2. 데이터 분석을 위한 기초

2.1 기초 통계량

통계에서는 일반적으로 평균값, 중앙값 , 최빈값, 최댓값, 최솟값, 범위를 사용해 데이터의 정보를 요약해 표현한다. 엑셀을 통해 이를 표현해보자.

2.1.1평균

average(데이터 범위) 함수를 이용하여 표현할 수 있다.

2.1.2 중앙값

MEDIAN(데이터 범우)함수를 이용하면 위의 복잡한 과정을 내부처리를 통해 우리는 결과만 볼 수 있따.

2.1.3 최빈값

MODE(데이터 범위) 를 이용하면 구할 수 있다.

2.1.4 최댓값, 최솟값, 범위

MAX(데이터 범위), MIN(데이터 범위), MAX(데이터 범위)-MIN(데이터 범위)로 각각 최댓값, 최솟값, 범위를 구할 수 있다.

평균은 데이터 전체를 대상으로 계산을 하며, 중앙값은 데이터의 중앙, 최빈값은 가장 많이 등장하는 숫자를 파악하는데 유용하다.

정확한 데이터 분석을 위해서는 데이터의 특성을 파악한 후에 거기에 맞는 대푯값을 이용하는 것이 중요하다.

데이터의 특성을 파악하는 가장 좋은 방법은 그래프를 그려보는 것이다. 기초 데이터의 추세나 히스토그램 등은 그린 후에 데이터의 분포, 이상점 등을 파악한 후 어떠한 데이터 분석을 할지 결정하는 것이 가장 중요하다. 많은 데이터 분석 과정에서 이 과정을 거치지 않고 바로 대푯값부터 시작해 데이터 분석을 했을 때 결과가 좋지 않은 경우가 많이 있다. 아무리 바쁘더라도 기초 데이터는 꼭 확인하고 넘어가자.

2.2 분산과 표준편차

간단하게 계산할 수 있는 기초 통계량과 함께 데이터의 특징을 표현하는 지표로 분산과 표준편차가 있다. 분산과 표준편차가 있다. 분산과 표준편차는 데이터가 평균을 기준으로 어느 정도 흐트러져 있는지를 알려주는 지표이다. 평균으로부터 먼 곳까지 데이터가 퍼저 있다면 분산과 표준편차의 값이 커지며, 대부분의 데이터가 평균 근처에 위치한다면 분산과 표준편차의 값이 작아진다.

2.2.1 표준오차

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$0 = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{식}\,(2.3)$$

식(2.3)을 살펴보면 각 데이터와 평균과의 차이를 모두 합하면 결과는 0이라는 것을 알 수 있다. 분산과 표준편차는 평균부터 데이터가 흐트러져 이쓴ㄴ 정도를 나타내는 값이라고 말했다. 그러한 식(2.3)과 같이 모든 차이의 합이 0이라면 계산이 불가하다. 어떻게 구할 수 있을까? 여러 가지 방법이 있을 수 있겠지마나 가장 쉽게 생각할 수 있는 방법은 절댓값과 제곱이다.

각 데이터와 평균과의 차이 정도를 표준오차(Standard Error, Se)라고 하면, 절댓값을 이용한 표준오차는 다음과 같이 정의할 수 있다.

$$Se = \sum_{i=1}^{n} |x_i - \overline{x}| = |x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}| \quad \cdots\cdots\cdots 식(2.4)$$

그리고 제곱을 이용한 표준오차는 다음과 같다.

$$Se = \sum_{i=1}^{n} (x_i - \overline{x})^2 = (x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2 \quad \cdots\cdots\cdots 식(2.5)$$

이 두가지 방법 중에서, 데이터 마이닝 기법에서는 절댓값보다 제곱을 이용한 표준오차를 사용하고 있다. 절댓값과 제곱을 이용한 표준오차 중에서 제곱을 사용하는 것이 데이터의 흐트러짐을 조금 더 잘 표현해 주기 때문이다.

두 데이터 x=[-2,-2,2,2]와 y=[-3,-1,0,4]를 갖고 비교해 보자. 두 데이터 x,y의 평균은 0이다. 절댓값을 이용해 x,y의 오차를 계산해 보면 다음과 같다.

$$Se_x = |-2-0| + |-2-0| + |2-0| + |2-0| = 8$$
$$Se_y = |-3-0| + |-1-0| + |0-0| + |4-0| = 8$$

x와 y의 오차는 모두 8로 계산되었다.

이제 제곱을 이용해 x와 y의 오차를 계산해보자.

$$Se_x = (-2-0)^2 + (-2-0)^2 + (2-0)^2 + (2-0)^2 = 16$$
$$Se_y = (-3-0)^2 + (-1-0)^2 + (0-0)^2 + (4-0)^2 = 26$$

데이터	절댓값으로 계산한 오차	제곱으로 계산한 오차
x	8	16
y	8	26

x와 y의 분포를 보면 평균 0으로 부터 x보다는 y의 데이터 들이 넓게 퍼져 있는 것을 알 수 있다.

절댓값으로 계산한 오차는 x와 y가 모두 8로 같은 값을 가지지만, 제곱으로 계산한 오차는 y가 26으로 x의 16보다 큰 값을 가진다.

이렇게 비슷한 두 방법으로 오차를 계산하지만 제곱을 이용해 계산한 오차가 데이터의 퍼진 정도를 좀 더 자세하게 표현할 수 있다. 그렇기에 분산과 표준편차도 제곱으로 계산한 오차를 사용해 계산한다.

2.2.2 분산

식 2.5와 같이 제곱을 이용한 분산은 다음과 같다.

$$S^2 = \frac{Se}{n-1} = \frac{1}{n-1}(\sum_{i=1}^{n}(x_i - \overline{x})^2) \cdots\cdots\cdots\cdots 식(2.6)$$

데이터의 오차 제곱 합인 표준오차 (Se)를 데이터의 개수로 나눈 것으로, 오차제곱합의 평균이라는 의미이다.

2.2.3 표준편차

식 (2.6)으로 계산한 분산은 각 데이터와 평균값의 차이를 제곱한 값이므로 원래 데이터와 단위가 다르게 된다. 예를 들면 원래의 데이터의 단위는 길이(m)지만 분산은 제곱을 했기에 넓이(m^2)로 단위가 변한 것으로 생각하면 이해하기 쉽다.

이렇게 바뀐 단위를 원래 데이터의 단위로 되돌리기 위해서는 분산에 제곱근을 취하면 된다. 이렇게 분산에 제곱근을 취한 것이 표준편차다. 표준편차는 원래 데이터와 단위가 같아서 데이터의 흐트러진 정도를 더 쉽게 직관적으로 이해하기 쉽다.

표준편차는 분산의 제곱근이므로 다음과 같이 구할 수 있다.

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}(\sum_{i=1}^{n}(x_i - \overline{x})^2)} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots 식\,(2.7)$$

분산과 표준편차는 편차들의 평균이라는 의미이므로 데이터의 개수로 나눠야 한다. 이때 통계책 혹은 웹사이트를 보면 분모는 n이나 (n-1)이 되는 것을 볼 수 있다. 이것은 가지고 있는 데이터가 모집단인지 표본집단인지에 의해 결정된다.

2.2.4 모집단의 개수와 표본집단의 개수

모집단에 대한 분산과 표준편차는 대상이 명확하므로, 분모를 개체 수 n으로 나누면 정확한 값이 계산 가능하게 된다. 하지만 표본집단에 대해서 n으로 나누게 되면 추정하기 위한 모집단의 값보다 작은 값이 나온다. 이 작은 값을 보정하기 위해 표본집단에 대한 분산과 표준편차는 n대신 (n-1)을 사용해 계산한다. n과 (n-1)로 나누는 일은 다른 것 같지만, 결국 표본집단의 개수 n이 무한히 커지면 같은 식이 된다.

2.2.5 엑셀로 분산과 표준편차를 계산하는 방법

모집단의 분산과 표준편차를 구하는 함수는 다음과 같다.

- **분산(σ^2) = VAR.P(데이터 범위)**
- **표준편차(σ) = STDEV.P(데이터 범위)**

표본집단의 분산과 표준편차를 구하는 함수는 다음과 같다.

- **분산(S^2) = VAR.S(데이터 범위)**
- **표준편차(S) = STDEV.S(데이터 범위)**

P는 모집단의 Population을 의미하며, S는 표본집단(Sample)을 의미한다.

2.2.6 정규분포

자연계의 많은 데이터를 히스토그램으로 그려보면 데이터의 분포가 종 모양(Bell-Shape)처럼 평균을 중심으로 좌우 대칭을 이루는 것을 볼 수 있는데, 이것을 정규분포 (normal distribution)이라고 한다.

정규분포 곡선(파란색)과 표준편차의 관계는 다음과 같다.

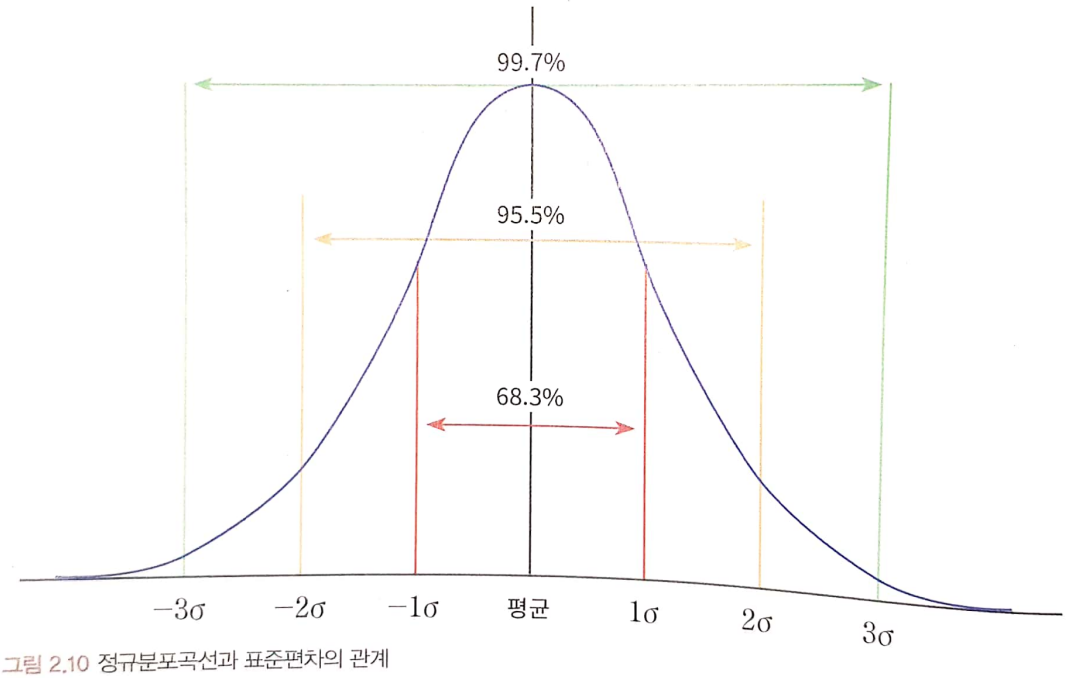


그림 2.10 정규분포곡선과 표준편차의 관계

- 총 데이터의 68.7%가 $\pm 1\sigma$ 범위 내에 존재한다

  데이터가 $\pm 1\sigma$ 범위 내에 포함될 확률은 68.2%이다.
- 총 데이터의 95.5%가 $\pm 2\sigma$ 범위 내에 존재한다

  데이터가 $\pm 2\sigma$ 범위 내에 포함될 확률은 68.2%이다.
- 총 데이터의 99.7%가 $\pm 3\sigma$ 범위 내에 존재한다

  데이터가 $\pm 3\sigma$ 범위 내에 포함될 확률은 99.7%이다.

2.2.7 정규분포를 이용한 이상점 검출

어떤 상품을 $\pm 3\sigma$로 품질 관리한다는 의미는 품질 조사 결과 $\pm 3\sigma$ 밖에 상품의 데이터가 존재할 확률은 0.3% 이기에 이 범위를 벗어나는 상품은 불량으로 간주한다는 의미이다. 이것을 3시크마 규칙이라고 한다.

실제로 이 규칙이 맞는지 실습 데이터의 정규분포와 시그마 시트의 데이터를 이용해 확인해보자. 이 데이터는 개수가 n=40개로 충분하지는 않지만 의미 파악은 가능하다.

먼저 안타부터 OPS까지 각 항목의 평균과 표준편차를 계산하자.

이제 1,2,3 시그마 범위 안에 포함되는 데이터의 개수를 찾아보자. 우선 조건을 만족시키는 셀의 개수를 반환하는 함수가 필요하다. 특히 우리는 $-\sigma$ 보다는 크고 $+\sigma$ 보다는 작은 조건을 이용해야 한다. 즉 두 가지 조건을 만족시키는 함수가 필요한데, 엑셀에서 제공하는 함수 =COUNTIFS()를 사용하면 된다.

=COUNTIFS(조건1의 데이터의 범위, "조건1",조건2의 데이터의 범위, "조건2 ", ···)

COUNTIFS : 두 가지 조건을 만족시키는 함수

여기서 조건1, 조건2는 따옴표를 사용해야 한다. =COUNTIFS()를 이용해서 안타의 개수를 기준으로 40명의 선수 중에 $\pm 1\sigma$에 포함되는 선수가 몇 명인지 알아보자.

선수가 $\pm 1\sigma$ 안에 포함될 조건을 정리하면 다음과 같다.

(평균 -1 X 표준편차) < = 데이터 AND 데이터 < = (평균 + 1 X 표준편차)

그러므로 이것을 COUNTIFS를 쓰면 다음과 같다.

=COUNTIFS(C8:C47, "<=" &평균+1x 표준편차, C8:C47, ">=" &평균-1x 표준편차)/40 *100

위 함수에서 결과를 백분율로 확인하기 위해 선수들의 수 40으로 나눈 후 100을 곱했다.

안타를 보면 1시그마는 52.5%, 2시그마는 100%는 3시그마는 100%이다. 예를 들어, **3시그마의 100%는 모든 데이터가 3시그마 안에 포함되어 있다는 의미이다.** 항목별로 1시그마의 경우를 보면 안타는 52.5%, 2루타는 60%, 3루타는 85%로 계산되었다.

1시그마 범위 안에 데이터가 포함될 확률은 68.3%라고 했지만 각각의 항목에 대해서 비슷한 항목도 있고 전혀 다른 항목도 있다. 결과가 다르다고 해서 3시그마 규칙이 틀리다고 단정할 수 없다. 13개씩 계산된 각 시그마들의 평균을 계산해보자.

1시그마는 68.1%, 2시그마는 95.0%, 3시그마는 99.8%로 3시그마 규칙과 비슷함을 알 수 있다.

데이터가 많아지면 많아질수록 점점 3시그마 규칙과 가까워진다.

마지막으로, 도루를 제외한 모두 3시그마 데이터가 100%로 $\pm 3\sigma$범위 안에 포함되어 있는 것을 알 수 있다. 확률로 알 수 있듯이 3시그마를 벗어나는 것이 이렇게 어렵다. 도루의 평균이 7.3개, 표준편차가 7.46으로 $\pm 3\sigma$의 범위는 다음과 같다.

ex) -15.1<=도루 개수 <= 29.7

오직 버나디나 선수만 32개로, $\pm 3\sigma$범위를 벗어난 것으로, 버나디나 선수의 도루 개수는 일반적이지 않은 것으로 판단할 수 있다.

일반적으로 데이터가 $\pm 3\sigma$의 범위를 벗어날 확률은 대단히 작으므로, 대부분의 데이터가 3시그마 규칙을 따른다는 것을 이해해두면, 품질관리. 이상감지 등 여러 가지 분야에 활용할 수 있다.

2.3 데이터 표준화

데이터끼리 비교하기 위해서는 서로 같은 기준이나 척도가 적용되어야 한다. 홈런 2-개를 친 타자와 안타 110개를 친 타자 중 어떤 타자가 잘했는지를 비교할 수 없는 것과 같다. 어떤 선수는 득점 능력이 뛰어나고, 어떤 선수는 도루 능력이 뛰어나는 등, 선수마다 뛰어난 능력이 다르다. 물론 모든 것을 다 잘하는 선수도 있다. 득점이나 도루를 어떻게 비교하면 좋을까? 이렇게, 서로 다른 기준이나 척도를 가진 데이터를 비굣하기 위해 사용하는 방법이 **데이터 표준화**이다.

$x = \{x_1, x_2, \cdots, x_n\}$ 일 때 표준화를 다음과 같이 정의할 수 있다.

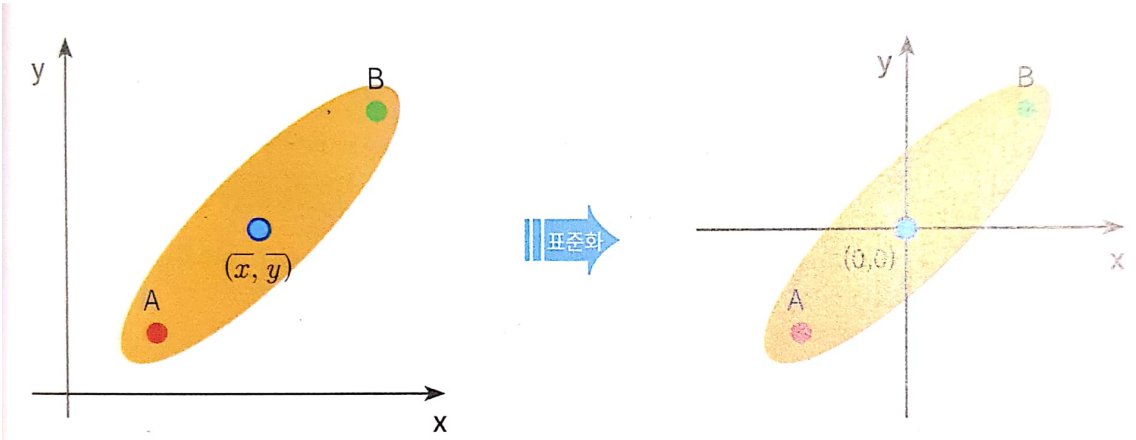
$$z_i = \frac{(x_i - \overline{x})}{\sigma} \quad \cdots\cdots\cdots\cdots\cdots \text{식}\,(2.8)$$

이 식은 각 데이터를 데이터의 평균과의 오차를 계산한 후 표준편차(σ)로 나누는 것이다. 데이터 표준화는 데이터를 표준편차로 나누어 단위를 없애는 이미지이다. 이 값을 Z값 이라고 하므로 z로 표시한다. 데이터가 타원형으로 분포하고 있다고 가정해 보자. 이 데이터에는 평균보다 작은 값을 가지는 점 A와 평균보다 큰 값을 가지는 점 B가 있고, 데이터 분포의 중심은 $(\overline{x}, \overline{y})$이다. 이때 식(2,8)을 이용해 표준화를 계산하면, 데이터 분포의 중심은 원점(0.0)으로 이동을 하게 된다.

이 과정은 식 (2.8)에 $x_i = \overline{x}, y_i = \overline{y}$를 대입해서 다음과 같이 계산해 볼 수 있다.

$$0 = \frac{(\overline{x} - \overline{x})}{\sigma}$$
$$0 = \frac{(\overline{y} - \overline{y})}{\sigma}$$

그리고 평균보다 작은 점 A는 음수(-) 값을 가지며, 평균보다 큰 점 B는 양수(+)가 된다.



(a) 원래 데이터 분포 (b) 표준화 후 데이터 분포

그림 2.13 표준화 전후 데이터 분포

표준화된 데이터 분포는 다음과 같은 특징을 가진다.

1. 평균 $\overline{x} = 0$
2. 표준편차 $\sigma = 1$
3. 단위없음

데이터 표준화 시트를 이용하여, 프로야구 선수들의 키와 체중을 표준화 해보자. 표준화를 하기 위해 평균과 표준편차를 미리 계산해야 한다.

하나의 식을 이용해서 여러 번 계산을 하는 경우에, 쉽게 하기 위해서 엑셀에서 식을 복사하는 방법에 대해 앞에서 배웠으며, 식을 복사하면 참조하는 셀도 같이 변하게 되는 것도 알고 있다. 표준화 식을 복사할 때 키와 몸무게의 값은 변해도 되지만, 평균과 표준편차는 고정 되어야 계산할 수 있다. 이때 필요한 것이 셀을 참조하는 방식이다.

엑셀에서는 상대 참조, 절대 참조, 열 고정, 행 고정과 같이 네 가지 식을 참조하는 방법을 제공한다.

상대 참조

상대 참조는 참조하는 주소가 복사해서 이동한 만큼 저절로 이동하는 것이다. 다음과 같이 A46 셀의 값을 E46 셀로 복사하는 식은 =A46이다. E46셀에 =46A을 입력하자. 그리고 옆 셀을 드래그 하면 주변 셀의 정보가 알아서 참조된다.

절대 참조

절대 참조는 참조하는 주소가 절대로 바뀌지 않고 고정된다. E46셀에 =A46을 입력한 후, [F4]키를 한번 눌러보면 =$A$46 와 같이 $가 열과 행에 추가된 것을 볼 수 있다. 이를 복사 붙이기 하면 상대 참조 했던 것과는 달리 고정된 값인 =$A$46이 가리키는 값을 참조한다. 이렇게 행과 열을 이동시켜 복사해도 참조 후소가 절대 변하지 않는 것을 **절대 참조**라고 한다.

행 고정

식을 복사해서 붙였을 때 행은 그대로 고정되지만 열이 이동하는 참조 방식을 행 고정이라고 한다. E46 셀에 =A46을 입력한 후, [F4] 키를 두 번 눌러보면 =A$46으로 바뀌는 것을 볼 수 있다. 행 번호 앞에 "$"이 추가 됐다. 그리고 상대 참조에서 했던 것과 마찬가지로 E46셀을 복사 한 후 범위 선택하여 붙여넣으면 행은 고정되고 열만 바뀌는 것을 볼 수 있다.

열 고정

행 고정과는 반대로 열이 고정되고 행만 이동하는 참조 방식이다.

f4 키를 세번 누르면 된다.

상대참조(A46) -> 절대참조($A$46) -> 행고정 (A$46) -> 열고정 ($A46) -> 상대참조(A46)

"$"가 붙은 행이나 열은 고정이 된다는 의미이다.

데이터를 표준화한 후에는 항상 평균과 표준편차를 계산해 정확하게 표준화가 되었는지 확인하는 과정이 매우 중요하다.표준화 과정은 데이터 분석의 가장 첫 단계이므로 여기서 잘못 계산하면 데이터 분석을 다시하는 경우도 많이 있으니 주의해야 한다.

2.4 공분산과 상관계수

이제 까지 하나의 데이터 평균, 분산, 표준편차, 표준화를 알아봤다. 이제 "습도가 높을 수록 짜증지수가 높아진다" 나 "공부에 집중한 시간과 상위권 대학의 합격률은 비례한다"와 같이 두 개의 데이터 간의 관계를 알아볼 수 있는 공분산(covariance)과 상관계수(correlation)에 대해서 알아보자.

2.4.1 공분산

데이터 $x = \{x_1, x_2, \cdots, x_n\},\ y = \{y_1, y_2, \cdots, y_n\}$가 있을 때 x, y에 대해서 공분산을 구하는 식은 다음과 같다.

$$S_{xy} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \cdots\cdots\cdots\cdots \text{식}\ (2.9)$$

S_{xy}는 x와 y의 공분산이라는 의미이다. 식 (2.9)에서 x의 평균과의 편차 $(x_i - \overline{x})$와 y의 평균과의 편차 $(y_i - \overline{y})$ 의 곱 $(x_i - \overline{x})(y_i - \overline{y})$ 은 다음과 같이 4가지 경우를 생각할 수 있다.

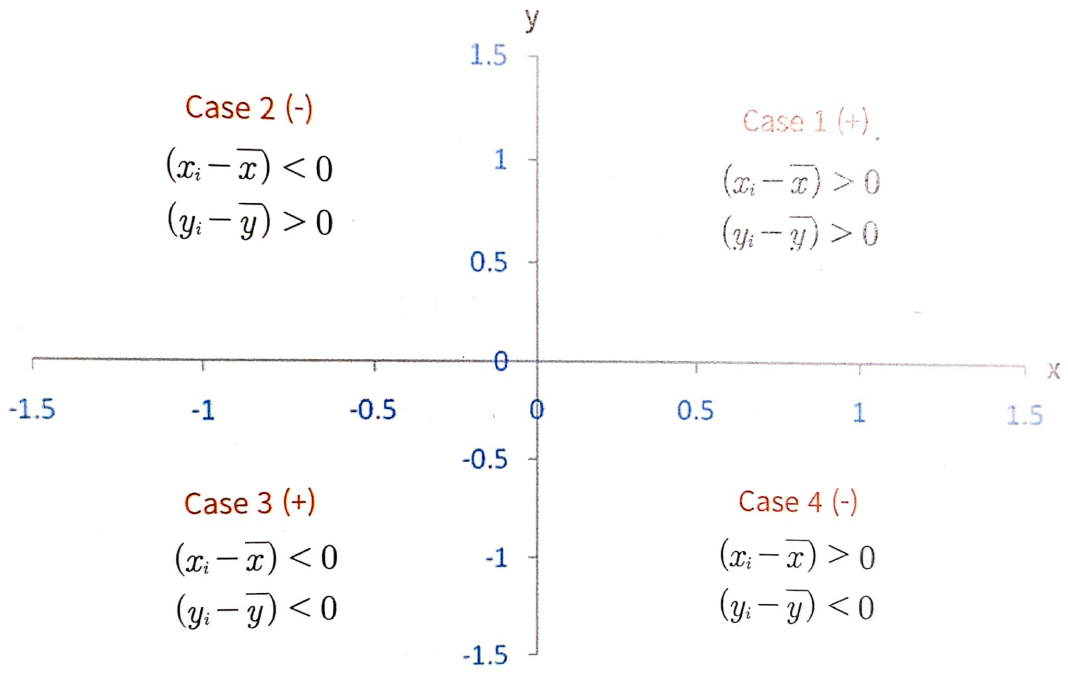


그림 2.22 Case별 데이터의 위치도

Case 1, 3에서 $(x_i - \overline{x})(y_i - \overline{y})$의 결과는 양수이다. 이는 x와 y가 비례관계에 있다는 의미이며, 같은 방향으로 같이 움직인다고 해석할 수 있다.

즉 x가 증가할 때 y도 증가하며 반대로 x가 감소할 때는 y도 감소한다.

하지만 Case 2,4에서 $(x_i - \overline{x})(y_i - \overline{y})$의 결과는 음수이다. 이것은 x가 증가할 때 y는 감소하며, x가 감소할 때는 y는 증가하는 서로 반대 방향으로 움직다고 해석할 수 있으며 반비례 관계이다.

그렇다면 공분산 시트를 사양해 실제 프로 야구선수의 키와 몸무게를 가지고 식 2.9에 맞춰 공분산을 계산해 보자.

지금 까지 식(2.9)를 기본으로 공분산을 계산했다. 엑셀에서는 모집단과 표본집단의 공분산 계산을 위해 두 가지 함수를 사용할 수 있다.

두 모집단의 공분산은 다음 함수를 사용해 계산하면 된다.

=COVARIANCE.P(데이터1의 범위, 데이터2의 범위)

또 두 표본집단에 대한 공분산은 다음 함수로 계산할 수 있다.

=COVARIANCE.S(데이터1의 범위, 데이터2의 범위)

ex) =COVARIANCE.S(B2:B21,C2:C21)

공분산에서 숫자는 그리 큰 의미가 없다. 다만 부호에 의미를 부여하는 것이 중요하다. 계산 결과가 양수(+)이므로 키와 몸무게는 비례관계에 있다고 말할 수 있다. 당연하지만 공분산 계산 결과로부터 프로야구 선수들의 키와 몸무게 사이에는 비례관계가 있다고 판단할 수 있다.

공분산은 데이터의 단위에 의존하므로, 값의 크기보다는 부호를 보고 결과를 판단해야 한다.

2.4.2 상관계수

지금까지, 두 개의 데이터가 비례관계인지 반비례관계인지에 대해서 알아볼 수 있는 공분산에 대해 알아봤다. 하지만 공분산은 데이터 간의 관계 (비례관계 또는 반비례)만 알려줄 뿐 그 관계가 어느 정도 인지에 대해서는 알려주지 않는다. 이 어느 정도를 알기 위한 것이 상관계수이다. 상관계수는 두 가지 데이터가 어느 정도의 관계를 가지고 있는지를 보여주는 계수이다.

상관계수는 다음과 같이 정의할 수 있다.

$$
\begin{aligned}
R_{xy} &= \frac{S_{xy}}{S_x S_y} \\
&= \frac{\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{(n-1)}}{\sqrt{\frac{\sum(x_i-\overline{x})^2 \sum(y_i-\overline{y})^2}{(n-1)(n-1)}}} \\
&= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \qquad \text{식}\,(2.10)
\end{aligned}
$$

R_{xy}는 x와 y의 상관계수라는 의미이다. 상관계수는 두 데이터의 공분산을 각각의 표준편차로 나누는 형태로 되어있다.

식 2.10을 표준화 데이터의 형태로 바꿔보자.

$$R_{xy} = \frac{\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{(n-1)}}{S_x S_y}$$

$$= \frac{1}{(n-1)} \sum_{i=1}^{n} (\frac{(x_i-\overline{x})(y_i-\overline{y})}{S_x S_y})$$

$$= \frac{1}{(n-1)} \sum_{i=1}^{n} (z_{x_i} z_{y_i}) \qquad \text{식 } 2.11$$

$$S_{xy} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i-\overline{x})(y_i-\overline{y}) \cdots\cdots\cdots\cdots \text{식 } (2.9)$$

식 2.9와 식 2.11을 비교해 보면 공분산은 각 데이터의 평균과의 편차간 곱의 합이지만 상관계수는 표준화된 데이터 간 곱의 합으로 이뤄져 있을 뿐 형식은 똑같다는 것을 알 수 있다.

표준화로 데이터의 척도를 없앴기에 데이터 간의 비교가 가능하게 됐고 위에서 말한 것처럼 어느 정도를 수치화할 수 있는 것이다 .실습을 해보자.

step1. 키와 몸무게를 표준화(z_x, z_y)

step2. 키와 몸무게의 곱(z_x, z_y) 계산

step3. 데이터간의 곱의 합 ($\sum_{i=1}^{n} z_{x_i}, z_{y_i}$)

step4. 데이터의 개수로 나눈다.

엑셀에서는 다른 함수들과 마찬가지로 상관계수에 대해서도 함수를 사용할 수 있다.

=CORREL()함수를 사용하면 상관계수 계산이 가능하다.

ex) =CORREL(B7:B21,C7:C21)

상관계수 계산 결과는 공분산 계산 결과와 달리 표준화된 데이터를 이용했기에, 단위가 무시되어 데이터의 단위가 달라도 상관계수의 결과는 0.62로 계산되는 것을 확인할 수 있다.

0~0.6	**약한 양 또는 음의 상관관계**
0.6~0.8	**보통 양 또는 음의 상관관계**
0.8~1	**강한 양 또는 음의 상관관계**

2.5 행렬

수학에서 데이터가 배열 형태로 구성되어 있을 때, 1차원 배열을 **벡터**라고 하며, 2차원 배열은 **행렬**이라고 한다. n개 데이터를 열 벡터로 표현하면 다음과 같이 표현할 수 있다.

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{bmatrix}$$

2.5.5 전치행렬

전치행렬(transpose matrix) 이란 행렬의 행과 열을 바꾼 행렬을 의미한다.

엑셀에서는 =TRANSPOSE() 함수를 사용하면 간단하게 전치행렬을 만들 수 있다.

하지만 배열 수식을 계산하기 위해서 수식의 입력이 Shift+Ctrl+Enter로 계산을 마무리해야 한다.

=TRANSPOSE() 함수를 사용해 전치행렬을 계산하는 방법을 보여준다.

1. 전치행렬이 들어갈 범위를 선택한다.
2. =TRANSPOSE(A3:C6)
3. Shift+Ctrl+Enter 을 누른다.

2.5.6 행렬 연산

배열수식으로 계산된 셀들을 지울 때에는 지울 때에는 범위를 지정해서 지워야 한다. 셀의 부분을 선택해서 지우면 오류가 생기니 주의해야 한다. 행렬 연산은 Shift + Ctrl + Enter 로 연산한다.

엑셀에서는 행렬의 곱셈을 지원하는 함수로 =MMULT()가 있다. =MMULT()의 사용법은 다음과 같다.

ex) =MMULT(A33:C34,E33:F35)

표본집단의 분산을 구하는 식을 표준화 데이터에 대한 분산을 구하는 식으로 바꾸면 다음과 같다.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_{x_i} - \bar{z}_i)^2 = \frac{1}{n-1}\sum_{n=1}^{n}(z_{x_i})^2 \quad ------식\ 2.15$$

왜냐하면 z_x의 평균 $\bar{z}_x$는 0이기 때문이다.

식 2.15의 각 데이터 제곱 $(z_{x_i})^2$을 행렬로 어떻게 표현하는지에 대해서 알아보자.

$$Z_x' Z = \begin{bmatrix} Z_{x_1} & Z_{x_2} & \cdots & Z_{x_n} \end{bmatrix} \begin{bmatrix} Z_{x_1} \\ Z_{x_2} \\ \cdot \\ \cdot \\ \cdot \\ Z_{x_n} \end{bmatrix}$$

$$= (Z_{x_1})^2 + (Z_{x_2})^2 + \cdots + (Z_{x_n})^2$$

$$= \sum_{i=1}^{n}(z_{x_i})^2 \qquad 식\ \ 2.16$$

$$V_{z_x} = \frac{1}{(n-1)} Z_x' Z_x \quad ----------식\ \ 2.18$$

엑셀로 분산을 계산해 본 결과는 다음과 같다.

행렬 분산 계산							
Zx		Zx'					
-1.11		-1.108	0.49	1.55	-0.58	0.75	-1.11
0.49							
1.55							
-0.58							
0.75		VAR.S 분산		1.2			
-1.11		행렬분산		1.2279			

먼저, 함수 =VAR.S()로 구한 분산 1.2이다. 그리고 행렬 Z_x의 전치행렬을 TRANSPOSE 함수로 구한 후 구한 행렬분산의 결과도 1.2로, 같다는 것을 확인하자.

다만 행렬분산은 =MMULT(R52:W52,P52:P57)/5 와 같이 5로 나눠야 한다.

2.6.3 공분산 행렬과 상관행렬

이제 공분산행렬과 상관행렬에 대해서 알아보자.

공분산행렬과 상관행렬이란 변수들 간의 공분산과 상관계수를 행렬로 나타낸 것이다.

변수 m개가 $x_1, x_2, x_3, \cdots, x_n$ 와 같이 있을 때 다음과 같이 정의할 수 있다.

$$\text{공분산행렬 } S = \begin{bmatrix} S_{x_1y_1} & S_{x_1y_2} \cdots S_{x_1y_m} \\ S_{x_2y_1} & S_{x_2y_2} \cdots S_{x_2y_m} \\ \cdot & \\ \cdot & \\ \cdot & \\ S_{x_my_1} & S_{x_my_2} \cdots S_{x_my_m} \end{bmatrix}$$

$$\text{상관행렬 } R = \begin{bmatrix} r_{x_1y_1} & r_{x_1y_2} \cdots r_{x_1y_m} \\ r_{x_2y_1} & r_{x_2y_2} \cdots r_{x_2y_m} \\ \cdot & \\ \cdot & \\ \cdot & \\ r_{x_my_1} & r_{x_my_2} \cdots r_{x_my_m} \end{bmatrix}$$

$S_{x_1x_1}$ 은 자기자신의 분산이고, $S_{x_1x_2}$ 에는 x_1, x_2의 공분산을 계산한 값이 들어간다. 공분산은 변수의 순서가 바뀌어도 상관이 없으므로 $S_{x_1x_2} = S_{x_2x_1}$ 이다.

상관행렬 R의 특징은 i 와 j 를 변수의 번호라고 했을 때 다음과 같이 정리할 수 있다.

1. ***상관행렬의 대각성분은 자기 자신에 대한 상관계수를 나타낸다.***

   자기 자신에 대한 상관계수는 1 이다.
2. ***대각 성분 이외의 성분들은 각 변수들 간의 상관계수를 나타낸다.***
3. ***대각선을 중심으로 대칭되는 성분의 값을 같다.***

   x_1 ***과*** x_2 ***상관계수*** $r_{x_1y_2}$ ***와*** x_2 ***과*** x_1 ***상관계수*** $r_{x_2y_1}$ ***는 같다.***

상관계수에서도 설명했지만, 표준화된 데이터를 이용해서 공분산행렬 S를 계산하면 상관행렬 R이 된다. 이제 공분산행렬과 상관행렬이 어떻게 계산되는지 알아보자.

$$S = \frac{1}{(n-1)} \begin{bmatrix} \sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{1i}-\overline{x}_1) & \sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2) \cdots \sum_{i=1}^{n}(x_{1i}-\overline{x}_1)(x_{mi}-\overline{x}_m) \\ \sum_{i=1}^{n}(x_{2i}-\overline{x}_2)(x_{1i}-\overline{x}_1) & \sum_{i=1}^{n}(x_{2i}-\overline{x}_2)(x_{2i}-\overline{x}_2) \cdots \sum_{i=1}^{n}(x_{2i}-\overline{x}_2)(x_{mi}-\overline{x}_m) \\ & \cdot \\ & \cdot \\ & \cdot \\ \sum_{i=1}^{n}(x_{mi}-\overline{x}_m)(x_{1i}-\overline{x}_1) & \sum_{i=1}^{n}(x_{mi}-\overline{x}_m)(x_{2i}-\overline{x}_2) \cdots \sum_{i=1}^{n}(x_{mi}-\overline{x}_m)(x_{mi}-\overline{x}_m) \end{bmatrix}$$

$$R = \frac{1}{(n-1)} \begin{bmatrix} \sum_{i=1}^{n} z_{x1i}z_{x1i} & \sum_{i=1}^{n} z_{x1i}z_{x2i} \cdots & \sum_{i=1}^{n} z_{x1i}z_{xmi} \\ \sum_{i=1}^{n} z_{x2i}z_{x1i} & \sum_{i=1}^{n} z_{x2i}z_{x2i} \cdots & \sum_{i=1}^{n} z_{x2i}z_{xmi} \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ \sum_{i=1}^{n} z_{xmi}z_{x1i} & \sum_{i=1}^{n} z_{xmi}z_{x2i} \cdots & \sum_{i=1}^{n} z_{xmi}z_{xmi} \end{bmatrix}$$

여기서 n은 데이터의 개수이다.

상관계수 행렬 R을 구해보자.

변수 x_1, x_2 은 n개의 데이터를 가지는 열 벡터이며 각 변수들의 표준화된 데이터를 다음과 같이 정의하자.

$$Z_{x_1} = \begin{bmatrix} Z_{x_{11}} \\ Z_{x_{21}} \\ \cdot \\ \cdot \\ \cdot \\ Z_{x_{n1}} \end{bmatrix} . Z_{x_2} = \begin{bmatrix} Z_{x_{12}} \\ Z_{x_{22}} \\ \cdot \\ \cdot \\ \cdot \\ Z_{x_{n2}} \end{bmatrix}$$

이 변수들의 집합을 Z_x 라고 정의하고 다음과 같이 행렬로 정의하자.

$$Z_x = \begin{bmatrix} z_{x_{11}} & z_{x_{12}} \\ z_{x_{21}} & z_{x_{22}} \\ \cdot & \\ \cdot & \\ \cdot & \\ z_{x_{n1}} & z_{x_{n2}} \end{bmatrix}$$

표준화된 데이터 Z_x를 이용해서 상관행렬을 계산하려면 중복 계산을 포함해서 (변수의 개수)^{2}만큼 계산해야 한다. 하지만 행렬을 이용해서 간단하게 계산할 수 있다.

위의 상관계수 행렬 R을 행렬로 표현하면 다음과 같다.

$$R = \frac{1}{(n-1)} Z_x{}' Z_x \quad \text{---------식} \quad 2.22$$

언뜻 보면, 분산을 구하는 식 2. 18과 비슷하지만 식 2.18은 열 벡터를 사용하고 있지만. 식 2.22은 행렬을 사용하고 있다.

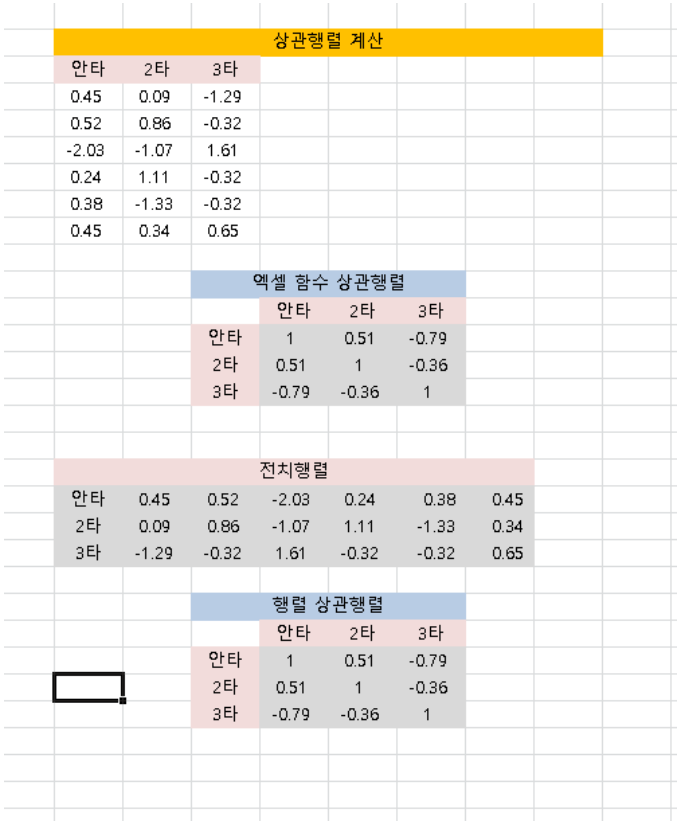
이를 이용해 풀어보자.

상관행렬 계산		
안타	2타	3타
0.45	0.09	-1.29
0.52	0.86	-0.32
-2.03	-1.07	1.61
0.24	1.11	-0.32
0.38	-1.33	-0.32
0.45	0.34	0.65

엑셀 함수 상관행렬			
	안타	2타	3타
안타	1	0.51	-0.79
2타	0.51	1	-0.36
3타	-0.79	-0.36	1

위 그림은 =CORREL() 함수를 이용해서 상관행렬을 계산한 결과이다.

전치행렬을 사용하면 이를 보다 쉽게 구할 수 있다.



상관행렬 계산

안타	2타	3타
0.45	0.09	-1.29
0.52	0.86	-0.32
-2.03	-1.07	1.61
0.24	1.11	-0.32
0.38	-1.33	-0.32
0.45	0.34	0.65

엑셀 함수 상관행렬

	안타	2타	3타
안타	1	0.51	-0.79
2타	0.51	1	-0.36
3타	-0.79	-0.36	1

전치행렬

안타	0.45	0.52	-2.03	0.24	0.38	0.45
2타	0.09	0.86	-1.07	1.11	-1.33	0.34
3타	-1.29	-0.32	1.61	-0.32	-0.32	0.65

행렬 상관행렬

	안타	2타	3타
안타	1	0.51	-0.79
2타	0.51	1	-0.36
3타	-0.79	-0.36	1

=MMULT(Q78:V80,P63:R68)/5 를 사용하여 전치행렬과 상관계수 행렬을 연산하면, 기존에 3번에 걸처 처리했던 것을 한 번의 연산으로 마무리 지을 수 있다!!

2.6.3 역행렬

엑셀에서 역행렬 연산은 =MINVERSE() 함수를 이용하면 된다.

2.7.2 좌표변환과 행렬

$$A = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \{\text{시계방향} : \theta > 0, \text{반시계방향} : \theta < 0\}$$

2.8 미분과 편미분

데이터 마이닝에서 미분은 아주 중요하다.

편미분이란 어떤 함수가 여러 가지 변수를 가지고 있을 때 각 변수에 대해 미분하는 방식을 말한다.

2.8.2 편미분

편미분의 중요성은 최근 딥러닝이 유명해지면서 다시 주목받고 있다. 딥러닝의 기본 구조인 신경망에는 무수히 많은 파라미터가 있으며, 오차역전파(back propagation) 라는 학습 방법을 이용한다.

이 오차역전파법은 신경망의 손실 함수 (loss function)가 최솟값을 가지도록 각 파라미터의 최적값을 찾는 학습방법이며 경사하강법(gradiant decent)를 이용해서 구현한다.

편미분은 경사하강법에서 각 파라미터들을 학습시킬 때 이용하는 방법으로 알고리즘의 핵심이라고 할 수 있다.

미분과 편미분은 데이터 분석에서 빠지지 않고 나오기에 개념을 확실히 이해하고 넘어가야 한다.