1. 데이터를 표에 나타낼때 표의 열은 특성을 대표한다. 이 특성을 variable 이라고 한다. 또한, 표의 행은 데이터이다. 이를 case또는 observational unit 이라고 한다. variable은 numerical 과 categorical로 나뉠수 있는데, numerical은 뜻대로 숫자들을 커버할 수 있다는 뜻이다. numerical variable은 더하기,빼기, 평균을 구하는 연산등이 가능하다. 반면에 categorical variable은 수학적 연산의 의미가 없다. 예를 들어서 수업시간에 예시로 드신 지역번호에 대해서 생각해보면, 그 지역번호들을 더하고,빼고, 평균연산을 해도 아무 의미가 없다. 이러한 variable 들을 categorical variable이라고 한다.
추가적으로 numerical variable은 연속적인(continuous) variable 과 불연속적인(discrete) variable로 나뉘고, categorical variable은 nominal(unordered) , ordinal(ordered) 로 나뉘어진다.
Continuous variable은 연속적이어야한다. 즉 예를들어서 고용률같이 데이터가 연속적으로 변할 수 있어야 한다. 반면에 인구 수는 음수가 될 수 없고, 연속 적이지 않다. 따라서 discrete variable 이라고 할 수 있다. 데이터가 연속적인가 불연속적인가에 따라서로 구분 할 수 있다.
Nominal 은 순서가 없다. 즉, 성별처럼 카테고리적으로는 분류될 수 있으나 순서가 없기 때문에 정렬할 기준이 없다. 반면에 ordinal은 순서가 있기 때문에 데이터를 시간순등 여러 가지 순서로 정렬할 수 있다.

2. 1.10연습문제
(a): 각 행은 UK residents의 smoking habits을 나타낸다.
(b): 이 설문조사에는 1691명의 참가자가 참여했다.
(c ): sex 는 categorical 이고 순서가 없기 때문에 nominal 이다.
Age 는 numerical 이고 연속적이지는 않기 때문에 discrete 이다.
Marital 은 categorical 그리고 순서가 없기 때문에 nominal이다.
GrossIncome 은 categorical이고, 돈의 양으로 순서를 나타낼 수 있기 때문에 ordinal 이다.
Smoke 는 categorical 이고, 순서가 없기 때문에 nominal이다.
AmtWeekends, amtWeekdays는 numerical 이고 연속적이지는 않기 때문에 discrete이다.

3. 1.14 연습문제

(a): 이 연구의 모집단은 5~15세의 어린이들이다.

일단 연구자들은 가르침을 준 아이들과 그렇지 않은 아이들로 나누었으므로, stratified sampling 의 정의인 divide-and-conquer 에 부합한다. 따라서 stratified sampling 한 표본집단들에 해당한다.

(b): 이 연구는 모집단에 일반화 될 수 있다고 생각한다. sample mean에서 표본집단이 많아질수록 평균은 모집단의 평균에 수렴하게된다. 이 추정은 완전하지 않을 수 있으나, 샘플집단이 좋다면,즉 골고루 잘 뽑았다면, 꽤 괜찮은 추정이다.

이 연구결과로 인과관계를 적립하기엔 무리라고 생각한다. 왜냐하면 이 어린아이들의 나이, 솔직함, 자기컨트롤능력등은 모두가 다 다르고, 애초에 변인들이 너무 다양하기 때문에 어떤 경우는 딱 이게 답, 또 다른 경우는 딱 이게 정답 이런 것처럼 답이 딱 떨어지는 것이아니고, 이 어린아이들중 평균적으로 몇 명은 거짓말을 할 것이고, 또 다른 몇 명은 거짓말을 안할 것이라는 것을 알 수 있을 뿐이지, 원인과 결과의 관계를 적립할 수는 없을 것 같다.