1. 비모수검정이란?

z, t, F등의 검정통계량을 활용하는 분석을 '모수검정'이라고 한다. 모수검정을 하기위해서는 몇 가지 가정을 만족해야한다. 첫째, 표본의 모집단이 정규분포를 따라야하고, 둘째, 등분산 가정을 만족해야하고 셋째, 독립변수는 등간척도나 비율척도로 측정되어야한다. 넷째, 최소 변수 한 개당 30개의 표본이 필요하다.

비모수검정은 이와 같은 모수검정의 가정을 충족하지 않는 표본을 대상으로 가설검정할 경우 활용하는 벙법이다.

2. 비모수검정 방법의 유형과 기본 개념

1) 유형

다음의 표는 비모수검정을 모수검정과 비교하여 보여준다.

목적	모수검정	비모수검정
단일 표본 분석	단일 표본 t검정	Kolmogorov-Smirnov test
두 표본 간 비교	독립표본 t검정 대응표본 t검정	Wilcoxon Rank-Sum test Wilcoxon Signed-Rank test
여러 표본 간 비교	분산분석(ANOVA)	Kruskal-Wallis test
상관관계의 분석	피어슨의 상관관계분석	스피어만의 순위 상관관계분석 켄달의 순위 상관관계분석

단일 표본 t검정은 표본의 특성이 모집단의 특성과 통계적으로 차이가 있는지를 검정하는 것인데 반하여 Kolmogorov-Smirnov test는 단일 표본 집단이 정규분포를 따르는지 여부를 검정하는 비모수검정 방법이다.

2) Kolmogorov-Smirnov test

Kolmogorov-Smirnov test는 단일 표본과 두 표본 모두 분석이 가능하다. 단일 표본의 경우 분석 대상이 되는 표본 집단이 정규분포(혹은 연구자가 지정한 분포)를 이루는지 여부를 분석하는 것이고, 두 표본의 경우는 분석 대상이 되는 두 표본 집단이 동일한 분포를 이루는지를 분석하는 것이다. 검정통계량은 D이다.

3) Wilcoxon Rank-Sum test

Wilcoxon Rank-Sum test는 독립표본 t검정과 마찬가지로 두 개의 표본 집단 간 차이를 통계적으로 규명하는 것이다. 독립표본 t검정과 달리 변수의 분포가 정규분포를 반드시 따를 필요가 없고, 비율변수가 아닌 서열변수로 구성된 집단의 특성을 분석하기에도 적합하다. 검정통계량은 W이다.

4) Wilcoxon Signed-Rank test

Wilcoxon Signed-Rank test는 대응표본 t검정과 매우 유사한 원리를 가진 비모수통계 방법이다. 이 방법은 평균이 0인지 아닌지를 검정하는 t검정과 달리 데이터의 순서를 활용하여 데이터의 중위소(median)이 0인지 아닌지를 검정하는 방법이다. 검정통계량은 V이다.

5) Kruskal-Wallis test

Kruskal-Wallis test는 Wilcoxon test와 유사하게 순위에 의한 집단 간 차이를 통계적으로 규명하는 비모수검정 방법이다. Wilcoxon test는 두 표본 집단 간의 차이만을 규명할수 있는데 비해 Kruskal-Wallis test는 세 집단 이상의 차이도 규명할 수 있다. 따라서 등분산을 가정할 수 없거나 변수가 정규분포를 따르지 않는 경우 Kruskal-Wallis test를 활용할 수 있다. 검정통계량은 H이다.

6) 스피어만 순위 상관관계분석

스피어만 상관계수는 데이터가 비율척도인 피어슨의 상관계수와는 달리 서열척도인 경우에 데이터값 대신에 이 데이터의 순위를 이용하는 상관계수이다. 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용하여 상관계수를 구하게 된다. 두 변수 간의 연관 관계가 있는지 없는지를 밝혀 주며 데이터에 이상값(outlier)이 있거나 표본 크기가 작을 때 유용하다.

스피어만 상관계수는 -1과 1사이의 값을 가지는데 두 변수 안의 순위가 완전히 일치하면 1, 완전히 반대이면 -1이다.

3. R 활용 비모수 방법

1) Kolmogorov-Smirnov test

Kolmogorov-Smirnov test를 진행하여 결과를 도출하는 과정은 다음과 같다.

```
> setwd("C:/Users/USER/Desktop/교육원초급/data/xlsx 파일")
> library(openxlsx)
> mydata11 = read.xlsx("kn_non.xlsx",1) # 파을을 불러와 mydata11로 저장
> attach(mydata11) # 불러온 데이터를 R과 연동
> View(mydata11) # 불러온 데이터를 R Studio에서 확인
```

cities	ID	birth_rate
경기 수원시	1	1.292
경기 성남시	2	1.159
경기 의정부시	3	1.104
경기 안양시	4	1.177
경기 부천시	5	1.072
경기 광명시	6	1.235
경기 평택시	7	1.469
경기 동두천시	8	1.292
경기 안산시	9	1.219
경기 고양시	10	1.161

```
> names(mydata11) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "cities"      "ID"          "birth_rate"
> ks.test(birth_rate,"pnorm") # birth_rate가 정규분포인지 여부를 검정

        One-sample Kolmogorov-Smirnov test

data:  birth_rate
D = 0.85814, p-value = 3.224e-13
alternative hypothesis: two-sided
```

결과를 통해 유의수준이 0.01에서 통계적으로 유의함을 알 수 있다. 따라서 표본이 정규분포를 따른다는 가설을 기각하여 정규분포를 따르지 않는다고 할 수 있다.

2) Wilcoxon Rank-Sum test

독립적인 두 표본이 하나의 모집단에서 추출되었는지 여부를 분석하고자 한다.

```
> mydata12 = read.xlsx("wil_indep.xlsx",1) # 파을을 불러와 mydata12로 저장
> attach(mydata12) # 불러온 데이터를 R과 연동
> View(mydata12) # 불러온 데이터를 R Studio에서 확인
```

cities	birth_rate	dummy
경기 수원시	1.292	1
경기 성남시	1.159	1
경기 의정부시	1.104	1
경기 안양시	1.177	1
경기 부천시	1.072	1
경기 광명시	1.235	1
경기 평택시	1.469	1
경기 동두천시	1.292	1
경기 안산시	1.219	1
경기 고양시	1.161	1

```
> names(mydata12) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "cities"     "birth_rate" "dummy"
> wilcox.test(birth_rate~dummy) # 종속변수가 앞에, 독립변수가 뒤에

        Wilcoxon rank sum test with continuity correction

data:  birth_rate by dummy
W = 570.5, p-value = 0.0001543
alternative hypothesis: true location shift is not equal to 0
```

결과를 통해 유의수준 0.01에서 통계적으로 유의하게 나타났다. 따라서 두 집단 간 차이가 있다고 할 수 있다.

3) Wilcoxon Signed-Rank test

대응포본 t검정의 비모수 버전인 Wilcoxon Signed-Rank test분석을 시행하여 결과를 도출하는 과정은 다음과 같다.

```
> mydata13 = read.xlsx("wil_paired.xlsx",1) # 파을을 불러와 mydata13로 저장
> attach(mydata13) # 불러온 데이터를 R과 연동
> View(mydata13) # 불러온 데이터를 R Studio에서 확인
```

cities	birth_rate_2015	birth_rate_2010
경기 수원시	1.292	1.226
경기 성남시	1.159	1.170
경기 의정부시	1.104	1.179
경기 과천시	1.099	1.148
경기 구리시	1.159	1.232
경기 남양주시	1.255	1.349
경기 오산시	1.458	1.589
경기 시흥시	1.306	1.478
전북 군산시	1.495	1.416
전북 익산시	1.341	1.337

```
> names(mydata13) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "cities"          "birth_rate_2015" "birth_rate_2010"
> wilcox.test(birth_rate_2015,birth_rate_2010) # 두 시점의 통계적 비교 시행

        Wilcoxon rank sum test with continuity correction

data:  birth_rate_2015 and birth_rate_2010
W = 438, p-value = 0.7915
alternative hypothesis: true location shift is not equal to 0
```

결과를 통해 유의확률은 통계적으로 유의하지 않다. 따라서 두 집단 간 차이가 없다는 귀무가설이 채택되었음을 알 수 있다.

4) Kruskal-Wallis test

일원분산분석의 비모수 버전인 Kruskal-Wallis test분석을 시행하여 결과를 도출하는 과정은 다음과 같다.

```
> mydata14 = read.xlsx("kw_non.xlsx",1) # 파을을 불러와 mydata14로 저장
> attach(mydata14) # 불러온 데이터를 R과 연동
> View(mydata14) # 불러온 데이터를 R Studio에서 확인
```

cities	birth_rate	ad_layer
경기 시흥시	1.306	자치시
경기 군포시	1.422	자치시
경기 의왕시	1.209	자치시
경기 하남시	1.076	자치시
경기 용인시	1.316	자치시
경기 파주시	1.443	자치시
경기 미천시	1.470	자치시
경기 안성시	1.335	자치시
경기 김포시	1.461	자치시
경기 화성시	1.555	자치시

```
> names(mydata14) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "cities"     "birth_rate" "ad_layer"
> mydata14$ad_layer=as.factor(ad_layer)
> kruskal.test(birth_rate~ad_layer,data=mydata14) # 종속변수 : birth_rate, 독립변수 : ad_layer인 kruskal-Wallis test 시행

        Kruskal-Wallis rank sum test

data:  birth_rate by ad_layer
Kruskal-Wallis chi-squared = 64.596, df = 2, p-value = 9.4e-15
```

유의확률이 9.4e-15로 유의수준 0.01에서 유의하다. 따라서 귀무가설이 기각되어 집단 간 차이가 있다는 대립가설이 채택되었음을 알 수 있다.

5) 스피어만 순위 상관관계분석

피어슨 상관관계분석의 비모수 버전인 스피어만의 순위 상관계수분석을 시행하여 결과를 도출하는 과정은 다음과 같다. 인구증가율과 합계출산율의 상관계수를 도출하는 것이 목적이다.

```
> mydata15 = read.xlsx("spearman.xlsx",1) # 파을을 불러와 mydata15로 저장
> attach(mydata15) # 불러온 데이터를 R과 연동
> View(mydata15) # 불러온 데이터를 R Studio에서 확인
```

군별	pop_growth	birth_rate
경기 수원시	1.06	1.292
경기 성남시	-0.26	1.159
경기 의정부시	0.69	1.104
경기 안양시	-0.47	1.177
경기 부천시	-0.64	1.072
경기 광명시	-0.88	1.235
경기 평택시	2.67	1.469
경기 동두천시	0.52	1.292
경기 안산시	-1.05	1.219
경기 고양시	2.14	1.161

```
> names(mydata15) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "군별"         "pop_growth" "birth_rate"
> cor.test(pop_growth,birth_rate,data=mydata15,method="spearman") # 스피어만 상관계수 도출

        Spearman's rank correlation rho

data:  pop_growth and birth_rate
S = 48572, p-value = 0.006974
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3090728
```

결과를 통해 rho가 약 0.31로 도출되었음을 알 수 있다.

```
> cor.test(pop_growth,birth_rate,data=mydata15) # method의 디폴트는 피어슨 상관계수 도출

        Pearson's product-moment correlation

data:  pop_growth and birth_rate
t = 1.7602, df = 73, p-value = 0.08256
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02638792  0.40996894
sample estimates:
      cor
0.2017824
```

같은 자료로 method만 수정하여 피어슨의 상관계수를 구하면 약 0.20이었다.

만약 데이터에 이상값이 존재하고 정규분포를 따르지 않는다면 스피어만

상관계수를 활용하고 평균이 이상값에 영향을 많이 받지 않고 정규분포를 따르고 있다면 피어슨 상관계수를 활용하는 것이 바람직하다.