<경태분 | 함수 총정리> (문자 : chr / 범주 : factor)

ok

· 데이터 파일이름 <— read_csv("file.csv", col_names = T, locale=locale('ko', encoding='euc-kr')) : file을 데이터 파일이름으로 불러와서 저장하기
☆ 우선 library(readr) 먼저하기

View(데이터) : 데이터 보여줌

str(데이터 or 데이터$변수) : 데이터 또는 데이터의 변수에 대해 다양한 정보 보여줌

summary(데이터) : 범주형척도에 대한 개별변수에 대한 수치적인 정보보여줌, NA가 어떤변수에 있는지 알려줌

데이터$변수 <— as.factor(데이터$변수) : 해당변수의 척도를 범주형으로 변경

glimpse(데이터) : 데이터의 다양한 정보 보여줌 (척도를 한눈에)
☆ 우선 library(dplyr) 먼저하기

table(데이터$변수) : 해당변수의 빈도수

freq(데이터$변수) : 해당변수의빈도수와 막대그래프
☆ 우선 library(descr) 먼저하기

qplot(data= 데이터, 변수) : 해당변수의 빈도수를 막대그래프로 표현

qplot(data= 데이터, 변수1, fill=변수2) : 변수1과 변수2의 조합된 빈도수를 막대그래프로 표현
(fill=변수2 → 색상구분)
☆ 우선 library(ggplot2) 먼저하기

• 변수에는 " " 안씀

mean(데이터$변수) : 해당변수의 평균

var(데이터$변수) : 해당변수의 분산

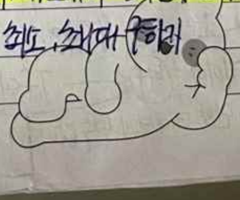
sd(데이터$변수) : 해당변수의 표준편차

mean/var/sd (데이터$변수, na.rm=T) : 데이터 변수에서 결측치(NA)를 제외하고 계산

freq(is.na(데이터$변수)) : 데이터의 변수에서 결측치가 있는지 여부를 구하고 있으면 빈도수 T로 표현
☆ 우선 library(descr) 먼저하기 / 즉 NA=결측치가 몇개인지

describe(데이터 or 데이터$변수) : 해당 데이터 또는 변수의 해당하는 정규분포표의 수치
(skew-왜도, kurtosis-첨도 ....)
☆ 우선 library(psych) 먼저하기

hist(데이터$변수, breaks = seq(N1, N2, K)) : 해당변수를 N1~N2까지 간격을 K인 히스토그램
· 변수의 척도가 수치형 / N1, N2는 summary(데이터$변수) 통해 최소, 최대 구하기



같다	==
같지않다	!=
크다	>
작다	<
이상	>=
이하	<=
그리고	&
또는	| (shift + ₩)

<패키지모음>

library(readr)
library(dplyr)
library(descr)
library(ggplot2)
library(psych)

table(데이터$변수 %in% c("a", "b", "c")) : 데이터의 변수가 a, b, c 중에 적어도 하나의 조건에 만족하는것들의 빈도수, (문자가 아니면 " " 필요X)

- | 연산자 사용과 동일

데이터$A <—— B : B라는값을 데이터의 새로운변수 'A'에 저장
└ 함수이거나 어떤데이터의 변수여도 OK

데이터$A <—— NULL : 데이터의 A라는변수삭제

weekdays(데이터$변수) : 해당변수(날짜)의 요일이 무엇인지 보여줌
└ 척도가 날짜(date) 여야함

데이터$변수 <—— as.Date(데이터$변수) : 해당변수의 척도를 날짜(date)로 변경

ex) weekdays(as.Date("2020-01-01")) : 수요일

★ ex) weather$요일 <—— weekdays(weather$일시) : weather의 변수 '일시'의 요일을 우한 값들은 데이터 weather의 새로운 변수 '요일'에 지정

데이터$변수 <—— factor(데이터$변수, levels = c("a", "b", "c"....))
(┌ filter와 같은 함수랑 사용가능)
: a, b, c순으로 변수를 배열하고 척도를 범주형으로 바꾼뒤 해당데이터의 변수에 지정

데이터 <—— 데이터 %>% rename(바꿀변수명 = 기존변수명) : 변수명 변경하기
└ ★ 우선 library(dplyr) 먼저하기 / " " 안써도됨

데이터$변수 <—— ifelse(데이터$변수 == "a", "b", 데이터$변수) : 만약 데이터의 변수값이 a이면 b로 바꾸고 아니면 데이터의 원래 변수값 그대로 두어라
└ 같은변수가 아니어도 OK

weather의 평균기압이 NA일경우는
is.na(weather$평균기압) 이지 weather$평균기압 == NA는 오답
is.na(데이터$변수): 해당변수가 결측치(NA)이다
" " 안써도됨, 단 사례에 문자는 " " 쓰기

★ 데이터전처리 함수 ★ 우선 library(dplyr) 먼저하기
데이터 %>% filter(변수 == 사례): 데이터 안에있는 변수에서 해당 사례만 추출
⊕ var 나 mean과 같은 기술통계함수를 이용하기위해서는 해당사례들만
추출한것들을 새로운 데이터에 assign해야함
round(어떤값, digits=k): 해당값에서 소수 k 번째 자리까지
└ 함수를이용해서 숫자값그대로 붙기
→ filter 함수안에 다양한 여러변수들을 기준으로 조건에 맞게설정가능
ex) exam %>% filter(class != 4, math >= 90 | history >= 95)
(4반이 아닌학생중에 수학이 90점이상이거나 역사가 95점이상)

데이터 %>% filter(변수 >= quantile(변수, probs = c(k))): 데이터의 해당변수값이 (변수의 값에서 k인값)
변수의 값중에서 k인값보다 이상인사례만 추출 (즉 k가 0.9이면 0.9이상인값, 변수가 상위 10%인
사례들만 추출 (변수의 값이 상위 1-k 인값들만추출)
데이터 %>% select(변수): 해당변수들만 추출 (여러변수들을 , 로 연결해서 추출가능)
데이터 %>% select(-변수): 해당변수를 제외하고 추출
데이터 %>% select(contains("특정단어")): 특정단어가 포함된 변수추출 ("특정단어"를 연쇄으로 X 연산자 | 사용하기)
print(n=Inf): 밑에 행까지 다보여주기
데이터 %>% arrange(변수): 변수에 대해서 오름차순정렬
데이터 %>% arrange(-변수): 변수에 대해서 내림차순정렬 (큰순서) } 두변수를 같이 사용가능 arrange(class, -math)
└ exam$math의 형태가아니라 math, -math 변수 그대로
└ summarise 쓸때 변수명 지정해야 arrange 이용가능
(반별로 오름차순 정렬한것들에서 수학점수 내림차순해서 정렬하기)

데이터 <- 데이터 %>% mutate(새변수 = □): □라는 과정을거친
새변수를 데이터에 지정, mutate함수로 동시에 여러변수생성가능, 이과정에
주로 ifelse 함수가 쓰이는데 ifelse 안에 또 ifelse가 들어갈수 있음
(변수가 n개로 구분되면 ifelse를 (n-1)회 사용)

☆ 데이터 <- 데이터 %>% mutate(새변수 = case_when(기준변수 < k ~ " ", 기준변수 > k ~ " "
ex) exam <- exam %>% mutate(test = case_when(total < 180 ~ "fail",
total ≥ 180 ~ "pass")
※ 변수측정값에 NA가 있다면 TRUE ~ 조건이 아닌 total ≥ 180 ~ "pass" 처럼 세로조건이용해서 쓰

데이터 ← 데이터 %>% relocate(변수, .before(혹은 .after)=변수2)
: 변수를 변수2 왼쪽아 오른쪽에 이동시키기
☆ ① .before: 변수를 변수2 왼쪽에(앞)
② .after: 변수를 변수2 오른쪽에(뒤)

데이터 ← 데이터 %>% relocate(where(is.character 혹은 is.factor))
: 문자형or혹은 범주형척도를 맨앞으로이동하기
데이터 ← 데이터 %>% relocate(where(is.factor), .before=where(is.character))
: 범주형 척도변수를 문자형 척도 변수 앞으로 이동하기

데이터 %>% group_by(변수) %>% summarise(mean 아Var아 sd ···)
: 변수별로 그룹을 나누고 기술통계량을 조건에 맞게 제시
☆ summarise을 보여주는 결과값 앞에 변수명 지정가능 ex) mean_math=mean(math)
mutate는 summarise을 이용하는 count=n() 빈도이용한 코드사용가능
[변명 함수코드]
count/sum(count): 이미 기준이된 변수로 그룹이 나누어진 '사례들끼리'의 전체 갯수에서 해당사례 비율
count/sum(데이터$count): 모든사례들(전체) 갯수에서 해당사례비율

n_distinct(데이터$변수): 변수의척도라는 당연하게 중복되는 값이아닌 고유의값이 몇가지 종류가 있는가 (변수에해당하는 사례가 중복없지 않고 몇가지나 있나?)

c(a:b): a부터 b까지가 연속된 숫자들의 배열
ex) c(1:7): 1~7
relocate(변수): 데이터상에서 변수가 제일앞으로
ex) relocate(ID): 데이터상에서 'ID'라는 변수가 제일앞으로

데이터 ← left_join(데이터, 데이터2, by="변수"): 데이터와 데이터 2을 데이터에 공통변수을 이용해서합침
데이터 ← left_join(데이터, 데이터2, by=c("변수1"="변수2"))
: 공통변수의 이름이 다를때 변수1의 이름과 변수2의 이름이 같다고 지정한뒤 데이터에 합치기

데이터 ← bind_rows(데이터, 데이터2) : 데이터와 데이터2을 합쳐서 사례수가 .
단, 변수명이 일치해도 척도가 다르면 오류발생 (척도가 동일해야함) / 모든변수척도동일
내용상 동일한 변수여도 변수명이 다르면 다른변수로 인식 / 모든변수 변수명동일
⇒ 1) 변수명 일치하게변경 : rename 함수
2) 척도 일치하게 변경 : as.factor 함수

데이터 ← 데이터 %>% distinct(변수, .keep_all=T) : 해당변수가 같은 사례들중에 한개만 남기고 다 제외하기

☆ ex) ID가 중복변수일때 동일한 사례인지 확인

```
exam %>% group_by(ID) %>% summarise(count=n()) %>%
arrange(-count)
```

↳ 결과가 ID가 30인것이 count가 2가 나옴 즉 동일된것이니까 distinct 써서 지워야함

데이터[a:b] : 데이터에서 a~b까지의 변수순서
ex) exam[5:8] : exam에서 5번째 ~ 8번째의 변수들 (5열에서 8열까지)

데이터$mean ± 2.57583x데이터$sd : 이기준을넘어가면 이상치처리

<t-검정하기>

① 데이터 프레임만들기

데이터파일이름 ← read.csv("file.csv")

② 보기좋게 변수위치조정 : 범주형 → 수치형 → 문자형

○ 데이터 ← 데이터 %>% relocate 함수이용

③ 이상치 검토및 빈도수확인

```
descr <- describe(ttest[6:10])
descr <- descr %>% mutate(LL=mean-2xsd, UL=mean+2xsd)
```

④ 이상치를제외한 데이터프레임형성

ex) ttest_new ← ttest %>% filter(변수≤k ······)
└ UL값

1) 가설수립
Ho : $M_{HO} = M_{CS}$ (귀무가설)
Ha : $M_{HO} \neq M_{CS}$ (대립가설)] 양측검정

2) 집단별로 데이터프레임 만들기
ex) Test_Ho <- Test_new %>% filter(customer == "Home office")
Test_CS <- Test_new %>% filter(customer == "Consumer")

3) 정규성 조건검토 (정규분포의 형태를 띠는가)
summary, hist 함수이용해서 모양확인
ex) summary(Test_ho$sales)
hist(Test_ho$sales, breaks=seq(0, 9000, 50))
=> 두 데이터 다 왼쪽으로 치우쳐진 히스토그램 모양이라 정규분포가 아님
shapiro.test(데이터$변수) : 데이터의 해당하는 변수에 대해 P-value구하기
*** P-value > 0.05 즉 P-value가 유의하지 않아야 정규성 조건을 만족하는데
위 함수를 이용한결과 P-value < 0.05 (α) 이기에 유의해서 정규성 조건에 부합X
=> 정규성 조건에 부합하게 왜곡된 sales를 자연로그로 변환
ex) Test_ho <- Test_ho %>% mutate(lnsales=log(sales))
└ 이후 hist, shapiro.test 이용하여 정규성 조건검토
└ 여기서 P-value값이 여전히 0에 가까워도 OK

4) 등분산성 조건검토
var.test 함수이용해서 P-value ≤ 0.05 면 이분산 t-검정 P-value ≥ 0.05 면 등분산 t-검정
var.test(Test_ho$lnsales, Test_CS$lnsales)
└ • F = 1.3083으로 1에 가까우니 표본분산이 비슷하고 P-value = 0.5536 > 0.05 (α) 이니 등분산

5) 독립표본 t-검정 일서 및 가설검정
t.test(Test_ho$lnsales, Test_CS$lnsales, alternative = "two.sided", var.equal=(F))
└ 만약 4)과정에서 이분산이면 F
계산하면 P-value = 0.2601 이라 0.05 (α) 보다 크고
mean of x (Test_ho$lnsales) = 5.976144
mean of y (Test_CS$lnsales) = 5.941587
이니 표본평균이 비슷하고 P-value도 유의하지 않고 Ho = $M_{HO} = M_{CS}$ 의 귀무가설채택

<고사범위> 0,05

one-way ANOVA - 일원분산분석.

<기말범위 패키지 모음>

```
library(readr)
library(dplyr)
library(psych)
library(forcats)
library(agricolae)
library(dunn.test)
library(car)
library(HH)
```

└> 한개번이하면 오류뜨니까 차근히하나씩

1단계: 가설도입

○ 이상치제거후 anova_new 데이터 프레임 만들기

```
library(psych)
descr <- describe(anova1$price)
descr <- descr %>% mutate(UL=mean+2xsd, LL=mean-2xsd)
anova1_new <- anova1 %>% filter(price <=descr$UL)
```

○ 새로운변수 price 만들기 → ctiptical과 high 통합

```
install.packages("forcats")
library(forcats)
anova_new <- anova_new %>% mutate(price = fct_collapse(price,
"high"=c("critical", "high")))
```

└ 새로운변수 (price)

○ 두가지 가설 수립

H_0: 4가지집단의 price 모평균이 같다

H_a: 적어도 한집단의 price 모평균은 다른집단과 다르다.

2단계: 집단간 데이터프레임 생성하기

```
anova1_H <- anova1_new %>% filter(price == "high")
```

* 각 price의 변수 4개 단의 프레임 동일하게 생성 *

3단계: 종속변수 정규성검토

```
summary(anova1_H$price)
hist(anova1_H$price, breaks=seq(0,600,10))
shapiro.test(anova1_H$price)
```

└> p-value 가 0 이하라 유의해서 정규성조건에부합하지 않지만 n의크기가 많아서 조건에 만족할것이라고 가정하고 이후분석진행

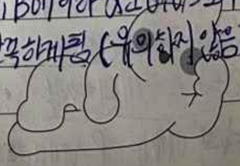
* 모든데이터프레임에서 다확인 *

4단계: 등분산성검토

```
install.packages("car")
library(car)
leveneTest(price~price, data=anova_new)
```

price → 종속변수, price → 독립변수, anova_new → 전체데이터프레임

☆ Pr(>F) = 0.1307이라 α인 0.05보다 커서 등분산조건을 만족하게됨 (유의하지 않음)



☆ 일원분산에서 대립이나 귀무냐 그냥 P-value와 α값 비교 (F분포)

5단계: ANOVA 검정 실시 (등분산성 조건이 만족할 때) /

```
anova_result <- aov(price~prior, data=anova_new)
summary(anova_result)
```

P-value가 0.133이니 α인 0.05가 크므로
P-value가 유의하지 않고 귀무가설을 채택
즉 4가지의 집단의 price의 모평균이 모두 같다.

만약 귀무가설이 아니라 실습2처럼 대립가설 채택이면 사후검정실시 (등분산조건에서 대립가설채택일 때)

6단계: 사후분석 Duncan test 실시

```
install.packages("agricolae")
library(agricolae)
duncan.test(anova2_result, "payment", console=T)
```

anova2_result: aov함수이용데이터 / "payment": 독립변수

- 모평균이 큰 집단부터 표시하고 같은 group으로 나타난 집단끼리의 모평균은 같다고 지정
- 신용카드 expense 모평균 > 계좌이체 expense 모평균 = 간편결제 expense 모평균 %

만약 α가 0.01이 아니라 0.05이면 5단계에서 등분산이 아니라 이분산조건을 실시해야함

```
oneway.test(expense~payment, data=anova2_new)
```

- P-value가 0에 가까워서 대립가설채택

이분산조건 사후분석 dunn.test 실시 (이분산조건에서 대립가설을 채택할 때)

```
install.packages("dunn.test")
library(dunn.test)
dunn.test(anova2_new$expense, anova2_new$payment, method="bonferroni")
```

anova2_new$expense: 데이터$종속변수 / anova2_new$payment: 데이터$독립변수

- 간편결제와 계좌이체의 P-value가 1이다. α/2보다 커서 유의하지 않고 모평균이 동일하다.
- 간편결제와 신용카드의 P-value가 0에 가까우니 α/2보다 작아서 유의하고 모평균 동일하지 않은데 신용카드가 더 크다 (간편결제 수치인 신용카드와 대응할 때 -7.301456이니)
  └ (계좌이체 - 신용카드 = -7.301456)
  └ 통계량값

계좌이체 = 간편결제
신용카드 > 간편결제
신용카드 > 계좌이체 %

☆ 여기서는 P-value와 α/2 (=0.025)와 비교)

two-way ANOVA 이원분산분석

1단계 : 가설수립

ㅇ 가설수립

Ho : 두 독립변수 간에 상호작용효과가 없다.

Ha : 두 독립변수 간에 상호작용효과가 있다.

ㅇ 이상치 검토 및 제거

```
library(psych)
descr <- describe(two_anova$expense)
descr <- descr %>% mutate(UL=mean+3*sd, LL=mean-3*sd)
two_anova_new <- two_anova %>% filter(expense <= descr$UL)
```

2단계 : 서브데이터 프레임 만들기

```
two_anova_male <- two_anova_new %>% filter(gender == "Male")
```

각 gender의 변수인 Female도 같이 만들어주기

3단계 : 정규성검토

```
summary(two_anova_male$expense)
hist(two_anova_male$expense, breaks= seq(0,2000,40))
shapiro.test(two_anova_male$expense)
```

→ P-value가 0보다 작아 유의해서 정규성 조건의 만족하기 힘들지만 사례가 많기에 (n) 그대로 진행

4단계 : 등분산성검토

```
library(car)
leveneTest(expense~gender, data=two_anova_new)
```

→ P-value가 0.001238이라 α인 0.05보다 작아서 유의함 등분산 조건을 만족하지 못한다.

5단계 : 이분산 가정 one way anova 시행

```
oneway.test(expense~gender, data=two_anova_new)
```

→ P-value가 0에 가까우기에 귀무가설을 기각하고 대립가설 채택, gender에 따라 집단을 두 개로 구분했을 때 종속변수 expense의 평균은 차이가 있다. → mfemale > mmale (553 > 359) ←

```
two_anova_new %>% group_by(gender) %>% summarise(mean(expense))
```

7단계: two-away anovA 시행 및 그래프

1) two_anova_result ← aov(expense ~ gender × os, data = two_anova_new)
└두 독립변수의 곱　　　　　전체데이터

summary(two_anova_result)
↳ 변수 gender의 P-value가 1.138e-12 ... 이므로 0보다 작고 유의하고 (5단계)
gender × os P-value가 0.0003172 ... 이므로 0보다 작고 유의하기에 (α=0.05보다 작고)
(Ha: 두 독립변수 간에 상호작용 효과가 있다) [대립가설 채택]
☆ 즉 IV1과 IV1과 IV2의 곱에 따른 P-value가 둘다 유의해야 하는 것이 조건 ☆.

2) two_anova_new$gender ← factor(two_anova_new$gender, levels = c("Male", "Female"))
↳ (one-way anovA에서 표본평균이 작은 순서대로 집단을 먼저 출력해야 함
따라서 작은 Male이 먼저 출력되게 순서 바꾸기)

install.packages("HH")
library(HH)
interaction2wt(expense ~ gender × os, data = two_anova_new)
↳ os와 무관하게 Male보다는 female일때 expense의 평균값이 증가한다.
↳ os가 IOS일때 Male에서 female로 갈때 expense의 평균차이가 더 커진다 (기울기 더 급함)
os가 Indroid일때 Male에서 female로 갈때 expense의 평균차이가 작다 (기울기 완만)
↳ 상호작용의 정도 차이

8단계: 추가분석) 집단을 네개로 구분해서 세분화한 뒤 one-way AnovA
☆ 집단 네개 MA, MI, FA, FI (남성, 여성 / Indroid, IOS)
two_anova_new ← two_anova_new %>% mutate(genderos = ifelse(gender == "Male" & os == "Android", "MA", ifelse(gender == "Male" & os == "IOS", "MI", ifelse(gender == "Female" & os == "Android", "FA", "FI"))))
☆ ┌ mutate와 ifelse 함수써서 새로운 변수 genderos 만드는데 네개의 집단 기준
└ 이후 one-way ANovA 1단계 ~ 6단계 실시

일원분산분석과 이원분산분석에서 대립가설, 귀무가설에서 어떤게 채택인지 확인하는 것은 P-value와 α (유의수준) 간에 비교하는 것이고. 이후 얼마만큼 차이가 있는지는 사후분석에서 ☆ P-value vs α이거나 P-value vs α/2 사이를 비교하면 된다.

(New)

모비율 비교와 독립성검정

모비율 차이에 대한 가설검정

① 데이터프레임 불러오기

```
library(readr)
proptest <- read_csv("proptest.csv")
```

② 빈도수 기준 교차표만들기

```
table(proptest$customer == 20 & proptest$trans == "Yes") = (48)
```

* customer와 trans의 모든 변수의 조합끼리 마해주기 *

행 \ 열	Yes	No	
20대	48	49	(97)
30대	30	73	(103)

```
prop <- matrix(c(48,49,30,73), nrow=2, ncol=2, byrow=T)
```

└ 행의 갯수 / └ 열의 갯수 / └ 행을 기준으로 나열

```
rownames(prop) <- c(20,30)
colnames(prop) <- c("Yes","No")
```

↳ 각 행과 열의 이름을 다시 지정해주기

③ 비율 기준 교차표만들기

```
prop.table(prop, margin=1)
```

└ prop 이라는 빈도수로 되있는것을 비율로 바꿔주는데 행마다 값의 합이 1로 해야

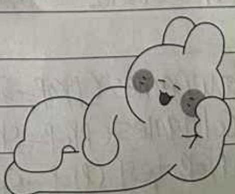
④ 가설검정

- $H_0: p_{20} - p_{30} = 0$
- $H_a: p_{20} - p_{30} \neq 0$

$\overline{p_{20}} - \overline{p_{30}}$ 이 정규분포를 그리기위한 조건인 np & $n(1-p) \geq 5$ 인지 확인을 우선해야함

- $97 \times 0.495 = 48.015 > 5$
- $97 \times 0.505 = 48.985 > 5$
- $103 \times 0.291 = 29.973 > 5$
- $103 \times 0.709 = 73.027 > 5$

모든경우가 5보다 크니 $\overline{p_{20}} - \overline{p_{30}}$ 은 정규분포를 그린다



prop.test(prop, alternative="two.sided", correct=T)
→ p-value의 값이 0.005031인데 α인 0.05보다 작으니 즉 유의하므로 대립가설 채택
(양측검정이므로 p20 > p30 (대립가설 채택)) prop1 0.1494845 > prop2 0.1291262

다양한 검정 가설검정

① 데이터프레임불러오기 (각각 과거, 현재에서 변수 check)
library(readr)
telecom ← read_csv("telecom.csv")
table(telecom$past) = skt:120 / kt:100 / LGt 80
table(telecom$current) : skt:149 / kt:85 / LGt 66

② 첫번째 가설검정: 과거시점에 3사의 시장점유율이 동일한지
chisq.test(c(120, 100, 80))
└ 각 변수들 넣기
→ p-value = 0.01832로 α인 0.05보다 작게 나왔으니 대립가설채택 즉
과거의 이동통신 3사의 시장점유율은 서로 같지않다.

③ 두번째 가설검정: 현재시점에 3사의 시장점유율이 동일한지
chisq.test(c(149, 85, 66))
→ p-value = 6.8e-09 이므로 α인 0.05보다 작게 나왔으니
마찬가지로 현재의 이동통신 3사의 시장점유율은 서로 같지않다.

④ 세번째 가설검정: 과거 시장점유율과 현재 시장점유율의 동일한지여부
chisq.test(c(149, 85, 66), p=c(0.4, 1/3, 4/15))
└ 현재 빈도수 └ 과거시점 해당통신사개수 / 과거시점 전체개수) ex) 200/500 (KT통신사 비율)
→ p-value = 0.002868이라 α인 0.05보다 작게 나왔으니 대립가설 채택하고
과거에 비해서 이동통신 3사의 시장점유율은 변화되었다.

⑤ 세번째 가설검정에 대한 사후분석
prop_skt ← matrix(c(120, 180, 149, 151), nrow=2, ncol=2, byrow=T)
└ skt가아닌 나머지 통신사들의 과거, 현재의 개수합
rownames(prop_skt) ← c("Past", "Current")
colnames(prop_skt) ← c("Skt", "not Skt")
prop.test(prop_skt, alternative="two.sided", correct=T) 과거보다 시장점유율이 더

올해 KT와 LG+도 똑같은 방식으로 해주면 되는데 Prop_KT는 2P-value=0.2154
로 α인 0.05보다 크기 때문에 유의하지 않고 KT의 시장점유율은 과거와 현재 비교해 변화가 없음
Prop_LG+도 2P-value가 0.2116이라 α인 0.05보다 크기 때문에 유의하지 않고
이 역시도 LG+의 시장점유율은 과거와 현재 비교해 변화가 없음.

독립성검정

① 데이터불러오기

library(readr)

tennis ← read_csv("tennis.csv")

② 데이터 전처리

1) names(tennis) ← tolower(names(tennis)) : tennis 데이터의 변수들의 이름을 다 소문자로
└변수들만 보여주는함수 └소문자 변경함수

2) tennis $surface ← tolower(tennis $surface) : tennis의 변수의 속성값을 소문자로

★ 독립변수 surface와 아까가지고 종속변수 result에드해주기★

3) surface의 속성값 clay와 clay (T)을 합치고 hard와 hard (T)을 합치기

a. library(forcats)
tennis $surface ← fct_collapse(tennis $surface, "clay" = c("clay", "clay (T)"))

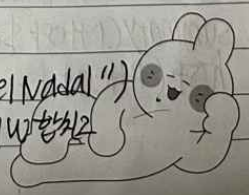
b. library(forcats)
library(dplyr)
tennis ← tennis %>% mutate(surface=fct_collapse(surface, "clay"=c("clay", "clay (T)")))

★ 변수의 속성값 hard, hard(T)도 같은방식으로합쳐줘★ / 두 방법아무거나 (a,b 중에 골라써넣음)

만약 surface의 속성값중에서 결측치가 존재하면 filter 이용해서 제거하고 새로운 데이터 만들기
tennis_new ← tennis %>% filter(!is.na(surface))

4) player가 Rafael Nadal인것만 추출해서 데이터만들기
tennis_nadal ← tennis_new %>% filter(player == "Rafael Nadal")

5) 위 3)의 코드를 이용해서 nadal의 result의 속성값중에서 l, lr, lw 합치고 w, wr, ww합치기



③ 독립성검정

(독립변수: surface, 종속변수: result)

```
xtabs(~surface+result, tennis_nadal)
xtabs(~surface+result, tennis_nadal), margin=1)
prop.table(xtabs(~surface+result, tennis_nadal))
chisq.test(xtabs(~surface+result, tennis_nadal))
```

└ p-value가 2.47e-09 이므로 사실상 0이고 surface와 result는 독립적이지 않다고하는 대립가설채택, 즉, surface가 result에 영향을 미침, clay코트에서 승률이 더 높다
└ prop.table에서 확인

대응표본 t-검정 (독립변수없이 두 종속변수끼리의 평균의차이)

① step 1 : 가설수립

Ho : Mm-Mw(Md)=0

Ha : Mm-Mw≠0

② step 2 : 파일불러오기 및 차이변수만들기

```
library(readr)
pttest <- read_csv("pttest.csv", locale=locale('ko', encoding="euc-kr"))
```

└> 데이터에 한글이있으면 이항목코드를사용

```
library(dplyr)
pttest <- pttest %>% mutate(d=morning-weekend)
```

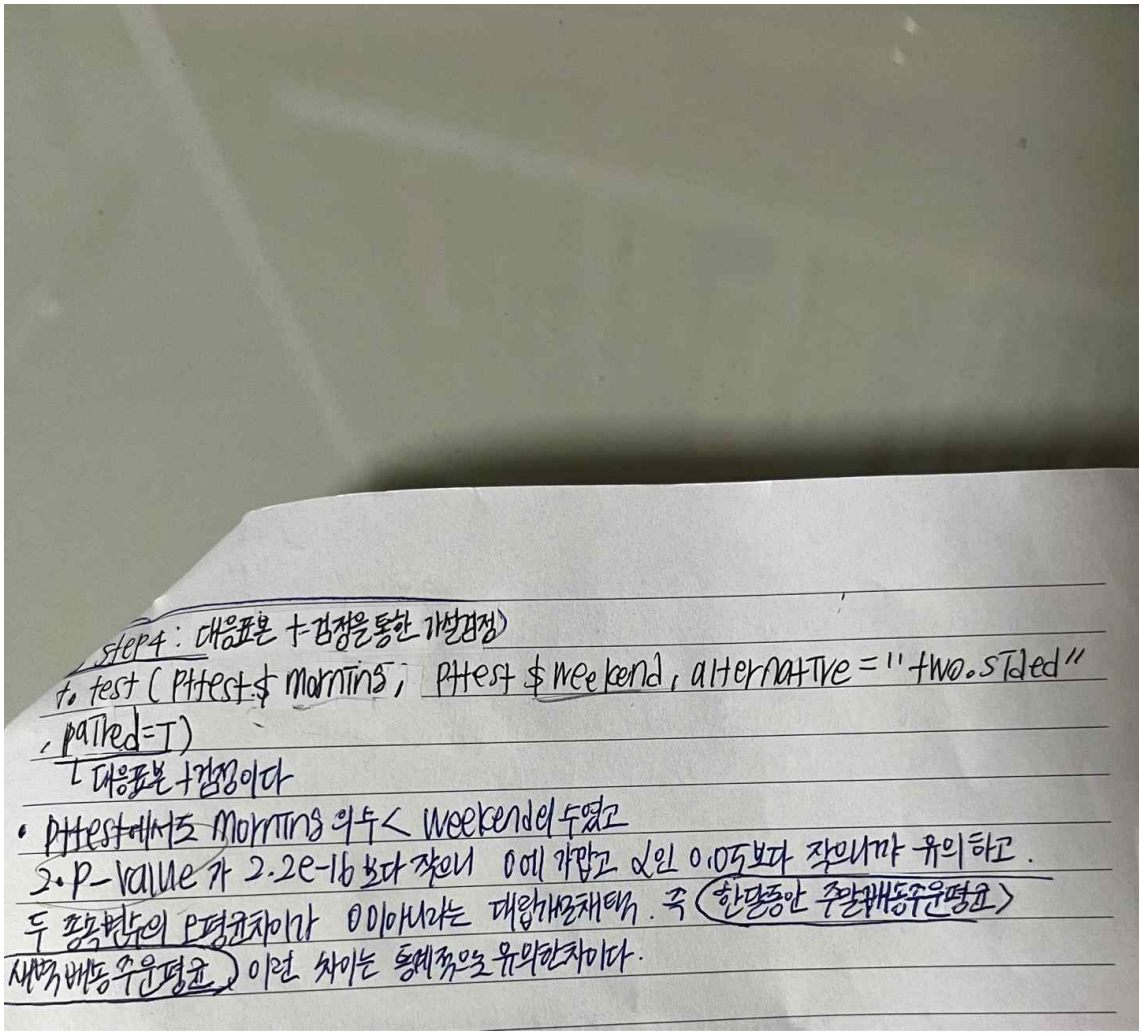
└ 두 종속변수의차이에 대한 변수(d) 생성

③ step 3 : d에 대한 : 정규성검토

```
shapiro.test(pttest$d)
```

└ p-value가 9.70e-12 라서 0에가까우나 유의해서 정규성조건에 완벽하게 어렵지만 앞에 강의안에서 중심극한정리에서 두 확률변수가 정규분포를 따르기에 표본평균의 차이도 정규분포를 따른다고 생각하고 넘기기

④
```
summary(pttest$d)
hist(pttest$d, breaks=seq(-13,8,1))
```



step4 : 대응표본 t-검정을 통한 가설검정

t.test(PMtest$mornings, PMtest$weekend, alternative="two.sided", paired=T)
└ 대응표본 t검정이다

- PMtest에서도 mornings의 수 < weekend의 수였고
2. p-value가 2.2e-16보다 작으니 0에 가깝고 α인 0.05보다 작으니까 유의하고.
두 종속변수의 t평균차이가 0이아니라는 대립가설채택. 즉 (한명당 주말체중운동평균 - 새벽기상체중운동평균) 이런 차이는 통계적으로 유의한차이다.