



# 의약품 빅데이터 분석 과정

- 6 차시 -

## 통계의 이해 및 사례 1

## 6차시. 통계의 이해 및 사례 1

### · 학습목표

1. 데이터와 변수를 통해 분석적인 관점에서 필요한 통계적인 개념에 대해 설명할 수 있다.
2. 상관관계와 인과관계를 통해 변수와 변수들의 관계에서 어떤 통계분석 방법이 있는지 파악할 수 있다.
3. 통계분석 방법을 통해서 가설에 대해 파악할 수 있다.

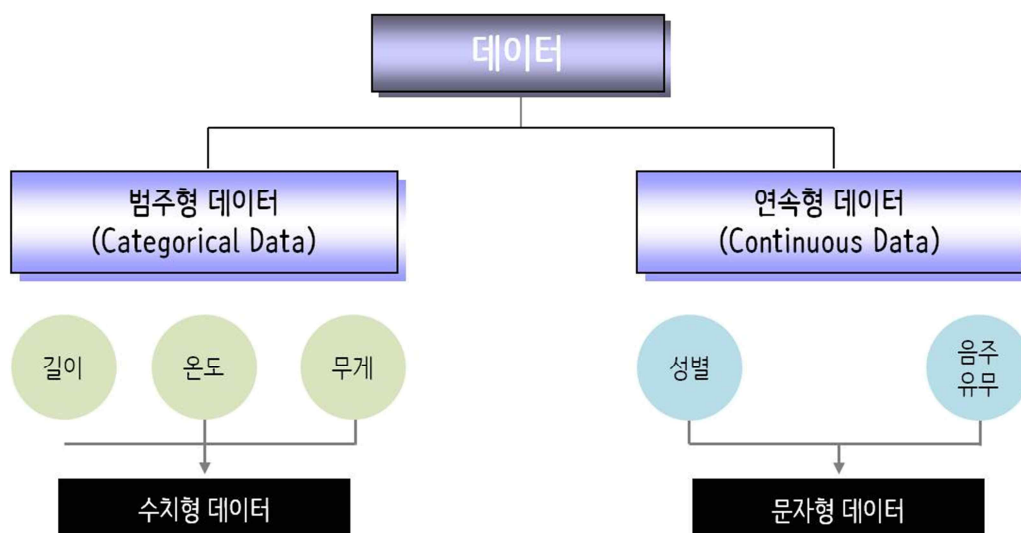
### · 학습하기

#### 1. 데이터와 변수

##### ■ 통계분석

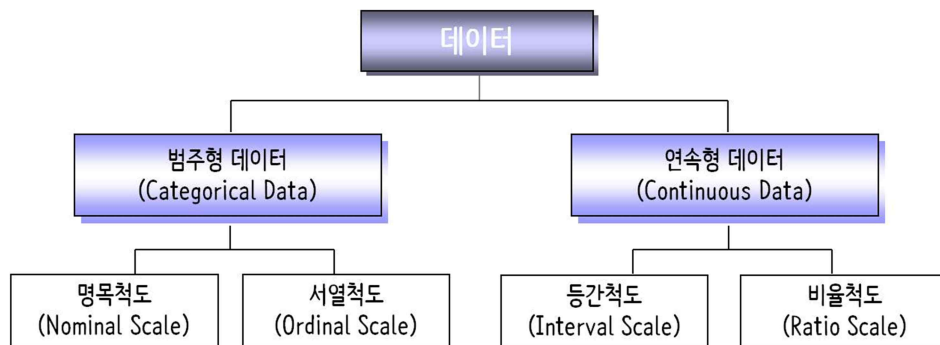
통계 분석은 데이터와 변수가 시작이라고 할 수 있다. 데이터는 값 하나 하나를 의미한다. 예를 들어 5명의 사람을 조사하는 경우를 살펴보면 먼저 그 5명의 성별을 조사했더니 남자, 여자, 여자, 남자, 여자 와 같이 나왔을 때 남자, 여자 와 같은 값 하나 하나를 데이터라고 하며, 그 데이터들의 모임인 성별을 변수라고 한다. 변수들은 이러한 데이터들의 변하는 것의 모임이다. 만약 5 명이 모두 남자였다면 데이터는 변하지 않습니다. 이러한 자료를 통계에서는 상수라고 한다.

##### ■ 데이터 유형



데이터 또는 변수는 그 성격에 따라 분류하는 방법이 여러 가지가 있다.

범주형 데이터와 연속형 데이터가 있다. 연속형 데이터는 길이, 온도, 무게 등과 같이 데이터가 숫자로 되어 있어서 수치형 데이터라고도 하며, 범주형 데이터는 남자, 여자와 같은 성별, 음주 유무와 같이 데이터가 문자로 되어 있어 문자형 데이터라고도 한다.

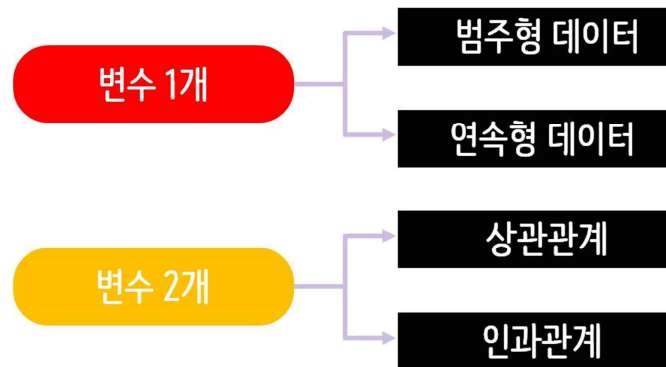


이러한 범주형과 연속형 데이터를 좀 더 세분화한 구분으로 명목, 서열, 등간, 비율 척도가 있다.

- 명목 척도는 남자, 여자와 같이 데이터 하나 하나에 의미가 있는 척도로 수학적 연산자로는 같다(=)와 다르다(≠)를 사용할 수 있다. 남자, 남자 같죠. 남자와 여자는 다르다.
- 서열 척도는 명목 척도의 성격에 높낮이가 추가된 개념이다. 대표적인 변수로는 학력, 계급, 직급 등이 있다. 초졸과 중졸은 다르죠. 명목척도이다. 그런데 초졸 보다는 중졸이 높다. 그래서 서열 척도이며 수학적 연산자로는 같다, 다르다와 높다, 낮다가 사용된다.
- 명목 척도와 서열 척도를 합쳐서 범주형 데이터라고 하며, 범주형 변수는 빈도와 백분율을 사용한다. 그래프로는 원그래프와 막대그래프가 사용된다.
- 등간 척도는 서열 척도에서 간격이 추가된 개념이다. 학력에서 초졸과 중졸 사이의 간격과 고졸과 대졸 사이의 간격은 다릅니다. 그래서 학력은 서열척도이다. 온도의 경우 1도와 2도 사이의 간격인 1도와 10도와 11도 사이의 간격인 1도는 동일하다.
- 비율 척도는 등간 척도에서 배, 비율의 추가된다. 5도와 10도를 비교했을 경우 "10도가 5도의 2배다." 라고 할 수 없다. 이에 비해 무게는 "10kg 은 5kg 의 2배다" 라고 할 수 있다.
- 등간 척도와 비율 척도를 합해서 연속형 데이터라고 하며, 연속형 데이터들로 이루어진 변수를 연속형 변수라고 한다. 연속형 변수는 평균과 표준편차의 통계량을 사용하며, 꺾은선 그래프, 막대 그래프, 상자 도표(Box plot) 을 사용한다.

- 등간 척도와 비율 척도를 완벽하게 구분하는 것은 까다로우며 실제 분석에서 이 둘을 굳이 구분해야 할 필요성은 없다. 하지만 명목 척도와 서열 척도는 구분하는 것이 좋습니다. 실제 통계 분석에서 명목 척도와 서열 척도에 따른 통계 분석 방법은 달라지는 경우가 많기 때문이다.

## 2. 상관관계와 인과관계



변수가 1개인 경우에 대해서는 바로 앞에서 살펴보았듯이 변수가 1개인 경우에는 그 변수가 범주형 데이터로 이루어진 범주형 변수인지 연속형 데이터로 이루어진 연속형 변수인지를 알면 된다. 이번에는 변수가 2개인 경우를 살펴봅니다.

변수가 2개인 경우 그 두 변수 사이의 관계를 생각해 볼 수 있다.

변수 사이의 관계는 상관관계와 인과관계로 구분하고 상관관계는 두 개의 변수가 서로 "주고 받는" 관계가 있을 때 두 변수 사이의 관계의 정도를 나타내는 분석으로 가장 대표적인 분석이 상관분석이다.

이에 비하여 인과관계는 어떤 하나의 변수가 다른 하나의 변수에 영향을 주는 관계를 규명하는 방법이다.

분 류	내 용
상관관계	
인과관계	
제 3의 변수와의 관계	

인과관계에서 다른 변수에 영향을 주는 변수를 독립변수라고 하며 X로 표시한다. 반대로 다른 변수에 영향을 받는 변수는 종속변수라 하며 Y로 표시한다.



독립변수 independent variable, 종속변수 dependent variable는 분석 방법마다 또 각 전공마다 여러 가지 용어로 부르고 있다. 회귀분석에서는 독립변수를 설명변수, 종속변수를 반응변수라 하며 구조방정식분석에서는 외생변수와 내생변수라고 한다. 공학에서는 인자와 특성값으로 부르기도 하며 산출변수와 투입변수라고도 한다. 이와 같이 통계의 용어들은 전공마다 분석 방법마다 여러 가지로 불러주기 때문에 용어의 혼동이 생길 수 있다.

통계학에서는 가장 많이 사용하는 표현은 독립변수와 종속변수이며 X와 Y로 표시하는 것이 대표적인 방법이다.

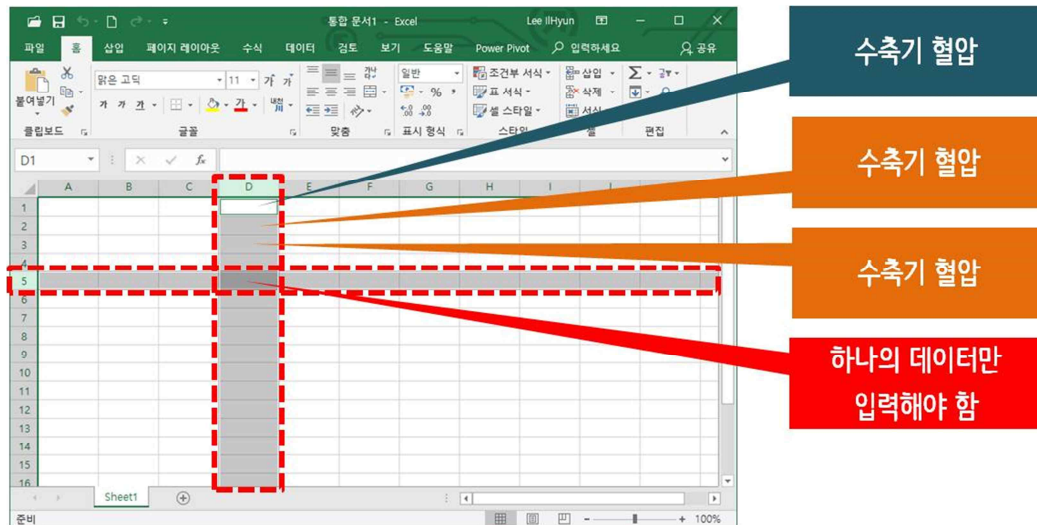
통계분석은 바로 두 변수 사이의 관계에서 시작한다. 상관관계와 인과관계에서 주로 사용하는 분석은 인과관계이며 인과 관계를 규명하는 분석 방법으로 t-test, ANOVA, 회귀분석, 로지스틱 회귀분석 등의 분석 방법이 있다.

### 3. 통계분석 방법

#### ■ 데이터 입력 규칙

가장 기본적인 데이터의 입력 규칙은 3가지이다.

- 첫째 한 사람의 자료는 한 줄에 입력한다.
- 둘째 동일한 항목은 한 칸에 입력한다.
- 셋째 한 셀에는 데이터는 1개만 들어가야 한다.

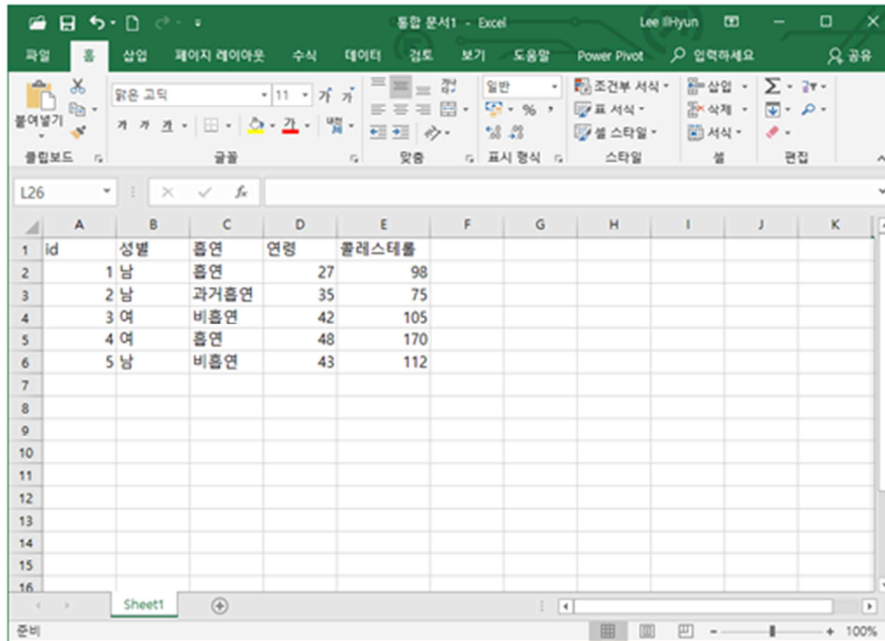


예를 들어 5줄에는 한 사람의 데이터가 모두 입력되어야 하며, D열에는 동일한 항목 여기서는 한 변수의 데이터만 입력해야 한다. 만약 D1에 수축기 혈압이 입력되었다면 D2, D3 등의 열에는 모두 수축기 혈압만이 들어가야 한다. 마지막 D5와 같은 한 셀에는 하나의 데이터 한 사람의 수축기 혈압만을 입력해야 한다. 한 셀에 수축기 혈압과 이완기 혈압같이 2개의 데이터를 입력하면 안 된다. 실제 서버에 입력된 자료나 고급 분석에서는 한 사람의 자료가 한 줄이 아니라 여러 줄에 나누어서 입력되는 경우도 있다. 예를 들어 건강보험공단 자료나 심평원 자료 같은 경우죠. 그와 같은 자료의 형태의 분석의 이번 과정에서는 다루지 않는다.

#### ■ 자료 통계 분석 방법

기본적인 규칙에 입력된 자료의 통계 분석 방법으로는 t-test, ANOVA, 회귀분석, 교차분석 로지스틱 회귀분석, 상관분석 등의 분석 방법이 있으며, 심평원 자료 등을 분석하는 분석 방법으로는 GEE의 통계분석과 빅데이터의 방법론인 딥러닝 등이 있다.

## ■ 데이터 입력 규칙



	A	B	C	D	E	F	G	H	I	J	K
1	id	성별	흡연	연령	콜레스테롤						
2	1	남	흡연	27	98						
3	2	남	과거흡연	35	75						
4	3	여	비흡연	42	105						
5	4	여	흡연	48	170						
6	5	남	비흡연	43	112						
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											

데이터 입력 규칙에 의해서 입력된 자료의 예를 보면 이와 같습니다. 1번째 환자는 흡연하는 남자이고 연령은 27세이다. 그리고 콜레스테롤은 98이다. 자료가 수집되면 통계분석을 하게 된다.

통계분석을 할 때 먼저 확인해야 하는 것은 사용된 변수가 무엇이 있는가이다. 현재 이 자료에서는 4개의 변수가 사용되었습니다. 성별, 흡연, 연령과 콜레스테롤이죠.

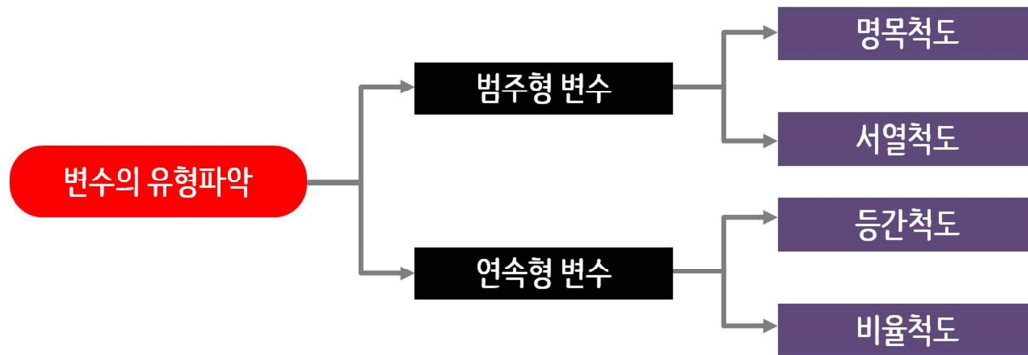
두 번째는 4개의 변수 중에서 종속변수가 있는가이다. 종속변수가 있다는 것은 독립변수도 있다는 것이므로 변수들 사이의 관계를 규명하는 분석은 인과관계 분석이 되는 것이죠.

4개의 변수 중에서 종속변수로 생각할 수 있는 변수는 콜레스테롤이다. 시간의 흐름상 4개의 변수 중 콜레스테롤이 가장 나중에 발생한다.

세 번째는 독립변수이다. 보통 인과관계에서 종속변수의 개수는 1개인 경우가 대부분이다. 물론 고급 분석에서는 종속변수가 여러 개인 경우도 있지만 통계분석을 배우는 과정에서는 종속변수가 1개인 경우를 우선 생각하게 된다.

그럼 위 자료에서는 종속변수는 콜레스테롤 1개이고, 독립변수는 성별, 흡연과 연령의 3개의 독립변수가 사용된 것이죠. 이 자료들의 분석에서는 보통 독립변수 각각과 종속변수와의 관계를 분석한다. 즉 성별과 콜레스테롤, 흡연과 콜레스테롤, 연령과 콜레스테롤의 관계에 대해서 각각 분석, 총 3번의 분석을 한다. 이렇게 독립변수 1개와 종속변수 1개의 관계를 규명하는 인과분석을 Univariate analysis 또는 Univariable analysis라고 한다. 이 분석에서의 목적은 각각의 분석에서 종속변수에 유의한 영향을 주는 변수를 찾는 것

이다.



네 번째는 종속변수와 독립변수의 유형 파악이다. 범주형 변수인가? 연속형 변수인가? 명목, 서열, 등간, 비율 척도에서 어떤 것인지를 파악하면 통계분석 방법을 찾을 수 있다.

### ■ 통계분석 방법

	A	B	C	D	E
1	id	성별	흡연	연령	콜레스테롤
2	1	남	흡연	27	98
3	2	남	과거흡연	35	75
4	3	여	비흡연	42	105
5	4	여	흡연	48	170
6	5	남	비흡연	43	112

성별, 흡연 상태, 연령, 콜레스테롤이 입력된 자료를 예를 들어보면 이 자료에서 사용된 변수는 성별, 흡연 상태, 연령, 콜레스테롤이다.

4개의 변수 중에서 종속변수는 콜레스테롤이다. 4개의 변수 중에서 시간의 흐름상 가장 마지막에 발생한 변수가 콜레스테롤이고, 성별, 흡연 상태, 연령은 콜레스테롤에 영향을 줄 것으로 생각된다. 그럼 나머지 변수인 성별, 흡연 상태, 연령은 독립변수이다.

통계분석 방법은 성별, 흡연 상태, 연령과 콜레스테롤을 모두 포함한 상태로 분석하는 것이 아니라 성별과 콜레스테롤, 흡연 상태와 콜레스테롤, 연령과 콜레스테롤의 관계에 대



해서 각각 분석을 한다. 이런 분석 후에 통계적으로 유의하게 나타난 독립변수들만을 선택해서 종속변수와 다시 분석을 하는 것이다.

예

id	성별	흡연	연령	콜레스테롤
1	남	흡연	2	98
2	남	과거흡연	3	75
3	여	비흡연	4	105
4	여	흡연	4	170
5	남	비흡연	4	112

독립변수

종속변수

연속형 변수로 등간이나 비율 척도가 됨

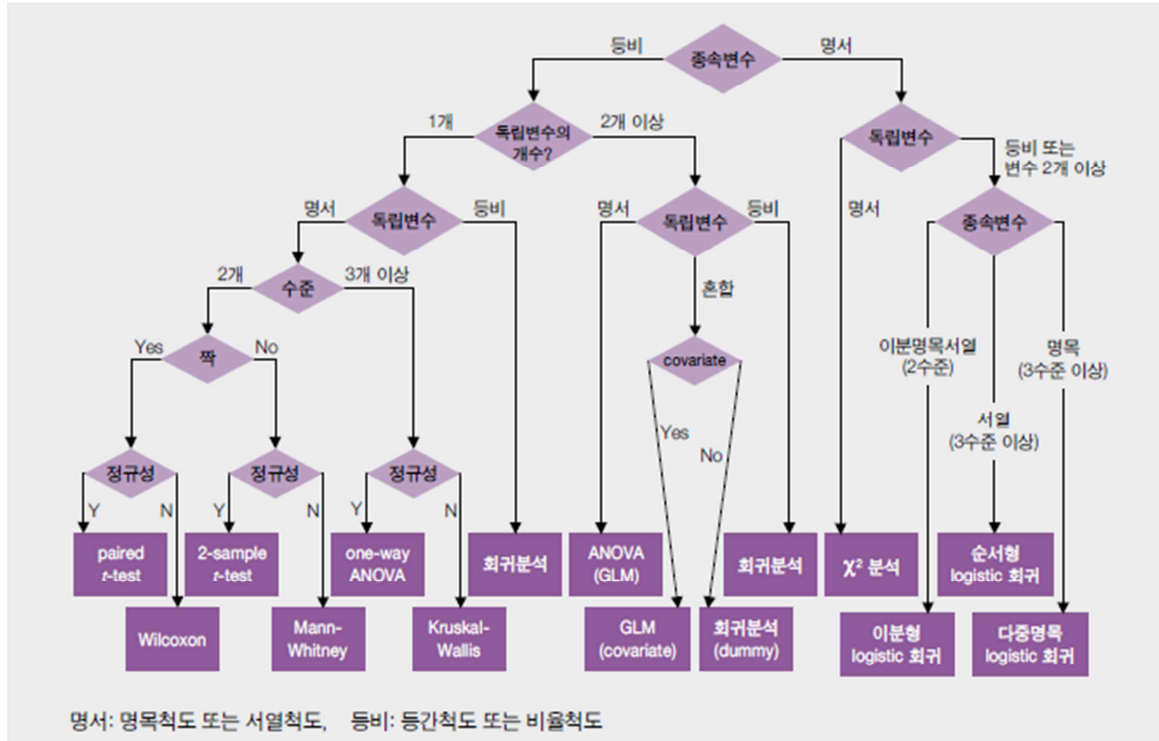
먼저, 성별과 콜레스테롤의 관계를 살펴보겠습니다.

2개의 변수 중 종속변수는 콜레스테롤이고 독립변수는 성별이다.

종속변수인 콜레스테롤의 데이터는 98, 75, 105 등과 같이 숫자로 되어 있기 때문에 연속형 변수이다. 따라서 등간이나 비율 척도가 된다.

독립변수인 성별은 남자, 여자와 같은 문자로 되어 있기 때문에 범주형 변수이다. 그리고 남자와 여자 데이터 사이에는 높낮이가 없기 때문에 명목 척도이다.

이에 통계분석 방법에 대한 Flowchart를 보면



먼저 종속변수는 콜레스테롤이며 등간이나 비율 척도이기 때문에 왼쪽으로 갑니다.

독립변수의 개수에 대해 살펴볼텐데 우리가 살펴본 자료에서 독립변수는 성별, 흡연 상태, 연령이었다. 하지만 현재 분석에서는 성별과 콜레스테롤의 관계만 보고 있기 때문에 독립변수는 성별 1개이기 때문에 왼쪽으로 갑니다. 독립변수인 성별은 명목 척도이므로 왼쪽으로 갑니다.

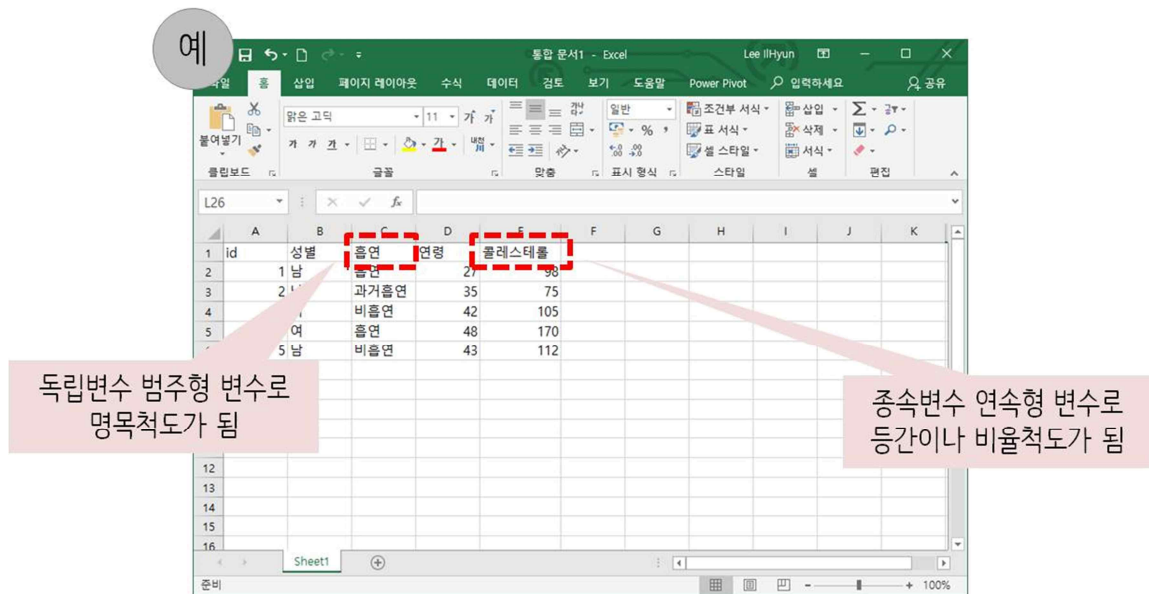
수준이라는 것은 범주형 독립변수가 갖고 있는 집단의 수를 의미한다. 성별이라는 독립변수는 남자와 여자의 2개 집단, 그룹을 가지고 있다. 따라서 수준은 2개 이므로 왼쪽으로 갑니다.

기초통계분석 나오는 개념에서 조금 어려운 개념이 바로 이 짝이라는 개념이다. 짝이라는 것은 한 사람에게서 동일한 데이터를 2개 얻었을 때 사용한다. 예를 들어 콜레스테롤을 측정하고 약을 복용한 후에 다시 콜레스테롤을 측정한다. 그럼 약 복용 전과 후의 콜레스테롤 데이터가 있으며 바로 이때 이 2개의 데이터는 짝을 이룬 자료라고 한다.

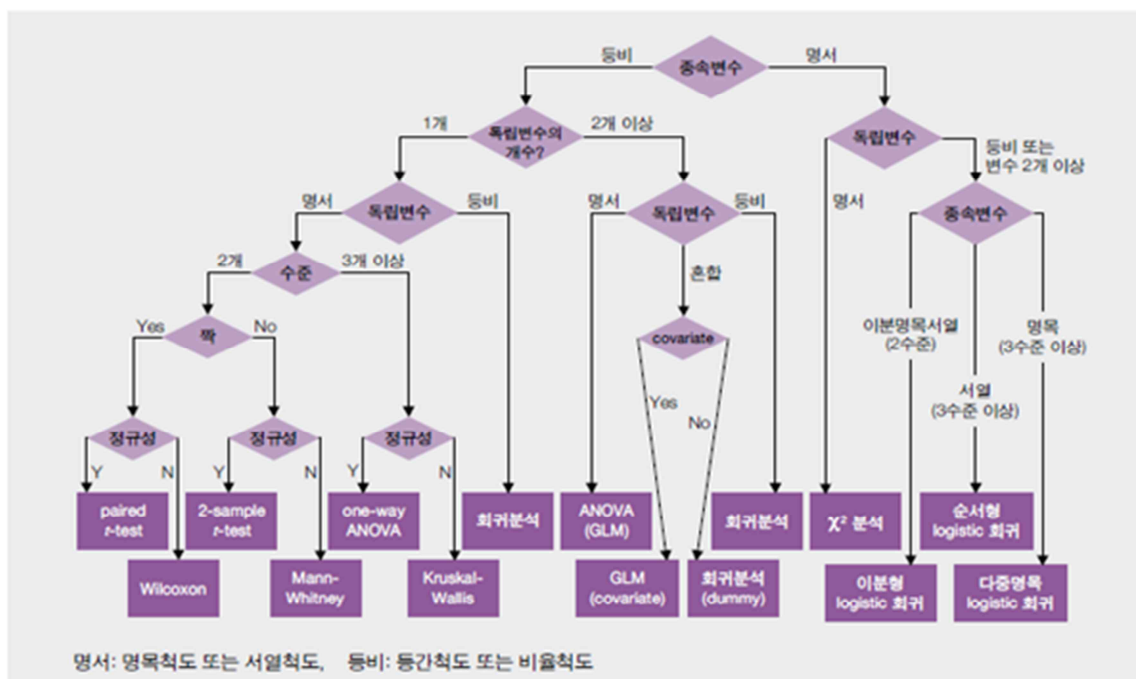
남자의 콜레스테롤과 여자의 콜레스테롤은 서로 관련성이 없죠. 그러므로 본 자료는 짝이 아니므로 오른쪽으로 갑니다.

정규성은 자료가 정규분포를 가지는 자료인가를 확인하는 통계분석 방법이다. 일반적으로 기초 통계분석에서 나오는 대부분의 통계분석 방법은 표본의 데이터가 정규분포이어야 한다는 가정이 있다. 만약 이 자료가 정규분포이라면 왼쪽으로 가서 2-sample t-test라는 분석이다. 이 분석은 독립표본 t 검정이라고도 하며 일반적으로 t-test 라고 한다.

만약 정규성 가정을 만족하지 못하면 비모수 검정인 Mann-Whitney 검정을 한다.



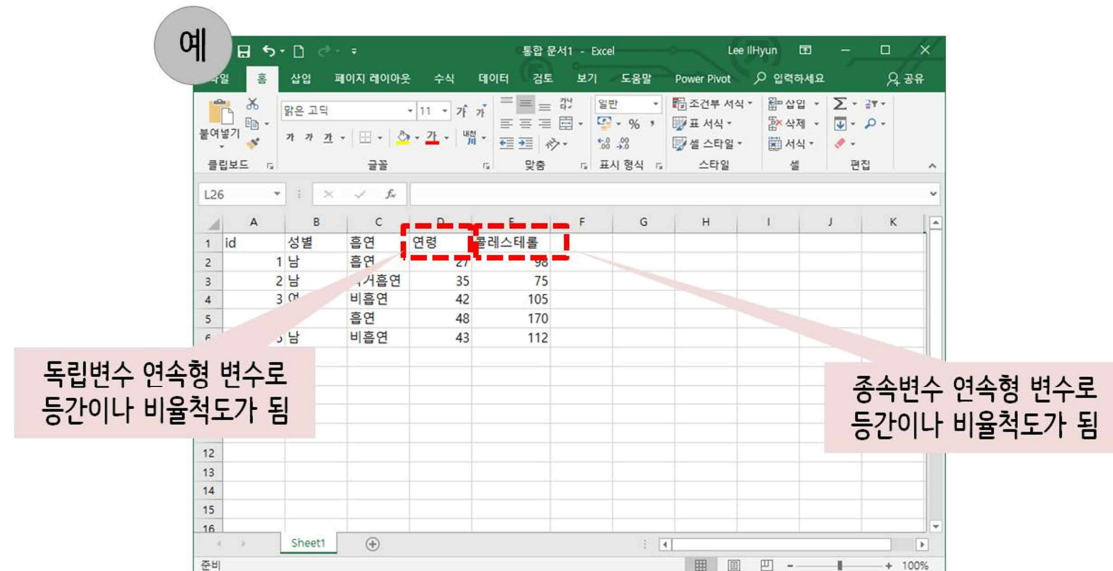
두 번째 예를 살펴보면 흡연 상태와 콜레스테롤의 관계이다. 종속변수는 콜레스테롤 연속형 변수인 등간이나 비율 척도이다. 독립변수는 흡연 상태로 가질 수 있는 값은 흡연, 비흡연, 과거흡연의 문자형 데이터이므로 범주형 변수이고 데이터 간에 높낮이가 없으므로 명목 척도이다. 또한 집단의 3집단이다.



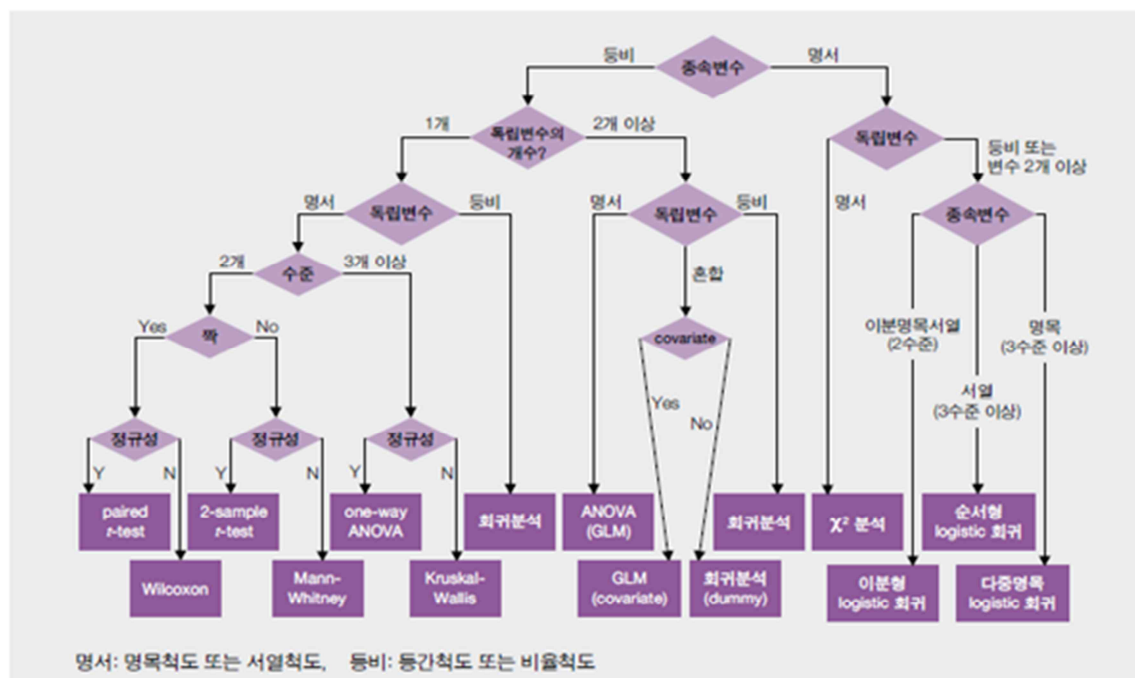
Flowchart를 통해 살펴 보면 종속변수는 콜레스테롤 등간 또는 비율 척도이므로 왼쪽으로 갑니다. 독립변수의 개수는 흡연 하나이므로 다시 왼쪽으로 갑니다.

독립변수는 흡연 상태로 명목 척도이기 때문에 왼쪽으로 갑니다.

흡연 상태는 흡연, 비흡연, 과거흡연의 3집단으로 수준은 3이므로 오른쪽으로 갑니다.  
다시 정규성이 나오는데 정규분포이면 왼쪽으로 one-way ANOVA, 일원배치 분산분석, 분산분석, 변량분석이라고도 하는 분석이다. 보통은 ANOVA 또는 분산분석이다.



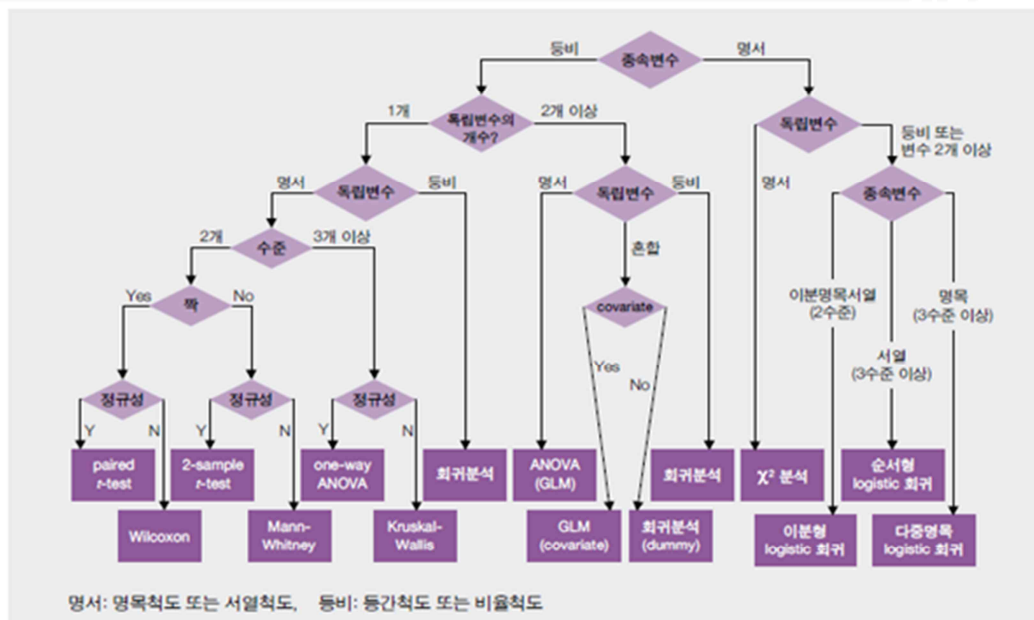
세 번째 예제는 연령과 콜레스테롤이다. 종속변수는 콜레스테롤 연속형 변수인 등간이나 비율 척도이다. 독립변수는 연령이며 연속형 변수인 등간이나 비율 척도이다.



Flowchart를 보면 종속변수는 콜레스테롤 등간 또는 비율척도이므로 왼쪽이다.  
독립변수의 개수는 연령 1개로 왼쪽으로 갑니다. 독립변수인 연령은 등간이나 비율 척도이므로 오른쪽으로 갑니다. 분석 방법은 회귀분석, 정확히는 단순 회귀분석이다.

세 번째 예제에서 연령을 다시 생각해 보면 연령은 27, 35, 42과 같은 연속형 변수이다. 상황에 따라서 20대, 30대, 40대와 같이 연령보다는 연령대가 더 의미 있는 경우가 있다. 연속형 변수인 등간이나 비율 척도를 20대, 30대와 같이 범주화하면 이제 연령대라는 변수는 서열 척도가 된다. 따라서 분석 방법도 달라진다.

Flowchart를 따라가서 확인하면 이때의 분석 방법은 ANOVA이다.



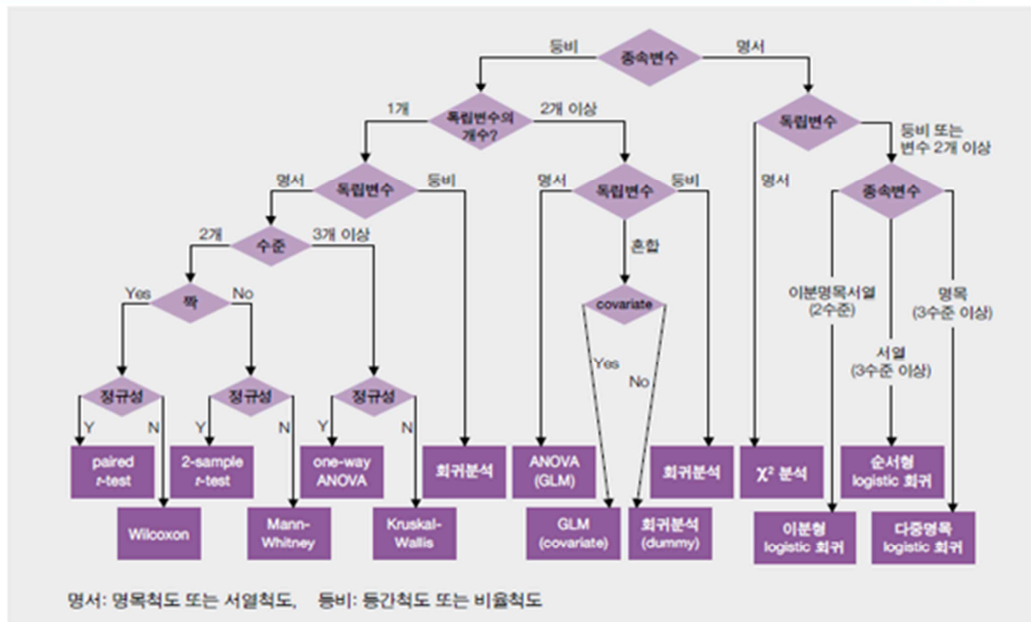
연령대를 59세 이하, 60세 이상으로 구분을 하면 어떻게 될까요?

서열 척도이기는 하지만 수준이 2개가 되고 분석 방법도 달라지게 되며 이때의 분석은 t-test가 된다. 통계분석 방법은 데이터의 성격에 따라서 분석 방법이 달라질 수 있다.

종속변수 콜레스테롤, 독립변수 연령이라 하더라도 독립변수의 유형에 따라서 회귀분석, ANOVA, t-test 등의 분석방법이 달라지는 것이다.

그럼 종속변수인 콜레스테롤의 형태가 달라지면 어떻게 될까요?

총콜레스테롤 기준이 200미만이 정상, 200 이상 위험군으로 구분을 한다면 서열척도가 된다.

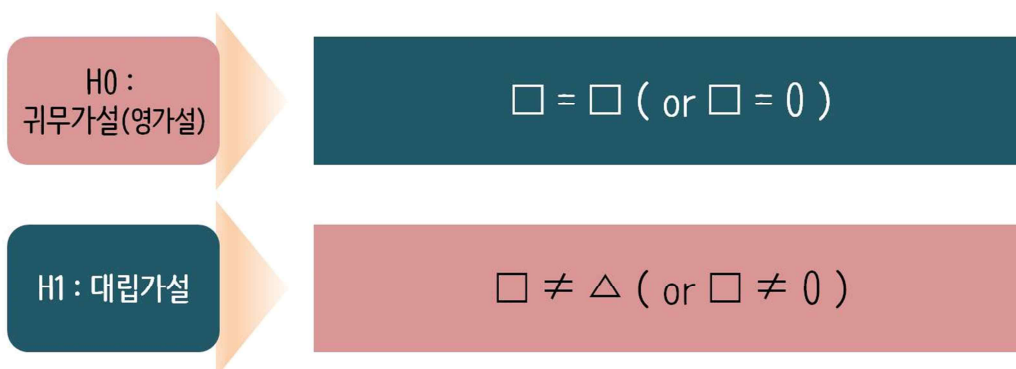


Flowchart를 보면 종속변수 콜레스테롤 서열 척도로 오른쪽이다.

독립변수가 연령을 연속형 변수인 등간이나 비율척도로 본다면 오른쪽이다.

다시 한 번 종속변수를 살펴보면 서열척도이기는 하지만 3수준 이상이라고 되어 있다. 200미만 정상, 200이상 위험군으로 2집단으로 구분을 했는데 이럴 때 통계학에서는 특별히 이분형 변수라고 부르며 분석 방법은 이분형 로지스틱 회귀분석이 된다.

## ■ 가설



가설은 H0 라는 귀무가설과 H1 이라는 대립가설로 되어 있다.

귀무가설과 대립가설은 설정하는 규칙이 있으며, 귀무가설의 경우 비교 대상이 있으며 같다, 비교 대상이 없으면 0이다와 같이 설정한다. 이에 반하여 대립가설은 다르다, 0이 아니므로 설정한다.

예를 들어 앞의 예 성별과 콜레스테롤의 경우를 살펴보면

남자라는 집단과 여자라는 집단이 있고, 각 집단의 콜레스테롤을 비교하는 것이다.

비교하는 값이 콜레스테롤은 연속형 변수이므로 평균과 표준편차를 사용할 수 있죠. 이때 우리가 관심 있는 것은 평균이 된다. 따라서 궁극적으로 남자 집단의 평균 콜레스테롤과 여자 집단의 평균 콜레스테롤을 비교하는 것이죠.

### ■ 통계분석

통계분석은 우리가 가지고 있는 데이터인 표본으로 모집단을 추정하는 것이다. 이때 추정할 때 가설이라는 것으로 사용하고 가설은  $H_0$ 와  $H_1$ 이 있는 것이다. 우리는 모집단 전체를 조사할 수 없기 때문에 두 집단의 평균이 같은지 다른지 알 수가 없다.

그래서 두 집단의 평균이 같을 경우인  $H_0$ 와 다를 경우인  $H_1$ 으로 나눈 것이고, 2개의 가설 중 어느 가설을 채택할 것인지 검정하는 것이고 바로 이것을 가설 검정이라고 한다.

가설 선택 \ 진실	진실	귀무가설 진실	대립가설 진실
$H_0$ 귀무가설 선택		옳은 결정 신뢰수준( $1-\alpha$ )	제 2종 오류 ( $\beta$ )
$H_1$ 대립가설 선택		제 1종 오류 유의수준( $\alpha$ )	옳은 결정 검정력( $1-\beta$ )

이때 가설 검정에서 문제가 발생할 수 있다.

$H_0$ 가 참인데  $H_0$ 를 선택하거나,  $H_1$ 이 참인데  $H_1$ 을 선택하는 경우에는 상관이 없죠. 옳게 결정한 것이다.

하지만  $H_0$ 가 참인데  $H_1$ 을 선택하거나 반대로  $H_1$ 이 참인데  $H_0$ 를 선택하는 경우 오류가 발생한다. 그리고 그 두 개의 오류를 각각 제 1종 오류와 제 2종 오류라고 한다.

그 동안 통계학에서는 제 1종 오류에 더 관심을 가지고 있었다.

제 1종 오류는 실제로는  $H_0$ 가 참인데  $H_0$ 를 기각하고  $H_1$ 을 잘못 선택할 확률, 즉 오류이다.

이때 오류는 작을수록 좋겠으며, 일반적 기준으로 5% 0.05를 주로 사용한다. 제 1종 오류의 최대 허용치를 유의수준이라고 한다. 즉 유의수준은 제 1종 오류를 범할 확률의 최대 허용값인 5%이다.

그리고 가설 검정인 통계분석에서 데이터들에서 실제 계산 되어진 값이 유의수준이며, 그 유의 수준을 보통 p-value라고 한다.

p-value는 유의수준 5% 보다 크다면 귀무가설을  $H_0$ 를 기각하지 못하는 것이고, 5%보다



도 작다면  $H_0$ 를 기각하기 때문에  $H_1$ 인 대립가설을 채택한다.

$H_1$  대립가설을 채택하면 유의한 또는 유의미한 차이가 있다. 라고 해석을 한다.