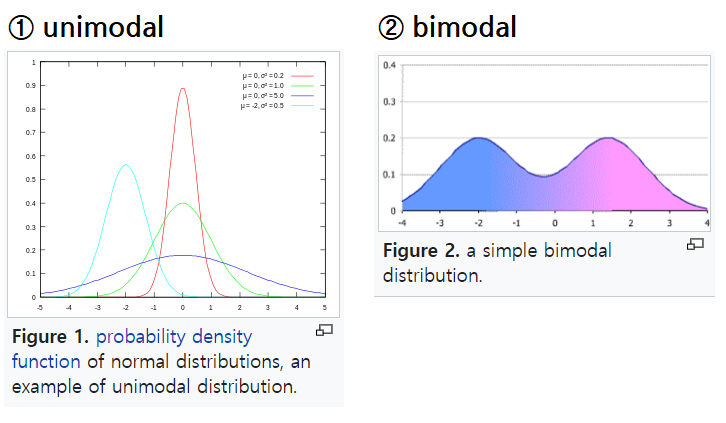
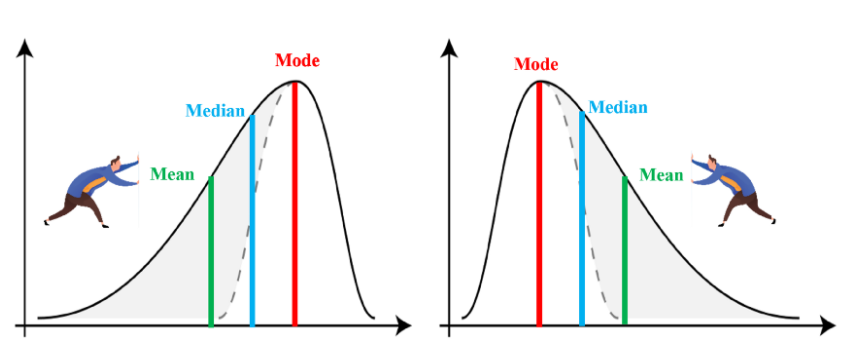
1. **Univariate Analysis (일변량 분석)**

**통계적 분석에서 단일변수만 분석하거나(예, 도수분포), 독립변수(independent variable, x)가 있어도 종속변수(dependent variable, y)가 하나인 분석유형이다. 일변량분석 이외의 분석은 다변량분석(multivariate analysis)이다.**

1. Numerical Variables (숫자변수)
   1. 중심경향치(Central Tendency)  
      중심 경향치는 통계학 및 수학에서 자료 데이터 분포의 중심을 보여주는 값으로서 자료 전체를 대표할 수 있는 값을 이르는 말이다. 대표적으로 산술평균(mean), 중앙값(median), 최빈값(mode)등이 있다.
      1. 최빈값 (mode)
         * 가장 큰 빈도수를 가진 관측치
         * 계산(computation)이 아닌 세기(counting)으로 계산
         * 주로 범주형 데이터에서 사용된다(most common category)
         * 하나의 최빈값을 가질때 unimodal, 2개의 최빈값을 가질 때 bimodal

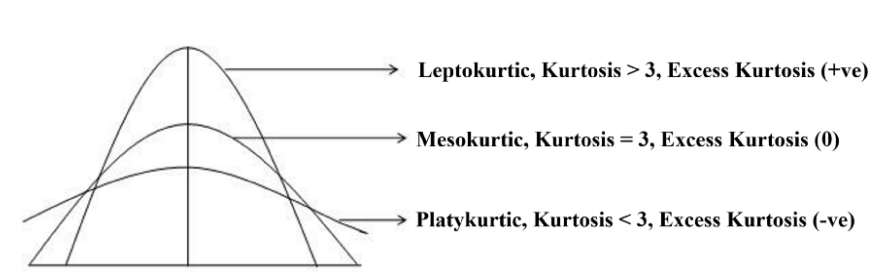


* + 1. 중앙값 (median)
       - 어떤 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미
       - 중앙값을 구하는 방법:
         1. 데이터가 홀수 개 일 때 : 중앙에 위치한 값
         2. 데이터가 짝수 개 일 때 : 중앙에 위치한 두 수의 평균
    2. 평균(mean)
       - 분포의 산술평균(arithmetic average)
       - 평균을 구하는 방법 : 값의 합을 전체 값의 개수로 나눈다.
    3. 왜도 (Skewness) = (Mean-Mode) / Standard Deviation  
       데이터 분포의 좌우 비대칭도를 표현하는 척도
       - 오른꼬리 분포, Positively skewed = right skewed = Mean > Median > Mode
       - 왼꼬리 분포, Negatively skewed = left skewed = Mean < Median < Mode



* + 1. 첨도 (Kurtosis)

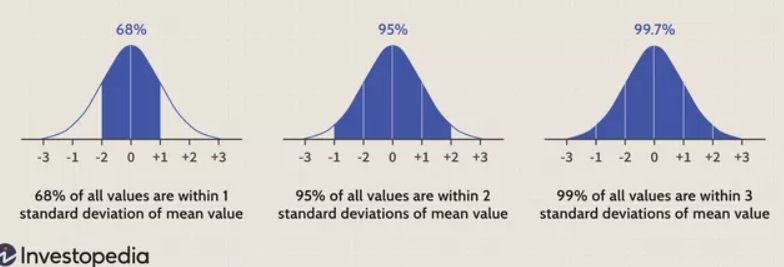
정규분포보다 얼마나 뾰족하거나 완만한지 정도를 나타내는 것. 데이터가 중심에 많이 몰려 있을수록 뾰족한 모양이 되고, 두루 퍼지면 구릉모양이 된다.



* 1. 분산치 (Variability)

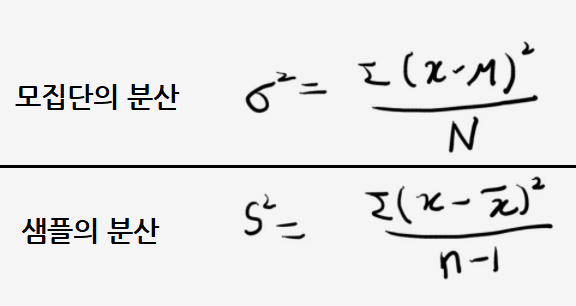
확률변수가 기댓값으로부터 얼마나 떨어진 곳에 분포하는지 가늠

1. Range = (MAX – MIN)+1
2. Variance and Standard Deviation (표준편차)

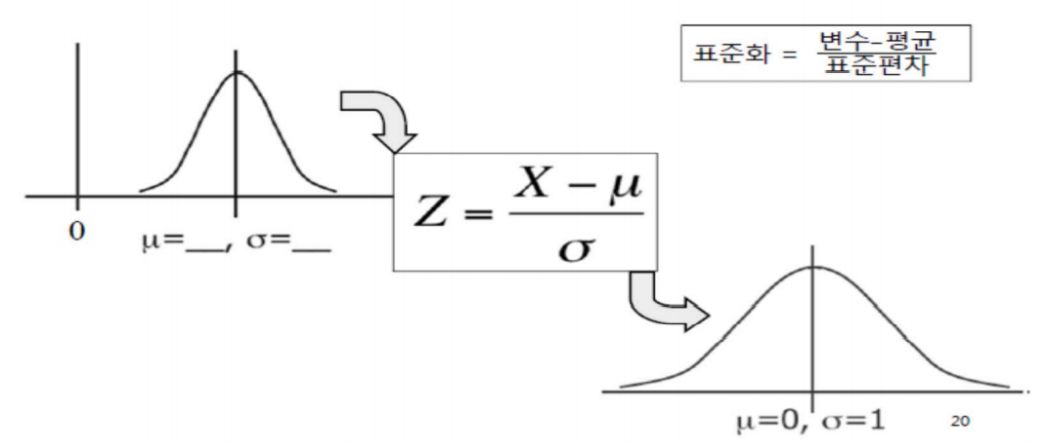


* + - * Variance (σ2, squared sigma) = the average of the squared differences from the Mean



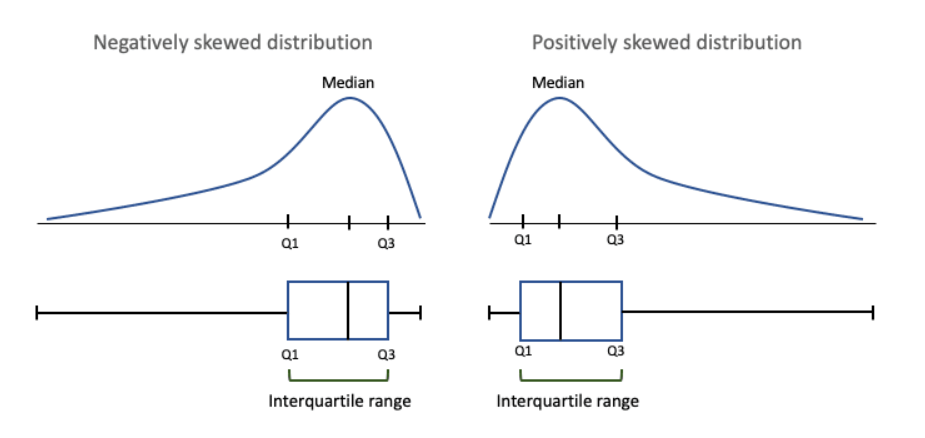
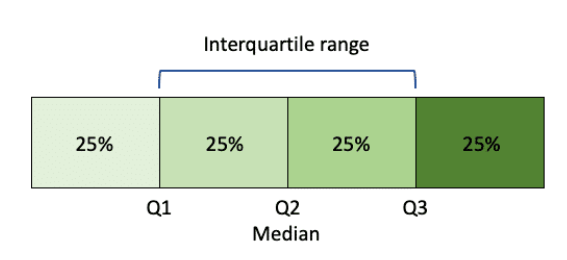


* + - * Standard Deviation (σ , sigma) = the square root of Variance: is a measure of how spread out numbers are
        1. Calculate the Mean: (1+2+3+4+5) / 5 = 2.4
        2. Calculate the each items difference from Mean and square it: (1 - 2.4)^ + (2 - 2.4)^ + (3 - 2.4)^ + (4 - 2.4)^ + (5 - 2.4)^
        3. Divide by number of items: [(1 - 2.4)^ + (2 - 2.4)^ + (3 - 2.4)^ + (4 - 2.4)^ + (5 - 2.4)^] / 5 = Variance
        4. Square root of Variance = Standard Deviation
        5. 68% of values are within plus or minus 1 standard deviation
        6. If SD is lower, less variety = bell curve is high. If SD is higher, more variety = bell curve lower and spread out



1. 사분위수 범위 (IQR)

자료 집합의 중간 50%에 포함되는 자료의 산포도를 나타낸다

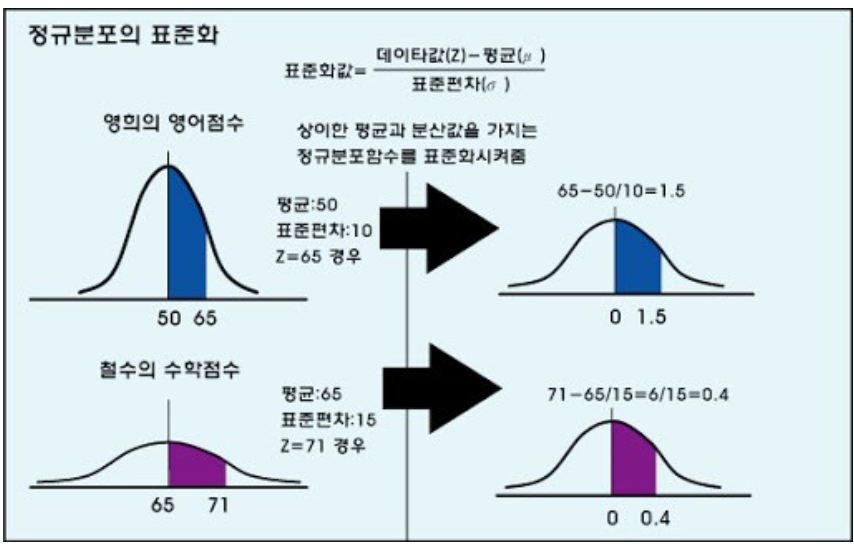


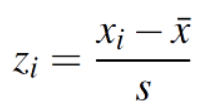
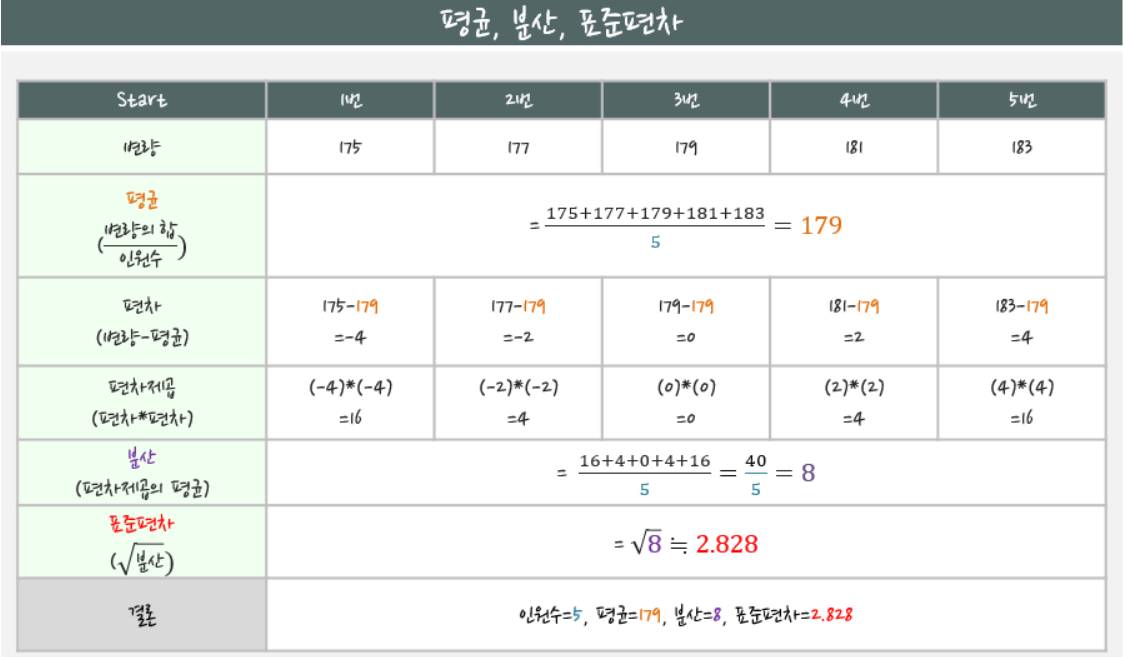
1. 표준화 (Standardization)

데이터 단위의 불일치를 해결할때 표준화와 정규화(Normalization)를 사용. 대표적으로 2개 이상의 대상이 단위가 다를 때 대상 데이터를 같은 기준으로 볼 수 있게 해준다.

* + - * 표준화 점수(Standard Scores) = Z score

평균을 중심으로 몇 SD(표준편차)만큼 떨어져 있는가를 확인. 서로 다른 분포를 가지고있는 값을 비교할때, 표준화 점수를 사용한다.

* + - * (측정값 – 평균) / 표준편차 = (salary\_2018 – average\_salary\_2018)/STD
      * MEAN값이 0으로 변하고 Z score로 평균에서 얼마나 떨어져있는지 확인한다. 

1. Categorical Variables (분류변수)

Count, Percentage, Mode, Pie Chart 또는 Bar Chart로 분석

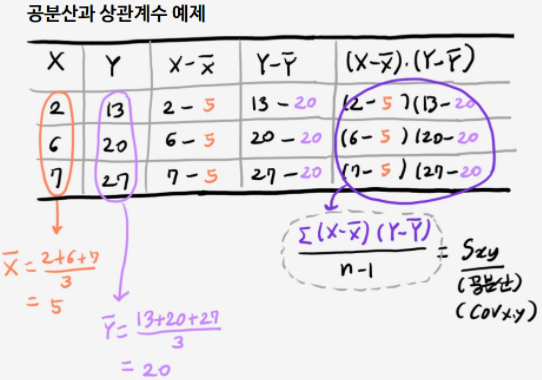
|  |  |  |
| --- | --- | --- |
| Type of variable | What does the data represent? | Examples |
| Binary variables (aka dichotomous variables) | Yes/no outcomes. | * Heads/tails in a coin flip * Win/lose in a football game |
| Nominal variables | Groups with no rank or order between them. | * Species names * Colors * Brands |
| Ordinal variables | Groups that are ranked in a specific order. | * Finishing place in a race * Rating scale responses in a survey\* |

1. **Bivariate Analysis**

**Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis.[1] It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical(실증적인) relationship between them.**

* 1. Numerical and Numerical
  2. Correlation Coefficient(상관관계)

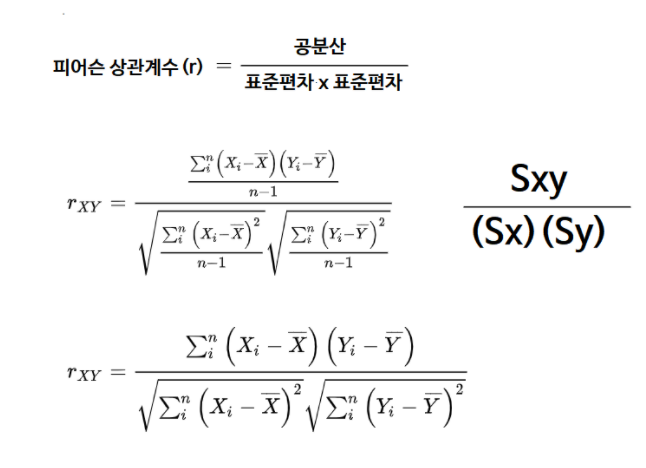
1. Covariance(공분산) is a measure of how much two random variables vary together. It’s similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.
   * + - 상관관계가 전혀 없을 경우 0, 두 변수가 비례하는 경우 양수, 반비례하는 경우 음수
       - Cov(X, Y) = Σ E((X – μ)\*(Y – ν)) / n-1
       - X is a random variable
       - E(X) = μ is the expected value (the mean) of the random variable X
       - E(Y) = ν is the expected value (the mean) of the random variable Y
       - n = the number of items in the data set
       - Σ = summation notation



* 1. Pearson Correlation Coefficient(PCC, 피어슨 상관계수)

A large covariance can mean a strong relationship between variables. However, you can’t compare variances over data sets with different scales (like pounds and inches). A weak covariance in one data set may be a strong one in a different data set with different scales.

The problem can be fixed by dividing the covariance by the standard deviation to get the correlation coefficient. r = Corr(X,Y) = Cov(X,Y) / σXσY

.