

<경태분 | 함수 총정리>

(문자: char
변수: factor)

OK

데이터 파일 이름 <- read CSV("file.csv", col.names = T, locale = locale("ko"), encoding = "euc-kr") ; file을 데이터 파일 이름으로 불러와서 저장하기

* 우선 ITbary(read) 먼저하기

View(데이터) : 데이터 보여줌

str(데이터) : 데이터 또는 데이터의 뿐에 대해 다양한 정보 보여줌

summary(데이터) : 범주형 변수에 대한 개별 변수에 대한 자세한 정보 보여줌, NA가 어떤 변수에 있는지 알려줌

데이터 \$변수 <- as.factor(데이터\$변수) : 해당 변수의 칙도를 범주형으로 변경

hist(데이터) : 데이터의 다양한 정보 보여줌 (선택할 때는)

* 우선 ITbary(dplyr) 먼저하기

table(데이터\$변수) : 해당 변수의 빈도수

freq(데이터\$변수) : 해당 변수의 빈도수와 막대그래프

* 우선 ITbary(desc) 먼저하기

qplot(data = 데이터, 변수) : 해당 변수의 빈도수 막대그래프로 표현

qplot(data = 데이터, 변수1, 변수2) : 변수1과 변수2의 조합된 빈도를 막대그래프로 표현

* 우선 ITbary(ggplot2) 먼저하기 | 박상우

* 변수에는 " " 안됨.

mean(데이터\$변수) : 해당 변수의 평균

var(데이터\$변수) : 해당 변수의 분산

sd(데이터\$변수) : 해당 변수의 표준편차

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

freq(TS.NA(데이터\$변수)) : 데이터의 변수에서 결측치가 있는지 여부를 구하고 있으면 변수 T로 표현

* 우선 ITbary(desc) 먼저하기 / 즉 NA=결측치가 몇개인지

descTab(데이터, 데이터\$변수) : 해당 데이터 또는 변수의 해당하는 정규분포표의 수치

(skew-왜도, kurtosis-첨도 ...)

* 우선 ITbary(psych) 먼저하기

hist(데이터\$변수, breaks = seq(N1, N2, K)) : 해당 변수를 N1~N2까지 간격으로 만화스토그램

* 변수의 칙도가 구식형 / N1, N2는 summary(데이터\$변수) 통해 확인, 최대 칙도

<파이썬으로>

같다 ==
같지 않다 !=
크다 >
작다 <
이상 >=
아하 <=
그리고 &
또는 | (shift + alt)

| Library (readf)
| Library (dplyr)
| Library (desc)
| Library (gaplot2)
| Library (psych)

table 인 데이터 \$변수 %in% c ("a", "b", "c") : 데이터의 변수가 a, b, c 중에
개어도 하나의 조건에 만족하는 것들의 번호, (문자가 아니면 "") 필요X
• 1 연산자 사용과 동일

데이터 \$A <- B : B라는 값을 데이터의 새로운 변수 'A'에 저장
└ 항목끼리 어떤 데이터의 번호여도 OK

데이터 \$A <- NULL : 데이터의 A라는 변수 삭제

weekdays(데이터 \$변수) : 해당 변수(날짜)의 요일이 무엇인지 보여줌
└ 해당 날짜(date) 여야함

데이터 \$변수 <- as.Date(데이터 \$변수) : 해당 변수의 첫째를 날짜(date)로 변경
ex) weekdays(as.Date("2020-01-01")) : 수요일

→ ex) Weather \$요일 <- weekdays(weather\$일시) : weather의 변수 '일시'의
요일을 구한 값을 데이터 weather의 새로운 변수 '요일'에 지정

데이터 \$변수 <- factor(데이터 \$변수, levels = c ("a", "b", "c", ...))
└ a, b, c 순으로 변수를 배열하고 첫째를 변수형으로 바꾼 뒤 해당 데이터의 변수에 지정

데이터 <- 데이터 %>% rename(별변수명 = 기존변수명) : 변수명 변경하기
└ 우선 library(dplyr) 설치하기 / " " 안써도됨

데이터 \$변수 <- Tfalse(데이터 \$변수) == "a", "b", 데이터 \$변수) :

만약 데이터의 변수값이 a 이면 b로 바꾸고 아니면 데이터의 원래 변수값 그대로 두어라
└ 같은 변수가 아니여도 이<

weather의 평균기압이 NA일 경우는

TS\$NA (weather\$평균기압) 이지 weather\$평균기압 == NA는 오답

TS\$NA (데이터 \$변수): 해당변수가 평균치 (NA)이다

→ " " 인식됨, 단사례에 문자는 " " 쓰기

* 데이터 전체의 핵심 \rightarrow 우선 (TidyData (tidyData) 먼저하기)

데이터 %>% filter(변수 == 사례): 데이터 안에 있는 변수에서 해당 사례만 추출

④ var나 mean과 같은 기술통계학을 이용해 위에서도 해당사례들만
주제한 것들을 새로운 데이터에 assort해 사용함

round(여러값, dTTS=k): 해당값에서 소수 k번째 자리까지
└ 칸법을 이용하거나 소수점 2자리로 봄

→ filter함수안에 다양한 여러 변수들을 기준으로 조건에 맞게 설정 가능

ex) exam %>% filter(class != 4, math >= 90 | hTStaY >= 95)

(4번이 아닌 학생들에 수학이 90점 이상이나 영어가 95점 이상)

데이터 %>% filter(변수) = quantile(변수, probs = c(k)): 데이터의 해당변수값이

변수의 값중에서 k인값보다 이상한 사례만 추출 (즉 k가 0.9이면 0.9이상인 값, 변수가 상위 10%인
사례들만 추출 (변수의 값이 상위 1-k인값들만 추출))

데이터 %>% select(변수): 해당변수만 추출 (여러변수들, 조연결해서 추출 가능)

데이터 %>% select(-변수): 해당변수 제외하고 추출

데이터 %>% select (contains("특정단어")): 특정단어가 포함된 변수추출 "특정단어"를 연속으로 X
PFTnt(n=Inf): 말이 끝까지 다보여주기 연산자 | 사용하기

데이터 %>% arrange(변수): 변수에 대해서 오름차순정렬 → 두 변수를 같이 사용 가능

데이터 %>% arrange(-변수): 변수에 대해서 내림차순정렬 (음수시) arrange(class, -math)

└ exam\$math의 형태가 아니라 math, -math 변수 그대로
summarise쓸 때 변수명 지정해야 arrange 이용 가능

데이터 <- 데이터 %>% mutate(새변수 = []): []라는 과정을 거친

새변수를 데이터에 저장, mutate함수로 동시에 여러변수 생성 가능, 미과정에

주로 ifelse 함수가 쓰는데 ifelse 안에 또 ifelse가 들어갈 수 있음

(변수가 1개로 구분되면 ifelse를 ())회 사용)

별별로 오름차순

정렬한 것들에서 수학정수

내림차순해서 정렬하기

데이터 <- 데이터 %>% mutate(새변수 = case_when(기준변수 < k ~ " ", 기준변수 >= k ~ " ")

ex) exam <- exam %>% mutate(test = case_when(total < 180 ~ " fail ",
total >= 180 ~ " pass ")

* 변수수정값에 NA가 있다면 TRUE인 조건이 아닌 total >= 180 ~ " pass " 같이 서로 조건이용해서 씀

데이터 ← 데이터 %>% relocate(변수, before(혹은 after)=변수2)

: 변수를 변수2 왼쪽에 오른쪽에 이동시키기

※ ①. before: 변수를 변수2 왼쪽에(앞)

②. after: 변수를 변수2 오른쪽에(뒤)

데이터 ← 데이터 %>% relocate (where (is.character 혹은 is.factor))

: 문자형으로 혹은 범주형으로 앤표로 이동하기

데이터 ← 데이터 %>% relocate(where(is.factor), before=where(is.character))

: 범주형으로 변수를 문자형으로 변환 앞으로 이동하기

데이터 %>% group_by(변수) %>% summarise (mean or var or sd ...)

: 변수별로 그룹을 나누고 기술통계량을 그룹에 맞게 제시

* summarise 를 보여주는 결과값 앞에 변수명 지정 가능 ex) mean_math = mean(math)

mutate는 summarise를 이용하는 count = n() 번도 이용한 코드 사용 가능

(변수)
[변수명 학교코드]

count / sum (count): 이미 기준이 된 변수 그룹이 되어 있고 나누어진 '사례들끼리'의 전체 갯수에서 해당사례 비율

count / sum (데이터 \$ count): 모든 사례들(전체) 갯수에서 해당사례비율

1 - distinct(데이터 \$ 변수): 변수의 첫 번째와는 유관하게 중복되는 값이 아닌 고유의 값이 몇 가지 중복이 있는가
(변수에 해당하는 사례가 중복하지 않고 몇 가지나 있나?)

c(a:b): a부터 b까지가 연속된 숫자들의 배열

ex) c(1:7): 1 ~ 7

relocate(변수): 데이터상에서 변수가 제일앞으로

ex) relocate(1:b): 데이터상에서 '1:b'라는 변수가 제일앞으로

데이터 ← left_join(데이터, 데이터2, by="변수"); 데이터와 데이터2를 데이터에
공통 변수를 이용해서 합침

데이터 ← left_join(데이터, 데이터2, by=c("변수1" = "변수2"))

: 공통변수의 이름이 다른 때 변수의 이름과 변수2의 이름이 같다고 지정한 뒤 데이터에 합침

데이터 \leftarrow bind_rows (데이터, 데이터2) : 데이터와 데이터2를 합쳐서 사례 추가.

단. 변수명이 일치하도록 하면 오류 발생 (번호가 중일 때 오류)

내용상 동일한 번호여도 변수명이 다르면 다른 번호로 인식 / 모든 번호 명 중일

\Rightarrow 1) 변수명 일치하게 변경 : rename 함수

2) 수도 일치하게 변경 : dcast 함수

데이터 \leftarrow 데이터 %>% dftinct(번호, .keep_all=T) : 해당 번호가 같은

사례들 중에 한 개만 남기고 다 제외하기

ex) ID가 중복될 때 동일한 사례인지 확인

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

결과가 ID가 3인 것의 count가 2가 나온 것을 통일화함

dftinct 서서 자유여행

데이터[ca:b] : 데이터에서 a:b까지의 번호들

ex) exam[5:8] : exam에서 5번째 ~ 8번째의 번호들 (5에서 8까지)

데이터\$mean \pm 2,575.63X데이터\$sd : 이 기준을 넘어가면 이상치 처리

<+ 경영학자>

① 데이터 프레임 만들기

데이터 파일 읽기 \leftarrow read.csv("file.csv")

② 보기 좋게 변수 위치 조정: 병렬화 \rightarrow 수직화 \rightarrow 원자화

③ 데이터 \leftarrow 데이터 %>% 헤드(10)에 한 번 이용

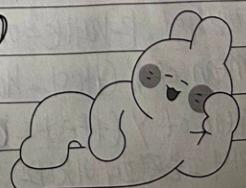
④ 이상치 검토 및 번호 확인

descrl \leftarrow describe(hest[6:10])

descrl \leftarrow descrl %>% mutate(LL=mean-2*sd, UL=mean+2*sd)

④ 이상치를 제외한 데이터 프레임 생성

ex) hest_new \leftarrow hest %>% filter(번호 < k ...)
LUL값



1) 가설수립

$H_0: M_{HO} = M_{CS}$ (구유기설)
 $H_A: M_{HO} \neq M_{CS}$ (대립가설)

2) 집단별로 데이터프레임 만들기

ex) $Htest - HO \leftarrow Htest - new \% > \% filter (customer == "Home office")$
 $Htest - CS \leftarrow Htest - new \% > \% filter (customer == "Consumer")$

3) 정규성 테스트(정규분포의 형태를 파는거)

Summary, $htst$ 함수 이용해서 정규화

ex) $summary (Htest - ho \$ sales)$

$htst (Htest - ho \$ sales, breaks = seq (0, 9000, 50))$

=> 두 데이터 다 왼쪽으로 치우쳐진 히스토그램 모양이라 정규분포가 아님

$shapiro.test$ (데이터변수): 데이터의 해당하는 봤에 대해 P-value 구하기

* P-value > 0.05 즉 P-value가 유의하지 않아야 정규성 조건을 만족하는데

위 함수를 이용한 결과 P-value < 0.05 (α) 이기에 유의해서 정규성 조건에 부합X

=> 정규성 조건에 부합해지 않으면 sales를 자연로그로 변환

ex) $Htest - ho \leftarrow Htest - ho \% > \% mutate (lnsales = log(sales))$

L 이후 $htst$, $shapiro.test$ 이용하여 정규성 테스트

L 여기서 P-value 값이 여전히 예기치에도 OK

4) 등분산성 테스트

$Var.ttest$ 함수 이용해서 $P-value \leq 0.05$ 면 이분산 + 검정 $P-value \geq 0.105$ 면 등분산 + 검정

$Var.ttest (Htest - ho \$ lnsales, Htest - CS \$ lnsales)$

L • $F = 1.3083$ 으로 예기치에 표본분산이 비슷하고 $P-value = 0.5536 > 0.05$ (α) 아니 등분산

5) 독립표본 + 검정 테스트 가설검정

$t.ttest (Htest - ho \$ lnsales, Htest - CS \$ lnsales, alternative = "two.sided")$

/ $Var.equal = T$

L 만약 4) 과정에서 이분산이면 F

계산하면 $P-value = 0.2601$ 이라 0.05 (α) 보다 크고

mean of X ($Htest - ho \$ lnsales$) = 5.976144

mean of Y ($Htest - CS \$ lnsales$) = 5.915587

이니 표본평균이 비슷하고 P-value도 유의하지 않고 $H_0: M_{HO} = M_{CS}$ 의 귀무가설채택

고4 범위 > 0,05

one-way ANOVA - 일원분산분석.

1 단계: 가설수립

0 이상치검정으로 ANOVA-new 데이터 프레임 만들기
library(plyr)

descr <- descr %>% filter(price <= descr\$UL)

descr <- descr %>% mutate(UL = mean + 2 * sd, LL = mean - 2 * sd)

anova_new <- anova %>% filter(price <= descr\$UL)

<기말영어파이지>

library(readr) library(car)

library(dplyr)

library(psych)

library(forcats)

library(collapse)

library(glm.test)

↳ 한개변수 하면 유도나까 차리하하나씩

(0 세출운행수 PRIor 만들기 → CHTT와 HTSh 통합)

install.packages("forcats")

library(forcats)

anova_new <- anova %>% mutate(PRIor = fct_collapse(PRIor, "HTSh"))
"HTSh" = c("CHTCA", "HTSh"))

↳ 세출운행

0 두 가지 가설 수립

H0: 4가지집단의 price 평균이 같다

H1: 적어도 한집단의 price 평균은 다른집단과 다르다.

1단계: 잡단간 데이터프레임 생성하기

anova_H <- anova_new %>% filter(PRIor == "HTSh")

* 각 PRIor의 변수 4개간의 프레임 통일하게 생성 *

2단계: 통계변수 정유성검토

summary(anova_H\$price)

htst(anova_H\$price, breaks=seq(0, 600, 10))

shapiro.test(anova_H\$price)

→ p-value가 0 아래로 유의수준에서 정규성조건에 부합하지 않지만

N의 크기가 많아서 그전에 만족할것이라고 가정하고 아주분석 진행

* 모든 데이터프레임에서 다를듯 *

3단계: 통분산성검토

install.packages("car")

library(car)

leveneTest(pricen_PRIor, data=anova_new)

통계변수

독립변수

전체데이터프레임

$P(F > F_0) = 0.107$ 이라 d의 차이가 0.5보다

거의 통분산조건을 만족하지 못함 (유의수준 0.05)

5단계: ANOVA 실행 완료 (통계신전이 안쪽할 때) / 알맞은 범위에서 대립이나 차이가 2단계 P-value가 유의하고 (F-test)

anova -> result <- aov(PITCE ~ PITCOPTR, data = anova_new)

summary(anova -> result)

PValue가 0.1193이니 0.05가 크고

P-value가 유의하지 않고 차이를 채택

즉 4개 카드의 PITCE의 모평균이 모두 같다..

만약 차이가 있다면 2단계 대립가설 채택시면 차후증정설사 (통계신전에서 대립가설을 채택)

6단계: 차후증정 Duncan test 완료

install.packages("DGTclae")

library(DGTclae)

duncan.test(anova -> result, "payment", console = T)

aov학습이용데이터

죽립변수

• 모평균이 큰 카드부터 표하고 같은 group으로 나눠서

집단끼리의 모평균은 같다고 지정

• 신용카드 expense 평균 > 계좌이체 expense 평균 = 간편결제 expense 평균 %

만약 차이가 있다면 0.05이면 5단계에서 통분산이 아니라 이분산전을 실시해야함

one way. test(expense ~ payment, data = anova_new)

• P-value가 0에 가까워 대립가설 채택

(이분산전 사용으로 duncan.test() (이분산전에서 대립가설을 채택할 때))

install.packages("dunn.test")

library(dunn.test)

dunn.test(anova_new\$expense, anova_new\$payment, method = "bonferroni")

데이터 \$ 풍족변수

데이터 \$ 죽립변수

• 간편결제와 계좌이체의 P-value가 0이다. 0/2보다 커서 유의하지 않고 모평균이 동일하다.

• 간편결제와 신용카드의 P-value가 0에 가까운 0/2보다 커서 유의하고 모평균은 같지 않은데

신용카드가 더 크다 (간편결제 두개인 신용카드와 대립할 때 (-7.3이 456이니))

(계좌이체 - 신용카드 = -7.3이 456)

+통계량 값

신용카드 > 간편결제

신용카드 > 계좌이체

대비하는 P-value와 0/2 (=0.025)와 비교

No-way ANOVA 아울렛 분석

1단계: 가설 설정

○ 가설 설정

H₀: 두 그룹 간에 상호작용 효과가 없다.

H_a: 두 그룹 간에 상호작용 효과가 있다.

○ 이상치 검토 및 제거

library(psych)

descr <- describe(two_anova\$expense)

descr <- descr %>% mutate(VL=mean+3*sd, LL=mean-3*sd)

two_anova_new <- two_anova %>% filter(expense <= descr\$VL)

2단계: 서브데이터 프레임 만들기

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

*↑ gender의 빠진 Females도 같이 만들어주기 *

3단계: 정규성 검토

summary(two_anova_male\$expense)

hist(two_anova_male\$expense, breaks = seq(0, 2000, 40))

shapiro.test(two_anova_male\$expense)

↳ P-value가 0.05보다 작아 유의해서 정규성 조건 위반

만족하지 못할 경우에만 사전에 정규화를 진행

4단계: 등분산성 검증

library(car)

leveneTest(expense ~ gender, data = two_anova_new)

↳ P-value가 0.001238이라 0.05보다 작아서 유의함

등분산 조건을 만족하지 못함

5단계: 이분산 가정 One way ANOVA 실행

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

↳ P-value가 0이기 때문에 구두기준을 귀류하고

대입가능 차이가 있다. ↳ gender에 따라 집단을 두개로 구분했을 때 평균 expense

평균 차이가 있다. → mFemale > mMale ($553 > 359$)

two_anova_new %>% group_by(gender) %>% summarise(means(expense))

제1: two-way ANOVA 시행 및 그림

1) two_anova_result ← aov(expense~gender*OS, data=two_anova_new)
[두 독립변수의 상호작용]

summary(two_anova_result)

↳ $F(2, 138) = 1.38, p = 0.12$... 이므로 모델 적합 유의하고 (5단계)
 $\text{gender} \times \text{OS } F(2, 138) = 0.003172$... 이므로 모델 적합 유의하지 않음 ($\alpha = 0.05$ 모델 적합)
(H_0 : 두 독립변수간에 상호작용 효과가 있다) [미래기준 체크]

즉 I_{V1} 과 I_{V2} 와 I_{V3} 의 차에 따른 $p-value$ 를 더 유의해야 하는 것임이 조건 ✕.

2) two_anova_new\$gender ← factor(two_anova_new\$gender, levels=c("Male", "Female"))

↳ (one-way ANOVA에서 표본평균이 작은 순서대로 집단을 먼저 출력하게 함
따라서 작은 Male이 먼저 출력되게 순서바꾸기)

install.packages("HH")

library(HH)

interaction2wt(expense~gender*OS, data=two_anova_new)

↳ OS와 OS간에 Male과 Female 양대 expense의 평균값이 증가한다.

↳ OS가 OS일 때 Male과 Female로 각각 expense의 평균차이가 더 커진다 (기울기 터짐)

↳ OS가 OS일 때 Male과 Female로 각각 expense의 평균차이가 커진다 (기울기 원란)

↳ 상호작용의 정도이다

8단계: 우기분석 집단을 네 개로 구분해서 세 분화한 뒤 one-way ANOVA

즉 집단 내부 MA, MT, FA, FT (병, 여성 / Android, iOS)

two_anova_new ← two_anova_new %>% mutate(OSGender = ifelse(gender == "Male" & OS == "Android", "MA", ifelse(gender == "Male" & OS == "iOS", "MT", ifelse(gender == "Female" & OS == "Android", "FA", "FT"))))

↳ mutate와 ifelse함수에서 사용한 변수 \$OSGender은 만드는데 예전의 접근기준
one-way ANOVA 단계는 별개로

양원분산분석과 이원분산분석에서 미안가설, 귀무가설에서 어떤 대목인지 확인하는 것은

P-value과 α (유의수준) 간에 비교하는 것이다. 이후 알아갈 차이가 있는지는

사후분석에서 $p-value$ vs $\alpha/2$ 나 $p-value$ vs $\alpha/2$ 사이를 비교하면 된다.

비교적 통계분석

모바일 사이트에 대한 개설정성

① 데이터프레임 불러오기

library (readr)

Proptest \leftarrow read_csv ("Proptest.csv")

② 번수 기준 표 만들기

table (Proptest \$ Customer == 20 & Proptest \$ Trans == "Yes") = (48)

* Customer와 Trans의 모든 변수의 조합끼리 차해주기 *

	Yes	No
20세	48	49 (97)
30세	30	73 (103)

prop \leftarrow matrix (c (48, 49, 30, 73), nrow = 2, ncol = 2, byrow = T)

└ 행의갯수 └ 열의갯수 └ 행을 기준으로 나누기

rownames (prop) \leftarrow c (20, 30)

colnames (prop) \leftarrow c ("Yes", "No")

└ 각 행과 열의 이름을 대체 지정해주기

③ 베이즈 기준 표 만들기

prop.table (prop, margin = 1)

└ prop이라는 번수로 해놓은것을 베이즈 비율로 바꿔주는 대행마다 값의 합이 1로 해야함

④ 개설정성

$$H_0: P_{20} - P_{30} = 0$$

$$H_a: P_{20} - P_{30} \neq 0$$

$\overline{P_{20}} - \overline{P_{30}}$ 이 정규분포를 12개 유한 조건으로 $|DP| \times |N(DP)| \geq 5$ 일치 확률을 우선해야 함

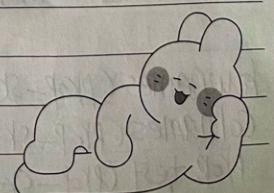
$$0.97 \times 0.1495 = 48, 015 > 5$$

$$0.97 \times 0.1505 = 48, 985 > 5$$

$$|0.97 \times 0.29| = 29, 973 > 5$$

$$|0.97 \times 0.109| = 73, 027 > 5$$

모든 경우가 5보다 크기 $\overline{P_{20}} - \overline{P_{30}}$ 은 정규분포를 만난다



Prop. test (prop, alternative = "two.sided", correct = T)

$P\text{-value}$ 의 값이 0.05보다 작거나 같으므로 유의수준 대립가설 채택
 (양측검정일 때) $P_{20} > P_{30}$ (대립가설 채택) | $\text{prop} / 0.14948434 > \text{prop} / 0.2912621$

다항모형은 적합성검정

- ① 데이터프레임 분리된 각각 과거, 현재에서 번호 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 번호를 넣기

$P\text{-value} = 0.01832$ 로 0.05보다 작거나 같으니 대립가설 채택 즉
과거에 이동통신 3사의 시장점유율은 서로 같지 않다.

- ③ 두번째 가설검정: 현재시점에 3사의 시장점유율이 동일한지

chisq.test(c(149, 85, 66))

$P\text{-value} = 6.03e-09$ 이므로 0.05보다 작거나 같으니
마찬가지로 현재의 이동통신 3사의 시장점유율은 서로 같지 않다.

- ④ 세번째 가설검정: 과거 시장점유율과 현재 시장점유율의 모두 동일한지 여부

chisq.test(c(120, 100, 80), p = c(0.4, 1/3, 1/3))

현재 번호

과거시점 해당통신사번호
과거시점 전체매수

100
300

(KT통신비율)

$P\text{-value} = 0.002868$ 이라 0.05보다 작거나 같으니 대립가설 채택하고
과거에 비해서 이동통신 3사의 시장점유율은 변화가 있다.

- ⑤ 세번째 가설검정에 대한 사용법

prop_skf <- matrix(c(c(120, 149, 66), nrow=2, ncol=2, byrow=T))

SKT가 아닌 나머지 통신사들의 과거, 현재의 개수합

rownames(prop_skf) <- c("Past", "Current")

colnames(prop_skf) <- c("SKT", "Not SKT")

prop.test(prop_skf, alternative = "two.sided", correct = T)

과거보다 시장점유율이 더
늘어나고 약화되는 경향

하여 KT와 LG+도 같은 방식으로 해석되는데 $P\text{op_kt}$ 는 $P\text{-value} = 0.2159$
즉 이인 이 0.2159보다 크기 때문에 유의하지 않고 KT의 시장점유율은 과거와 현재 비교해 변화가 없음
 $P\text{op_lg+}$ 도 $P\text{-value}$ 가 0.2161이라 이인 이 0.2159보다 크기 때문에 유의하지 않고
마찬가지로 LG+의 시장점유율은 과거와 현재 비교해 변화가 없음.

독립성 검정

① 데이터 불러오기

`library(readr)`

`tennis <- read_csv("tennis.csv")`

② 데이터 전처리

1) `names(tennis) <- tolower(names(tennis))`: tennis 데이터의 변수들의 이름을 다 소문자로
 └ 변수들만 보여주는 방식 └ 소문자 표기 방식

2) `tennis$surface <- tolower(tennis$surface)`: tennis의 변수의 속성을 같은 소문자로
 └ 풀링변수 surface 와 마찬가지로 풀링변수 result에도 해주기 *

3) surface의 속성을 clay와 clay(T)를 합쳐고 hard와 hard(T)를 합쳐기

`library(forcats)`

4) `tennis$surface <- fct_collapse(tennis$surface, "clay" = c("clay", "clay(T)"))`

`library(gaplyr)`

`tennis <- tennis %>% mutate(surface = fct_collapse(surface, "clay" = c("clay", "clay(T)")))`

* 변수의 속성을 hard, hard(T)로 같은 방식으로 합쳐기 * / 두 방법 다 가능 (A, B 중에 끌어서 실행)

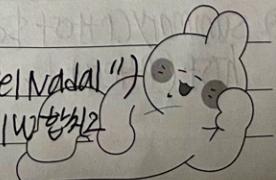
만약 surface의 속성을 풀 때마다 끌어내면 filter 이용해서 제외하고 새로운 데이터 만들기

`tennis_new <- tennis %>% filter(! Ts_na(surface))`

4) Player가 Rafael Nadal인 것만 끌어내면 데이터 만들기

`tennis_nadal <- tennis_new %>% filter(Player == "Rafael Nadal")`

5) 4), 3)의 코드를 이용해서 nadal의 result의 속성을 중에서 1, 1T, 1W로 합치기
 W, Wt, Wn 합치기



③ 독립성검정

Xtabs (~surface + result, tennis - nadal)

Prop.table (Xtabs (~surface + result, tennis - nadal), margin = 1)

chisq.test (Xtabs (~ surface + result, tennis - nadal))

└ P-value가 2.477e-09 이므로 사방향이이고 Surface와 result는 종속적이지 않다고 하는 대립가설에 대해 꼭, surface가 result에 영향을 미침.

영향을 미침, C언어코드에서 승률이 더 높다
└ Prop.table에서 확인

대수표본 + -검정 (종류변수없이 두 종류변수끼리의 오평균의 차이)

① Step 1 : 개설구성

$$H_0 : M_m - M_w : (Md) = 0$$

$$H_a : M_m - M_w \neq 0$$

② Step 2 : 파일 불러오기 및 카이벌류 만들기

library (readr)

pptest <- read_csv ("pptest.csv", locale = locale ("ko", encoding = "euc-kr"))

└ 데이터에 한글이 있으면 이 항목코드를 사용

library (dplyr)

pptest <- pptest %>% mutate (d = morning - weekend)

└ 두 종류변수의 차이에 대한 변수 (d) 생성

③ Step 3 : d에 대한 정규성검증

shapiro.test (pptest \$ d)

└ P-value가 0.709e-12 라서 0에 가까워 유의해서 정규성검증에

안맞아서 어렵지만 앞에 같은 단위에서 중심주한정리에서 두 확률변수가 정규분포를 띠기기에.

표본 평균의 차이로 정규분포를 엿대고 생각하고 넘기기

④ summary (pptest \$ d)

htst (pptest \$ d, breaks = seq (-15, 11))

step4: 대응표본 + 검정을 통한 가설검정)

to test (ptest \$ morning, ptest \$ weekend, alternative = "two.sided"
, paired = T)

└ 대응표본 + 검정이다

• ptest에서도 morning의 두 < weekend의 두 염고

2. P-value 가 2.2e-16 보다 작으니 0에 가깝고 0이 0.05보다 작으니까 유의하고.

두 풍속변수의 평균차이가 0이 아님라는 대립가설을 채택. 즉 (한달동안 주말비행주운평균)
비행주운평균) 이런 차이는 통계적으로 유의한 차이이다.