본 공분산 (Sample Covariance)**

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

◆ 선형관계의 방향

- $s_{xy} > 0$: 양의 선형 관계, 비례관계
- $s_{xy} < 0$: 음의 선형 관계, 반비례관계

◆ 선형관계의 강도

- $-s_x s_y \le s_{xy} \le s_x s_y$ by Cauchy-Schwarz 부등식

◆ 표본 공분산은 x와 y 의 측정 단위에 의존하는 지표임.

- $x' = ax + b$ 이고 $y' = cy + d$인 경우에, $s_{x'y'} = ac \cdot s_{xy}$

수치적 기술통계

- **표본 상관계수 (Sample Correlation)**

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- $-1 \leq r_{xy} \leq 1$, by Cauchy-Schwarz 부등식

◆ 선형관계의 방향

- $r_{xy} > 0$: 양의 선형관계
- $r_{xy} < 0$: 음의 선형관계

◆ 선형관계의 강도

- $|r_{xy}| \approx 0$: 강도가 약함.
- $|r_{xy}| \approx 1$: 강도가 강함.

◆ 표본 상관계수는 x와 y 의 측정 단위에 의존하지 않음.

- $x' = ax + b,\ y' = cy + d$이고, $ac > 0$ 인 경우에, $r_{x'y'} = r_{xy}$

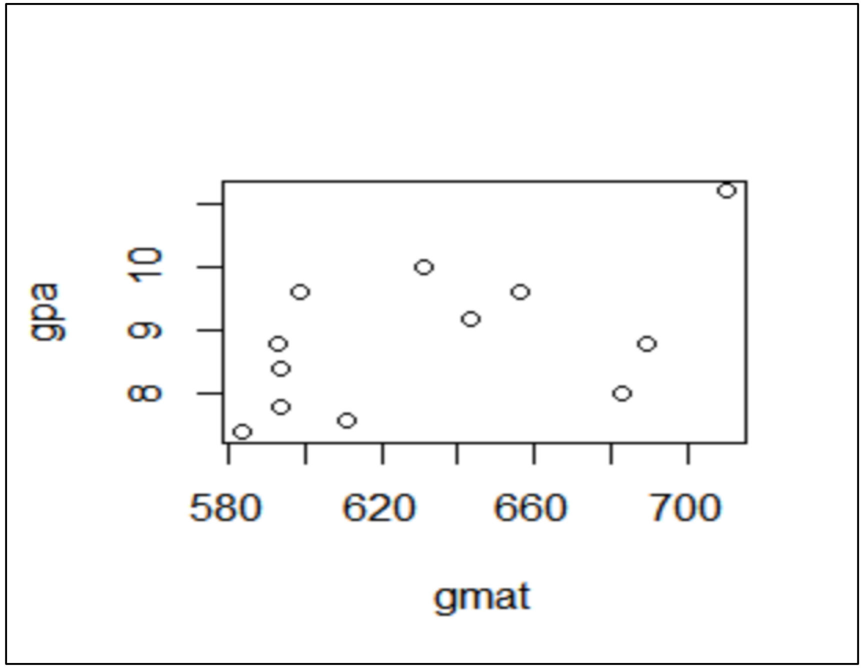
수치적 기술통계

- **MBA 학생들의 GMAT과 GPA 예제**
  - 어느 학교는 MBA 프로그램을 3년 째 운영 중이다. 이 학교에서는 입학생들의 GMAT 점수가 MBA 성과를 얼마나 잘 예측하는지 판단하기 위해 12명의 졸업생을 대상으로 각 졸업생들의 학기 평균 GPA 점수(0에서 12 사이의 값)와 입학 시 제출한 GMAT 점수(200에서 800 사이의 값)를 조사해 보았다. 이 자료(GMAT_and_GPA_scores_for_MBA_students.csv)를 이용하여 GMAT 점수와 GPA 점수 간의 표본 공분산과 표본 상관관계를 도출하고 그 결과를 해석하여라.

```
> scores <- read.csv('GMAT_and_GPA_scores_for_MBA_students.csv')
> head( scores )
  GMAT  GPA
1  599  9.6
2  689  8.8
3  584  7.4
4  631 10.0
5  594  7.8
6  643  9.2
> gmat <- scores$GMAT
> gpa <- scores$GPA
```

수치적 기술통계

```
> plot( gmat, gpa )
```

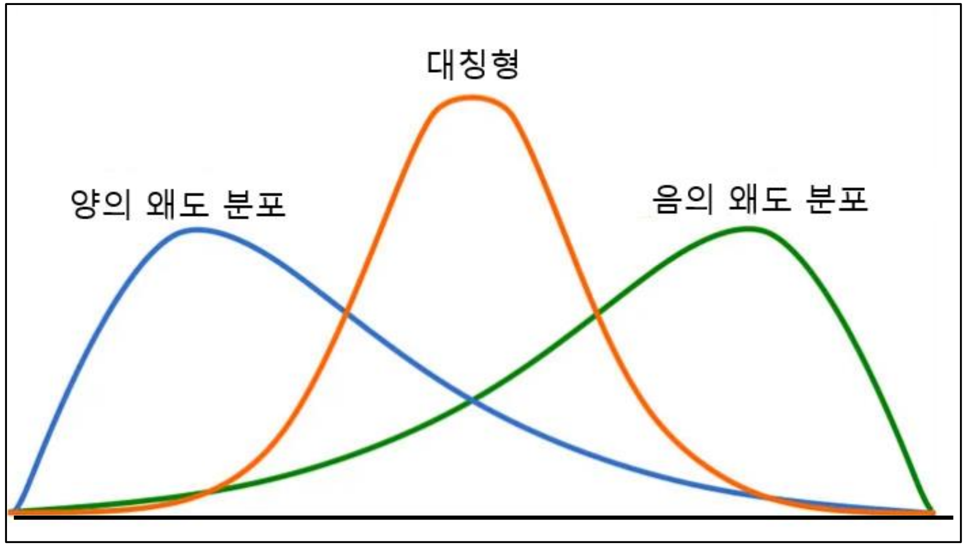



```
> cov( gmat, gpa )
[1] 26.16364
> cor( gmat, gpa )
[1] 0.5364827
```

수치적 기술통계

- **분포의 형태에 관한 척도**
  - 왜도 계수(Coefficient of Skewness): 분포의 **비대칭** 정도를 나타내는 척도.
    - 양의 왜도 : 오른쪽 꼬리가 길게 늘어진 형태
    - 음의 왜도 : 왼쪽 꼬리가 길게 늘어진 형태

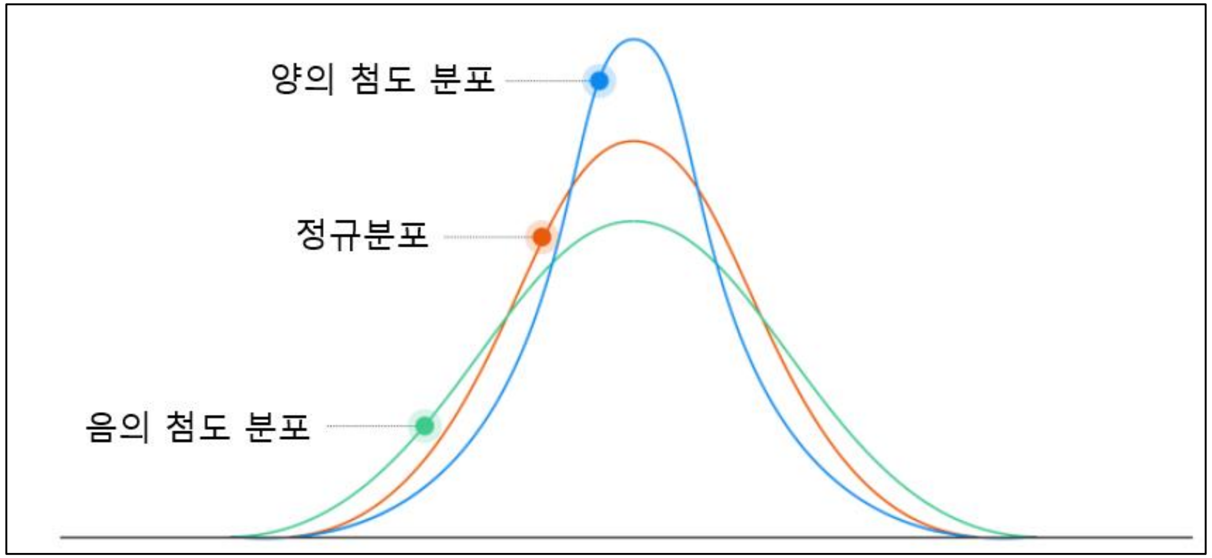
$$skewness = \frac{n}{(n-1)(n-2)}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}\right]$$



수치적 기술통계

- 첨도계수 (Coefficient of Kurtosis) : 분포의 **뾰족함** 정도를 나타내는 척도.
  - 양의 첨도 : 정규분포에 비해 꼬리가 얇고 봉우리가 뾰족한 형태
  - 음의 첨도 : 정규분포에 비해 꼬리가 두껍고 봉우리가 뭉툭한 형태

$$kurtosis = \frac{n(n-1)}{(n-1)(n-2)(n-3)}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}\right] - 3$$



수치적 기술통계

- **통신요금 예제**
  - 통신요금 데이터(telephone_bills.csv)을 이용하여 월 청구액에 대한 다음의 기술통계량 값을 구하여라.
    - 왜도 계수, 첨도 계수

```
> billdata <- read.csv('telephone_bills.csv')
> bills <- billdata$Bills
> install.packages('moments')
...

package ‘moments’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in ...
> library( moments )
> skewness( bills )
[1] 0.5373048
> kurtosis( bills )
[1] 1.710315
```