DF - 스터디 모임 3차

Chapter 0. 스터디 일정

> 일시: 2018.09.19
> 장소: 미정
> 주제: 독립표본 T-Test 검정에 관한 기본 통계이론 리뷰

Chapter 1. Intro

- 통계에서 평균을 비교하는 방법에는 크게 4가지가 있음

> z-test
> t-test (single sample) / 단일표본
> t-test (dependent) / 대응표본
> t-test (independent) / 독립표본

1. 통계에서 말하는 z와 t는 무엇을 말할까?

$$z = \frac{Observed \ - \ Expected}{Standard \ Error(SE)}$$

$$t = \frac{Observed \ - \ Expected}{Standard \ Error(SE)}$$

2. 언제 z-test 와 t-test를 사용해야할까?

구분	모평균과 샘플 평균 비교?	전체 모수의 표준편차를 아는가?
z-test	YES	YES
Single sample t-test	YES	NO

- 빅데이터라고 해서 전체 모수를 알 수 있는 건 아님. 따라서 “표본”의 의미를 정확히 이해하고 가는 것이 중요

3. Dependent T-Test VS Independent T-Test

(1) Dependent t-test

- 같은 그룹을 두번 측정하는 경우 평균의 변화 및 차이점 확인

(2) Independent t-test

- 두 독립적인 그룹간의 차이점 확인
- E.g. men vs women, drug vs placebo, iphone vs Android, AWS vs GCP

4. Observed, Expected, and SE

구분	Observed	Expected	SE

z-test	샘플 평균	모수 평균	평균의 표준오차
단일표본	샘플 평균	모수 평균	평균의 표준오차
대응표본	차이에 대한 샘플 평균	차이에 대한 모평균	평균 차이의 오차
독립표본	두 그룹간 표본 평균의 차이	두 그룹간 모 평균의 차이	두 그룹간 평균 차이의 오차

- From DataCamp

구분	Observed	Expected	SE
z-test	Sample Mean	Population Mean	SE of the mean
단일표본	Sample Mean	Population Mean	SE of the mean
대응표본	Sample mean of difference scores	Population mean of difference scores	SE of the mean difference
독립표본	Difference between two sample means	Difference between two population means	SE of the difference between means

5.1. Sampling Distribution (1)

- Hypothetical Distribution of Summary Statistic
- Multiple Samples of the same size
- e.g. 샘플링 에러 leads to slightly different means
- Distribution of sample means will have a mean to close to population mean
- In the case of z-test,
    - Mean of sampling distribution equal to mean of population
    - SD of sampling distribution is the SE

$$Standard\ Error = \frac{Population\ SD}{Sq.root\ of\ sample\ size}$$

5.2. Sampling Distribution (2)

- Significance Tests
- Null Hypothesis Significance Testing (NHST)
    - Null Hypothesis = “no effect” or “no difference”
    - 샘플과 모수와의 평균 차이가 없음 (Expected Value = 0)

$$Standard\ Error = \frac{Population\ SD}{Sq.\,root\ of\ sample\ size}$$

6.1. Probability Values

- Conditional Probability (assumes null hypothesis is true)
- As or more extreme than observed value

구분	Expected	Actual	p-value
z-value	0	2	< 0.05

- P-value with t-tests
    - For large samples:
        - Very similar to sampling distribution of z-test
        - t-value of 2 yields statistically significant result
    - For small samples:
        - Sampling distribution wider
        - t-value slightly greater than 2 required

7. 질문

> 특정 그룹 사람들이 지능 테스트를 어떻게 잘 수행하는지 알고 싶습니다. 당신이 이 특정 지역의 사람들이 그 지역의 모든 사람들의 평균 지능보다 높은지 아닌지를 알고 싶습니다. 영가설은 “특정 지역 사람들의 지능과 전체 사람들의 지능에는 차이가 없다”입니다.
>
> 당신은 특정 그룹에서 10 명의 피실험자 중 무작위로 표본을 추출했는데 10명의 실험결과, 전체 지역의 표준오차보다 1.5 높은 것을 알게 되었습니다. 신뢰도 95% 구간 안에서 당신은 어떻게 결론을 내리겠습니까? (Recall that the critical value for the z-distribution at a significance level of 5% is 1.96.)

- [ ] 결론을 내리기에는 정보가 부족하다
- [ ] 이 특정 그룹의 평균 IQ는 전체 그룹의 IQ와 유의하게 다르지 않다
- [ ] 이 특정 그룹의 평균 IQ는 전체 그룹의 IQ와 유의하게 다르다
- [ ] 이 특정 그룹의 평균 IQ는 전체 그룹의 평균 IQ와 동일하다

- 참고

$$z = \frac{M + 1.5 \times SE - M}{SE} = \frac{1.5 \times SE}{SE} = 1.5$$

정답은

- [x] 이 특정 그룹의 평균 IQ는 전체 그룹의 IQ와 유의하게 다르지 않다

이유는?

Chapter 2. T-Distribution에 대한 이해

1. t-value 구하는 공식

$$t = \frac{X - M}{SE}$$

> X is the observed value, M is the expected value under the null hypothesis (or population mean), and SE is the standard error. Once you've computed the t-statistic, you then compare it to the so-called critical value, which comes from the relevant t-distribution.
> The shape of a t-distribution, and thus the critical value, is determined entirely by its degrees of freedom. To demonstrate this, let's draw some density plots for t-distributions using different degrees of freedom.

2. t-distribution 그래프 및 해석

```
Sys.setlocale("LC_ALL", "en_US.UTF-8")
# install.packages("extrafont")
library(extrafont) ## package 불러오기!
font_import() ## 설치된 모든 폰트 가져오기
par(family="AppleGothic")

# t 분포 그래프 함수화
t_distribution <- function (df1 = df1, df2 = df2, df3 = df3) {

# Generate a vector of 100 values between -4 and 4
x <- seq(-4, 4, length = 1000)

# Simulate the t-distribution
y_1 <- dt(x, df = df1)
y_2 <- dt(x, df = df2)
y_3 <- dt(x, df = df3)

# Plot the t-distributions
plot(x, y_1, type = "l", lwd = 2, xlab = "t-value", ylab = "Densi
ty",
       main = paste0("t-분포의 비교, df = ", as.character(df3)), col
= "black", ylim = c(0, 0.4))
lines(x, y_2, col = "red")
```

```
  lines(x, y_3, col = "blue")

  # Add a legend
  legend("topright", c(paste0("df = ", as.character(df1)),
                       paste0("df = ", as.character(df2)),
                       paste0("df = ", as.character(df3))
                       ),
           col = c("black", "red", "blue"),
           title = "t-분포", lty = 1)
    return(print("Done!!"))
  }

  # 동적 비교 위해 for-loop 구현
  t_lengths <- c(6, 10, 20, 30, 40, 50)
  for (i in t_lengths) {
    t_distribution(df1 = 2, df2 = 6, df3 = i)
    Sys.sleep(3)
  }

  # df3 그래프가 어떻게 움직이는지 확인하시고, 그 이유가 무엇인지 해석해보세요~
```

Chapter 3. Dependent t-test

1. 기본이론

- 대응표본 t-test
- 같은 그룹을 두번 테스트 한다 (A/B 테스트, 광고 전후 실적 비교 등)
- 실험 전후로 차이 Score 확인
- 차이에 대해 평균 (ex. 예시)

Subject	pre	post	score
x1	0	3	3-0 = 3
x2	1	1	1-1 = 0

xn	xn.value1	xn.value2	xn.value2 - xn.value1
			average(x1 + x2 + .. + xn)

- 3가지 분석을 참고해야 함
- t-value, p-value, & Cohen's d (effect size)
- 대응표본 t-value
  t = (Observed - Expected) / SE
    - Observed: Mean of difference scores
    - Expected: Zero ("no effect")

- 효과의 크기 (Cohen's D)
    - 유의성 검사 시 샘플 크기에 대해 편향적일 수 있음
    - 효과의 크기에 대해 표준화 된 지표를 말하며
    - 효과크기가 0이라는 것은 비교집단들 사이의 차이(혹은 연관성)가 없다는 것을 의미
    - 예를 들면, 남녀의 비율이 53:47이라고 하면 효과크기는 3% or 평균 IQ가 100 / IQ 105 효과크기는 5임
    - 좀더 쉬운 예시를 들면, 남녀가 결혼을 해야 하는데, 100명의 인구 안에 53명은 남자 47명은 여자 남자 3명은 결혼 못함 <— 우울함!! 남자와 여자의 두 독립적인 그룹이 서로 연관성이 있다는 있는데 3만큼 있다.

> 실험수행방법
>
> - step 1. 귀무가설과 대립가설 수립 (실험전후간 그룹 간 차이 없음)
> - step 2. p-value 유의수준 level α 정의 (95% or 99%)
> - step 3. 관측된 t-value 구하기
> - step 4. 결정계수(Critical Value) 찾기 (Find)
> - step 5. 관측된 t-value와 critical value 비교하기
> - 이 실험을 수행하기 위한 결정계수는 0.05 (1.96, 95% 신뢰구간)

2. observed t-value 구하는 공식

$$t = \frac{x_D}{s_D \div \sqrt{n}}$$

> n = 샘플
> x_D는 전후 실험 스코어 차이값의 평균
> s_D는 전후 실험 스코어 차이값에 대한 표준편차

3. 표준편차 구하는 공식

$$s_D = \sqrt{\frac{\sum (x_D - \bar{x}_D)^2}{n-1}}$$

> 여기에서 s_D & x_D는 실험전후 각각의 스코어 (Individual difference scores)
> $\bar{x}_D$는 the mean of the difference scores

4. R 스크립트로 구현하기

```r
library(psych)
wm <- read.csv("wm.csv")

# Take a look at the dataset
wm

# Create training subset of wm
wm_t <- subset(wm, wm$train == 1)

# Summary statistics
describe(wm_t)

# Create a boxplot with pre- and post-training groups
boxplot(wm_t$pre, wm_t$post, main = "Boxplot",
        xlab = "Pre- and Post-Training", ylab = "Intelligence Sco
```

```
re",
        col = c("red", "green"))

## The training subset, wm_t,
wm_t

# Define the sample size
n <- nrow(wm_t)

# Mean of the difference scores
mean_diff <- sum(wm_t$gain) / n

# Standard deviation of the difference scores
sd_diff <- sqrt(sum((mean_diff - wm_t$gain)^2) / (n-1))

# Observed t-value
t_obs <- mean_diff / (sd_diff / sqrt(n))

# Print observed t-value
t_obs

# Compute the critical value
t_crit <- qt(0.975, df = 79)

# Print the critical value
t_crit

# Print the observed t-value to compare
t_obs

# Compute Cohen's d
cohens_d <- mean_diff / sd_diff

# View Cohen's d
cohens_d
```

- 대응표본의 결과, 통계적으로 유의한가?

> t_crit의 값은 1.99045
> t_obs의 값은 14.49238

(5) R function 소개

```
# Apply the t.test function
t.test(wm_t$post, wm_t$pre, paired = TRUE)

# Calculate Cohen's d
cohensD(wm_t$post, wm_t$pre, method = "paired")
# 결과값 1.6은
```

- 해석

```
    Paired t-test

data:  wm_t$post and wm_t$pre
t = 14.492, df = 79, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 3.008511 3.966489
sample estimates:
mean of the differences
                 3.4875
```

Chapter 4. Independent t-test

(1) 기본이론

- 두개의 독립된 그룹으로부터 평균을 비교하는 것

- 예시
    - 남자 VS 여자
    - Treatment vs control
    - Patient vs healthy

- 실험방법
    - Training and Control Groups
    - Pre-test / post-test design
    - Increase in intelligence after training?

- t-value를 계산하는 방법
    - t = (Observed - Expected) / SE
    - t = (M1-M2) / SE
    - 그룹간 차이에 대한 표준오차
        - SE = (SE1 + SE2) / 2
        - SE는 샘플링 에러에 대한 평균적인 값

- 대응표본과 마찬가지로 t-test는 샘플의 크기에 따라 편향적일 수 있음(biased)
- Cohen's D

$$d = \frac{M1 - M2}{SD_{pooled}}$$

$$SD_{pooled} = \frac{(SD_1 + SD_2)}{2}$$

(2) 등분산 가정 (Variance Assumption)

- Homogeneity(동종) of Variance Assumption (분산 가정)
- 만약에 그룹간 분산이 똑같다면 SD_{pooled} 만 적정할 것임
- 그렇지 않으면, SE, sampling distribution, 그리고 p-value 모두 부적합함
- Levene's test
    - 그룹간 분산을 비교

- 만약에 유의한 결과가 나올 시, 등분산성은 위배될 것임
- 반대로, 유의한 결과가 나오지 않는다면, 중요한 정보가 데이터에 있음 (t-test 결과는 유의함)
- 유의하다는 건, 두 그룹간 평균의 차이가 있음

```
# View the wm_t dataset
wm_t

# Create subsets for each training time
wm_t08 <- subset(wm_t, wm_t$cond == "t08")
wm_t12 <- subset(wm_t, wm_t$cond == "t12")
wm_t17 <- subset(wm_t, wm_t$cond == "t17")
wm_t19 <- subset(wm_t, wm_t$cond == "t19")

# Summary statistics for the change in training scores before and
after training
describe(wm_t08)
describe(wm_t12)
describe(wm_t17)
describe(wm_t19)

# Create a boxplot of the different training times
ggplot(wm_t, aes(x = cond, y = gain, fill = cond)) + geom_boxplot
()

# Levene's test
leveneTest(wm_t$gain ~ wm_t$cond)
```

(3) 독립표본 t-test에서 t-value 구하는 공식

> (1) $\bar{x}_1$, $\bar{x}_2$의 값은 그룹 1, 그룹 2의 평균

$$t = \frac{(\bar{x}_1 + \bar{x}_2)}{se_p}$$

(2) se_p는 se_{pooled} 전체 표준오차(standard error)이며 구하는 공식은 아래와 같음

$$se_p = \sqrt{\frac{var_1}{n_1} + \frac{var_2}{n_2}}$$

(4) R Script 공식 구하기

```
# View the wm_t dataset
wm_t

# Create subsets for each training time
wm_t08 <- subset(wm_t, wm_t$cond == "t08")
wm_t12 <- subset(wm_t, wm_t$cond == "t12")
wm_t17 <- subset(wm_t, wm_t$cond == "t17")
wm_t19 <- subset(wm_t, wm_t$cond == "t19")

# Summary statistics for the change in training scores before and
after training
describe(wm_t08)
describe(wm_t12)
describe(wm_t17)
describe(wm_t19)

# Create a boxplot of the different training times
# install.packages("tidyverse", dependencies = T)
library(tidyverse)
ggplot(wm_t, aes(x = cond, y = gain, fill = cond)) + geom_boxplot
()

# Levene's test
# install.packages("car", dependencies = T)
library(car)
leveneTest(wm_t$gain ~ wm_t$cond)
```

```r
# The subsets wm_t08 and wm_t19 are still loaded in the console

# Find the mean intelligence gain for both the 8 and 19 training
day group
mean_t08 <- mean(wm_t08$gain)
mean_t19 <- mean(wm_t19$gain)

mean_t08
mean_t19
# Calculate mean difference by subtracting t08 by t19
mean_diff <- mean_t19 - mean_t08

# Determine the number of subjects in each sample
n_t08 <- nrow(wm_t08)
n_t19 <- nrow(wm_t19)

# Calculate degrees of freedom
df <- (n_t08 + n_t19) - 2

# Calculate variance for each group
var_t08 <- var(wm_t08$gain)
var_t19 <- var(wm_t19$gain)

# Compute pooled standard error
se_pooled <- sqrt(var_t08/n_t08 + var_t19/n_t19)
se_pooled

## All variables from the previous exercises are preloaded in you
r workspace

# Calculate the t-value
t_value <- mean_diff / se_pooled
t_value

# Calculate p-value
p_value <- 2 * (1 - pt(t_value, df = df))

# Calculate standard deviations
sd_t08 <- sd(wm_t08$gain)
sd_t19 <- sd(wm_t19$gain)
```

```
# Calculate the pooled stanard deviation
pooled_sd <- (sd_t08 + sd_t19) / 2

# Calculate Cohen's d
cohens_d <- (mean_t19 - mean_t08) / pooled_sd
```

- p_value의 값: 6.443468e-11 해석은?? 통계적으로 유의한가?
- cohens_d의 값: 2.876648, 해석은?? 효과의 크기는 작은가?

(5) R function 소개

```
# Conduct an independent t-test
t.test(wm_t08$gain, wm_t19$gain, var.equal = TRUE)

# Calculate Cohen's d
cohensD(wm_t19$gain, wm_t08$gain, method = "pooled")
```

Chapter 5. 쉬운 예제 적용 (독립 표본 t-test)

```
#### independent t-test example ####
# Data in two numeric vectors
women_weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8,
48.5)
men_weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
# Create a data frame
my_data <- data.frame(
  group = rep(c("Woman", "Man"), each = 9),
  weight = c(women_weight,  men_weight)
)

# Print all data
print(my_data)
```

```r
# 연구주제는 다음과 같음
# 실제로 여성의 몸무게가 남성의 몸무게와 다른가? 비교하고 싶음
group_by(weight, group) %>%
  summarise(
    count = n(),
    mean = mean(weight, na.rm = TRUE),
    sd = sd(weight, na.rm = TRUE)
  )

ggplot(my_data, aes(x = group, y = weight, colour = group)) + geo
m_boxplot()

# t-test를 하기 위해서는 몇가지 사전 검증이 필요함
# 1. 남자와 여자가 독립성을 이루는가?
# 첫번째 대답은?
# 2. 두 그룹 모두 정규분포를 이루는가? (Normal Distribution)
# shapiro.test()

# Shapiro-Wilk normality test for Men's weights
with(my_data, shapiro.test(weight[group == "Man"]))# p = 0.1

# Shapiro-Wilk normality test for Women's weights
with(my_data, shapiro.test(weight[group == "Woman"])) # p = 0.6

# 결론, 두 그룹 모두 정규성으로부터 통계적으로 유의하게 다르지 않다 = 두 그룹 모두
통계적으로 정규분포를 따른다

# 3. 두 그룹 모두 등분산성을 이루는가?
leveneTest(weight ~ group, data = my_data)
# p-value가 0.05보다 높다는 것은 두 그룹 분산사이의 차이가 통계적으로 유의하게 다
르지 않다 = 두 그룹 모두 통계적으로 등분산을 이룬다.

# 위 3가지 조건을 만족할 때 t.test를 사용할 수 있다.
# 만약 그렇지 않다면? 검색해보세요~~

# Compute t-test
res <- t.test(weight ~ group, data = my_data, var.equal = TRUE)
res

# t는 t-value (t = 2.784),
```

```
# df is 자유도 (df= 16) (왜 16일까???)
# p-value is the significance level of the t-test (p-value = 0.01
327).
# conf.int is 평균의 신뢰구간 95% (conf.int = [4.0298, 29.748]).
```