R통계 과제2 교차분석

조성우_201823869

2020년 10월 10일

교차분석

교차분석은 명목(또는 범주형)자료 및 순위・등간 데이터의 성향을 갖고 있는 두 개 혹은 그 이상의 변수에 대한 내용을 분석하기 위해 사용되는 통계적 분석방법입니다.

교차 분석은 한 변수에 속한 빈도 수와 다른 변수에 속한 빈도 수를 함께 교차로 분석합니다. 이 방법은 두 변수 사이에 관련이 있는지, 있다면 관련의 정도가 어느 정도인지를 카이 제곱 검정을 실시합니다.

1.개요

목적:

'국가고시 유형 별로 남녀의 합격자 수가 다를것이다' 라는 대립가설(H1)을 해당 데이터를 교차분석하여 검정해 보고자 합니다.

활용 데이터셋:

'9급공채합격자2019'

상세:

해당 데이터에서 9급공무원 '직렬'을 대립가설에서 제시한 '유형'으로 여깁니다.

2.본문

2.1 데이터 정제

불러오기

```
#불러오기
setwd("C:WWUsersWWJSWWWDesktopWW아주대학교WW1_학기WW3학년 2학기_2020_2WW강의자료WW1_R통계WWR통계 과제WW과제2")
rawdata = read.csv("9급공채합격자2019.csv",header=T)
str(rawdata)
```

```
## 'data.frame':    120 obs. of  11 variables:
##  $ X2019년.국가공무원.9급.공개경쟁채용시험.합격선..최종합격자.및.여성비율.: chr  "" "" "구분" "전 모집단위 합계" ...
##  $ X                                                                    : chr  "" "" "선발Wn예정인원" " 4,987 " ...
##  $ X.1                                                                  : chr  "" "" "출원인원" " 195,322 " ...
##  $ X.2                                                                  : chr  "" "" "응시인원" " 154,331 " ...
##  $ X.3                                                                  : chr  "" "" "합격선" " - " ...
##  $ X.4                                                                  : chr  "" "" "필기Wn합격인원" " 6,914 " ...
##  $ X.5                                                                  : chr  "" "" "추가합격Wn(면접미달)" " 8 " ...
##  $ X.6                                                                  : chr  "" "" "최종Wn합격인원" " 5,067 " ...
##  $ X.7                                                                  : chr  "" "" "여성Wn합격인원" " 2,907 " ...
##  $ X.8                                                                  : chr  "" "" "여성비율Wn(%)" "57.4%" ...
##  $ X.9                                                                  : logi  NA NA NA NA NA NA ...
```

데이터처리

```
#전처리_필요한 변수만 추출

rawdata = rawdata[-c(1,2),]

colnames(rawdata) = rawdata[1,]

rawdata = rawdata[-1,]
str(rawdata)
```

```
## 'data.frame':    117 obs. of  11 variables:
##  $ 구            분    : chr  "전 모집단위 합계" "일반모집 계" "행정(일반행정 전국:일반)"
"행정(일반행정 지역:일반)" ...
##  $ 선발
## 예정인원      : chr  " 4,987 " " 4,567 " " 294 " " 123 " ...
##  $ 출원인원            : chr  " 195,322 " " 189,180 " " 33,539 " " 14,184 " ...
##  $ 응시인원            : chr  " 154,331 " " 149,543 " " 25,204 " " 10,941 " ...
##  $ 합격선              : chr  " - " "-" "407.37 " "-" ...
##  $ 필      기
## 합격인원: chr  " 6,914 " " 6,380 " " 410 " " 170 " ...
##  $ 추가합격
## (면접미달): chr  " 8 " "8 " "0 " "0 " ...
##  $ 최      종
## 합격인원: chr  " 5,067 " "4,671 " "295" "126" ...
##  $ 여      성
## 합격인원: chr  " 2,907 " "2,758 " "188 " "78 " ...
##  $ 여성비율
## (%)       : chr  "57.4%" "59.0%" "63.7%" "61.9%" ...
##  $ NA                  : logi  NA NA NA NA NA NA ...
```

```
rawdata = rawdata[c(3:4,14,31:32,33,38,40,42:52),]
p_data = rawdata

colnames(p_data) = c('구분','선발예정인원','출원인원','응시인원','합격선','필기합격인원','추가
합격','최종합격인원','여성합격인원','여성비율')

#
p_data = data.frame(p_data$최종합격인원,p_data$여성합격인원)

colnames(p_data) = c('최종합격인원','여성합격인원')

 #전처리_타입변환
p_data$최종합격인원 = as.numeric(p_data$최종합격인원)
p_data$여성합격인원 = as.numeric(p_data$여성합격인원)

 #전처리_남성합격인원 변수추가
p_data$남성합격인원 = p_data$최종합격인원 - p_data$여성합격인원
 #전처리_record 이름 직렬분류명으로 바꾸기
row.names(p_data) = rawdata[,1]
 #전처리_직렬별 record 통합
p_data[1:3,]
```

```
##                                 최종합격인원 여성합격인원 남성합격인원
## 행정(일반행정 전국:일반)              295          188          107
## 행정(일반행정 지역:일반)              126           78           48
## 행정(우정사업본부 지역:일반)          606          418          188
```

```
attach(p_data)
행정 = p_data[1,] + p_data[2,]+ p_data[3,]
세무 = p_data[4,] + p_data[5,]
통계 = p_data[6,]
검찰 = p_data[7,]
기술 = p_data[9,] + p_data[10,] + p_data[11,] + p_data[12,] + p_data[13,] + p_data[14,] + p_dat
a[15,] + p_data[16,] + p_data[17,] + p_data[18,] + p_data[19,]

 #전처리_새롭게 구성한 record들을 rbind를 통해 새로운 데이터프레임으로 재구축
p_data2 = rbind(행정,세무,통계,검찰,기술)

row.names(p_data2) = c('행정직','세무직','통계직','검찰직','기술직')


 # 데이터3로 갈아탐 : 최종합격인원 변수를 제외시키고

p_data3 = p_data2[,-1]

p_data3
```

```
##        여성합격인원 남성합격인원
## 행정직          684          343
## 세무직          726          326
## 통계직           48           30
## 검찰직          148          102
## 기술직          275          353
```

전처리가 끝나서 활용할 변수만 남은 최종적 데이터프레임의 모습

원데이터 재구성

Table 구성 및 교차분석을 위해 불러들여 전처리한 가공데이터로부터 원데이터를 재구성한다.

```
# Cross Table 및 교차분석을 위해 앞에서 처리한 요약통계량으로부터 원데이터를 구성하기

rep.row<-function(x,n){ #전처리해둔 요약정보로부터 Cross table을 재구성하기위한 rawdata 만들기

  m <- matrix(rep(x,each=n),nrow=n)

  return(m)

}

df <-  data.frame(rbind(rep.row(c("남성합격자", "행정직"), p_data3[1,2] ),

                        rep.row(c("남성합격자", "세무직"), p_data3[1,2] ),

                        rep.row(c("남성합격자", "통계직"), p_data3[3,2] ),

                        rep.row(c("남성합격자", "검찰직"), p_data3[4,2] ),

                        rep.row(c("남성합격자", "기술직"), p_data3[5,2] ),

                        rep.row(c("여성합격자", "행정직"), p_data3[1,1] ),
                        rep.row(c("여성합격자", "세무직"), p_data3[2,1]),
                        rep.row(c("여성합격자", "통계직"), p_data3[3,1]),
                        rep.row(c("여성합격자", "검찰직"), p_data3[4,1]),
                        rep.row(c("여성합격자", "기술직"), p_data3[5,1])
                        ))

names(df) <- c("합격자성별", "직렬") #새로 생성한 rawdata의 x,y축에 이름붙이기

head(df,10)
```

```
##    합격자성별   직렬
## 1  남성합격자 행정직
## 2  남성합격자 행정직
## 3  남성합격자 행정직
## 4  남성합격자 행정직
## 5  남성합격자 행정직
## 6  남성합격자 행정직
## 7  남성합격자 행정직
## 8  남성합격자 행정직
## 9  남성합격자 행정직
## 10 남성합격자 행정직
```

교차분석을 위해 처리한 데이터로부터 원데이터를 재구성한 모습

2.2 교차분석 실시

2.2.1 내장함수 교차표 출력

table()함수 등 내장함수를 활용하여 빈도표 및 교차표를 출력합니다.

```
###### <1: table 함수 등 내장함수를 이용한 교차분석>
df_table = table(df) #빈도표 출력 및 저장
df_table
```

```
##             직렬
## 합격자성별   검찰직 기술직 세무직 통계직 행정직
##   남성합격자    102    353    343     30    343
##   여성합격자    148    275    726     48    684
```

```
margin.table(df_table)
```

```
## [1] 3052
```

```
margin.table(df_table,1) # 성별(첫번째 변수)을 기준으로 주변합계 출력
```

```
## 합격자성별
## 남성합격자 여성합격자
##       1171       1881
```

```
margin.table(df_table,2) # 직렬(두번째 변수) 을 기준으로 주변합계 출력
```

```
## 직렬
## 검찰직 기술직 세무직 통계직 행정직
##    250    628   1069     78   1027
```

```
   # 성별과 직렬 변수에 따른 백분율 계산 및 출력
round(prop.table(df_table)*100,2)    #셀의 비율 산출
```

```
##             직렬
## 합격자성별   검찰직 기술직 세무직 통계직 행정직
##   남성합격자   3.34  11.57  11.24   0.98  11.24
##   여성합격자   4.85   9.01  23.79   1.57  22.41
```

```
round(prop.table(df_table,2)*100,2)  #열의 비율 산출
```

```
##             직렬
## 합격자성별   검찰직 기술직 세무직 통계직 행정직
##   남성합격자  40.80  56.21  32.09  38.46  33.40
##   여성합격자  59.20  43.79  67.91  61.54  66.60
```

```
round(margin.table(prop.table(df_table),1)*100,2) #행의 주변합계 비율 산출
```

```
## 합격자성별
## 남성합격자 여성합격자
##      38.37      61.63
```

```
chisq.test(p_data3) # 내장함수를 이용한 카이제곱검정
```

```
##
##  Pearson's Chi-squared test
##
## data:  p_data3
## X-squared = 120.38, df = 4, p-value < 2.2e-16
```

내장함수를 이용한 교차표 및 카이제곱검정 결과가 출력된 모습

2.2.2 gmodels 패키지

gmodels 패키지의 함수를 사용하여 교차표를 출력하고 결과자료를 분석하여 가설을 검증합니다.

```
###### <2: gmodels 패키지를 사용한 교차분석>

  #gmodels 패키지를 사용하여 교차표를 작성하기
library(gmodels)
CrossTable(df_table,digits=2,prop.c=FALSE,prop.r=TRUE,prop.t=FALSE,prop.chisq=FALSE,chisq=TRUE)
# Option: 소숫점2자리,행의비율(성별기준) 출력o, 카이제곱값 출력 / 열의비율,전체사례수 대비 비율,전체 카이제곱 대비 셀별 카이제곱 들은 출력X
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  3052
##
##
##              | 직렬
##   합격자성별 |    검찰직 |    기술직 |    세무직 |    통계직 |    행정직 | Row Total |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##   남성합격자 |       102 |       353 |       343 |        30 |       343 |      1171 |
##              |      0.09 |      0.30 |      0.29 |      0.03 |      0.29 |      0.38 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##   여성합격자 |       148 |       275 |       726 |        48 |       684 |      1881 |
##              |      0.08 |      0.15 |      0.39 |      0.03 |      0.36 |      0.62 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
## Column Total |       250 |       628 |      1069 |        78 |      1027 |      3052 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  113.7353     d.f. =  4     p =  1.161797e-23
##
##
##
```

```
        # X2(카이제곱) = 120.3838 , 유의확률(p-value) = 4.422701e-25 로
         # ' 9급 공채중 직렬별로, 남녀의 합격자 수의 분포가 다르다는 연구가설(대립가설)이 채택
된다.
```

2.2.2 sjPlot 패키지

sjPlot 패키지의 함수를 사용하여 교차표를 출력하고 결과자료를 분석하여 가설을 검증합니다.

```
###### <3: sjPlot 패키지를 사용한 교차분석>

  #sjPlot 패키지를 이용하여 교차표를 출력하고 가설을 검증한다.
       #Viewer에 직접 출력하는 방법
library(sjPlot)

sjt.xtab(df$합격자성별,df$직렬,
         show.col.prc=T, # 열 백분율을 출력하도록함
         show.exp=T,
         var.labels=c("합격자성별","직렬"),
         value.labels=list(c("남성합격자","여성합격자"),c("검찰직","기술직","세무직","통계직",
"행정직"))
         ,encoding="EUC-KR")
```

합격자성별	*직렬*					***Total***
	검찰직	기술직	세무직	통계직	행정직	
남성합격자	102	353	343	30	343	1171
	96	241	410	30	394	1171
	40.8 %	56.2 %	32.1 %	38.5 %	33.4 %	38.4 %
여성합격자	148	275	726	48	684	1881
	154	387	659	48	633	1881
	59.2 %	43.8 %	67.9 %	61.5 %	66.6 %	61.6 %
Total	250	628	1069	78	1027	3052
	250	628	1069	78	1027	3052
	100 %	100 %	100 %	100 %	100 %	100 %

χ^2=113.735 · df=4 · Cramer's V=0.193 · p=0.000

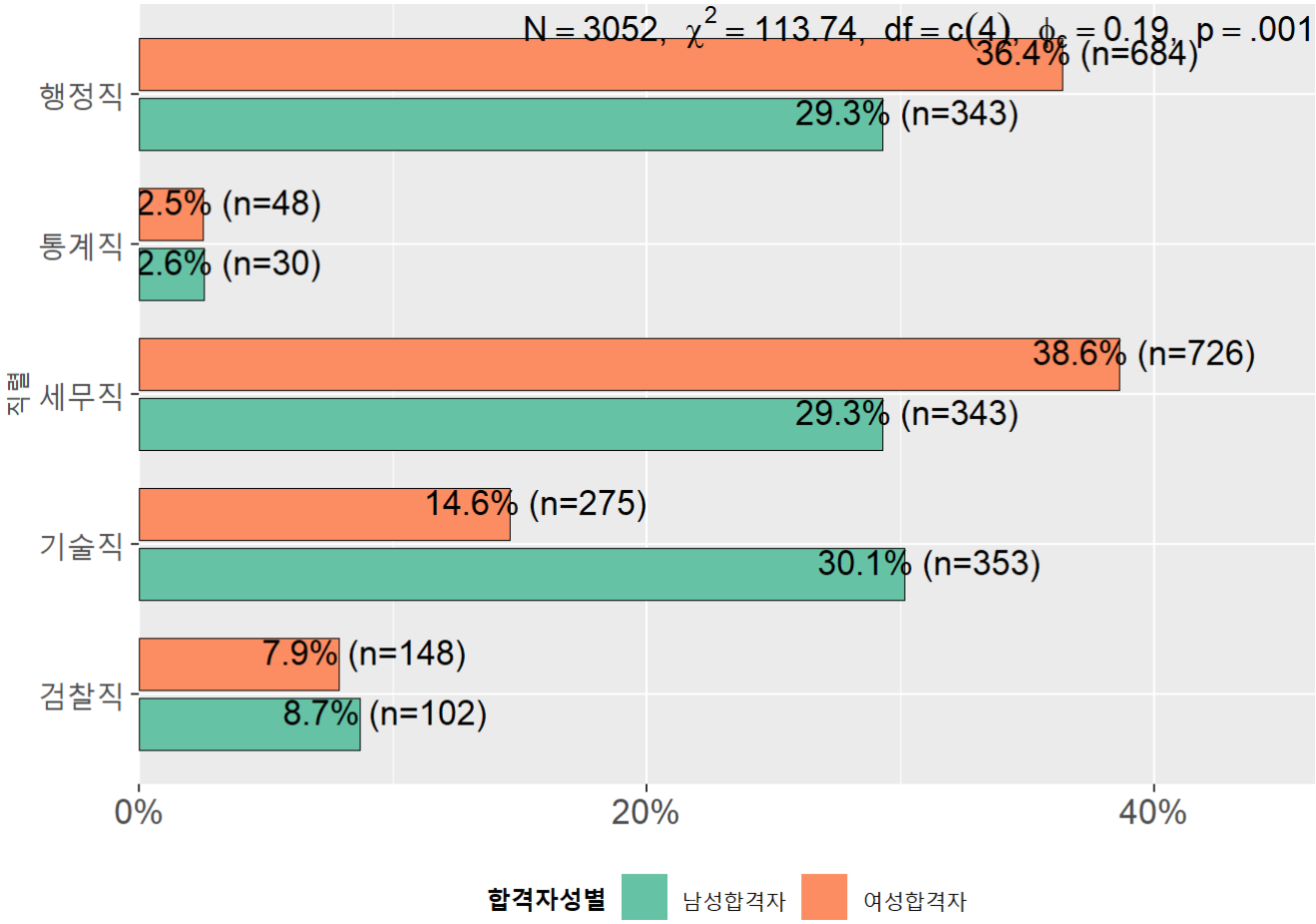
 # 결과해석 : 카이제곱값은 120.384이고 유의확률은 p-value<0.001로 영가설을 기각하거나 채택할 수 있는 기준인 0.05보다 낮기 때문에, 9급 공무원 공채의 직렬에 따라 남녀 합격자 수의 분포가 다르다는 연구가설(대립가설)을 채택함

2.3 시각화

2.3.1 sjPlot : 바차트

sjPlot 패키지를 활용하여 앞에서 만든 교차표를 교차분석에 특화된 다차원의 막대그래프로 시각화합니다.

```
  #sjPlot 패키지를 이용한 도표 출력
set_theme(geom.label.size = 4.5, axis.textsize =1.1,
          legend.pos ='bottom')

plot_xtab(df$직렬,df$합격자성별,
          type="bar",
          y.offset =0.01,
          margin = "col",coord.flip =T, wrap.labels= 7,
          geom.colors = "Set2",show.summary=T,show.total=F,
          axis.titles="직렬",
          axis.labels=c("검찰직","기술직","세무직","통계직","행정직"),
          legend.title="합격자성별",
          legend.labels=c('남성합격자','여성합격자'))
```



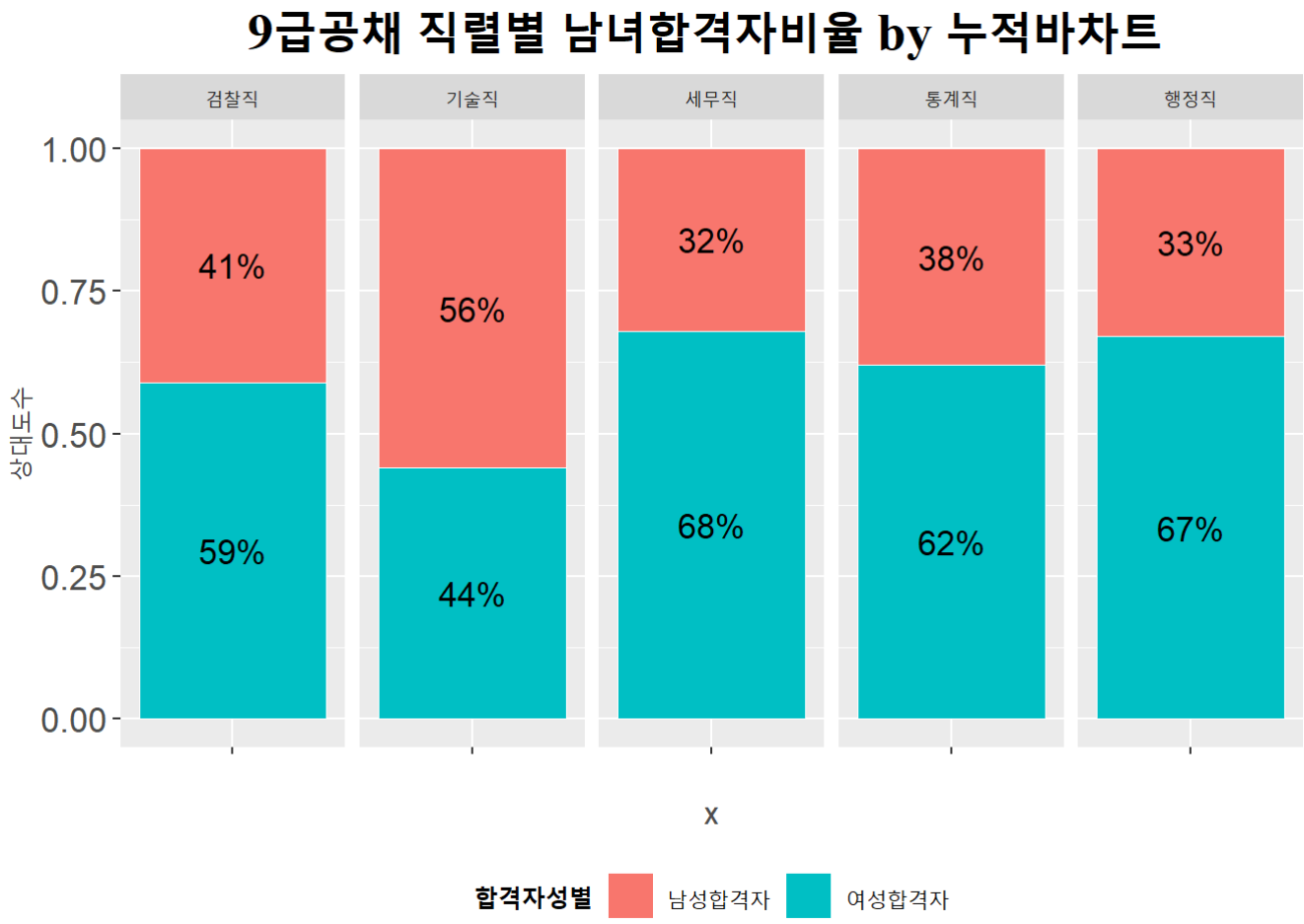
2.3.2 ggplot2 : 누적바차트

ggplot2 패키지를 활용하여 누적바차트를 만들기 위해 맞춤형으로 데이터를 재가공하고 현 보고서의 교차분석에 알맞는 다차원의 누적바차트로 시각화합니다.

```
    #ggplot으로 그리기 위한 데이터 재가공
str(df_table)
```

```
##  'table' int [1:2, 1:5] 102 148 353 275 343 726 30 48 343 684
##  - attr(*, "dimnames")=List of 2
##   ..$ 합격자성별: chr [1:2] "남성합격자" "여성합격자"
##   ..$ 직렬      : chr [1:5] "검찰직" "기술직" "세무직" "통계직" ...
```

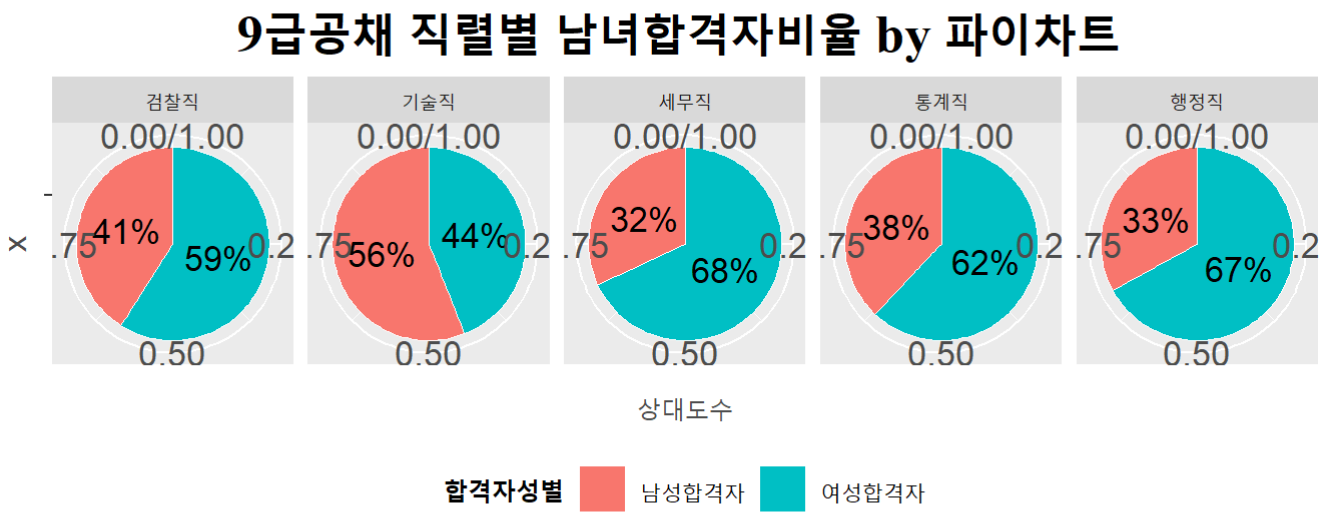
```
df2 = data.frame(df_table)


df2_검찰직 = df2[c(1,2),]
df2_기술직 = df2[c(3,4),]
df2_세무직 = df2[5:6,]
df2_통계직 = df2[7:8,]
df2_행정직 = df2[9:10,]


df2_검찰직$상대도수 = round(df2_검찰직$Freq/sum(df2_검찰직$Freq),2)
df2_기술직$상대도수 = round(df2_기술직$Freq/sum(df2_기술직$Freq),2)
df2_세무직$상대도수 = round(df2_세무직$Freq/sum(df2_세무직$Freq),2)
df2_통계직$상대도수 = round(df2_통계직$Freq/sum(df2_통계직$Freq),2)
df2_행정직$상대도수 = round(df2_행정직$Freq/sum(df2_행정직$Freq),2)

df2_상대도수 = rbind(df2_검찰직,df2_기술직,df2_세무직,df2_통계직,df2_행정직)

   #ggplot / 직렬로 분류된 상대누적막대 차트 그리기
library(ggplot2)

ggplot(df2_상대도수,aes(x="",y= 상대도수,fill=합격자성별))+
        geom_bar(width=1,stat= 'identity',color='white')+
  facet_grid(facets = .~직렬)+
  ggtitle("9급공채 직렬별 남녀합격자비율 by 누적바차트 ")+
  theme(plot.title = element_text(family='serif',face="bold",hjust=0.5,size=20,color='black'))+
  geom_text(aes(label =paste0(round(상대도수*100,1),"%")),
            position = position_stack(vjust = 0.5))
```



2.3.3 ggplot2 : 파이그래프

ggplot2 패키지를 활용하여 현 보고서의 교차분석에 알맞는 다차원의 파이차트로 시각화합니다.

```
#ggplot /
ggplot(df2_상대도수,aes(x="",y= 상대도수,fill=합격자성별))+
           geom_bar(width=1,stat= 'identity',color='white')+
           facet_grid(facets = .~직렬)+
  ggtitle("9급공채 직렬별 남녀합격자비율 by 파이차트 ")+
  theme(plot.title = element_text(family='serif',face="bold",hjust=0.5,size=20,color='Black'))+
         coord_polar('y') +
           geom_text(aes(label =paste0(round(상대도수*100,1),"%")),
                     position = position_stack(vjust = 0.5))
```



3. 결론

해당 보고서상에서 시행된 각종 카이제곱 검정 및 시각화를 통해 카이제곱값은 120.384이고 유의확률은 p-value<0.001로 영가설을 기각하거나 채택할 수 있는 기준인 0.05보다 낮기 때문에,

‘9급 공채중 직렬별로, 남녀의 합격자 수의 분포가 다를것이다’ 라는 대립가설(H1)이 채택된다.