One-way ANOVA test

Youngjin\_Lee

2022 2 13

Table of Contents

# 일원 분산분석(One-way ANOVA)

통계분석 차이검정 : 일원 분산분석(One-way ANOVA) test 문제 별다방을 프렌차이즈 운영하는 K회사는 강남, 강동, 강서, 강북에 위치한 매장의 서비스에 대한 고객조사를 실시하였다. 과연 4곳 매장의 고객만족도는 차이가 있는지? 있다면 어느 레스토랑의 서비스 만족도가 가장 안 좋은가 확인해보자

# 1.기본 package 설정, library 로드

# 2.데이터 불러오기

owa\_tb <- read\_csv('data\\owa.csv',   
 col\_names = TRUE,  
 locale=locale('ko', encoding='euc-kr'), # 한글  
 na=".") %>%  
 round(2) %>% # 소수점 2자리로 반올림  
 mutate\_if(is.character, as.factor) %>%  
 mutate(매장명 = factor(매장명,  
 levels=c(1:4),  
 labels=c("강남","강서","강동","강북")))

## Rows: 139 Columns: 3  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (3): 번호, 매장명, 만족도  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

str(owa\_tb)

## spec\_tbl\_df [139 x 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ 번호 : num [1:139] 1 2 3 4 5 6 7 8 9 10 ...  
## $ 매장명: Factor w/ 4 levels "강남","강서",..: 1 2 3 4 1 2 3 4 1 2 ...  
## $ 만족도: num [1:139] 85 82 87 88 93 83 86 85 94 93 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. 번호 = col\_double(),  
## .. 매장명 = col\_double(),  
## .. 만족도 = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

owa\_tb

## # A tibble: 139 x 3  
## 번호 매장명 만족도  
## <dbl> <fct> <dbl>  
## 1 1 강남 85  
## 2 2 강서 82  
## 3 3 강동 87  
## 4 4 강북 88  
## 5 5 강남 93  
## 6 6 강서 83  
## 7 7 강동 86  
## 8 8 강북 85  
## 9 9 강남 94  
## 10 10 강서 93  
## # ... with 129 more rows

# 3.기술통계분석

skim(owa\_tb)

Data summary

|  |  |
| --- | --- |
| Name | owa\_tb |
| Number of rows | 139 |
| Number of columns | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| 매장명 | 0 | 1 | FALSE | 4 | 강북: 39, 강남: 38, 강서: 32, 강동: 30 |

**Variable type: numeric**

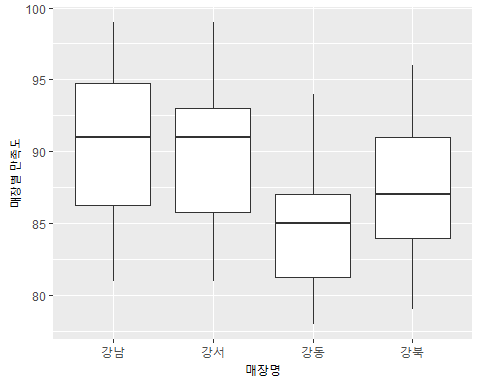
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 번호 | 0 | 1 | 70.00 | 40.27 | 1 | 35.5 | 70 | 104.5 | 139 | ▇▇▇▇▇ |
| 만족도 | 0 | 1 | 88.09 | 5.23 | 78 | 84.0 | 87 | 92.0 | 99 | ▅▇▇▇▃ |

owa\_tb %>%   
 group\_by(매장명) %>%  
 get\_summary\_stats(만족도)

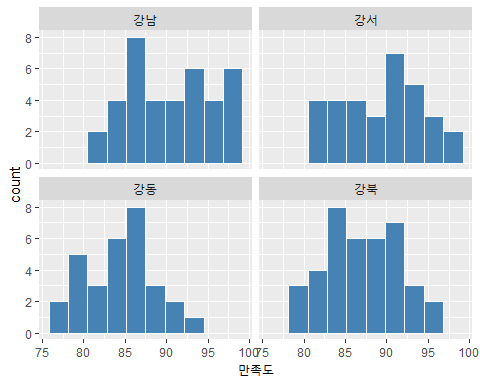
## # A tibble: 4 x 14  
## 매장명 variable n min max median q1 q3 iqr mad mean sd  
## <fct> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 강남 만족도 38 81 99 91 86.2 94.8 8.5 5.93 90.6 5.12  
## 2 강서 만족도 32 81 99 91 85.8 93 7.25 5.93 89.6 5.03  
## 3 강동 만족도 30 78 94 85 81.2 87 5.75 4.45 84.5 4.25  
## 4 강북 만족도 39 79 96 87 84 91 7 5.93 87.2 4.50  
## # ... with 2 more variables: se <dbl>, ci <dbl>

# 4.그래프 그리기(박스그래프,히스토그램)

owa\_tb %>%   
 ggplot(mapping = aes(x = 매장명,  
 y = 만족도)) +  
 geom\_boxplot() +  
 labs(y = "매장별 만족도")



owa\_tb %>%   
 ggplot(mapping = aes(x = 만족도)) +  
 geom\_histogram(bins = 10, # binwidth=1,000: 값차이  
 color = "white",   
 fill = "steelblue") +  
 facet\_wrap(~ 매장명) # 그룹별 분리



# 5.이상치 제거, 이상치 확인

owa\_tb %>%  
 group\_by(매장명) %>%  
 identify\_outliers(만족도)

## [1] 매장명 번호 만족도 is.outlier is.extreme  
## <0 행> <또는 row.names의 길이가 0입니다>

# 6.정규분포 검정

owa\_tb %>%  
 group\_by(매장명) %>%  
 shapiro\_test(만족도)

## # A tibble: 4 x 4  
## 매장명 variable statistic p  
## <fct> <chr> <dbl> <dbl>  
## 1 강남 만족도 0.956 0.137  
## 2 강서 만족도 0.957 0.231  
## 3 강동 만족도 0.959 0.299  
## 4 강북 만족도 0.973 0.454

# 7.등분산 검정

이분산일때는 하단의 Welch’s ANOVA test 참조

owa\_tb %>%  
 levene\_test(만족도 ~ 매장명)

## # A tibble: 1 x 4  
## df1 df2 statistic p  
## <int> <int> <dbl> <dbl>  
## 1 3 135 1.06 0.369

# 8.등분산 ANOVA

등분산일때 ANOVA분석

owa\_result <- aov(만족도 ~ 매장명,   
 data=owa\_tb)  
tidy(owa\_result)

## # A tibble: 2 x 6  
## term df sumsq meansq statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 매장명 3 734. 245. 10.8 0.00000201  
## 2 Residuals 135 3047. 22.6 NA NA

owa\_tb %>%  
 anova\_test(만족도 ~ 매장명)

## Coefficient covariances computed by hccm()

## ANOVA Table (type II tests)  
##   
## Effect DFn DFd F p p<.05 ges  
## 1 매장명 3 135 10.833 2.01e-06 \* 0.194

Cohen’s d(effect size) 0.2 (small effect), 0.5 (moderate effect) and 0.8 (large effect)

owa\_tb %>%   
 cohens\_d(formula = 만족도 ~ 매장명,   
 var.equal = TRUE)

## # A tibble: 6 x 7  
## .y. group1 group2 effsize n1 n2 magnitude   
## \* <chr> <chr> <chr> <dbl> <int> <int> <ord>   
## 1 만족도 강남 강서 0.199 38 32 negligible  
## 2 만족도 강남 강동 1.28 38 30 large   
## 3 만족도 강남 강북 0.717 38 39 moderate   
## 4 만족도 강서 강동 1.09 32 30 large   
## 5 만족도 강서 강북 0.514 32 39 moderate   
## 6 만족도 강동 강북 -0.604 30 39 moderate

# 9.사후검정(Multicamparison test )

## Fisher LSD

owa\_tb %>%   
 pairwise\_t\_test(만족도 ~ 매장명,   
 p.adj="non") %>% # 아무것도 없다(non = Fisher)   
 mutate\_if(is.numeric, round, 5)

## # A tibble: 6 x 9  
## .y. group1 group2 n1 n2 p p.signif p.adj p.adj.signif  
## \* <chr> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>   
## 1 만족도 강남 강서 38 32 0.376 ns 0.376 ns   
## 2 만족도 강남 강동 38 30 0 \*\*\*\* 0 \*\*\*\*   
## 3 만족도 강서 강동 32 30 0.00004 \*\*\*\* 0.00004 \*\*\*\*   
## 4 만족도 강남 강북 38 39 0.00179 \*\* 0.00179 \*\*   
## 5 만족도 강서 강북 32 39 0.0331 \* 0.0331 \*   
## 6 만족도 강동 강북 30 39 0.023 \* 0.023 \*

## Bonferroni

owa\_tb %>%   
 pairwise\_t\_test(만족도 ~ 매장명,   
 p.adj="bonferroni") %>%  
 mutate\_if(is.numeric, round, 5)

## # A tibble: 6 x 9  
## .y. group1 group2 n1 n2 p p.signif p.adj p.adj.signif  
## \* <chr> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>   
## 1 만족도 강남 강서 38 32 0.376 ns 1 ns   
## 2 만족도 강남 강동 38 30 0 \*\*\*\* 0 \*\*\*\*   
## 3 만족도 강서 강동 32 30 0.00004 \*\*\*\* 0.00027 \*\*\*   
## 4 만족도 강남 강북 38 39 0.00179 \*\* 0.0107 \*   
## 5 만족도 강서 강북 32 39 0.0331 \* 0.199 ns   
## 6 만족도 강동 강북 30 39 0.023 \* 0.138 ns

## Tukey HSD

owa\_result %>%   
 TukeyHSD() %>%   
 tidy()

## # A tibble: 6 x 7  
## term contrast null.value estimate conf.low conf.high adj.p.value  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 매장명 강서-강남 0 -1.01 -3.98 1.95 0.811   
## 2 매장명 강동-강남 0 -6.11 -9.12 -3.09 0.00000325  
## 3 매장명 강북-강남 0 -3.45 -6.27 -0.634 0.00956   
## 4 매장명 강동-강서 0 -5.09 -8.23 -1.95 0.000259   
## 5 매장명 강북-강서 0 -2.44 -5.39 0.508 0.142   
## 6 매장명 강북-강동 0 2.65 -0.348 5.66 0.103

## 매장명으로 표현

console=TRUE: 결과를 화면에 표시 매장명=TRUE: 그룹으로 묶어서 표시, FALSE: 1:1로 비교 LSD.test, duncan.test, scheffe.test로 구분 LSD.test

owa\_result %>%  
 LSD.test("매장명",  
 group=TRUE,   
 console = TRUE)

##   
## Study: . ~ "매장명"  
##   
## LSD t Test for 만족도   
##   
## Mean Square Error: 22.57315   
##   
## 매장명, means and individual ( 95 %) CI  
##   
## 만족도 std r LCL UCL Min Max  
## 강남 90.60526 5.123024 38 89.08099 92.12954 81 99  
## 강동 84.50000 4.248732 30 82.78449 86.21551 78 94  
## 강북 87.15385 4.498763 39 85.64924 88.65845 79 96  
## 강서 89.59375 5.028046 32 87.93271 91.25479 81 99  
##   
## Alpha: 0.05 ; DF Error: 135  
## Critical Value of t: 1.977692   
##   
## Groups according to probability of means differences and alpha level( 0.05 )  
##   
## Treatments with the same letter are not significantly different.  
##   
## 만족도 groups  
## 강남 90.60526 a  
## 강서 89.59375 a  
## 강북 87.15385 b  
## 강동 84.50000 c

duncan.test

owa\_result %>%  
 duncan.test("매장명",  
 group=TRUE,   
 console = TRUE)

##   
## Study: . ~ "매장명"  
##   
## Duncan's new multiple range test  
## for 만족도   
##   
## Mean Square Error: 22.57315   
##   
## 매장명, means  
##   
## 만족도 std r Min Max  
## 강남 90.60526 5.123024 38 81 99  
## 강동 84.50000 4.248732 30 78 94  
## 강북 87.15385 4.498763 39 79 96  
## 강서 89.59375 5.028046 32 81 99  
##   
## Groups according to probability of means differences and alpha level( 0.05 )  
##   
## Means with the same letter are not significantly different.  
##   
## 만족도 groups  
## 강남 90.60526 a  
## 강서 89.59375 a  
## 강북 87.15385 b  
## 강동 84.50000 c

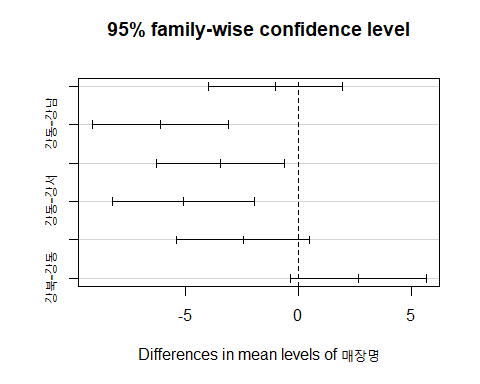
scheffe.test

owa\_result %>%  
 scheffe.test("매장명",  
 group=TRUE,   
 console = TRUE)

##   
## Study: . ~ "매장명"  
##   
## Scheffe Test for 만족도   
##   
## Mean Square Error : 22.57315   
##   
## 매장명, means  
##   
## 만족도 std r Min Max  
## 강남 90.60526 5.123024 38 81 99  
## 강동 84.50000 4.248732 30 78 94  
## 강북 87.15385 4.498763 39 79 96  
## 강서 89.59375 5.028046 32 81 99  
##   
## Alpha: 0.05 ; DF Error: 135   
## Critical Value of F: 2.671676   
##   
## Groups according to probability of means differences and alpha level( 0.05 )  
##   
## Means with the same letter are not significantly different.  
##   
## 만족도 groups  
## 강남 90.60526 a  
## 강서 89.59375 ab  
## 강북 87.15385 bc  
## 강동 84.50000 c

## 다중비교 그래프

owa\_result %>%   
 TukeyHSD() %>%   
 plot()



# 10.추론(infer)을 이용한 가설검정 및 그래프

## f\_cal 계산

f\_cal <- owa\_tb %>%  
 specify(formula = 만족도 ~ 매장명) %>% # hypothesize 없음  
 calculate(stat = "F") %>%  
 print()

## Response: 만족도 (numeric)  
## Explanatory: 매장명 (factor)  
## # A tibble: 1 x 1  
## stat  
## <dbl>  
## 1 10.8

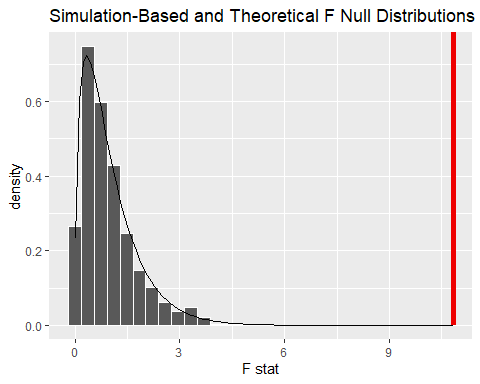
## Bootstrapping을 이용한 귀무가설 분포 생성

set.seed(123)   
null\_dist\_f <- owa\_tb %>%  
 specify(formula = 만족도 ~ 매장명) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "F") %>%  
 print()

## Response: 만족도 (numeric)  
## Explanatory: 매장명 (factor)  
## Null Hypothesis: independence  
## # A tibble: 1,000 x 2  
## replicate stat  
## <int> <dbl>  
## 1 1 1.15   
## 2 2 3.40   
## 3 3 1.04   
## 4 4 0.288  
## 5 5 0.238  
## 6 6 0.845  
## 7 7 1.15   
## 8 8 0.773  
## 9 9 0.468  
## 10 10 0.107  
## # ... with 990 more rows

null\_dist\_f %>%  
 visualize(method = "both") + #method = "both": 이론분포+boot분포  
 shade\_p\_value(obs\_stat = f\_cal,  
 direction = "greater")

## Warning: Check to make sure the conditions have been met for the theoretical  
## method. {infer} currently does not check these for you.

 ## p\_value

null\_dist\_f %>%  
 get\_p\_value(obs\_stat = f\_cal,   
 direction = "greater") %>%  
 print()

## Warning: Please be cautious in reporting a p-value of 0. This result is an  
## approximation based on the number of `reps` chosen in the `generate()` step. See  
## `?get\_p\_value()` for more information.

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0

# 부록: 이분산(levene\_test < 0.05) 일때 Welch’s ANOVA test

owa\_tb %>%   
 welch\_anova\_test(만족도 ~ 매장명)

## # A tibble: 1 x 7  
## .y. n statistic DFn DFd p method   
## \* <chr> <int> <dbl> <dbl> <dbl> <dbl> <chr>   
## 1 만족도 139 11.2 3 73.5 0.00000373 Welch ANOVA

owa\_tb %>%   
 pairwise\_t\_test(만족도 ~ 매장명,   
 pool.sd = FALSE, # 등분산 가정 안함  
 p.adj="bonferroni")

## # A tibble: 6 x 10  
## .y. group1 group2 n1 n2 statistic df p p.adj p.adj.signif  
## \* <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <chr>   
## 1 만족도 강남 강서 38 32 0.831 66.4 4.09e-1 1 e+0 ns   
## 2 만족도 강남 강동 38 30 5.37 65.8 1.1 e-6 6.6 e-6 \*\*\*\*   
## 3 만족도 강남 강북 38 39 3.14 73.2 2 e-3 1.5 e-2 \*   
## 4 만족도 강서 강동 32 30 4.32 59.4 6.07e-5 3.64e-4 \*\*\*   
## 5 만족도 강서 강북 32 39 2.13 62.9 3.7 e-2 2.21e-1 ns   
## 6 만족도 강동 강북 30 39 -2.51 64.2 1.5 e-2 8.8 e-2 ns

# 부록: 정규분포 가정 문제가 있으면 shapiro\_test < 0.05

비모수통계분석 Kruskal Wallis H test

owa\_tb %>%  
 kruskal\_test(만족도 ~ 매장명)

## # A tibble: 1 x 6  
## .y. n statistic df p method   
## \* <chr> <int> <dbl> <int> <dbl> <chr>   
## 1 만족도 139 24.9 3 0.0000166 Kruskal-Wallis

비모수 다중비교 Dunn test

owa\_tb %>%   
 dunn\_test(만족도 ~ 매장명)

## # A tibble: 6 x 9  
## .y. group1 group2 n1 n2 statistic p p.adj p.adj.signif  
## \* <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <chr>   
## 1 만족도 강남 강서 38 32 -0.734 0.463 0.463 ns   
## 2 만족도 강남 강동 38 30 -4.59 0.00000437 0.0000262 \*\*\*\*   
## 3 만족도 강남 강북 38 39 -2.75 0.00591 0.0237 \*   
## 4 만족도 강서 강동 32 30 -3.72 0.000198 0.000991 \*\*\*   
## 5 만족도 강서 강북 32 39 -1.89 0.0584 0.125 ns   
## 6 만족도 강동 강북 30 39 2.04 0.0418 0.125 ns