

# R Giriş ve Varsayım kontrolleri- Ders 2

Nihat Tak

2023-03-09

## Parametrik varsayımların kontrolü

İstatistik analizlerinde uygulanacak testlerin seçilmesi yani parametrik veya parametrik olmayan testlerden hangisinin uygulanacağına karar vermek için analiz edilen veriye ait açıklayıcı istatistiklerin incelenmesi ve testlere ilgili varsayımların geçerliliklerinin varsayımlarının kontrol edilmesi gereklidir. **not:** Uygulamada testler ile ilgili varsayımların kontrolünün yapılmaması sıkça yapılan bir hatadır. Unutulmamalıdır ki, parametrik istatistik yöntemleri mutlaka belli başlı varsayımlara dayanmaktadır. Bu varsayımların ihlali söz konusu olduğu zamanlarda, bu testler güvenilir olmayan sonuçlar verecektir. Örneğin, bir araştırmaya göre kişilik ve sosyal psikoloji dergilerindeki 72 makale gözden geçirilmiş ve bu makalelerin sadece %19'unun çok değişkenli normallik varsayımından bahseettiğini bildirmiştir.

## Veri tanıtımı

İki farklı lokasyonda yetiştirilen yerli bir kiraz çeşidinde giberallik asit uygulamalarının meyve özelliklerine etkilerinin araştırıldığı bir çalışmada püskürtme yoluyla 20 ppm ve 50 ppm dozlarında giberallik asit dozları ile hiç asit uygulanmayan (0 ppm) kontrol uygulaması karşılaştırılmıştır. Her bir uygulama 10 kiraz ağacında denenmiştir. kiraz1 ile adlandırılan veri 1. lokasyondaki deneme verisi ve kiraz 2 ile adlandırılan veri 2. lokasyondaki deneme verisidir. "magr" ile kodlanan sütun meyve ağırlıklarını, "grup" ile kodlanan sütun asit dozlarını göstermekte olup, kontrol grubu (0 ppm) K ile, 20 ppm grubu A ile ve 50 ppm grubu B ile kodlanmıştır.

```
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"
```

## Varsayımların kontrolü

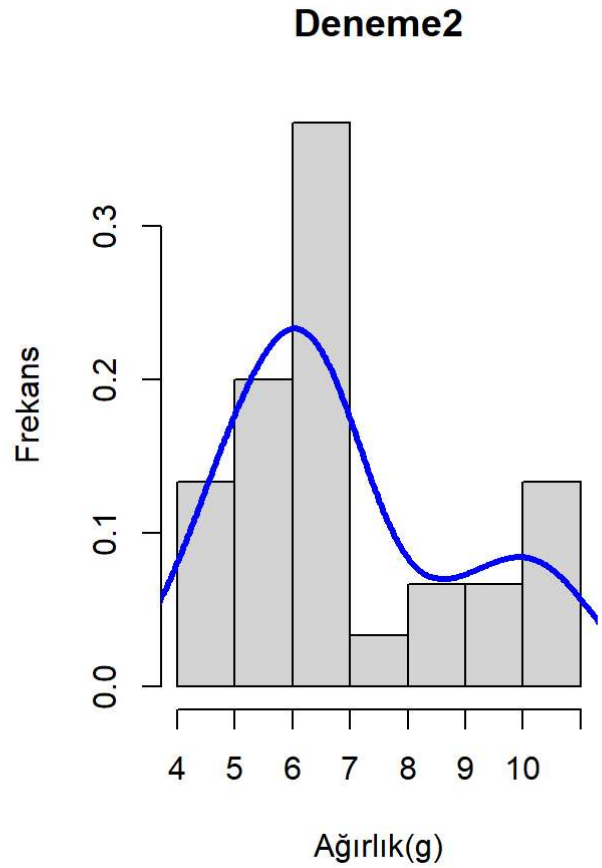
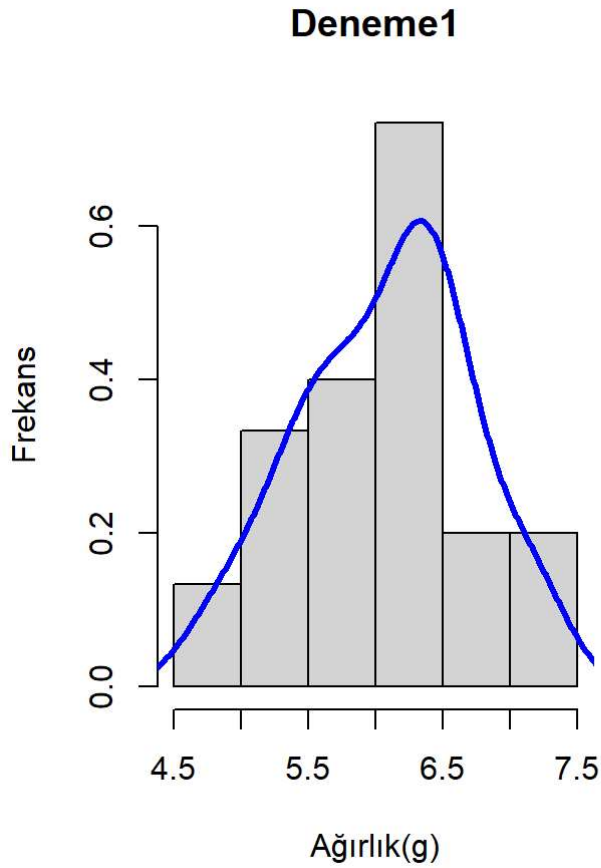
Parametrik ve paarametrik olmayan istatistik yöntemlerle ilgili varsayımların geçerli olup olmadıklarını anlamak üzere yapılacak işlemler aşağıdaki gibi gruplanabilir.

- Görsel değerlendirme
- Açıklayıcı istatistikleri inceleme
- İstatistiksel testler

### Görsel değerlendirme

#### Histogram

```
oldpar<-par(mfrow=c(1,2))
hist(kiraz1$magr,prob=T,xlab="Ağırlık(g)", ylab="Frekans",main="Deneme1")
lines(density(kiraz1$magr),lwd=3,col="blue")
hist(kiraz2$magr,prob=T,xlab="Ağırlık(g)", ylab="Frekans",main="Deneme2")
lines(density(kiraz2$magr),lwd=3,col="blue")
```



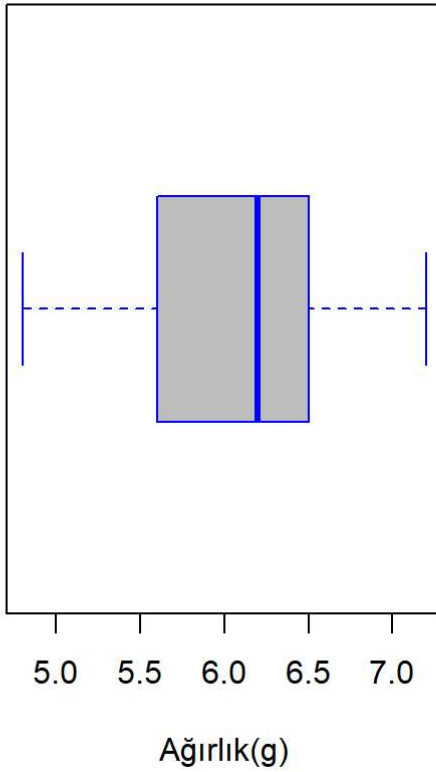
2

denemeye ait histogramlar incelendiğinde, kiraz1 denemesine ait histogram verinin normal dağılıma yakın bir dağılım olduğunu gösterirken, kiraz2 denemesine ait histogram verinin normal dağılımdan farklı bir dağılıma sahip olduğunu ifade eder.

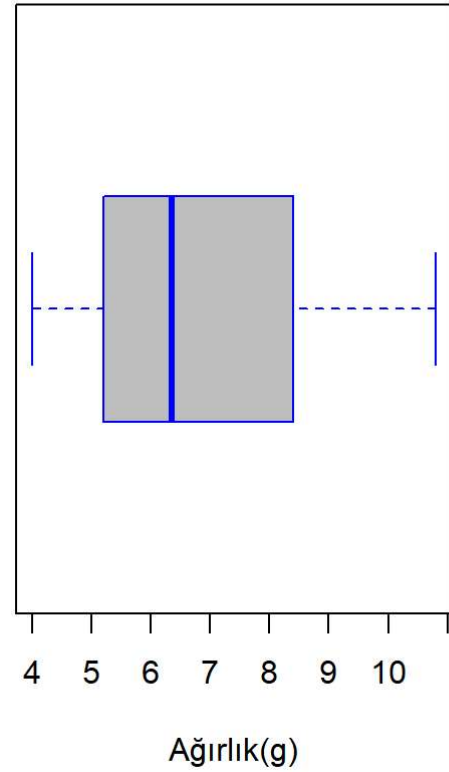
### Kutu Grafiği (Box-Plot)

```
oldpar<-par(mfrow=c(1,2))
boxplot(kiraz1$magr, horizontal = T, col="gray",border = "blue", main="Deneme 1 Kutu Grafiği",xlab="Ağırlık(g)", ylab="")
boxplot(kiraz2$magr, horizontal = T, col="gray",border = "blue", main="Deneme 2 Kutu Grafiği",xlab="Ağırlık(g)", ylab="")
```

Deneme 1 Kutu Grafiği



Deneme 1 Kutu Grafiği



```
par(oldpar)
```

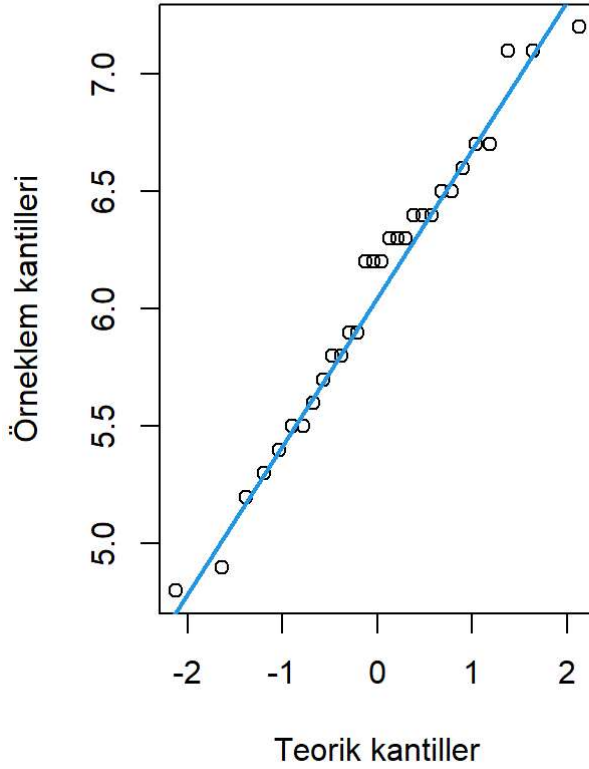
Yukarıdaki kutu grafiği incelendiğinde sol taraftaki kiraz1 denemesine ait kutunun yerleşimi ve kutu içindeki ortanca çizginin normal dağılışa yakın olduğu söylenebilir. Sağ taraftaki kiraz2 denemesine ait kutu ise normal dağılımda farklı olduğu söylenebilir çünkü hem kutu sola çok yakın hem de kutuda ortanca çizgisi kutunun soluna yaklaşmış durumdadır. Dolayısıyla kutu grafikleri normallik kontrolünde oldukça açıklayıcı bilgiler sağlamaktadır.

## QQ-plot

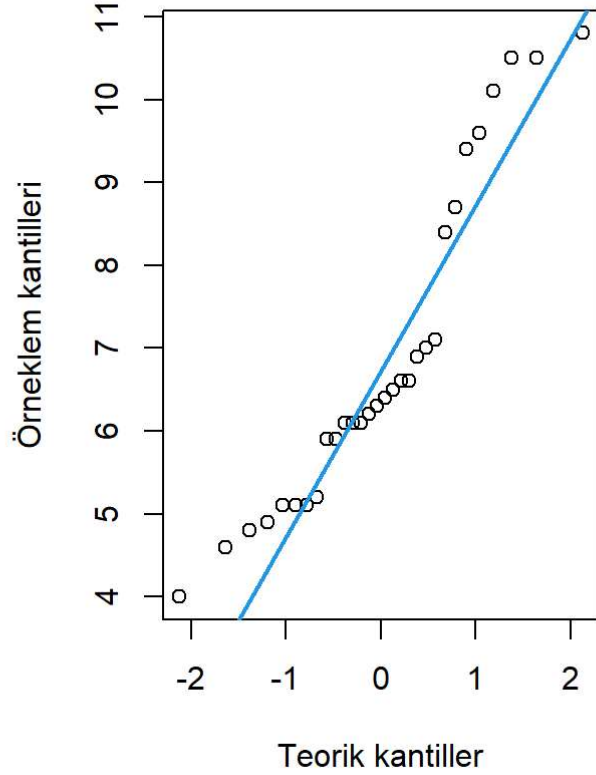
Histogramlar dağılışın şekli hakkında bilgi versede normalliğin incelenmesi için istatistikte Q-Q grafiği olarak bilinen Kantil-Kantil grafiklerinden faydalanılır. Q-Q grafiği, örneklem kantilleri ile ortalama ve standart sapmayı kullanarak hesaplanan normal dağılışın beklenen kantillerine (teorik kantiller) hangi oranda uyduğunu kıyaslamak için kullanılır.

```
oldpar<-par(mfrow=c(1,2))
qqnorm(kiraz1$magr,xlab="Teorik kantiller", ylab="Örneklem kantilleri", main="Deneme1 Q-Q grafiği")
qqline(kiraz1$magr,col=4,lwd=2)
qqnorm(kiraz2$magr,xlab="Teorik kantiller", ylab="Örneklem kantilleri", main="Deneme2 Q-Q grafiği")
qqline(kiraz2$magr,col=4,lwd=2)
```

Deneme1 Q-Q grafiği



Deneme2 Q-Q grafiği



Gözlemlerin köşegende uzanan doğru üstünde veya doğruya yakın olmaları normalliğin bir göstergesidir. Yukarıdaki r kodunda qqnorm fonksiyonu ile değişkenlere ait kantil çiftlerinin serpilmesi ve qqline fonksiyonu ile regresyon doğrusunun çizimi yapılmaktadır. Deneme1'e ait Q-Q grafiği incelendiğinde örneklem kantil çiftlerinin neredeyse mavi renkli regresyon doğrusu üzerinde bulundukları ; sadece alt ve üst uçlarda birkaç gözlemin uzak olduğu görülmektedir. Bu durumda kiraz1 verisine ait *magr* değişkeni normal yada normal olarak kabul edilebilecek bir dağılımdan geldiği söylenebilir. Deneme2 yani kiraz2 verisi göz önünde bulundurulduğunda, bir çok kantil çiftinin alt ve üst uçlarla birlikte regresyon doğrusunun orta kısımlarında da bulunmadığı grafikte gözlemlendiğinden, kiraz2 verisindeki *magr* değişkeninin normal dağılım göstermediği söylenebilir.

*not:* Histogramlar ve Q-Q grafikleri normalliğin incelenmesinde önemli bilgiler sağlasada, çok sayıda değişkenden oluşan bir veri seti için görsel inceleme çok zaman alıcıdır. Ayrıca, normal dağılımdan çok az ayrılan dağılımlarda normal dağılım olarak değerlendirme riskide söz konusu olabilir. Bu nedenle sadece grafiklere bağlı inceleme yapma yerine kada kesin sonuçlar için açıklayıcı istatistikler incelenmeli ve bunlara dayalı testlere başvurulmalıdır.

## Açıklayıcı İstatistiklerin İncelenmesi

Açıklayıcı istatistikler normallik hakkında yeterli sayılabilecek bilgi sağlarlar. Normal dağılan bir veride aritmetik ortalama, mod ve medyan değerleri birbirine eşit veya birbirine çok yakın değerler olmalıdır. R' de değişkenlerin açıklayıcı istatistiklerini hesaplamak için çok sayıda fonksiyon vardır. Bunlardan en yaygın olarak kullanılan **stats** paketindeki *summary*' dir.

Aşağıdaki örneklerde kiraz1 ve kiraz2 verisetlerindeki *magr* değişkeni için açıklayıcı istatistikler verilmiştir.

```
summary(kiraz1$magr)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  4.800   5.625   6.200   6.080   6.475   7.200
```

```
summary(kiraz2$magr)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  4.000   5.375   6.350   6.883   8.075  10.800
```

Değişkenlerin açıklayıcı istatistikleri için **psych** paketindeki “*describe*” ve “*describeBy*” fonksiyonlarından da faydalınabilir.

```
if(!require(psych)){install.packages("psych");require(psych)}
describe(kiraz1$magr)
```

```
##      vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1      1 30 6.08 0.63   6.2   6.09 0.59 4.8 7.2   2.4 -0.19   -0.76 0.11
```

```
describe(kiraz2$magr)
```

```
##      vars  n mean   sd median trimmed  mad min  max range skew kurtosis   se
## X1      1 30 6.88 1.95   6.35   6.72 1.78   4 10.8   6.8 0.69   -0.78 0.36
```

```
describeBy(kiraz1$magr,group = kiraz1$grup)
```

```
##
## Descriptive statistics by group
## group: A
##      vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1      1 10 5.67 0.59   5.75   5.65 0.59 4.8 6.7   1.9 0.09   -1.16 0.19
## -----
## group: B
##      vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1      1 10 6.52 0.41   6.45   6.53 0.3 5.8 7.2   1.4 0.14   -0.91 0.13
## -----
## group: K
##      vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1      1 10 6.05 0.58   6.2   6.03 0.67 5.2 7.1   1.9 0.14   -1.26 0.18
```

```
describeBy(kiraz2$magr,group = kiraz2$grup)
```

```
##
## Descriptive statistics by group
## group: A
##   vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 10 6.42 0.42   6.4    6.4 0.37 5.9 7.1   1.2 0.27   -1.41 0.13
## -----
## group: B
##   vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 10  9.1 1.59   9.5    9.26 1.48 6.1 10.8   4.7 -0.66   -1.08 0.5
## -----
## group: K
##   vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 10 5.13 0.69   5.1    5.11 0.37  4 6.4   2.4  0.4    -0.7 0.22
```

Değişkenlere ilişkin başka özetleyici istatistikleri veren paketlerde mevcuttur. Bunlardan biri **Rmisc** paketidir. Bu paketdeki summarySE fonksiyonu ilgili değişkene ilişkin grup ortalamalarını, standart sapmalarını(sd), standart hatalarını(se) ve güven aralıklarını(ci) hesaplar.

```
if(!require(Rmisc)){install.packages("Rmisc");require(Rmisc)}
summarySE(data=kiraz1,"magr",groupvars="grup",conf.interval=0.95)
```

```
##   grup  N magr      sd      se      ci
## 1    A 10 5.67 0.5907622 0.1868154 0.4226058
## 2    B 10 6.52 0.4131182 0.1306395 0.2955270
## 3    K 10 6.05 0.5835714 0.1845415 0.4174618
```

```
summarySE(data=kiraz2,"magr",groupvars="grup",conf.interval=0.95)
```

```
##   grup  N magr      sd      se      ci
## 1    A 10 6.42 0.4184628 0.1323296 0.2993503
## 2    B 10 9.10 1.5860503 0.5015531 1.1345920
## 3    K 10 5.13 0.6896859 0.2180978 0.4933716
```

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

```
if(!require(FSA)){install.packages("FSA");require(FSA)}
summarySE(data=kiraz1,"magr",groupvars="grup",conf.interval=0.95)
```

```
##   grup  N magr      sd      se      ci
## 1    A 10 5.67 0.5907622 0.1868154 0.4226058
## 2    B 10 6.52 0.4131182 0.1306395 0.2955270
## 3    K 10 6.05 0.5835714 0.1845415 0.4174618
```

```
summarySE(data=kiraz2,"magr",groupvars="grup",conf.interval=0.95)
```

```
##   grup  N magr      sd      se      ci
## 1    A 10 6.42 0.4184628 0.1323296 0.2993503
## 2    B 10 9.10 1.5860503 0.5015531 1.1345920
## 3    K 10 5.13 0.6896859 0.2180978 0.4933716
```

Açıklayıcı istatistiklerden çarpıklık ve basıklık katsayıları dağılımın normalliği açısından bilgi verici ölçülerdir. Normal dağılıma sahip bir verinin çarpıklık katsayısının  $[-2,2]$  aralığında ve basıklık katsayısının da  $[-7,7]$  aralığında olması beklenir.

# İstatistiksel Testler

## Normallik Testleri

Normallik testleri, değişkenlerin normal dağılım gösterip göstermediğini kontrol etmek için kullanılan testlerdir. Kolmogorov-Smirnov testi, Jarque-Bera testi ve Shapiro-Wilk testi normallik testleri içerisinde en yaygın kullanılan testlerdir.

- Kolmogorov-Smirnov testi: dağılımın kuyruklarındaki sorunlara duyarlı değil ve  $n > 50$  olduğunda uygun görülen bir testtir.
- Shapiro-Wilk testi: veride aynı değerler olduğunda yani gözlemler arasında bir bağ olduğunda iyi sonuç vermezler ve  $n < 50$  örneklem büyüklüğü olduğum durumlarda tercih edilir.
- Jarque-Bera testi: Basıklık ve çarpıklık ölçülerini kullandığından normallik için iyi bir test olarak kabul edilir.

**not:** Tek örneklem t-testi ve z-testi için normallik kontrolü direk olarak bağımlı değişkene uygulanabilir ancak diğer testlerde doğrusal modele ait kalıntıların normallik gösterip göstermediği araştırılmalıdır. R’de normallik testi için **stats** paketinde “shapiro.test” fonksiyonundan faydalanılabilir.

Shapiro-Wilk testinde hipotezler:

$H_0$ :Veri setinin dağılımı normaldir.

$H_1$ :Veri setinin dağılımı normal değildir.

Eğer  $p\text{-value} < \text{cut-off}(0.10 \text{ veya } 0.05 \text{ veya } 0.01, \text{ etc})$  ise  $H_0$  reddedilir.

```
shapiro.test(kiraz1$magr)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  kiraz1$magr
## W = 0.97324, p-value = 0.6312
```

```
shapiro.test(kiraz2$magr)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  kiraz2$magr
## W = 0.89644, p-value = 0.006879
```

Dolayısıyla yukarıdaki örneğe göre, kiraz1 veri setindeki *magr* değişkeninin dağılımı  $p\text{-value}=0.6312 > 0.01$  olduğundan normaldir denir. Benzer şekilde, kiraz2 veri setindeki *magr* değişkeninin dağılımı  $p\text{-value}=0.006879 < 0.01$  olduğundan normal değildir denir

Bir diğer normallik testi fBasics paketindeki jarqueberaTest fonksiyonu ile yapılan Jarque-Bera testidir.



```
model1<-lm(magr~grup,data=kiraz1)
model2<-lm(magr~grup,data=kiraz2)
shapiro.test(model1$residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residual
## W = 0.97405, p-value = 0.6547
```

```
shapiro.test(model2$residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residual
## W = 0.92221, p-value = 0.03062
```

```
if(!require(fBasics)){install.packages("fBasics");require(fBasics)}
jarqueberaTest(model1$res)
```

```
##
## Title:
##  Jarque - Bera Normalality Test
##
## Test Results:
##  STATISTIC:
##    X-squared: 0.4954
##    P VALUE:
##    Asymptotic p Value: 0.7806
```

```
jarqueberaTest(model2$res)
```

```
##
## Title:
##  Jarque - Bera Normalality Test
##
## Test Results:
##  STATISTIC:
##    X-squared: 7.7014
##    P VALUE:
##    Asymptotic p Value: 0.02126
```

## 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 seviyesinden 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 istatistiksel 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

Örnekleme 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 da tanımlanabilir. Bu durumun testip edilebilmesi için kullanılabilecek birçok test literatürde mevcuttur. Bunlar arasında 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 0.972
## Alternative hypothesis: rho != 0
```

## Referanslar

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