심혈관 질환에 영향을 주는 요인

2021320322 윤민서

2021.11.24

CONTENTS

1.1 탐구 동기와 목표

탐구 동기
한국인의 사망원인 중 암 다음으로 많은 비율을 차지하는 심혈관 질환을 예방하기 위해서 어떤 부분을 집중적으로 관리를 해야 할지 알아보기 위함

탐구 목표
신체적 요인, 신체에 영향을 미치는 요인들이 심혈관 질환의 발생 여부와 관련이 있을지 밝혀내는 것

1.2 데이터 설명

```
> sum(is.na(car))
[1] 0
```

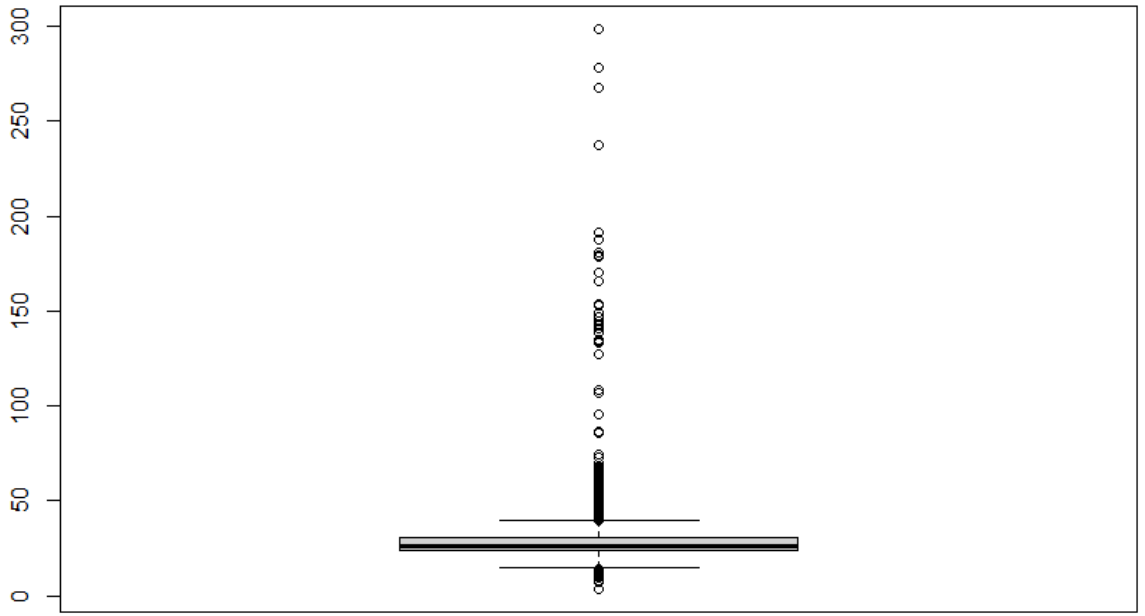
결측치 없음

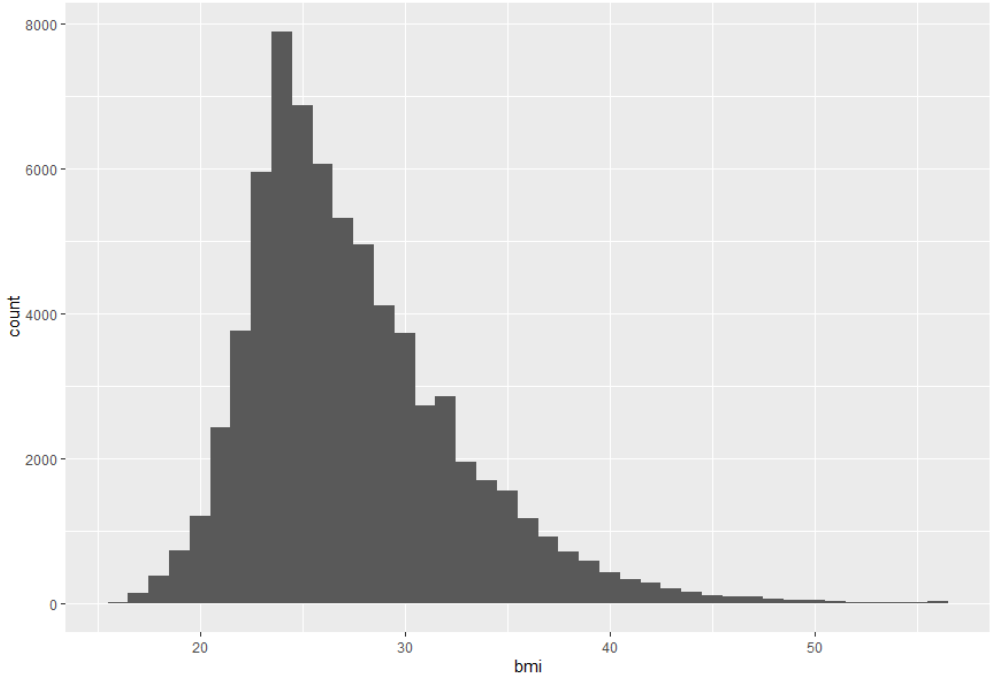
Kaggle에 기재된 Cardiovascular Disease Dataset

11개의 요소와 심혈관 질환 여부 데이터를
70000명의 환자로부터 수집

변수
age: 나이(일) gender: 성별, 1은 여자, 2는 남자
height: 키(cm) weight: 몸무게(kg)
ap_hi: 수축기 혈압 ap_lo: 확장기 혈압
cholesterol: 콜레스테롤, 1은 정상, 2는 정상보다 높음, 3은 정상보다 매우 높음
gluc: 포도당, 1은 정상, 2는 정상보다 높음, 3은 정상보다 매우 높음
smoke: 흡연 여부, alco: 음주 여부, active: 활동성, cardio: 심혈관 질환 여부

2.1 BMI 데이터 분포





weight / (height^2 / 10000)의 공식을 사용하여 bmi 열을 새로 만들어 주었고 전체적인 분포는 왼쪽 그림과 같다.
비현실적인 특잇값을 제거하기 위해 bmi로 정렬한 기준 상위 100개, 하위 100개의 데이터를 제거하였고
제거한 후의 bmi 분포는 오른쪽 그림과 같다.
예) bmi 최댓값의 경우 키 75cm, 몸무게 168kg으로 현실적으로 존재하기 어려운 수치이기 때문에 제거

2.2 BMI와 심혈관 질환

```
> car.0 <- car %>% filter(cardio == 0)
> car.1 <- car %>% filter(cardio == 1)
> t.test(car.1$bmi, car.0$bmi, alternative = "greater")

        Welch Two Sample t-test

data:  car.1$bmi and car.0$bmi
t = 51.65, df = 68368, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.916059      Inf
sample estimates:
mean of x mean of y
 28.49961  26.52052
```

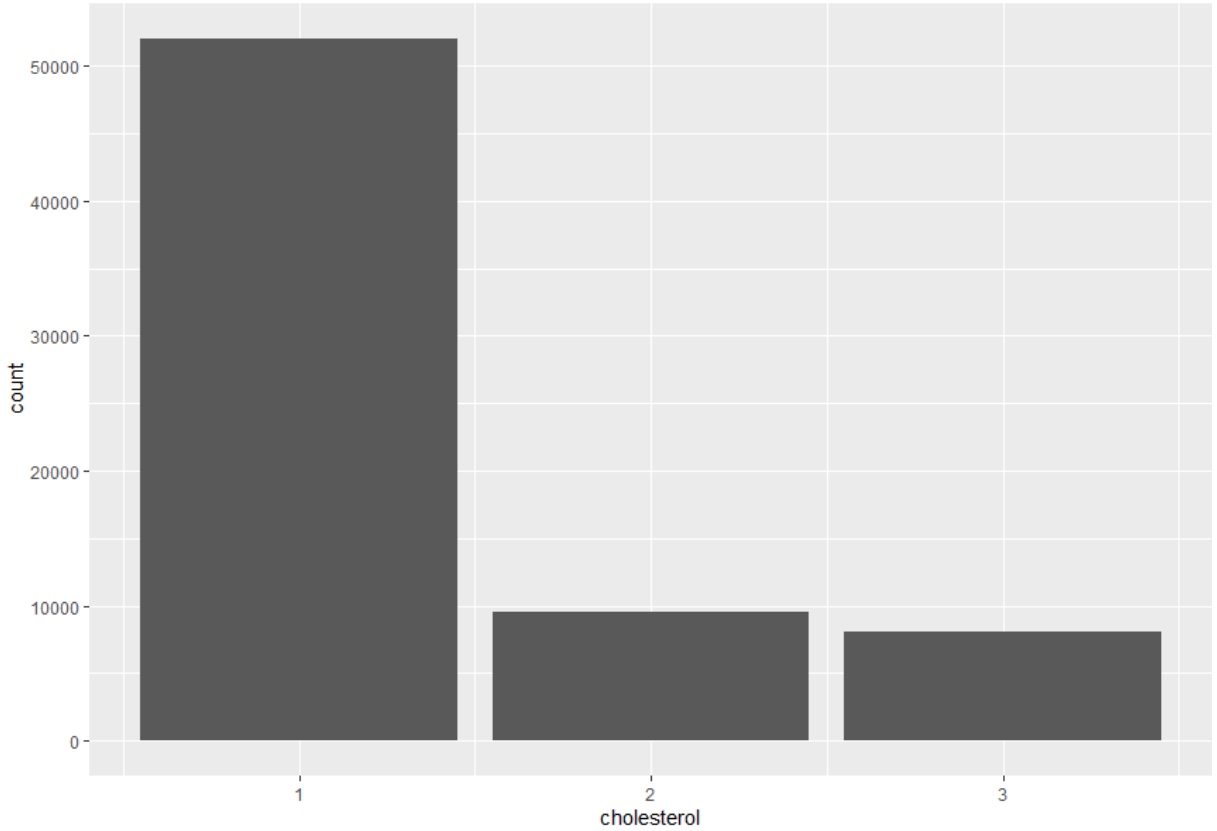
```
> var.test(car.1$bmi, car.0$bmi)

        F test to compare two variances

data:  car.1$bmi and car.0$bmi
F = 1.3156, num df = 34837, denom df = 34761, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.288216 1.343510
sample estimates:
ratio of variances
          1.315573
```

var.test 결과에 따라 분산은 다르다고 가정한다.

H_0 : 심혈관 질환이 있는 집단의 BMI 평균과 심혈관 질환이 없는 집단의 BMI 평균이 동일하다.
H_1 : 심혈관 질환이 있는 집단의 BMI 평균이 심혈관 질환이 없는 집단의 BMI 평균보다 크다.

심혈관 질환이 있는 집단(car.1)의 BMI 평균이 심혈관 질환이 없는 집단(car.0)의 BMI 평균보다 큰지 검정하기 위해 유의수준 5% t-test를 진행했다.
p-value < 2.2e-16으로 H_0를 기각할 수 있다.
즉 심혈관 질환이 있는 집단의 BMI 평균이 심혈관 질환이 없는 집단의 BMI 평균보다 크다.

3.1 콜레스테롤 분포



69800명 중
콜레스테롤 수치가 정상 범주인 사람이 52226명
정상보다 높은 범주의 사람이 9522명
정상보다 매우 높은 범주의 사람이 8052명

3.2 콜레스테롤과 심혈관 질환

```
> chol.1 <- car %>% filter(cholesterol == 1)
> sum(chol.1$cardio)
[1] 22995
> sum(chol.1$cholesterol) / 1
[1] 52226
> chol.2 <- car %>% filter(cholesterol == 2)
> sum(chol.2$cardio)
[1] 5735
> sum(chol.2$cholesterol) / 2
[1] 9522
> chol.3 <- car %>% filter(cholesterol == 3)
> sum(chol.3$cardio)
[1] 6162
> sum(chol.3$cholesterol) / 3
[1] 8052
```

콜레스테롤 기준에 따른 상대적인 수치에 따라서
1인 사람 52226명 중 22995명이 심혈관 질환을 가지고 있다.
2인 사람 9522명 중 5735명이 심혈관 질환을 가지고 있다.
3인 사람 8052명 중 6162명이 심혈관 질환을 가지고 있다.

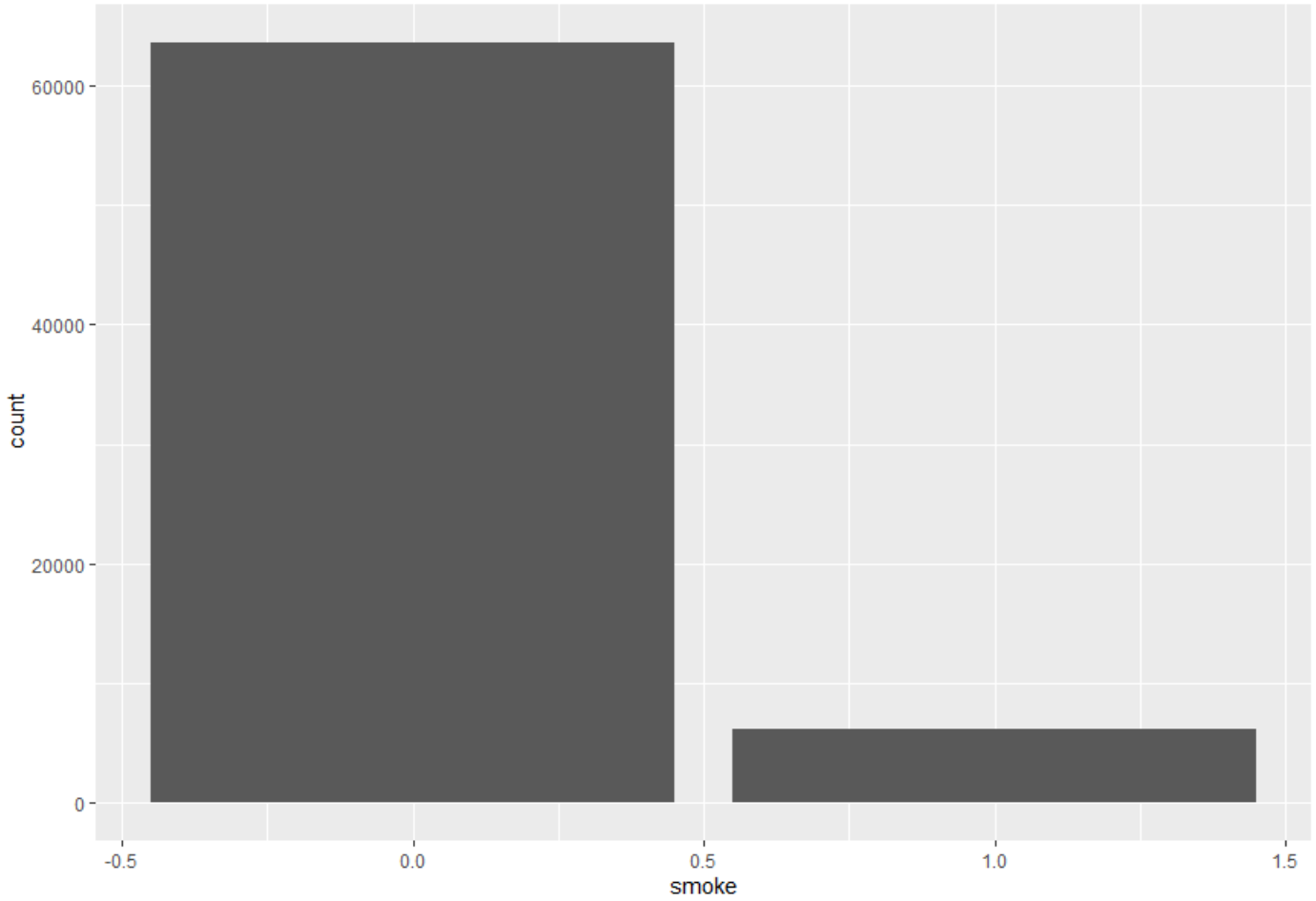
H_0 : 세 집단에서 심혈관 질환을 가지고 있는 사람의 비율이 같다.
H_1 : 세 집단에서 심혈관 질환을 가지고 있는 사람의 비율이 같지 않다.

```
> prop.test(c(22995, 5735, 6162), c(52226, 9522, 8052))

        3-sample test for equality of proportions without continuity correction

data:  c(22995, 5735, 6162) out of c(52226, 9522, 8052)
X-squared = 3409.6, df = 2, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.4402979 0.6022894 0.7652757
```

콜레스테롤 수치에 따른 심혈관 질환 보유의 비율을 검정하기 위해
유의수준 5%의 모비율 검정을 진행했다.
p-value < 2.2e-16으로 H_0를 기각할 수 있다.
콜레스테롤 상대 수치 1인 집단의 심혈관 질환 보유의 비율은 약 0.44,
2인 집단에서는 약 0.60, 3인 집단에서는 약 0.77로 측정됐다.

prop.test 함수 특성 상 집단이 3개 이상이기 때문에 양측 검정으로 진행했다.
콜레스테롤 수치가 높을수록 심혈관 질환에 걸릴 확률이 높아진다고 할 수 있다.

4.1 흡연 분포



69800명 중
비흡연자 63645명
흡연자 6155명

4.2 흡연과 심혈관 질환

```
> smoke.0 <- car %>% filter(smoke == 0)
> sum(smoke.0$cardio)
[1] 31968
> 69800 - sum(car$smoke)
[1] 63645
> smoke.1 <- car %>% filter(smoke == 1)
> sum(smoke.1$cardio)
[1] 2924
> sum(car$smoke)
[1] 6155
```

비흡연자 63645명 중 31968명이 심혈관 질환을 가지고 있다.
흡연자 6155명 중 2924명이 심혈관 질환을 가지고 있다.

```
> prop.test(c(31968, 2924), c(63645, 6155), alternative = "less")

        2-sample test for equality of proportions with continuity correction

data:  c(31968, 2924) out of c(63645, 6155)
X-squared = 16.531, df = 1, p-value = 1
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.03827996
sample estimates:
   prop 1    prop 2
0.5022861 0.4750609
```

H_0 : 비흡연자 중 심혈관 질환을 가지고 있는 사람의 비율이
흡연자 중 심혈관 질환을 가지고 있는 사람의 비율과 같다.
H_1 : 비흡연자 중 심혈관 질환을 가지고 있는 사람의 비율이
흡연자 중 심혈관 질환을 가지고 있는 사람의 비율보다 작다.

흡연 여부에 따른 심혈관 질환 보유의 비율을 검정하기 위해
유의수준 5%의 모비율 검정을 진행했다.
p-value = 1으로 H_0를 기각할 수 없다.
비흡연자 집단의 심혈관 질환 보유의 비율은 약 0.50,
흡연자 집단의 심혈관 질환 보유의 비율은 약 0.48로 측정됐다.

단순히 흡연 여부만을 따졌을 때 흡연이 심혈관 질환의 위험성을 높인다고 단정지을 수 없다.

5 결론

요약
- BMI가 높을수록 심혈관 질환에 걸릴 가능성이 높아진다.
- 콜레스테롤 수치가 높을수록 심혈관 질환에 걸릴 가능성이 높아진다.
- 흡연을 한다고 해서 심혈관 질환에 걸릴 가능성이 높아진다고 단정지을 수 없다.

토의
- 흡연자가 비흡연자에 비해 심장마비 등으로 사망할 위험이 크다고 흔히 알려져 있다.
  하지만 위와 같은 결과가 도출된 이유는 무엇일까?
- 심혈관 질환 이외에 폐암, 구강암 등의 다른 요인으로 인해 일찍 사망한 사람 등을 고려해야 할 것이다.
- 다른 외부적 요인에는 무엇이 있을까?

시사점
- 심혈관 질환을 예방하기 위해서 지속적인 체중 관리가 필요하고 육식 위주의 식단 역시 심혈관 질환 예방에 좋지 않으며 지나친 흡연 또한 자제해야 한다.

※ 참고문헌 및 부록

Dataset: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
Cardiovascular Disease dataset
The dataset consists of 70 000 records of patients data, 11 features + target.
Svetlana Ulianova

Font
한글: 나눔스퀘어라운드 Regular (본문)
영어, 숫자, 특수문자: Arial (본문)

R Code

```
#기본 설정
setwd("C:/Users/윤민서/OneDrive/Desktop/통계계산프로그래밍")
library(dplyr)
library(ggplot2)

# 결측치 확인
car <- read.csv("data/cardio_train.csv", header = T, sep = ';')
car <- car[ ,-1]
sum(is.na(car))

# BMI와 심혈관 질환
car <- car %>% mutate(bmi = weight / (height^2 / 10000)) %>% arrange(bmi)
boxplot(car$bmi)
car <- car[101:69900,]
ggplot(car, aes(x = bmi)) + geom_histogram(binwidth = 1)
car.0 <- car %>% filter(cardio == 0)
car.1 <- car %>% filter(cardio == 1)
var.test(car.1$bmi, car.0$bmi)
t.test(car.1$bmi, car.0$bmi, alternative = "greater")

#콜레스테롤과 심혈관 질환
ggplot(car, aes(x = cholesterol)) + geom_bar()
chol.1 <- car %>% filter(cholesterol == 1)
sum(chol.1$cardio)
sum(chol.1$cholesterol) / 1
chol.2 <- car %>% filter(cholesterol == 2)
sum(chol.2$cardio)
sum(chol.2$cholesterol) / 2
chol.3 <- car %>% filter(cholesterol == 3)
sum(chol.3$cardio)
sum(chol.3$cholesterol) / 3
prop.test(c(22995, 5735, 6162), c(52226, 9522, 8052))

#흡연과 심혈관 질환
ggplot(car, aes(x = smoke)) + geom_bar()
smoke.0 <- car %>% filter(smoke == 0)
sum(smoke.0$cardio)
69800 - sum(car$smoke)
smoke.1 <- car %>% filter(smoke == 1)
sum(smoke.1$cardio)
sum(car$smoke)
prop.test(c(31968, 2924), c(63645, 6155), alternative = "less")
```