통계적 가설 검정이란?

[참조] 쉽게 배우는 R 데이터 분석: 13장(P299)

기술통계(Descriptive Statistics)
데이터를 요약해 설명하는 통계 기법
사람들이 받는 월급을 집계한 전체 월급 평균

추론통계(Inferential Statistics)
단순히 숫자를 요약하는 것을 넘어 어떤 값이 발생할 확률을 계산하는 통계 기법
수집된 데이터에서 성별에 따라 월급에 차이가 있는 것으로 나타났을 때, 이런 차이가 **우연히** 발생할 확률을 계산하여 이런 차이가 우연히 나타날 확률이 작다면 성별에 따른 월급 차이가 **통계적으로 유의하다(Statistically Significant)고 결론** 내린다
반대로 이런 차이가 **우연히** 나타날 확률이 크다면 성별에 따른 월급 차이가 **통계적으로 유의하지(not statistically significant) 않다고 결론** 내린다.

유의확률의 필요성
일반적으로 통계 분석을 수행해다는 것은 **추론 통계를 이용해 가설검정을** 했다는 의미이다.
기술 통계 분석에서 집단 간 차이가 있는 것으로 나타났더라도 이는 **우연에 의한 차이**일 수 있다.
데이터를 이용해 신뢰할 수 있는 결론을 내리려면 **유의확률을 계산하는 통계적 가설 검정** 절차를 거쳐야 한다.

통계적 가설 검정
유의 확률을 이용해 가설을 검정하는 방법을 통계적 가설 검정(Statistical hypothesis test)이라 한다.
유의확률(Significance probability, p-value)은 실제로는 집단 간 차이가 없는 우연히 차이가 있는 데이터가 추출될 확률을 의미한다.
통계 분석을 실시한 결과 **유의확률이 크게 나타났다면** '**집단 간 차이가 통계적으로 유의하지 한다**'고 해석한다. 이는 실제로 차이가 없더라도 우연에 의해 이 정도의 차이가 관찰 될 가능성이 크다는 의미이다.
반대로 **유의 확률이 작다면** '**집단 간 차이가 통계적으로 유의하다**'고 해석한다. 이는 실제로 차이가 없는데 우연히 이 정도의 차이가 관찰될 가능성이 작다. 즉 **우연이라고 보기 힘들다**는 의미이다.

t 검정(t-test)

t 검정은 두 집단의 평균에 통계적으로 유의한 차이가 있는지 알아볼 때 사용하는 통계 기법이다.

R언어 : t.test()

t검정은 비교하는 **집단의 분산**(값이 퍼져 있는 정도)이 같은지 여부에 따라 적용하는 공식이 다르다. 여기서 집단 간 분산이 같다고 가정하고 **var.equal에 T**를 지정했다.

cty: 도시연비
class(자동자 종류)

(R 코드)

```
mpg <- as.data.frame(ggplot2::mpg)
library(dplyr)

mpg_diff <- mpg %>%
  select(class, cty) %>%
  filter(class %in% c("compact", "suv"))

table(msg_diff$class)

t.test(data=mpg_diff, cty ~ class, var.equal = T)
```

(R 결과)

```
p-value < 2.2e-16
sample estimates:
mean in group compact: 20.12766
mean in group suv: 13.50000
```

유의확률(p-value)이 5%를 판단 기준으로 삼고, p-value가 0.05 미만이면 '집단 간 차이가 통계적으로 유의하다'고 해석한다.

p-value(2.2e-16) 0.05보다 작기 때문에 '**compact와 suv 간 평균 도시 연비차이가 통계적으로 유의하다**'고 해석할 수 있다.

일반휘발유(Regular). 고급 휘발유(Premium)를 사용하는 자동차 간 도시 연비 차이가 통계적으로 유의한가?

(R 코드)

```
mpg <- as.data.frame(ggplot2::mpg)
library(dplyr)

mpg_diff2 <- mpg %>%
  select(f1, cty) %>%
  filter(f1 %in% c("r", "p")) # r:regular, p:premium

table(msg_diff2$f1)

t.test(data=mpg_diff2, cty ~ f1, var.equal = T)
```

(R 결과)

```
p-value = 0.2875
sample estimates:
mean in group p: 17.36538
mean in group r: 16.73810
```

유의확률(p-value)이 5%를 판단 기준으로 삼고, p-value가 0.05 미만이면 '집단 간 차이가 통계적으로 유의하다'고 해석한다.

p-value(0.2875)로서 0.05보다 크다.
실제로는 차이가 없는데 우연에 의해 이런 차이가 관찰될 확률이 28.75%라는 의미이다.
따라서 '**일반 휘발유와 고급 휘발유를 사용하는 자동차 간 도시 연비 차이가 통계적으로 유의하지 않다**'고 결론 내릴 수 있다.
고급 휘발유 자동차의 도시 연비 평균이 0.6 정도 높지만 이런 정도의 차이는 유연히 발생했을 가능성이 크다고 해석하는 것이다.