|  |  |
| --- | --- |
| 교육 제목 | **Git 설치와 Github 사용법** |
| 교육 일시 | 2021년 10월 26일 |
| 교육 장소 | 비대면 수업 |
| **교육 내용** | |
| 내용 | 평균과 표준편차, corelation -> 기초 통계학     1. 연속형 자료 ( 회귀의 문제, regression -> AI 예측형 )    1. 등간척도 : 온도 -> 0이 존재하지않고 곱셈이 적용안됨.    2. 비율척도 : 키,몸무게,혈압 2. 범주형자료 (이산형 분류, classification )    1. 명목척도(Nominal) : 속성간에 순위가 없음. ex. 남/녀, 국어/영어/수학    2. 순서척도(Ordinal) : 속성간에 순위개념이 있음. ex. 상/중/하, 좋음 / 보통 / 나쁨 3. 자료의 축약 빅데이타는 모집단으로부터 표본집단을 예측하는 것임.    1. 전체 고등학생 평균키, 전체 고혈압환자 (모집단)    2. 추출된 고등학생 100명, 연구에 참여한 고혈압 환자 50명 (표본집단)    3. 표본에서 평균은 X, 모집단 평균을 M이라할때, 일반적으로 X -> M^을 추정함    4. 표본과 모집단의 오차를 낮추는 것이 기술임. 모자(^)가 있으면 추정량(Estimator) 의미함.    5. AI에서 다루는 빅데이터는 거의 모집단에 가깝기 때문에, 주로 다루게 되는 것은 모집단이다. 따라서 확률이나 오차를 고려할 일은 별로 없다.    6. 모집단의 예 : 인구총조사    7. Quiz : 1시간 마다 전화 1통이 올 경우, 1시간안에 전화가 올 확률은? : 0.36% 확률분포 vs 정규분포 4. 범주형 자료의 요약    1. 자료는 그림으로 전달하는 것이 효과적이고 정보량이 많다. -> 시각화가 가장 중요하다.    2. 막대 그래프가 일반적으로 가장 효과적인 시각화 수단이다. (막대그래프 > 파이차트 > 분할표) 5. 연속형 자료의 요약    1. 도수분포표 (Frequency table)       1. 각 범주에 속하는 관측값의 개수를 그 범주의 도수라고 함.       2. 이 도수를 전체자료의 개수로 나눈 값을 상대도수라고 함.       3. 각 범주에 대응되는 도수와 상대도수를 나타낸 표    2. 히스토그램 (Histogram) : 데이터 개수 자체는 의미가 없고, 구간에 대한 확률(분포)을 나타냄.       1. (막대그래프는 X축이 이산형 자료 / 히스토그램은 X축이 구간자료임)       2. 선이 그어지는데, 선은 확률밀도함수라고 함. (PDF : Probablity density fuction)    3. 상자수염 그림 (Boxplot,대세 자주사용)       1. 정보량을 한번에 전달할 수 있음.       2. 인수종류 : 수염 -  whisker / 상자 : box / 꼭대기 : staple       3. 막대기의 끝과 끝은 최소와 최대값임. 박스의 아래(Q1)는 25% 구간, 중앙(Median)은 Median 값, 박스 위 쪽(Q3)은 75% 구간 ( 다섯숫자 요약 : min, Q1, median, Q3, max)       4. 만약 boxplot이 정규분포를 따른다면 박스는 중앙에 위치하게된다.    4. 바이올린 그림 (Violin plot, 대세 자주사용)       1. 데이터가 가진 분포를 한눈에 볼 수 있다.       2. violin plot에서 분포가 정규분포가 나온다는 것은 강력한 증거다. 6. 그래프를 통한 자료 요약 : 자료를 한눈에 볼 수 있으나, 모양에 따라 해석이 달라질 수 있음. 따라서 정량적 정보전달을 위해 수치를 통한 자료의 요약이 꼭 필요하다. 7. 수치를 통한 연속형 자료의 요약    1. 중심위치의 측도       1. 표본평균(sample mean) : 자료의 무게를 나타냄. 정규분포시 좋은 추정량. 그러나 자료의 이상치(outlier)에 영향을 많이 받음 (해결책 : Trimmed mean이라고 하며, 기존 자료와 차이가 크면 제외시키는 방식).       2. 중앙값(median) : 전체 관측값을 크기순으로 나열했을때, 중앙에 위치한 값. 연구분야에서 많이 사용되는 방법임. (1, 3, 5 = 3 / 1, 6, 3 ,2 = 2.5) 짝수의 경우에는 두 관측값의 평균으로 정한다. 자료의 이상치(outlier)에 영향이 거의 없다. 데이터의 편향이 없다.      1. 최빈값(mode) : 관측값 가운데 가장 많이 관측된 값. 2. P(X) = Probablity 함수. 3. 평균을 사용한다는 것은 그 이면에 정규분포를 따른다는 가정을 하는 것이고, 그것이 아니라면 중앙값이 더 좋은 통계량이 될 수 있다. -> F(x) ~ N(m,q^2) [m= 평균, q^2= 분산] 4. 예제      1. 표본평균은 이해하기쉽고 이론적 전개가 용이. 2. 표본평균은 전체 관측치를 반영하지만 이상치(outlier)에 영향을 받음, 3. 중앙값 중앙 부분의 관측치에 영향을 받고 이상치(outlier)에 영향을 받지않음. 4. 이상치들이 있는 경우 표본평균과 중앙값을 적절히 사용한다. Q. 무엇이 더 데이터를 대표하는 또는 의미있는 값인가?? 5. ex.평균연봉의 함정 : 사장의 연봉(이상치)때문에 평균연봉이 높아보이게됨. 따라서 정확한 연봉을 보고싶다면 연봉의 중앙값(Median)을 물어봐야한다. 6. 데이터 퍼짐의 정도    1. 분산과 표준편차(Variance and Standard deviation, SD)       1. 자료가 중심으로부터 얼마나 퍼져있는 지를 표현한 수치       2. S = 표준편차       3. 범위(Range) : 최대값 - 최소값       4. 사분위수범위(InterQuartile Range, IQR) : 3사분위수 - 1사분위수       5. 사분위수 범위는 자주 사용하고 아주 중요한 개념이다.       6. 이상치(Outlier)의 기준 : Q1 - 1.5\*IQR 이하, Q3 + 1.5\*IQR 이상의 값은 최대값이나 최소값으로 간주하지않는다. (하지만 해당 데이터에 대한 충분한 지식이 있는 사람이 충분한 근거가 있다면 이상치를 데이터에 포함시킬 수 있다.)       7. 분산과 표준편차는 평균값과 동일하게 그 이면에는 데이터가 정규분포를 따른다는 가정이 있다. 즉, 지수분포일 경우 적절하지 않을 수 있다. (전혀 다른 모양임..)       8. 지수분포 (Exponential)       9. 예제 : [2, 3, 5, 7, 9, 10] 표본분산, 표본표준편차, 범위, 사분위수를 구하여라   범위 : 10 - 2 = 8  사분위수 : median(0.5) = 6, Q1 =3.5, Q3 = 8.5, min = 2, max = 10  분산(S^2) : [ (2-6)^2 + (3-6)^2 + (5-6)^2 + (7-6)^2 + (9-6)^2 + (10-6)^2 ] / 6-1 = 10.4  표준편차(S) : 3.225  정규분포 : N(6,10.4)   1. Q. 내가 8이상을 뽑을 확률은?  H. **정규분포를 따른다고 가정한다.**     데이터가 정규분포를 따를때, 정규분포(확률밀도함수) 전체를 적분했을때 1이 된다.  적분을 이용하면 확률을 구할 수 있다.   1. 데이터 중심을 찾고 (평균 or 중앙값) 데이터의 퍼짐(분산, 표준편차)을 통해 분포를 파악하여 예측. 2. 두 변수간의 자료 요약    1. 두 변수의 상관성을 보는 것이 목적 (eg. 수학 성적 & 과학 성적 / 집의 평수 & 집의 가격)       1. eg. 연비와 기통수는 상관이 있다. -> 연비 & 자동 및 수동 미션은 상관이 있다 -> 자동 및 수동 미션과 기통수는 독립이 아닐 경우, 연비 & 자동 및 수동미션 상관성은 확신할 수 없다. **즉, 변수들이 서로 독립인가 아닌가에 따라서 상관성은 매우 복잡해질 수 있다.**      1. 상관 분석    1. 두 연속형 변수간에 선형적 연관관계가 있는 지를 분석하는 통계적 방법.    2. 두 연속형 변수간의 연관된 정도를 나타내는 척도이며, **인과관계를 설명하는 것이 아님.**       1. 수학과 과학성적은 상관성 / 수학성적과 공부시간은 인과관계가 있다고 판단한다.       2. 차량의 무게와 차량의 연비는 상관성이 있고 인과관계가 있다고 판단한다.       3. 인과성을 나타내는 분석으로는 “**회귀분석**”이 있다.    3. 상관분석의 정량적 지표 : 피어슨 상관계수(Pearson’s R)       1. 두 연속형 변수가 정규분포를 따르는 경우에 사용함.       2. r 값이 + 이면 양의 상관관계, - 값이면 음의 상관관계를 의미.       3. r 값의 범위는 -1에서 +1까지 분포하며 절대값이 1에 가까울 수록 높은 상관성을 의미.       4. r 값의 해석 (절대적 기준은 아님)          1. 절대값이 0.7 ~ 0.9 : 높은 상관관계          2. 절대값이 0.4 ~ 0.7 : 중등도 상관관계      1. 수식의 해석 : 상관계수 r의 값은 각 변수의 평균으로부터 움직이는 패턴이 같다면, 값이 올라간다. 공식의 분모는 단순히 값이 무한히 커지는 것을 막기위한 역할임.      1. 스피어만 상관계수    1. 두 변수간에 상관관계를 나타내는 비모수적 방법       1. 데이터가 선형성을 따르지않는다면 피어슨 상관계수를 대신 사용할 수 있는 방법이다.       2. eg. 수학점수와 영어 점수와의 관계 등 등       3. 1번째 = 스피어만 / 2번째 = 피어슨 / 3번째 = 둘다 X          1. 데이터가 정규분포를 따름 -> 피어슨의 상관계수          2. 데이터가 정규분포를 따르지않거나 표본이 10개 이하. -> 스피어만의 상관계수 |
| 요약 정리 | * **도수 분포표 : 각 범주에 속하는 관측값의 개수를 도수라 하며, 이 도수를 전체자료의 개수로 나눈 값을 상대도수라 한다. 각 범주에 대응되는 도수와 상대도수를 나타낸 표** * **히스토그램(Histogram) : 구간에 대한 확률 분포를 나타냄.** * **Boxplot : 막대기의 끝은 최소와 최대값이며, 박스는 Q1(25%), Median(50%), Q3(75%)로 구성된다.** * **바이올린 그림(violin plot) : 데이터가 가진 분포를 한눈에 볼 수 있다. 정규분포를 따르는지도 확인 할 수 있다.** * **평균(mean) : 정규분포시 좋은 추정량이나, 이상치에 대한 영향이 큼.** * **중앙값(median) : 전체 관측값을 정렬하였을 때, 중앙에 위치한 관측값. 이상치에 대한 영향이 거의 없음.** * **최빈값(mode) : 관측값 가운데 가장 많이 관측된 값.** * **분산과 표준편차(Variation and Standard deviation, SD) : 각 관측값에서 평균을 빼고난 값을 제곱한뒤, n-1 만큼 나눈 값이 분산이고, 이를 루트씌우면 표준편차가 된다.** * **이상치의 기준 : 일반적으로 Q1,3 +- 1.5\*IQR 값의 범위를 초과하면 이상치로 간주한다.** * **상관성 : 두 연속형 변수간에 선형적 연관관계를 상관성이라 함. 상관성이 있다고해서 인과관계가 설명되는 것은 아니다.** * **상관계수 : 데이터가 정규분포를 따르면 피어슨의 상관계수를 사용하고, 정규분포를 따르지않으면 스피어만의 상관계수를 사용한다. 만약 선형적 상관성이 보이지않으면 상관계수를 사용할 수 없다.** |