Varsayım Kontrolü ve Sıra atama İşlemleri

Doç. Dr. Nihat Tak

2023-03-16

Veri(Hatırlatma)

```
setwd("C:/Users/nihattak/Desktop/Verilen Dersler-Marmara/nonparametric")
kiraz1<-read.table("kiraz1.txt",head=TRUE,sep = "\t")
kiraz1$grup<-as.factor(kiraz1$grup)
kiraz2<-read.table("kiraz2.txt",head=TRUE,sep = "\t")
kiraz2$grup<-as.factor(kiraz2$grup)
str(kiraz1)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ magr: num  5.8 5.7 5.4 4.9 6.7 4.8 6.3 5.9 5.3 5.9 ...
##  $ grup: Factor w/ 3 levels "A","B","K": 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(kiraz1)
```

```
##   magr grup
## 1  5.8    A
## 2  5.7    A
## 3  5.4    A
## 4  4.9    A
## 5  6.7    A
## 6  4.8    A
```

```
tail(kiraz1)
```

```
##    magr grup
## 25  5.5    K
## 26  5.6    K
## 27  6.2    K
## 28  5.5    K
## 29  6.3    K
## 30  6.4    K
```

```
levels(kiraz2$grup)
```

```
## [1] "A" "B" "K"
```

Düzeltme

Ayrıca FSA paketindeki Summarize fonksiyonu da grup istatistiklerini hesaplamak için bir başka seçenektir.

```
if(!require(FSA)){install.packages("FSA");require(FSA)}
Summarize(magr~grup,data=kiraz1)
```

```
##   grup  n mean        sd min    Q1 median    Q3 max
```

```
## 1    A 10 5.67 0.5907622 4.8 5.325   5.75 5.900 6.7
## 2    B 10 6.52 0.4131182 5.8 6.325   6.45 6.675 7.2
## 3    K 10 6.05 0.5835714 5.2 5.525   6.20 6.375 7.1
```

```
Summarize(magr~grup,data=kiraz1)
```

```
##   grup  n mean        sd min    Q1 median    Q3 max
## 1    A 10 5.67 0.5907622 4.8 5.325   5.75 5.900 6.7
## 2    B 10 6.52 0.4131182 5.8 6.325   6.45 6.675 7.2
## 3    K 10 6.05 0.5835714 5.2 5.525   6.20 6.375 7.1
```

Varyansların Homojenliğinin Kontrolü

Karşılaştırılan gruplara ait varyansların benzer olmalarına varyansların homojenliği yada homoscedasticity denir. Eğer gruplara ait varyansların eşitliği söz konusu değil ise parametrik testler güvenilir sonuç vermezler. Bu durumda, gruplara ait varyansların homojen olmasına yönelik dönüşümlerden veya parametrik olmayan istatistikten yararlanılır. Gruplara ait varyansların homojen olup olmadığını anlamak için Levene testi, Brown-Forsythe testi, veya Barlett K-kare testinden faydalanılır.

Levene Testi Aşağıda car paketindeki leveneTest fonksiyonu ile kiraz1 ve kiraz2 verisindeki magr değişkenin gruplara göre varyans homojenliği testi yapılmıştır.

```
if(!require(car)){install.packages("car");require(car)}
leveneTest(magr~grup,data=kiraz1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2   0.647 0.5316
##       27
```

```
leveneTest(magr~grup,data=kiraz2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  4.7202 0.01746 *
##       27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene testinde H0 hipotezi gruplara ait varyansların homojen olduğunu söylerken, H1 hipotezi gruplara ait varyansların homojen olmadığını ifade eder. P-değerinin araştırmacı tarafından belirlenen önemlilik seviyisinden küçük olması gruplar arası varyansın istatistiksel olarak homojen olmadığını işaret eder. Bu durumda kiraz1 verisi için gruplar arası varyansların homojen olduğu gözlemlenirken, kiraz2 verisi için bu homojenlik istatiksel olarak gözlemlenmemiştir.

Brown-Forsyth Testi HH paketindeki hovBF fonksiyonu ile Brown-Forsyth testi gruplara ait varyansların homojenliğini araştırmakta kullanılan testlerden bir diğeridir.

```
if(!require(HH)){install.packages("HH");}
hovBF(magr~grup,data=kiraz1)
```

```
##
##  hov: Brown-Forsyth
##
## data:  magr
```

```
## F = 0.64696, df:grup = 2, df:Residuals = 27, p-value = 0.5316
## alternative hypothesis: variances are not identical
```

```
hovBF(magr~grup,data=kiraz2)
```

```
##
##  hov: Brown-Forsyth
##
## data:  magr
## F = 4.7202, df:grup = 2, df:Residuals = 27, p-value = 0.01746
## alternative hypothesis: variances are not identical
```

Brown-Forsyth testi ile elde edilen sonuçlar kiraz1 verisinde grup varyanslarının homojen (p=0.5316) iken kiraz2 verisi için grup varyansları heterojen (p=0.01401) bulunmuştur.

Bağımsızlık Testi Örneklem Grupları içinde yer alan her bir gözlemin diğerlerinden bağımsız olması gerekir yani ilişkili olmaması gereklidir. Bağımsızlık, örneklem içinde yer alan bir örnekte ölçülen/gözlenen değerin diğer örneklem birimindeki ölçülen/gözlenen değerleri etkilememesi olarak daa tanımlanabilir. Bu durumun testip edilebilmesi için kullanılabilecek birçok test literatürde mevcuttur. Bunlar arasından en yaygın olanı Durbin-Watson testidir. Bu testi R' de car paketinde bulunan durbinWatsonTest fonksiyonu ile test edebiliriz.

```
model1<-lm(magr~grup,data=kiraz1)
durbinWatsonTest(model1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.3863859      2.754766   0.066
##  Alternative hypothesis: rho != 0
```

```
model2<-lm(magr~grup,data=kiraz2)
durbinWatsonTest(model2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       -0.110759       2.15543   0.976
##  Alternative hypothesis: rho != 0
```

Bu sonuçlara göre kiraz1 verisinde kalıntıların otekorelasyonunun negatif ve orta düzey(-0.386) ancak önemsiz olduğu (p=0.084), kiraz2 verisinde ise çok küçük (0.03) ve aynı zamanda önemsiz olduğu (p=0.504) bulunmuştur. Bu sonuç her iki verisetinde de modele ait hataların dolayısıyla örneklem gruplarındaki birimlerin birbirinden bağımsız olduklarını göstermektedir.

Sıra Atama İşlemleri

Parametrik olamayan istatistik testlerin çoğunda orjinal gözlem değerleri yerine sıra değerleri (rank) kullanılır. X rastlantısal değişkeni değerlerinin küçükten büyüğe veya büyükten küçüğe sıraya konulmasına değişkenin sıralanması (sorting) veya sıra düzenine sokulması(ordering) olarak adlandırılır. Sıralanmış değerlere konumlarına göre yada dizideki yerlerine göre üstünlüklerini yada derecelerini gösteren değer atama işlemine sıra numarası verme veya sıra değeri atama (ranking) denir. Bu bir derecelendirme yada rütbe belirleme işlemidir. Gözlemin konum veya yerine göre aldı değer sıra değeri (rank) olarak adlandırılır.

Sıra Değeri Atama Yöntemleri

Sıra değeri atama,n gözlemden oluşan bir rastlantısal değişkenin sıralanmış gözlemlerine siradaki yerlerine ya da konumlarına göre 1,2,. . . ,n arasında sıra numarası atanarak yapılır. Örneğin 6 gözlemden oluşan bir X rastlantı değişkeni;

$x : 18, 14, 16, 12, 11, 20$

olduğunda, küçükten büyüğe sıralanırsa;

$x' : 11, 13, 14, 16, 18, 20$

elde edilir. Yukarıda sıralanmış gözlem değerlerine bulundukları konuma göre sıra numarası verilirse;

$r' : 1, 2, 3, 4, 5, 6$ elde edilir.

Dolayısıyla orjinal (x) gözlemleri (sıralanmamış) düşünüldüğünde, 18 değerinin sırası 5, 14 değerinin sırası 3, etc. olduğu açıktır. Bu örnekte birbirine eşit olan gözlem değerleri olmadığından sıralama ve sıra sayısı atama işlemleri kolaylıkla yapılmıştır.

Ancak gerçek uygulamaların bir çoğunda bir rastlantı değişkenin gözlemlenen değerleri birbirine eşit olabilir. Örneğin X rastgele değişkeni aşağıdaki gözlemlerde oluşuyorsa;

$x : 14, 10, 11, 14, 15, 19, 18, 17, 15, 17, 20, 14, 13, 16, 19$

Bu değişkende, 3 tane 14 değeri, iki tane 17 değeri ve 2 tane 19 değeri bulunmaktadır. Bu gibi durumlarda sıra numaralarının belirlenmesi için özel tekniklere gereksinim vardır. Yukarıdaki örnekte olduğu gibi, birbirine eşit değerlere sahip gözlemlere sıra numarası atanırken çeşitli yöntemler vardır. Bunlardan bazıları;

- Öndeki kazanır yöntemi
- Sondaki kazanır yöntemi
- Minimum yöntemi
- Maksimum yöntemi
- Ortalama yöntemi
- Rastlantısal yöntem

Yukarıdaki değişkenin gözlemlerinin küçükten büyüğe sıralanarak sıralanmış gözlem değerleri aşağıdaki gibi elde edilir.

$x' : 10, 11, 13, 14, 14, 14, 15, 15, 16, 17, 17, 18, 19, 19, 20$

Öndeki kazanır (first wins) yöntemi genel olarak en basit yöntemdir. Sıralanmış gözlemlere sıralamadaki konumları sıra değeri olarak atanır.

x':	10	11	13	14	14	14	15	15	16	17	17	18	19	19	20
r:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Sondaki kazanır (last wins) yönteminde sıralanmış gözlemlerden aynı değerde olanlardan en sondaki önceliği alır.

x':	10	11	13	14	14	14	15	15	16	17	17	18	19	19	20
r:	1	2	3	**6**	**5**	**4**	**8**	**7**	9	**11**	**10**	12	**14**	**13**	15

Minimum yönteminde, eş değerli gözlemlerin hepsine ilkinin konumu sıra değeri olarak atanır. Bu yöntem spor ve sanat etkinliklerinde sıkça kullanılan yöntemdir.

x':	10	11	13	14	14	14	15	15	16	17	17	18	19	19	20
r:	1	2	3	**4**	**4**	**4**	**7**	**7**	9	**10**	**10**	12	**13**	**13**	15

Maksimum yöntemi minimum yönteminin tersidir. Eş değerli gözlemlerin sıra değeri olarak sonuncusunun

konumu atanır.

x':	10	11	13	14	14	14	15	15	16	17	17	18	19	19	20
r:	1	2	3	**6**	**6**	**6**	**8**	**8**	9	**11**	**11**	12	**14**	**14**	15

Ortalama yöntemi (average) , minimum ve maksimum yöntemlerinin ortalamasını alır. Eş değerli gözlemlerin hepsine sıra değeri olarak konumlarının ortalaması atanır.

x':	10	11	13	14	14	14	15	15	16	17	17	18	19	19	20
r:	1	2	3	**5**	**5**	**5**	**7.5**	**7.5**	9	**10.5**	**10.5**	12	**13.5**	**13.5**	15

Rastlantısal yöntemde, eş değerli gözlemlerin sıra değeri rastgele olarak yada kura ile belirlenir.

x':	10	11	13	14	14	14	15	15	16	17	17	18	19	19	20
r:	1	2	3	**5**	**4**	**6**	**8**	**7**	9	**10**	**11**	12	**14**	**13**	15

not: iki yada daha fazla değişkene dayalık istastik analizlerde hangi sıra atama yöntemi kullanılırsa kullanılsın, analize konu olan değişkenlerin hepsine aynı yöntem uygulanmalıdır.

not: Parametrik olmayan analizlerin çoğu ortalama yöntemine dayanan sıra değeri atama yöntemini kullanır.

R'de Sıra Değeri Belirleme R' de sıra değerlerinin belirlenmesinde temel ***base*** paketindeki ***rank*** fonksiyonu kullanılır. Bu fonksiyon;

rank(x, na.last=TRUE, ties.method=c("average", "first", "last", "random", "max", "min")

burada ;

x: Sayısal, kompleks sayı, karakter dizesi veya mantıksal türden veri vektörünün adıdır

na.last: Kayıp gözlemleri(NA) işleme seçeneğini tanımlar. Varsayılan değer TRUE olup verideki NA' lar en sona atıldıktan sonra sıra değeri atanır. FALSE ise en öne getirilirler. Argümana NA atanırsa kayıp gözlemler veriden atılırken, keep atanırsa sıra değeri olarak NA atanarak oldukları konumda bırakılırlar.

```
x<-c(14,10,11,14,16,19,18,17,15,17,20,14,13,16,19)
x
```

```
##  [1] 14 10 11 14 16 19 18 17 15 17 20 14 13 16 19
```

```
rank_x<- rank(x,ties.method = "average")
rank_x
```

```
##  [1]  5.0  1.0  2.0  5.0  8.5 13.5 12.0 10.5  7.0 10.5 15.0  5.0  3.0  8.5 13.5
```

```
sorted_x<-sort(x)
rank(sorted_x,ties.method = "average")
```

```
##  [1]  1.0  2.0  3.0  5.0  5.0  5.0  7.0  8.5  8.5 10.5 10.5 12.0 13.5 13.5 15.0
```

```
rank(sorted_x,ties.method = "first")
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

```
rank(sorted_x,ties.method = "last")
```

```
##  [1]  1  2  3  6  5  4  7  9  8 11 10 12 14 13 15
```

```
rank(sorted_x,ties.method = "min")
```

```
##  [1]  1  2  3  4  4  4  7  8  8 10 10 12 13 13 15
```

```
rank(sorted_x,ties.method = "max")
```

```
##  [1]  1  2  3  6  6  6  7  9  9 11 11 12 14 14 15
```

‘‘‘

```
rank(sorted_x,ties.method = "random")
```

```
##  [1]  1  2  3  4  6  5  7  8  9 10 11 12 14 13 15
```

```
rank(sorted_x,ties.method = "random")
```

```
##  [1]  1  2  3  5  6  4  7  9  8 10 11 12 13 14 15
```

Şimdi veriye iki tane NA gözlemi ekleyelim.

```
x<-c(14,10,11,NA,14,16,19,18,17,15,17,20,14,13,NA,16,19)
x
```

```
##  [1] 14 10 11 NA 14 16 19 18 17 15 17 20 14 13 NA 16 19
```

```
rank(x,ties.method = "min")
```

```
##  [1]  4  1  2 16  4  8 13 12 10  7 10 15  4  3 17  8 13
```

```
rank(x,ties.method = "min",na.last = FALSE)
```

```
##  [1]  6  3  4  1  6 10 15 14 12  9 12 17  6  5  2 10 15
```

```
rank(x,ties.method = "min",na.last = "keep")
```

```
##  [1]  4  1  2 NA  4  8 13 12 10  7 10 15  4  3 NA  8 13
```

```
rank(x,ties.method = "min",na.last = NA)
```

```
##  [1]  4  1  2  4  8 13 12 10  7 10 15  4  3  8 13
```

Kaynaklar

1. Z. Cebeci , R ile Parametrik Olmayan İstatistik Analiz, Abaküs Yayın Evi, 2019