출석수업 과제물(평가결과물) 표지(온라인제출용)

**교과목명 : R 데이터 분석 (2022년 2학기)**

**학 번 : 202135-368864**

**성 명 : 홍원표**

**강 의 실 : 경기-성남 (화상강의)**

**연 락 처 : 010-5343-4341**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

- 이하 과제 작성

**R datarium 패키지에 내장된 performance 데이터셋은 남녀 60명의 성 별(gender), 스트레스 그룹(stress), 연구 시작 시점에서의 성과점수(t1), 연구 종료 시점에서의 성과점수(t2)의 데이터를 포함하고 있다. 이 데이 터를 이용하여 다음의 문항에 답하시오.**

library(datarium)  
library(tidyverse)

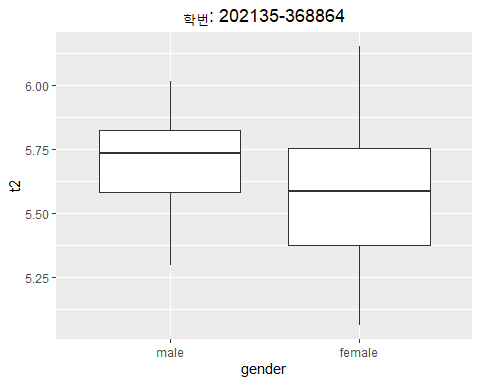
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

performance %>%   
 tail

## # A tibble: 6 x 5  
## id gender stress t1 t2  
## <int> <fct> <fct> <dbl> <dbl>  
## 1 55 female high 5.27 5.25  
## 2 56 female high 5.37 5.77  
## 3 57 female high 5.49 5.62  
## 4 58 female high 5.77 5.43  
## 5 59 female high 5.58 5.33  
## 6 60 female high 5.42 5.08

**1. 연구 종료 시점에서의 성과점수(t2)의 상자그림을 성별로 나란히 그리 시오. 그래프의 제목으로 본인의 학번을 넣으시오. (3점)**

ggplot(data=performance) +  
 geom\_boxplot(aes(x=gender, y=t2)) + # 상자그림을 그린다.  
 labs(title="학번: 202135-368864") + # 타이틀을 출력한다.  
 theme(plot.title = element\_text(hjust = 0.5))# 타이틀 중간으로 이동



**2. 성별에 따라 연구 종료 시점에서의 성과점수(t2)의 평균에 차이가 있 는지 검정하시오. (3점)**

# t 검정 이용  
t.test(t2~gender,data = performance)

##   
## Welch Two Sample t-test  
##   
## data: t2 by gender  
## t = 2.0417, df = 47.396, p-value = 0.04677  
## alternative hypothesis: true difference in means between group male and group female is not equal to 0  
## 95 percent confidence interval:  
## 0.001829799 0.243899622  
## sample estimates:  
## mean in group male mean in group female   
## 5.701607 5.578743

* p-value가(0.04677)로 **0.05**보다 작으므로 귀무가설을 기각한다. 성별에 따른 성과점수(t2)의 평균에 유의미한 차이가 있다고 할 수 있다.

**3. 스트레스 그룹 별로 연구 종료 시점에서의 성과점수(t2)의 평균을 구 하시오. (3점)**

summary(performance)

## id gender stress t1 t2   
## Min. : 1.00 male :30 low :20 Min. :5.233 Min. :5.061   
## 1st Qu.:15.75 female:30 moderate:20 1st Qu.:5.506 1st Qu.:5.480   
## Median :30.50 high :20 Median :5.608 Median :5.676   
## Mean :30.50 Mean :5.648 Mean :5.640   
## 3rd Qu.:45.25 3rd Qu.:5.766 3rd Qu.:5.795   
## Max. :60.00 Max. :6.147 Max. :6.151

(dat3<-performance %>%  
 group\_by(stress) %>% # 스트레스 그룹별   
 summarise(평균 = mean(t2, na.rm=TRUE))) # 그룹별로 평균 요약

## # A tibble: 3 x 2  
## stress 평균  
## <fct> <dbl>  
## 1 low 5.74  
## 2 moderate 5.69  
## 3 high 5.49

**4. 스트레스 그룹 별로 연구 종료 시점에서의 성과점수(t2)의 평균에 차 이가 있는지 검정하시오. (3점)**

# 일원배치분산분석 - oneway.test() 함수를 사용  
oneway.test(t2~stress, data=performance)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: t2 and stress  
## F = 6.9315, num df = 2.000, denom df = 36.262, p-value = 0.002823

* p-value = 0.002823가 **0.05**보다 작기 때문에 귀무가설을 기각하고 대립가설을 채택한다.  
  성과점수(t2)의 평균에 차이가 있다고 할 수 있다.

**5. 연구 종료 시점에서의 성과점수(t2) 값에 따라 ①5.5 이하, ② 5.5 초 과 5.8 이하, ③ 5.8 초과 세 그룹으로 나누는 범주형 변수를 만드시 오. (3점)**

# mutate와 cut 함수를 이용하여 구한다.  
# t2의 최소값은 5.061이고 최대값은 6.151이므로 5 ~ 7구간에 모두 포함된다.  
(dat4<-performance %>%   
 mutate(t2.group=as.factor(cut(t2, breaks=c(5, 5.5, 5.8, 7)))) )

## # A tibble: 60 x 6  
## id gender stress t1 t2 t2.group   
## <int> <fct> <fct> <dbl> <dbl> <fct>   
## 1 1 male low 5.96 5.58 (5.5,5.8]  
## 2 2 male low 5.51 5.82 (5.8,7]   
## 3 3 male low 5.63 5.47 (5,5.5]   
## 4 4 male low 5.71 5.79 (5.5,5.8]  
## 5 5 male low 5.74 5.72 (5.5,5.8]  
## 6 6 male low 5.62 5.68 (5.5,5.8]  
## 7 7 male low 5.48 5.60 (5.5,5.8]  
## 8 8 male low 5.64 5.92 (5.8,7]   
## 9 9 male low 6.06 5.84 (5.8,7]   
## 10 10 male low 5.86 5.57 (5.5,5.8]  
## # ... with 50 more rows

**6. 5번에서 만든 범주형 변수와 스트레스 그룹 변수 간에 연관성이 있는 지 검정하시오. (3점)**

chisq.test(dat4$t2.group, as.factor(performance$stress))

##   
## Pearson's Chi-squared test  
##   
## data: dat4$t2.group and as.factor(performance$stress)  
## X-squared = 13.627, df = 4, p-value = 0.008587

* p-value = 0.008587 이므로 **0.05**보다 작기 때문에 귀무가설을 기각하고 대립가설을 채택한다.  
  : 성과점수(t2)와 스트레스 그룹 변수는 연관이 없다.  
  : 성과점수(t2)와 스트레스 그룹 변수는 연관이 있다.  
  따라서 성과점수(t2)와 스트레스 그룹 변수간에는 연관성이 있다.

**7. 연구 종료 시점에서의 성과점수(t2)를 종속변수로, 연구 시작 시점에 서의 성과점수(t1)를 독립변수로 하는 단순선형회귀분석을 수행하여 회귀직선의 기울기와 절편을 구하시오. (3점)**

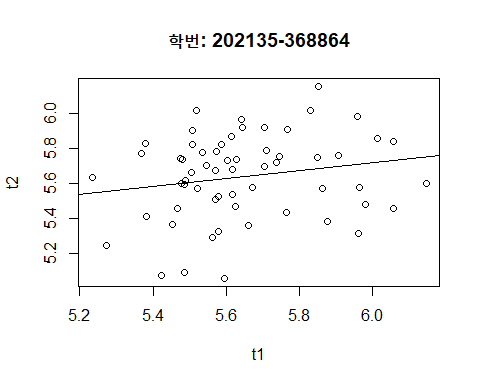
# 선형회귀분석  
obj = lm(t2~t1, data=performance)  
summary(obj)

##   
## Call:  
## lm(formula = t2 ~ t1, data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56676 -0.15440 0.05065 0.15849 0.46519   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.3801 0.8568 5.112 3.74e-06 \*\*\*  
## t1 0.2231 0.1516 1.472 0.147   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2369 on 58 degrees of freedom  
## Multiple R-squared: 0.036, Adjusted R-squared: 0.01938   
## F-statistic: 2.166 on 1 and 58 DF, p-value: 0.1465

* 회기식 :
  + 기울기: 0.2231
  + 절편 : 4.3801

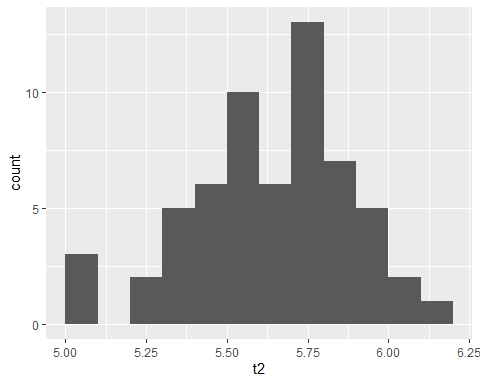
**8. 연구 종료 시점에서의 성과점수(t2)를 세로축으로, 연구 시작 시점에 서의 성과점수(t1)를 가로축으로 하는 산점도를 그리고, 그 위에 7번 에서 구한 회귀직선을 그리시오. 그래프의 제목으로 본인의 학번을 넣으시오. (3점)**

plot(t2~t1, data = performance, main="학번: 202135-368864")  
abline(obj)



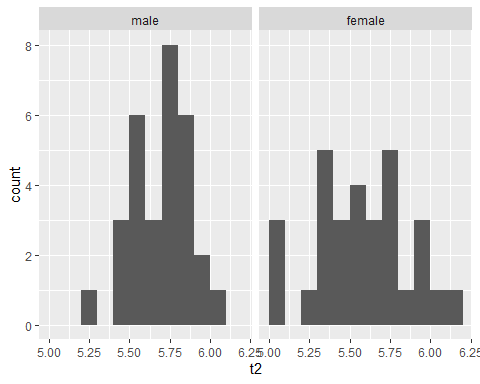
1. 연구 종료 시점에서의 성과점수(t2)의 히스토그램을 그리시오. (3점)

ggplot(data=performance) +  
 geom\_histogram(aes(x=t2), breaks=seq(5.0, 6.2, 0.1)) # t2의 최소값:5.061, 최대값:6.151



1. 연구 종료 시점에서의 성과점수(t2)의 히스토그램을, 성별로 따로 그 리시오. (3점)

# facet\_wrap() 함수를 붙여서 나눠서 그려준다.  
ggplot(data=performance) +  
 geom\_histogram(aes(x=t2), breaks=seq(5.0, 6.2, 0.1)) +  
 facet\_wrap(~gender, nrow=1) # 성별로 나누어 그래프를 출력력



-- 2022년 2학기 R 데이터 분석 출석수업 과제 끝 --