Univariate Descriptive Statistics

2016314726 정영준

Univariate Descriptive Statistics은 한국어로 단변량 기술통계량입니다. Univariate(단변량)은 종속변수가 1개임을 의미합니다. 종속변수가 여러 개일 경우 이를 Multivariate(다변량)이라고 합니다. 단변량 기술통계량 (일변량 기술통계량)에 해당하는 하나씩 살펴보겠습니다. 먼저 Central tendency에 해당하는 기술통계량을 설명하겠습니다. Central tendency는 표본이 어느 위치에 집중되어 있는지를 나타내는 척도입니다. 표본 평균, 표본 중앙값, 표본 최빈값이 이에 해당합니다. 먼저 평균에 대해 알아보겠습니다. 표본 평균은 산술 평균이라고도 불리며 기댓값을 나타냅니다. 모든 관측값을 더하고 이를 변량의 개수로 나누는 것이 표본 평균을 구하는 공식입니다. 표본 평균은 데이터에 대한 가장 직관적인 기술통계량으로 구하기가 쉽지만 하나의 이상치에 대해 크게 영향을 받는것이 단점입니다. 표본 평균만을 이용하여 자료를 설명하면 위와 같은 이상치에 의한 오류가 발생하게 됩니다. 예를 들어 수강생의 한달 평균 수입을 조사할 경우, 한달에 3000만원을 버는 한명의 수강생으로 표본 평균이 크게 올라가게 됩니다. 다음으로 표본 중앙값입니다. 중앙값은 중위수 라고도 하며 순서대로 표본을 오름차순으로 정리하였을 때 가장 중앙에 위치한 값을 의미합니다. 표본의 개수가 짝수일 때 중앙값은 2개가 되는데 이와 같은 경우에는 두 값의 평균을 중앙값으로 합니다. 표본 평균에서의 예시와 같은 상황에서 표본 중앙값이 표본 평균값보다 더 유용한 값이 됩니다. 지나치게 큰 값 또는 지나치게 작은 값에 대한 영향을 적게 받기 때문입니다. 다음으로 표본 최빈값입니다. 가장 많이 나타나는 자료의 값을 표본 최빈값이라고 합니다. 최빈값은 2개이상이 될 수있으며 주어진 관측치가 모두 다를경우에는 존재하지 않습니다. 다음으로 Degree of scattering(산포도)에 해당하는 최댓값, 최솟값, 범위, 사분위 편차, 분산, 표준편차 등이 있습니다. 사분위 편차란 중위수를 기반으로 하는 산포도 계산법입니다. 제 1사분위수, 제 2사분위수, 제3 사분위수는 각각 25백분위, 중앙값, 75백분위 수를 나타냅니다. 분산은 변수가 얼마나 흩어져있는지를 나타내는 지표로 편차 제곱의 합을 (변량의 개수 – 1)로 나눈 값입니다. 왜 변량의 개수가 아닌 변량의 개수 -1 로 나누는지는 자유도와 관련있습니다. 평균을 알고있는 가정하에 마지막 1개의 자료의 값은 평균과 지금까지 자료의 합을 통해 값이 고정되기 때문입니다. 표준 편차는 편차의 제곱의 평균으로 나타나는 분산의 스케일을 조정하기 위해 분산에 제곱근을 구한 값입니다. 변동 계수는 분산이 각 자료의 scale에 영향을 받기 때문에 분산의 상대적 비교를 위해 사용되는 값으로 표본 표준 편차를 표본 평균으로 나눈값입니다. 마지막으로 분포를 나타내는 일변량 기술통계량은 첨도와 왜도가 있습니다. 첨도는 한자 그대로 확률분포의 뾰족한 정도를 나타냅니다. 편차 4제곱의 평균을 포준편차의 4제곱으로 나누값이 첨도이며 3을 빼서 나타내기도 합니다. 3을 빼기전 첨도의 값이 3보다 크면 이는 정규분포보다 뾰족한 분포로 생각할 수 있습니다. 왜도는 확률분포의 비대칭도를 나타내며 첨도에서 4제곱을 3제곱으로 바꾸면 구할 수 있으며 이는 3차 표준 모멘트와 같다. 이 값이 양수이면 정규분포보다 왼쪽으로 분포가 치우쳐져 있으면 반대의 경우 오른쪽으로 분포가 치우쳐져 있다.