
R ile
Uygulamalı
İstatistik
Editör
ÇİĞDEM ARICIGİL ÇİLAN
BİDGE

BİDGE Yayınları

R İLE UYGULAMALI İSTATİSTİK

Editör: ÇİĞDEM ARICIGİL ÇİLAN

ISBN: 978-625-372-839-7

1. Baskı

Sayfa Düzeni: Gözde YÜCEL

Yayınlama Tarihi: 15.11.2025

BİDGE Yayınları

Sertifika No: 71374
Yayın hakları © BİDGE Yayınları
www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

İÇİNDEKİLER

R'A GİRİŞ, TEMEL HESAPLAMALAR, VEKTÖRLER VE MATRİSLER

ÇİĞDEM ARICIGİL ÇİLAN[1]

Giriş

R ücretsiz, açık kaynak kodlu bir programlama dilidir. R ile çalışılabilmesi için R ve R Studio olmak üzere iki yazılımın en güncel versiyonlarının bilgisayara indirilmesi gerekmektedir. R, yazılım geliştirme ve veri analizi için kullanılmaktadır. Bu kitapta da R, istatistik yöntemler temel alınarak veri analizi amacıyla kullanılacaktır.

Başlangıçta R bir hesap makinesi gibi düşünülebilir. Buna ilişkin aşağıda basit örnekler verilmiştir.

Tablo 1 R'da Toplama, Çıkarma , Çarpma ve Bölme İşlemleri

İşlemler	R'da	Sonuç
4+2	4+2	6
4-2	4-2	2
4.2	4*2	8
4/2	4/2	2

[1] Prof.Dr., İstanbul Üniversitesi İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı, Orcid: 0000-0002-7862-7028

Tablo 2 Üssel İşlemler

İşlemler	R'da	Sonuç
$\mathbf{3^2}$	3^2	9
$\mathbf{2^{(-3)}}$	2^(-3)	0,125
$\mathbf{100^{(\frac{1}{2})}}$	100^(1/2)	10
$\mathbf{\sqrt{100}}$	sqrt(100)	10

Tablo 3 Matematikte Sabitler

İşlemler	R'da	Sonuç
π	pi	3,1415927
e	exp(1)	2,7182818

Tablo 4 Logaritmik İşlemler

İşlemler	R'da	Sonuç
log(e)	log(exp(1))	1
$\mathbf{log_{10}(1000)}$	log10(1000)	3
$\mathbf{log_2(8)}$	log2(8)	3

1.1. Paketlerin Yüklenmesi- Veri Tipi ve Yapıları

R programı birçok fonksiyonu ve veri setini içermektedir. Ancak R'ın en güçlü yanlarından biri farklı yöntemlere ilişkin paketlerden oluşmasıdır. Çalışılmak istenen yöntem ile ilgili bir paketi indirip kendi veri setinize uyarlamanız mümkündür. Paket önce indirilmeli ve üzerinde çalışılacağı zaman çağırılmalıdır. Örneğin Gizli Sınıf Analizi(Latent Class Analysis) yapmak için poLCA paketini aşağıdaki gibi indirmek ve çağırmak gerekir.

```
install.packages("poLCA")

library(poLCA)
```

R'da veri tipleri; sürekli-nümerik (25,7) , kesikli-integer (3, 15 gibi), mantıksel veri (TRUE, FALSE, NA) ve karakterlerden oluşan veriler ("c", "kadın", "bir artı iki" gibi) olarak sınıflandırılabilir.

Veriler; aynı veri tipinden oluşuyorsa homojen, farklı veri tiplerinden oluşuyorsa heterojen veri olarak adlandırılmaktadır.

Vektörler, matrisler homojen verileri temsil ederken, listeler (Lists) ve farklı veri tiplerinden oluşan data setleri (Data Frame) heterojen verilerdir.

1.2. Vektörler

R da vektör oluşturmanın en basit yolu c( ) fonksiyonunu kullanmaktır. R'da aşağıdaki vektörü yazınız.

```
c(1,2,3,4,5,6,7)
```

Entere bastığınızda aşağıdaki gösterimle karşılaşılır.

```
##[1] 1 2 3 4 5 6 7
```

Bu vektör x değişkeni olarak adlandırıldığında; (<- yerine = işareti de kullanılabilir)

```
x<-c(1,2,3,4,5,6,7)
x
##[1] 1 2 3 4 5 6 7
```

sonucu elde edilir.

Vektörler birden fazla veri tipi içerebilir.

```
c(45, "kadin", TRUE)
##[1] "45" "kadin" "TRUE"


c(45, TRUE)
##[1] 45  1
```

Vektörler ile ardışık sayılardan oluşan bir sayı dizisi de tanımlanabilir. Aşağıda 1'den 100'e kadar ardışık tam sayılar tanımlanmıştır:

```
c(1:100)
## [1]   1   2   3   4   5   6   7   8   9  10  11  12
13  14  15  16  17  18

##[19]  19  20  21  22  23  24  25  26  27  28  29  30
31  32  33  34  35  36
```

```
##[37]  37  38  39  40  41  42  43  44  45  46  47  48
49  50  51  52  53  54

##[55]  55  56  57  58  59  60  61  62  63  64  65  66
67  68  69  70  71  72

## [73]  73  74  75  76  77  78  79  80  81  82  83  8
4  85  86  87  88  89  90

##[91]  91  92  93  94  95  96  97  98  99 100
```

R'da sayılar yazılırken bir elemanlı vektör olarak tanımlanır.

```
17
```

```
##[1] 17
```

R'da tamsayı olmayan belirli bir sabit ile artış gösteren ardışık sayı dizileri de tanımlanabilir.

```
seq(from=5,to=10, by=0.5 )
```

```
## [1]  5.0  5.5  6.0  6.5  7.0  7.5  8.0  8.5  9.0  9
.5 10.0
```

Bu sayı dizisini oluşturan fonksiyon aşağıdaki gibi de tanımlanabilir:

```
seq(5,10,0.5)
```

```
##[1]  5.0  5.5  6.0  6.5  7.0  7.5  8.0  8.5  9.0  9.
5 10.0
```

Aynı sayı veya karakteri birden fazla tekrar edilerek yazdırılmak istendiğinde, kaç kez tekrar edileceği aşağıdaki işlemde olduğu gibi belirtilir.

```
rep("Y",times=10)
```

```
##[1] "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y"
```

Bir vektörün elemanları da tekrar yazdırılabilir. Yukarıda tanımladığımız x vektörünü 3 kez yazdırmak için aşağıdaki fonksiyon yazılır.

```
x<-c(1,2,3,4,5,6,7)
```

```
x
```

```
##[1] 1 2 3 4 5 6 7
```

```
rep(x,times=3)
```

```
##[1] 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7
```

Vektör oluşturmak için kullanılan fonksiyonlar birlikte kullanılarak yeni vektör tanımlamaları da yapılabilir.

```
c(x,rep(seq(2,8,2),2),c(1,2,3),45,1:3)
```

```
##[1]  1  2  3  4  5  6  7  2  4  6  8  2  4  6  8  1
2  3 45  1  2  3
```

Vektörlerdeki eleman sayıları length() fonksiyonu ile elde edilebilmektedir. Örneğin x vektöründeki eleman sayısını elde etmek için;

```
length(x)
```

```
##[1] 7
```

Aşağıda tanımlanan y vektörünün eleman sayısı da length() fonksiyonu ile elde edilir.

```
y<-c(1:100)
length((y))
```

```
##[1] 100
```

-Vektörlerin Alt Kümelerinin Oluşturulması

Vektörlerin alt kümeleri tanımlanırken köşeli parantez [ ] kullanılır.

```
x
```

```
[1] 1 2 3 4 5 6 7
```

X vektörünün 1. elemanını çağırmak için x[1] yazılır.

```
x[1]
```

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

x vektörünün 3. elemanını çağırmak için x[3] yazılır.

```
x[3]
```

```
##[1] 3
```

Vektördeki elemanları çıkarmak için yine köşeli parantez [ ] kullanılırken kaçıncı elemanın çıkarılacağı eksi işareti ile belirlenir.

x vektörünün 2. elemanı çıkarılırken [-2] işlemi uygulanır.

```
x[-2]
##[1] 1 3 4 5 6 7
```

x vektörünün 1. 2. ve 3. elemanını seçmek için x[1:3] işlemi yapılır.

```
x[1:3]
##[1] 1 2 3
```

Vektörlerin alt kümelerini seçerken mantıksal verilerden oluşan vektörler kullanılabilir.

```
z<-c(TRUE,TRUE,FALSE,TRUE,TRUE,FALSE)
x[z]
##[1] 1 2 4 5 7
```

Burada x vektörünün TRUE olan elemanlarını seç komutu verilmektedir.

Vektörlere uygulanan işlemler yeni bir x vektörü tanımlayarak aşağıda örneklendirilmiştir.

```
x=1:5
x
##[1] 1 2 3 4 5
x+1
##[1] 2 3 4 5 6
2*x
##[1]  2  4  6  8 10
x^2
##[1]  1  4  9 16 25
```

```
sqrt(x)
```

```
##[1] 1.000000 1.414214 1.732051 2.000000 2.236068
```

```
log(x)
```

```
##[1] 0.0000000 0.6931472 1.0986123 1.3862944 1.609437
9
```

Tablo 5 Mantıksal Operatörler

Operatör	Açıklama	Örnek	Sonuç
x==y	x y'ye eşittir	5==10	FALSE
x<y	x y'den küçüktür.	3<45	TRUE
x>y	x y'den büyüktür.	3>45	FALSE
x=<y	x y'den küçüktür veya eşittir	3<=45	TRUE
x>=y	x y'den büyük veya eşittir	3>=45	FALSE
x!=y	x y'ye eşit değildir.	3!=45	TRUE
!x	x değildir.	!(3>45)	TRUE
x|y	x veya y	(3>45)|TRUE	TRUE
x&y	x ve y	(2<5)&(45>3)	TRUE

Aşağıda yeni bir x vektörü tanımlanarak mantıksal işlemler uygulanacaktır:

```
x=c(1,3,5,7,9,11)
x>3
```

```
##[1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
x<3
```

```
##[1]  TRUE FALSE FALSE FALSE FALSE FALSE
```

```
x==3
```

```
##[1] FALSE  TRUE FALSE FALSE FALSE FALSE
```

```
x!=3
```

```
##[1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```

```
x==3&x!=3
```

```
##[1] FALSE FALSE FALSE FALSE FALSE FALSE
```

```
x==3|x!=3
```

```
##[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

Alt kümeleri oluştururken aşağıdaki işlemler de oldukça kullanışlıdır. Öncelikle x vektöründeki elemanlar arasından değeri 3'ten büyük olanları seçilsin:

```
x=c(1,3,5,7,9,11)
x
```

```
##[1]  1  3  5  7  9 11
```

```
x[x>3]
##[1]  5  7  9 11
```

x vektöründe 3 değerine sahip olmayan sayıları seçilsin.

```
x[x!=3]
##[1]  1  5  7  9 11
```

which() fonksiyonu ise şöyle kullanılır. Örneğin which (x>2); x vektöründe 2'den büyük sayıların vektörde kaçıncı sırada olduklarını gösterir.

```
x
##[1]  1  3  5  7  9 11
which(x>2)
##[1] 2 3 4 5 6
x[which(x>2)]
##[1]  3  5  7  9 11
```

x vektöründe en büyük değeri elde etmek için max( ) operatörü kullanılır.

```
max(x)
      ##[1] 11
```

Vektördeki sayıların toplamı için sum( ) operatörü kullanılır.

```
sum(x)
##[1] 36
```

Aşağıda da x vektörü ile 6 tane 2 den oluşan bir vektörle toplama işlemi gösterilmektedir.

```
x+rep(2,6)
##[1]  3  5  7  9 11 13
```

Yine mantıksal bir fonksiyona şöyle bir örnek verilebilir:

```
x>rep(2,6)
##[1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
```

1.3. Matrisler

Bir x vektörü tanımlansın.

```
x=c(3:11)
x
## [1]  3  4  5  6  7  8  9 10 11
X=matrix(x,nrow=3,ncol=3)
X
 ##           [,1] [,2] [,3]
 ##      [1,]    3    6    9
 ##      [2,]    4    7   10
 ##      [3,]    5    8   11
```

x vektöründen sütun öncelikli, oluşturulan bu matris yerine satır öncelikli matris aşağıdaki işlemle oluşturulur:

```
Y=matrix(x,nrow=3, ncol = 3,byrow = TRUE)
Y
  ##          [,1] [,2] [,3]
  ##    [1,]    3    4    5
  ##    [2,]    6    7    8
  ##    [3,]    9   10   11
```

Yine tüm elemanları sabit bir sayıdan oluşan bir matris de oluşturulabilir.

```
Z=matrix(1,2,4)
Z
     ##       [,1] [,2][,3] [,4]
     ##[1,]    1    1    1    1
     ##[2,]    1    1    1    1
```

Aynı vektörlerde olduğu gibi matrisinde alt setlerini elde etmek için köşeli parantez kullanmak gerekmektedir. Örneğin X matrisinin 2.satır 3.sütundaki elemanı seçilsin.

```
X
##     [,1] [,2] [,3]
##[1,]    3    6    9
##[2,]    4    7   10
##[3,]    5    8   11
X[2,3]
##[1] 10
```

X matrisinin 2.satır elemanlarını çağırmak için aşağıdaki işlem uygulanır.

```
X[2,]
##[1]  4  7 10
```

X matrisinin 3.sütununu seçilsin.

```
X[,3]
##[1]  9 10 11
```

Şimdi de 2. satırın 1. ve 2. sütunlarında yer alan elemanlar seçilsin.

```
X[2,c(1,2)]
##[1] 4 7
```

Vektörleri satırda tanımlayarak matris oluşturan fonksiyon rbind, yine tanımlanan vektörleri sütunlarda tanımlayarak matris oluşturulmasını sağlayan fonksiyon ise cbind fonksiyonudur.

Aşağıda bir x vektörü, x vektörünün tersi rev(x) ve 9 tane 2'den oluşan vektörler oluşturulmuş ve bu vektörler satırları oluşturacak şekilde bir matris elde edilmiştir.

```
x=(2:10)
```

```
rev(x)
[1] 10  9  8  7  6  5  4  3  2
rep(2,9)
[1] 2 2 2 2 2 2 2 2 2
rbind(x,rev(x),rep(2,9))
##     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##x      2    3    4    5    6    7    8    9   10
##      10    9    8    7    6    5    4    3    2
##       2    2    2    2    2    2    2    2    2
```

Aynı vektörler sütun vektörleri olarak tanımlandığında aşağıdaki matris edilir.

```
cbind(x,rev(x),rep(2,9))
          x
 ##[1,]  2 10 2
 ##[2,]  3  9 2
 ##[3,]  4  8 2
 ##[4,]  5  7 2
 ##[5,]  6  6 2
 ##[6,]  7  5 2
 ##[7,]  8  4 2
 ##[8,]  9  3 2
 ##[9,] 10  2 2
```

-Matris İşlemler

X ve Y matrisleri tanımlansın.

```
x=2:10
y=10:2
matrix(x,3,3)
##      [,1] [,2] [,3]
```

```
##[1,]    2    5    8
##[2,]    3    6    9
##[3,]    4    7   10
matrix(y,3,3)
##      [,1] [,2] [,3]
##[1,]   10    7    4
##[2,]    9    6    3
##[3,]    8    5    2


X=matrix(x,3,3)
Y=matrix(y,3,3)
X+Y
##     [,1] [,2] [,3]
##[1,]   12   12   12
##[2,]   12   12   12
##[3,]   12   12   12
X-Y
##     [,1] [,2] [,3]
##[1,]   -8   -2    4
##[2,]   -6    0    6
##[3,]   -4    2    8
X*Y
##     [,1] [,2] [,3]
##[1,]   20   35   32
##[2,]   27   36   27
##[3,]   32   35   20
```

X*Y sadece matris elemanlarının çarpımını sağlamaktadır. Matrislerin çarpımında %*% işlemi kullanılır.

```
X%*%Y
```

```
##      [,1] [,2] [,3]
##[1,]  129   84   39
##[2,]  156  102   48
##[3,]  183  120   57
```

```
X/Y
```

```
##         [,1]     [,2]      [,3]
##[1,] 0.2000000 0.7142857    2
##[2,] 0.3333333 1.0000000    3
##[3,] 0.5000000 1.4000000    5
```

t( ) işlemi bir matrisin tanspozesinin alınmasını sağlarken, solve ( ) işlemi ise kare bir matrisin tersinin alınmasında kullanılır. Kare bir matris tanımlansın.

```
Z=matrix(c(11,4,-5,4,6,-4,-5,-4,18),3,byrow = TRUE)
Z
```

```
##      [,1] [,2] [,3]
##[1,]   11    4   -5
##[2,]    4    6   -4
##[3,]   -5   -4   18
```

```
t(Z)
```

```
##      [,1] [,2] [,3]
##[1,]   11    4   -5
##[2,]    4    6   -4
##[3,]   -5   -4   18
```

```
solve(Z)
```

```
##             [,1]        [,2]       [,3]
##[1,]  0.12534060 -0.07084469 0.01907357
##[2,] -0.07084469  0.23569482 0.03269755
##[3,]  0.01907357  0.03269755 0.06811989
```

Birim matrisin formülü $Z^{-1}Z = I$ olduğu hatırlandığında Z matrisi ile birim matris hesaplansın.

```
solve(Z)%*%Z
```

```
              [,1]          [,2]          [,3]
##[1,] 1.000000e+00 -1.526557e-16  1.665335e-16
##[2,] 5.551115e-17  1.000000e+00 -1.110223e-16
##[3,] 0.000000e+00  0.000000e+00  1.000000e+00
```

R daki hesaplamalar nedeniyle köşegen dışındaki elemanlar tam olarak sıfıra eşit görülmemektedir. Ancak bu sonucun 3 boyutlu bir diagonal matrise eşit olup olmadığı all.equal( ) fonksiyonu ile araştırılabilir.

```
I=solve(Z)%*%Z
I
```

```
              [,1]          [,2]          [,3]
##[1,] 1.000000e+00 -1.526557e-16  1.665335e-16
##[2,] 5.551115e-17  1.000000e+00 -1.110223e-16
##[3,] 0.000000e+00  0.000000e+00  1.000000e+00
```

```
all.equal(I,diag(3))
```

```
##[1] TRUE
```

R'da bir X matrisinin boyutu dim( ), satır toplamı rowSums( ), sütun toplamları colSums( ), satır ortalamaları rowMeans(sütun ortalamaları), colMeans( ) işlemleri ile hesaplanabilir:

```
X=matrix(5:10,2,3)
X
```

```
##     [,1] [,2] [,3]
##[1,]    5    7    9
##[2,]    6    8   10
```

```
dim(X)
```

```
##[1] 2 3
```

```
rowSums(X)
##[1] 21 24
colSums(X)
##[1] 11 15 19
rowMeans(X)
##[1] 7 8
colMeans(X)
##[1] 5.5 7.5 9.5
```

Bir B matrisi tanımlansın. Bu matrisin köşegen elemanları diag( ) fonksiyonu ile elde edilir.

```
B=matrix(1:4,2,2)
B
##     [,1] [,2]
##[1,]    1    3
##[2,]    2    4
diag(B)
##[1] 1 4
```

Köşegen elemanları 1,2 ve 3'ten oluşan bir köşegen matris tanımlansın.

```
diag(1:3)
##     [,1] [,2] [,3]
##[1,]    1    0    0
##[2,]    0    2    0
##[3,]    0    0    3
```

4*4 boyutlu bir birim matris elde edilsin.

```
diag(4)
##     [,1] [,2] [,3] [,4]
```

```
##[1,]    1    0    0    0
##[2,]    0    1    0    0
##[3,]    0    0    1    0
##[4,]    0    0    0    1
```

-Matris ve Vektörler İle Hesaplamalar

İstatistik yöntemlerin uygulanmasında sıklıkla vektör/matris çarpımları yapılmaktadır. x ve y vektörleri tanımlansın.

```
x=c(1,2,3)
y=c(2,2,2)
```

Bu iki vektörün iç çarpımında aşağıdaki fonksiyon kullanılır:

```
x%*%y
##      [,1]
##[1,]   12
```

Bu çarpım crossprod( ) işlemi ile de yapılabilmektedir.

Veri Çerçevesinin Oluşturulması (Data Frames)

Veri çerçevesi oluşturulurken her bir değişkenin aynı uzunlukta olması gerekir. Aşağıda x, y ve z değişkenlerinin 10'ar birimi olduğu görülmektedir.

```
ornek<-data.frame(x=c(1,2,3,4,5,6,7,8,9,10),y=c(rep("k
adin",9),"erkek"),
z=rep(c(TRUE,FALSE),5))
ornek
##    x     y     z
##1   1 kadin  TRUE
##2   2 kadin FALSE
##3   3 kadin  TRUE
##4   4 kadin FALSE
##5   5 kadin  TRUE
```

```
##6   6 kadin FALSE
##7   7 kadin  TRUE
##8   8 kadin FALSE
str(ornek)
##'data.frame':       10 obs. of  3 variables:
##$ x: num  1 2 3 4 5 6 7 8 9 10
##$ y: chr  "kadin" "kadin" "kadin" "kadin" ...
##$ z: logi  TRUE FALSE TRUE FALSE TRUE FALSE ...
nrow(ornek)
##[1] 10
ncol(ornek)
## [1] 3
dim(ornek)
## [1] 10  3
```

data.frame( ) işleminin dışında R'da tibble paketi çalıştırılarak da veri seti oluşturulabilmektedir.

```
 install.packages("tibble")
 library(tibble)
 ornek=as_tibble(ornek)
 ornek
 ##A tibble: 10 x 3
##   x      y     z
 ## <dbl> <chr> <lgl>
 ##1     1 kadin TRUE
 ##2     2 kadin FALSE
 ##3     3 kadin TRUE
 ##4     4 kadin FALSE
 ##5     5 kadin TRUE
```

```
##6      6 kadin FALSE
##7      7 kadin TRUE
##8      8 kadin FALSE
##9      9 kadin TRUE
       ##10    10 erkek FALSE
```

head( ) komutu veri çerçevesinin (datas etinin) değişkenlerinin ve verilerinin görüntülemesini sağlarken, str( ) komutu değişkenlerin veri tipleri hakkında bilgi vermektedir. "ggplot2" paketinde yer alan "mpg" verisi ile çalışılsın. Veri setinin değişkenleri ve açıklamaları şöyledir: manufacturer(marka), model(arabanın modeli),displ(motor hacmi), year(üretim yılı), cyl(silindir sayısı), trans(vites tipi), drv(çekiş tipi), cty(şehiriçi tüketim), hwy(otoyol tüketimi), fl(yakıt tipi), class(araç sınıfı)

Öncelikle "ggplot2" paketinin R'a indirilmesi gerekmektedir.

```
install.packages("ggplot2")
library(ggplot2)
head(mpg, n = 10)
```

```
##A tibble: 6 × 11
## manufacturer model displ  year   cyl trans      drv     cty   hw
y fl    class
## <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int
> <chr> <chr>
##1 audi         a4      1.8  1999     4 auto(l5)   f        18
29 p     compact
##2 audi         a4      1.8  1999     4 manual(m5) f        21
29 p     compact
##3 audi         a4      2    2008     4 manual(m6) f        20
31 p     compact
```

```
##4 audi          a4      2    2008     4 auto(av)   f        21
30 p     compact
##5 audi          a4      2.8  1999     6 auto(l5)   f        16
26 p     compact
##6 audi          a4      2.8  1999     6 manual(m5) f        18
26 p     compact
```

Sadece verinin adı girildiğinde de ilk 10 satır gösterilmektedir.

```
mpg
```

```
##A tibble: 234 × 11
##  manufacturer model      displ  year   cyl trans  drv     cty   h
wy fl    class
##   <chr>        <chr>      <dbl> <int> <int> <chr>  <chr> <int> <
int> <chr> <chr>
## 1 audi          a4            1.8  1999     4 auto(… f         18
29 p     comp…
## 2 audi          a4            1.8  1999     4 manua… f         21
29 p     comp…
## 3 audi          a4            2    2008     4 manua… f         20
31 p     comp…
## 4 audi          a4            2    2008     4 auto(… f         21
30 p     comp…
```

```
## 5 audi          a4            2.8  1999     6 auto(… f         16
26 p     comp…

## 6 audi          a4            2.8  1999     6 manua… f         18
26 p     comp…

## 7 audi          a4            3.1  2008     6 auto(… f         18
27 p     comp…

## 8 audi          a4 quattro    1.8  1999     4 manua… 4         18
26 p     comp…

## 9 audi          a4 quattro    1.8  1999     4 auto(… 4         16
25 p     comp…

##10 audi          a4 quattro    2    2008     4 manua… 4         20
28 p     comp…
```

str( ) komutu ile de verinin yapısını görmek mümkündür. Bu komut hangi değişkenlerin nicel hangilerinin nitel olduğunu gösterir.

```
str(mpg)
```

```
 ##tibble [234 × 11] (S3: tbl_df/tbl/data.frame)

 ##$ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" .
..

 ##$ model       : chr [1:234] "a4" "a4" "a4" "a4" ...

 ##$ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1
.8 2 ...
```

```
 ##$ year        : int [1:234] 1999 1999 2008 2008 1999 1999
2008 1999 1999 2008 ...##$ cyl         : int [1:234] 4 4 4 4
6 6 6 4 4 4 ...

 ##$ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manu
al(m6)" "auto(av)" ...

 ##$ drv         : chr [1:234] "f" "f" "f" "f" ...

 ##$ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20
...

##$ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28
...

 ##$ fl          : chr [1:234] "p" "p" "p" "p" ...

 ##$ class       : chr [1:234] "compact" "compact" "compact"
"compact" ...
```

Değişken isimleri names( ) komutu ile görülebilir. Aşağıda mpg datasının değişkenleri bu komut ile görülmektdir.

```
names(mpg)
```

```
 ##[1] "manufacturer" "model"         "displ"         "year"
"cyl"

 ##[6] "trans"        "drv"           "cty"           "hwy"
"fl"

 ##[11] "class"
```

Data setindeki değişkenlerden birini vektör olarak çağırmak için $ operatörü kullanılır. Örneğin mpg datasında yıl değişkeninin çağrılmasında aşağıdaki komut kullanılır.

```
mpg$year
```

```
##[1] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 2008 1999 1
999

##[14] 2008 2008 1999 2008 2008 2008 2008 2008 1999 2008 1999 1999 2
008

## [27] 2008 2008 2008 2008 1999 1999 1999 2008 1999 2008 2008 1999
1999

##[40] 1999 1999 2008 2008 2008 1999 1999 2008 2008 2008 2008 1999 1
999

##[53] 2008 2008 2008 1999 1999 1999 2008 2008 2008 1999 2008 1999 2
008

##[66] 2008 2008 2008 2008 2008 1999 1999 2008 1999 1999 1999 2008 1
999

##[79] 1999 1999 2008 2008 1999 1999 1999 1999 1999 2008 1999 2008 1
999

## [92] 1999 2008 2008 1999 1999 2008 2008 2008 1999 1999 1999 1999
1999

##[105] 2008 2008 2008 2008 1999 1999 2008 2008 1999 1999 2008 1999
1999

##[118] 2008 2008 2008 2008 2008 2008 2008 1999 1999 2008 2008 2008
2008

##[131] 1999 2008 2008 1999 1999 1999 2008 1999 2008 2008 1999 1999
1999

##[144] 2008 2008 2008 2008 1999 1999 2008 1999 1999 2008 2008 1999
1999

##[157] 1999 2008 2008 1999 1999 2008 2008 2008 2008 1999 1999 1999
1999

##[170] 2008 2008 2008 2008 1999 1999 1999 1999 2008 2008 1999 1999
2008

##[183] 2008 1999 1999 2008 1999 1999 2008 2008 1999 1999 2008 1999
1999

##[196] 1999 2008 2008 1999 2008 1999 1999 2008 1999 1999 2008 2008
1999

##[209] 1999 2008 2008 1999 1999 1999 1999 2008 2008 2008 2008 1999
1999

##[222] 1999 1999 1999 1999 2008 2008 1999 1999 2008 2008 1999 1999
2008
```

Matris olarak tanımlanan data setinin boyutunu görebilmek için dim ( ) komutu kullanılır.

"mpg" datasının boyutu 234×11'dir (234 satır,11 sütun).

```
dim (mpg)
##[1] 234  11
```

Yine datanın sadece satır sayısını görmek için nrow( ) komutu ve sadece sütun sayısını görmek için ncol() komutu kullanılır.

```
nrow(mpg)
##[1] 234
ncol(mpg)
##[1] 11
```

Bir data setinde subset( ) komutu kullanarak filtreleme yapmak da mümkündür. Aşağıda data setinden otoyol tüketimi 20'den fazla olan marka ve modeller seçilmiştir.

```
subset(mpg, subset = hwy > 20, select = c("manufacture
r", "model"))
## A tibble: 145 x 2
##   manufacturer model
##   <chr>        <chr>
## 1 audi         a4
## 2 audi         a4
## 3 audi         a4
## 4 audi         a4
## 5 audi         a4
      ## 6 audi         a4
## 7 audi         a4
## 8 audi         a4 quattro
```

```
## 9 audi          a4 quattro
##10 audi          a4 quattro
## ... with 135 more rows
```

R'da if/else komutunun nasıl çalıştığı aşağıda örneklendirilmiştir.

```
x = 2
y = 5
if (x > y) {
z = x * y
print("Hayir")
 } else {
z = x * 3 * y
print("Evet")
}
```

```
##[1] "Evet"
z
##[1] 30
```

R da ifelse() komutu da bu tür problemlerin çözümünde kolaylık sağlamaktadır.

```
ifelse(8< 10, 1, 0)
```
8<10 eşitsizliği doğruysa sonucu 1 yanlış ise 0 olarak belirle

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

Yine programlamada çok sık kullanılan for next döngüsüne de bir örnek verilsin.

```
x =6:10
for (i in 1:5) {
x[i] = x[i] * 2}
x
##[1] 22 24 26 28 30
```

Bu döngü yerine daha basit bir şekilde aynı sonuca ulaşılabilmektedir.

```
x=6:10
x*2
```

```
##[1] 12 14 16 18 20
```

Bu bölümde R ve RStudio tanıtılmış, temel hesaplamalar, veri tipleri, vektörler ve matrisler ele alınmıştır. Uygulamalarda doğrudan çalıştırılabilecek örnek R kodları verilmiş, böylece okuyucular kendi verileri üzerinde bu kodları uygulayabileceklerdir.

Kaynakça

Dalgaard, P. (2008). *Introductory statistics with R*. Springer.

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using R*. SAGE.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

RStudio Team (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC.

Verzani, J. (2014). *Using R for introductory statistics*. CRC Press.

Wickham, H. (2019). *Advanced R*. CRC Press.

TANIMSAL İSTATİSTİK ÖLÇÜLER VE GRAFİKLERİ

ÇİĞDEM ARICIGİL ÇİLAN[2]

Giriş

Bu bölümde verilerin özetlenmesinde kullanılan grafikler ve tanımsal istatistik ölçüler R uygulamaları ile ele alınacaktır. Tanımsal İstatistik kendi içerisinde merkezi eğilim ölçüleri (ortalamalar), değişkenlik ölçüleri ve dağılım ölçüleri olmak üzere 3 alt başlıkta incelenecektir.

2.1. Merkezi Eğilim Ölçüleri (Ortalamalar)

Merkezi eğilim ölçüleri verinin merkezini, lokasyonunu gösteren ve verinin tek bir değerle özetlenmesini sağlayan ölçülerdir. Bu ölçüler; hesaplanmasında tüm gözlem değerlerinin kullanıldığı aritmetik ortalama, geometrik ortalama, harmonik ortalama ve kareli ortalama ile tüm gözlem değerlerinin hesaba dahil edilmediği mod ve medyandan oluşmaktadır. Ayrıca kartillerde bu başlık altında anlatılmıştır.

[2] Prof.Dr., İstanbul Üniversitesi İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı, Orcid: 0000-0002-7862-7028

2.1.1. Aritmetik Ortalama

Aritmetik ortalama günlük hayatta en sık kullanılan ortalama tipidir. Ancak aritmetik ortalamanın aykırı gözlemlere (outlier) karşı hassasiyeti nicel değişkenlerin özetlenmesinde problem olabilmektedir. Yani veriler aritmetik ortalama hesaplanarak tek bir değerle özetlendiğinde veriyi temsil etme sorunu ortaya çıkabilir. Bununla birlikte aritmetik ortalama için geçerli olan $\sum(x - \bar{x}) = 0$ kuralı da tüm nicel veriler için geçerlidir ve bu kural aritmetik ortalamaya dayanan birçok istatistik yöntemin uygulanmasında kullanılmaktadır. Örnek ve anakütlede aritmetik ortalama formülleri aşağıdaki gibidir.

Örnek verisinde $\bar{x} = \frac{\sum x}{n}$ (1)

Anakütle verisinde $\mu = \frac{\sum X}{N}$ (2)

Aritmetik ortalama R'da mean ( ) komutu ile hesaplanmaktadır.

Aritmetik ortalama hesabında ggplot2 paketindeki mpg verisi ile çalışılsın. Bilindiği gibi mpg verisi 234 birimi ve 11 değişkeni olan bir data setidir. Bu data setinden cty değişkeni seçilsin ve aritmetik ortalaması hesaplansın.

```
library(ggplot2)
mean(mpg$cty)
##[1] 16.85897
```

Bir x vektörü tanımlansın:

```
x<-c(20,25,30,45,55)
mean(x)
##[1] 35
```

Hiçbir paket indirmeden de aritmetik ortalama hesabı yapmak mümkündür.

```
x<-c(20,25,30,45,55)
mean(x)
##[1] 35
```

Şimdi de formülü uygulayarak aritmetik ortalama hesaplansın;

```
x<-c(20,25,30,45,55)
a.o<-sum(x)/length(x)
a.o
##[1] 35
```

$\sum(x - \bar{x}) = 0$ kuralının tüm veri büyüklükleri için çalışan bir kural olduğu küçük bir örnekle gösterilsin.

```
x<-c(40,50,60)
mean(x)
##[1] 50
fark<-(x-mean(x))
sum(fark)
##[1] 0
```

Klasik istatistik kitaplarında frekans serisinin ve gruplanmış serinin de aritmetik ortalamaları aşağıdaki formüller ile hesaplanmaktadır.

$$\bar{x} = \frac{\sum fx}{\sum f} \tag{3}$$

(Frekans serilerinde aritmetik ortalama hesabında kullanılan formül)

$$\bar{x} = \frac{\sum fm}{\sum f} \tag{4}$$

(Gruplanmış serilerde aritmetik ortalama hesabında kullanılan formül, burada m grupların orta noktalarını göstermektedir).

Frekans serisi ve gruplanmış serilerde aritmetik ortalama hesapları aşağıdaki komutlar ile yapılabilir:

-Frekans Serisinde Aritmetik Ortalama

Tablo 1 200 Hanenin Çocuk Sayılarına Göre Dağılımı

Çocuk Sayısı (x)	*Hane Sayısı (f)*
0	22
1	57
2	68
3	32
4	15
5	6
	200

```
x<-c(0,1,2,3,4,5)
f<-c(22,57,68,32,15,6)
pay<-x*f
paytop<-sum(pay)
payda<-sum(f)
a.o<-paytop/payda
a.o
##[1] 1.895
```

-Gruplanmış Seride Aritmetik Ortalama

Bir şirketteki 1930 çalışan bir kongreye katılacaklardır. Bilet ücretleri dolar cinsinden aşağıdaki gibi dağılmıştır. Bir adet biletin ortalama fiyatı kaç dolardır? (Gruplanmış serilerde hesaplanan aritmetik yaklaşık bir ortalama değeridir.)

Tablo 2 Bilet Fiyatlarının Dağılımı

Bilet Fiyatları (Dolar)	*Bilet Adedi (f)*
80-85	50
85-90	180
90-95	600
95-100	900
100-105	200
	1930

```
altlimit<-c(80,85,90,95,100)
ustlimit<-c(85,90,95,100,105)
m<-(altlimit+ustlimit)/2
m
##[1]  82.5  87.5  92.5  97.5 102.5
f<-c(50,180,600,900,200)
pay<-sum(f*m)
payda<-sum(f)
a.o<-pay/payda
a.o
##[1] 95.14249
```

Kareli ortalama, tartılı ortalama, medyan ve mod hesaplamalarında da aynı frekans serisi ve gruplanmış seri kullanılacaktır.

2.1.2. Kırpılmış-Düzeltilmiş Aritmetik Ortalama (Trimmed Mean)

Kırpılmış ortalama, verilerin en küçük ve en büyük değerlerini belirli bir oranda atarak kalan veriler ile aritmetik ortalamanın hesaplanmasına dayanır. Özellikle aykırı değerler (outlier) olduğunda aritmetik ortalamanın veriyi temsil gücünü koruması amacıyla kullanılmaktadır. Kırpılmış ortalama aynı zamanda verinin değişkenliğinin düşürülmesinde etkilidir. Kırpılmış aritmetik ortalama yeterince veri olduğunda kullanılmalıdır.

```
veri<-c(1,6,7,8,9,10,5,12,5,30)
mean(veri)
##[1] 9.3
mean(veri, trim=0.10)
##[1] 7.75
mean(veri, trim=0.20)
```

```
##[1] 7.5
```

2.1.3. Kareli Ortalama

Özellikle eksi değerlerin bulunduğu verilerin ortalaması alınırken değerlerin karelerinin ortalaması alınarak bir ortalama hesabı yapılabilir. Ancak hesaplanan bu ortalamanın üzerinde kare etkisi olduğundan karekökü alınarak yorumlanır. Uygulamada en çok standart sapmanın hesabında kullanılır. Kareli ortalama;

$$\text{Örnek verisinde } k.o = \sqrt{\frac{\sum x^2}{n}} \tag{5}$$

$$\text{Anakütle verisinde } K.O = \sqrt{\frac{\sum X^2}{N}} \tag{6}$$

formülleri ile hesaplanır. R'da verinin kareli ortalama hesabı şöyledir:

```
veri<-c(1,6,7,8,9,10,5,12,5,30)
xkare<-veri^2
xkare
##[1]   1  36  49  64  81 100  25 144  25 900
sum(xkare)
##[1] 1425
length(veri)
##[1] 10
K.O=sqrt(sum(xkare)/length(veri))
K.O
##[1] 11.93734
```

Frekans serisi ve gruplanmış serilerde kareli ortalama hesapları aşağıdaki komutlar ile yapılabilir:

-Frekans Serisi İle Kareli Ortalama Hesabı;

$$k.o = \sqrt{\frac{\sum fx^2}{\sum f}} \qquad (7)$$

200 hanenin çocuk sayılarına göre dağılım örneği ile R'da kareli ortalama hesabı şöyledir:

```
x<-c(0,1,2,3,4,5)
 f<-c(22,57,68,32,15,6)
 xkare<-x^2
 pay<-sum(xkare*f)
 payda<-sum(f)
 karek.o<-pay/payda
 sqrt(karek.o)
```

```
##[1] 2.243881
```

Gruplanmış seride kareli ortalama hesabında kullanılan formül aşağıdaki gibidir:

$$k.o = \sqrt{\frac{\sum fm^2}{\sum f}} \qquad (8)$$

```
altlimit<-c(80,85,90,95,100)
ustlimit<-c(85,90,95,100,105)
m<-(altlimit+ustlimit)/2
f<-c(50,180,600,900,200)
mkare<-m^2
pay<-sum(f*mkare)
payda<-sum(f)
sqrt(pay/payda)
```

```
##[1] 95.24733
```

2.1.4. Harmonik Ortalama

Oranların aritmetik ortalaması bazen yanıltıcı sonuçlar verebilmektedir. Bu durumda Harmonik Ortalama kullanılabilmektedir.

Örnek verisinde $h.o = \frac{n}{\sum 1/x^2}$ (9)

Anakütle verisinde $H.O = \frac{N}{\sum 1/X^2}$ (10)

R ile harmonik ortalama komutu "psych" paketinde yer almaktadır.

```
install.packages("psych")
library(psych)
veri<-c(1,6,7,8,9,10,5,12,5,30)
harmonic.mean(veri)
##[1] 4.624702
```

Üç şirketin hisse senetlerinin verileri aşağıdaki gibi olsun. Burada (Fiyat/Kazanç) ların aritmetik ortalamasını hesaplamak, aritmetik ortalamanın farklı kazançları dikkate almamasından yanıltıcı olacaktır. Bu durumda harmonik ortalamanın kullanılması önerilmektedir.

Tablo 3 Şirketlerin Fiyat/Kazanç Değerleri

Şirket	Hisse Fiyatı	Hisse Başına Kazanç	Fiyat/Kazanç
A	20	2	10
B	30	5	6
C	40	10	4

```
veri<-c(10,6,4)
harmonic.mean(veri)
[1] 5.806452
```

2.1.5. Geometrik Ortalama

Genellikle belirli bir zaman periyodunda ortalama büyüme oranı, getiri oranı, büyüme hızı, artış oranının hesabında kullanılan bir ortalama türüdür. Aykırı gözlemlerin (outlier) olduğu verilerin ortalama hesabında da aritmetik ortalama yerine de kullanılmaktadır.

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n} \tag{11}$$

(orjinal değerlerin ortalaması)

$$logG = \frac{\sum logX}{N} \tag{12}$$

(Geometrik ortalama e^{logG} değeridir)

$$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1 \tag{13}$$

(büyüme, artış oranlarının ortalamasının hesabında)

R da geometrik ortalama, "EnvStats" ve "psych" paketleri kullanılarak doğrudan hesaplanabilir.

```
install.packages("EnvStats")
library(EnvStats)
veri<-c(1,6,7,8,9,10,5,12,5,30)
geoMean(veri)
##[1] 6.974008
library(psych)
 geometric.mean(veri)
##[1] 6.974008
```

Şimdi de (12) denklemi kullanılarak geometrik ortalama hesaplanacaktır.

```
x<-c(20,25,30,45,55)
lx<-log(x)
```

```
lx
##[1] 2.995732 3.218876 3.401197 3.806662 4.007333
sum(lx)
##[1] 17.4298
n<-length(x)
loggo<-sum(lx)/n
##[1] 2.995732 3.218876 3.401197 3.806662 4.007333
sum(lx)
##[1] 17.4298
n<-length(x)
loggo<-sum(lx)/n
loggo
##[1] 3.48596
exp(loggo)
##[1] 32.65377
```

-Frekans Serisi İle Geometrik Ortalama

$$logG = \frac{\sum flogx}{\sum f} \quad (14)$$

$$G.O=e^{logG} \quad (15)$$

Bir sektörde faaliyet gösteren şirketlerin karlılıklarına göre dağılımı aşağıdaki gibidir. Sektörün ortalama karlılığı geometrik ortalama kullanılarak R da hesaplanacaktır.

Tablo 4 Şirketlerin Karlılık Dağılımları

Karlılık (%)	Şirket Sayısı (f)
5	100
7,5	120
10	150
15	170
20	180
	720

Öncelikle frekans tablosu tanımlansın:

```
x<-c(5,7.5,10,15,20)
f<-c(100,120,150,170,180)
frekans_serisi<-data.frame(x,f)
frekans_serisi
```

```
##    x   f
##1  5.0 100
##2  7.5 120
##3 10.0 150
##4 15.0 170
##5 20.0 180
```

```
logx<-log(x)
pay<-sum(f*logx)
```

```
payda<-sum(f)
logG<-pay/payda
logG
##[1] 2.427389
G.O<-exp(logG)
G.O
##[1] 11.32927
```

-Gruplanmış Seri İle Geometrik Ortalama

$$logG = \frac{\sum f logm}{\sum f} \quad (16)$$

$$G.O = e^{logG.} \quad (17)$$

formülü ile hesaplanır. Bir frekans serisi tanımlanarak R da geometrik ortalama hesaplansın.

Bir şirkette aynı dil seviyesinde sahip 100 kişiye 4 hafta boyunca dil eğitimi verilmiştir. Eğitim sonrası çalışanların dil seviyelerine göre dağılımı aşağıdaki gruplanmış seride verilmektedir. Buna göre şirket çalışanlarının ortalama yabancı dil seviyesinin geometrik ortalaması R ile hesaplansın.

Tablo 5 Çalışanların Yabancı Dil Seviyeleri

Yabancı Dil Seviyesi	Çalışanların Sayısı (f)
20-40	30
40-60	40
60-80	20
80-100	10
	100

```
altlimit<-c(20,40,60,80)
ustlimit<-c(40,60,80,100)
m<-(altlimit+ustlimit)/2
```

```
m
##[1] 30 50 70 90
f<-c(30,40,20,10)
f
##[1] 30 40 20 10
logx<-log(m)
pay<-sum(f*logx)
payda<-sum(f)
logG<-pay/payda
logG
##[1] 3.884848
exp(logG)
##[1] 48.65957
```

Tüm değerlerin hesaba katıldığı ortalamaların (analitik ortalamalar) dışında ortalamanın değişkendeki tek bir değer ile özetlendiği ortalamalar (analitik olmayan ortalamalar) da merkezi eğilim ölçüsü olarak kullanılmaktadır.

2.1.6. Medyan

Medyan, değerler küçükten büyüğe sıralandıktan sonra $\frac{(n+1)}{2}$. yani tam ortadaki değerdir. R da "modeest" paketi ile medyan ve mod hesaplanabilmektedir. Medyan hem nicel hem de özellikle nitel-ordinal ölçekli değişkenler için bir merkezi eğilim ölçüsüdür.

```
install.packages("modeest")
library(modeest)
veri<-c(1,6,7,8,9,10,5,12,5,30)
median(veri)
##[1] 7
```

Frekans Serisinde Medyan Hesabı

Frekans serilerinde medyan hesabı yapabilmek için kümülatif frekansların hesaplanması gerekmektedir. Kümülatif frekanslarda $\frac{N}{2}$ değerini kapsayan kümülatif frekansa ait x değeri medyan değerini verir. Aşağıdaki örnekte $\frac{200}{2}$ sonucu 100'dür. 100'ü kapsayan birikimli frekans 147'dir. 147'ye karşılık gelen x değeri 2'dir. Medyan 2'dir.

Tablo 6 Çocuk Sayılarının Hane Sayısına Göre Dağılımı

Çocuk Sayısı (X)	Hane Sayısı (F)	Kümülatif Frekans
0	22	22
1	57	79
2	68	**147**
3	32	179
4	15	194
5	6	200

```
tab <- data.frame(value=c(0,1,2,3,4,5), freq=c(22,57,68,32,15,6))
tab
```

```
## value freq
##1     0   22
##2     1   57
##3     2   68
##4     3   32
##5     4   15
##6     5    6
```

Medyan hesabı için veriler R da tanımlasın.

```
data<-
c(rep(0,22),rep(1,57),rep(2,68),rep(3,32),rep(4,15),re
p(5,6))

data

median(data)

##[1] 2
```

Kümülatif frekansların hesabında cusum( ) fonksiyonu kullanılır

```
freq=c(22,57,68,32,15,6)

kumulatif<-cumsum(freq)

kumulatif

##[1]  22  79 147 179 194 200
```

Gruplanmış Serilerde Medyan Hesabı

Gruplanmış serilerde medyan hesabında yine birikimli frekanslar hesaplanır.

$$Medyan = altlimit + \frac{\frac{N}{2} - f_b}{f_{med}} \times i \qquad (18)$$

Burada

f_{med}: medyan grubunun frekansı

f_b: medyan grubu birikimli frekansından bir önceki birikimli frekans

$altlimit$: Medyan grubunun alt limiti

i: Medyan grup aralığı

Tablo 7 Çalışanların Gruplanmış Yabancı Dil Seviyeleri

Yabancı Dil Seviyesi	Çalışanların Sayısı (f)	Kümülatif Frekans
20-40	30	30
40-60	40	**70**
60-80	20	90
80-100	10	100

Gruplanmış serilerde medyan hesabında R ile sadece kümülatif frekans hesaplamaları ve gruplanmış serilerde medyan formülündeki aritmetik işlemler de yapılabilir.

```
f<-c(30,40,20,10)
kümülatif_frekans<-cumsum(f)
kümülatif_frekans
##[1]  30  70  90 100
```

Bu örnekte kümülatif frekanslardan 70, N/2. değer olan 50'yi kapsamaktadır. Bu nedenle 2. grup medyan grubudur. Formül uygulandığında medyanın 50 olduğu görülmektedir.

```
medyan<-40+((50-30)/40*20)
medyan
##[1] 50
```

2.1.7. Mod

Bir değişkende en çok tekrar eden değerdir. Hem nicel hem de nitel değişkenler için hesaplanabilir. Tekrarlı değerlerden oluşan bir x vektörü tanımlansın.

```
x<-c(100,100,100,100,300,2000000,500)
```

x vektörünün frekans tablosu oluşturulur. "modeest" paketindeki mfv( ) komutu ile mod hesaplanır.

```
as.data.frame(table(x))
##      x  Freq
##1   100    4
##2   300    1
##3   500    1
```

```
##4 2e+06    1
```

```
install.packages("modeest")
```

```
library(modeest)
```

```
mfv(x)
```

```
##[1] 100
```

Bir nitel serinin mod değerinin hesaplanması için cinsiyet değişkenin şıklarına göre dağılımı tanımlansın.

```
tab       <-       data.frame(value=c("kadin","erkek"),
freq=c(3,17))
```

```
tab
```

```
##  value freq
##1 kadin    3
##2 erkek   17
```

```
vec <- rep(tab$value, tab$freq)
```

```
vec
```

```
## [1] "kadin" "kadin" "kadin" "erkek" "erkek" "erkek" "erkek"
"erkek"
##[9] "erkek" "erkek" "erkek" "erkek" "erkek" "erkek" "erkek"
"erkek"
##[17] "erkek" "erkek" "erkek" "erkek"
```

```
mfv(vec)
```

```
##[1] "erkek"
```

Gruplanmış Serilerde Mod Hesabı:

$$Mod = altlimit + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}) \times i \qquad (19)$$

Burada;

f_m: Mod grubunun frekansı

f_{m-1}: Mod grubunun frekansından bir önceki frekans

f_{m+1}: Mod grubunun frekansından bir sonraki frekans

$altlimit$: Mod grubunun alt limiti

i: Mod grubunun grup aralığı

Aşağıda bir işletmenin çalışanları için satın aldığı yolcu biletlerinin fiyat aralıkları ve satın alınan bilet adetleri gruplanmış bir seride gösterilmektedir. Bilet fiyatlarının modu hesaplansın.

Tablo 8 Gruplanmış Bilet Fiyatları

Bilet Fiyatları (Dolar)	Bilet Adedi
80-85	50
85-90	180
90-95	600
95-100	**900**
100-105	200
	1930

Frekansı 900 olan grubun mod grubu olduğu görülmektedir. Gruplanmış serilerde mod formülü R'da uygulandığında mod değeri 96,36 elde edilmektedir.

```
Mod<-95+((300/(300+700))*5)
Mod
##[1] 96.5
```

Unutulmamalıdır ki gruplanmış serilerde elde edilen istatistik ölçüler tahmini değerlerdir.

2.1.8. Kartiller ve Persantiller

Değişken değerleri küçükten büyüğe sıralandıktan sonra veriyi 4 eşit parçaya bölen istatistik ölçüler kartiller (Q_1, Q_2, Q_3) olarak adlandırılır. R'da summary() komutu ile data setinin kartilleri elde edilebilir. Aynı zamanda bir x değişkenin kartilleri quantile() komutuyla da elde edilebilir.

```
x<-
c(18,18,16,18,20,18,6,10,12,16,18,20,22,24,26,22,20)
```

```
quantile(x, prob=c(.25,.5,.75))
##25% 50% 75%
## 16  18  20
summary(x)
##Min.  1st Qu.  Median  Mean   3rd Qu.   Max.
##6.00   16.00   18.00   17.88   20.00   26.00
```

Q_1=16, değerler küçükten büyüğe sıralandığında verilerin %25'lik kısmının maksimum değeridir. Q_2=18 (medyan) sıralanmış verilerin ilk %50 kısmının maksimum değeri ve verilerin en büyük değerlerinin%50'sinin minimum değeridir. Q_3=20 sıralanmış verilerin %75'inin en büyük değeri ve en büyük verilerin %25'inin minimum değeridir.

Persantilleri küçükten büyüğe sıralanan veriyi 100 eşit parçaya ayıran değerlerdir. R da hesaplanmalarında yine quantile ( ) komutu kullanılır. Örneğin x vektörünün 95. ve 90. persantil değerleri aşağıdaki komutla hesaplanır.

```
quantile(x,prob=c(.90,.95))
##90%  95%
##22.8 24.4
```

Tertiller(Tertiles) veri küçükten büyüğe sıralandığında veriyi 3 eşit parçaya ayıran değerlerdir.

```
quantile(x, prob=c(.33,.66))
##33% 66%
##18  20
```

2.1.9. Nitel Değişkenlerin Tanımlanması

Nitel değişkenlerden bilgi üretilmesi genellikle nitel seriler oluşturularak değişkenin şıklarına göre dağılımının belirlenmesi ile

yapılmaktadır. Nitel serilerin frekansları ve kısmi frekansları hesaplanarak yorumlanır.

"ggplot2" paketinde yer alan "mpg" verisi bir yakıt tasarrufu araştırmasında 1999 ve 2008 yıllarında üretilen 38 popüler otomobil modeli ve özelliklerini içeren bir veri setidir. Değişkenlerden biri "fl" yakıt tipidir. "fl" değişkeni 5 şıkkı olan bir nitel değişkendir.

```
library(ggplot2)
table(mpg$fl)
## c   d   e   p   r
## 1   5   8  52 168
table(mpg$fl) / nrow(mpg)
##   c           d           e           p           r
##0.004273504  0.021367521  0.034188034  0.222222222
0.717948718
```

Bu otomobil modellerinde en çok 0,718 oranla "r" tipi yakıt kullanıldığı saptanmıştır.

Nitel değişkenlerin özetlenmesinde oranlar kullanıldığı gibi nitel-nominal ölçekli değişkenlerde mod, nitel-ordinal değişkenlerde medyan hesaplamaları da yapılabilir. Bu örnekte "fl" nominal ölçekli bir değişken olduğundan mod hesabı da değişkenin özetlenmesinde kullanılabilir.

```
library(modeest)
mfv(mpg$fl)
##[1] "r"
```

2.2. Değişkenlik Ölçüleri

Ortalamaların temsil gücünü ve verilerin değişkenliğini gösteren ölçüler olan değişkenlik ölçülerinden; Aralık, Varyans,

Standart Sapma ve Kartiller Arası Aralık R programı kullanılarak uygulanacaktır.

2.2.1. Aralık Ölçüsü

Aralık ölçüsünün hesaplanmasında, sadece verinin maksimum ve minimum değerleri kullanıldığından basit bir değişkenlik ölçüsüdür. Aralık ölçüsü aykırı değerlere karşı hassastır ve değeri en az sıfır olabilir. Ancak maksimum değeri olmadığından bir büyüklük ölçüsü değildir. Farklı verilerin değişkenliğinin karşılaştırılmasında kullanılır. Böylece hangi verinin nispi olarak diğerlerine göre daha az değişken (homojen) olduğuna karar verilir. Bir başka deyimle hangi veri grubunun aritmetik ortalamasının daha temsili olduğuna karar verilmesini sağlar.

$$Aralık = Maksimum\ değer - Minium\ değer \qquad (20)$$

Aralık ölçüsünün hangi koşullarda nasıl kullanılacağı hipotetik bir örnek ile açıklansın:

2 veri grubunun değerleri x=c(30,40,50) ve y=c(10,20,90) olsun. Her iki veri grubunun aritmetik ortalaması 40'tır. Aralık ölçüleri ise sırasıyla 20 ve 80'dir. Aralık ölçüsü burada ortalamaları aynı olan iki grubun hangisinin aritmetik ortalamasının daha temsili olduğunun belirlenmesini sağlar. Başka bir ifadeyle hangi veri grubunun diğerine göre daha homojen olduğunu belirler. 20<80 olduğundan x veri grubunun y'ye göre daha homojen olduğu söylenebilir. Veri gruplarının değişkenliklerinin karşılaştırılabilmesi için grupların aynı birimde, benzer örnek büyüklüklerinde ve benzer değer aralıklarında olmaları gerekir.

R'da range( ) komutu veri grubunun minimum ve maksimum değerlerini vermektedir.

```
x=c(30,40,50)
y=(10,20,90)
```

```
range(x)
##[1] 30 50
range(y)
##[1] 10 90
```

x verisinin Aralık ölçüsü formülle hesaplansın.

```
maksimum<-max(x,na.rm = FALSE)
minimum<-min(x,na.rm=FALSE)
R=maksimum-minimum
R
##[1] 20
```

y verisinin Aralık Ölçüsü formülle hesaplansın.

```
maksimum<-max(y,na.rm = FALSE)
minimum<-min(y,na.rm=FALSE)
R=maksimum-minimum
R
##[1] 80
```

2.2.2. Varyans

Varyans teknik olarak birimlerin ortalamadan farklarının karelerinin ortalamasıdır. Çıktıların yorumlanmasında çok zorlanılan bu istatistik ölçü yerine genellikle standart sapma yani varyansın karekökü kullanılır. Aşağıda anakütle ve örnek verileri için formüller sırasıyla yer almaktadır:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{21}$$

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} \tag{22}$$

Varyans R da var() komutu ile hesaplanır.

```
x=
c(18,18,16,18,20,18,6,10,12,16,18,20,22,24,26,22,20)
var(x)
##[1] 24.73529
```

Burada varyans değeri örnek verisi varsayılarak hesaplanmaktadır. "mpg" verisinde yer alan "cty" değişkeni niceldir ve galon başına mil miktarını göstermektedir. Aşağıda "cty" nicel değişkenin varyansı hesaplanmaktadır.

```
library(ggplot2)
var(mpg$cty)
[1] 18.11307
```

2.2.3. Standart Sapma

Analiz çıktılarında genellikle aritmetik ortalama ile birlikte yorumlanan değişkenlik ölçüsüdür. Veriden tesadüfen seçilen herhangi bir birimin değerinin aritmetik ortalamadan ortalama olarak ne kadar saptığını gösterir. Aritmetik ortalama ile aynı birime sahiptir. Aşağıda anakütle ve örnek verileri için formüller sırasıyla yer almaktadır:

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}} \tag{23}$$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \tag{24}$$

Hipotetik bir örnekle standart sapmayı açıklamaya çalışalım. Bir sınıftaki 6 öğrencinin notları aşağıdaki gibi olsun.

```
x=c(30,50,60,70,80,95)
```

Bu durumda öğrencilerin aritmetik ortalaması ve standart sapması aşağıdaki gibi hesaplanır:

```
mean(x)
```

```
##[1] 64.16667
```

```
sd(x)
```

```
##[1] 22.89469
```

Burada standart sapma; öğrencilerden herhangi biri tesadüfen seçildiğinde ve notu bilinmediğinde, öğrencinin notunun ortalama not olan 64,16'dan ortalama olarak 22,89 puan sapabileceği şeklinde yorumlanır.

"mpg" verisinde yer alan "cty" değişkeninin standart sapması R da şöyle hesaplanır:

```
sd(mpg$cty)
```

```
##[1] 4.255946
```

2.2.4. Kartiller Arası Değişkenlik (KAD)

Değişkenlik ölçüleri arasında son olarak ele alacağımız ölçü Kartiller Arası Değişkenlik ölçüsüdür. Hesaplanmasında 1. Kartil(Q_1) ve 3. Kartil (Q_3) kullanılar.

$$KAD = Q_3 - Q_1 \tag{25}$$

Kartiller Arası Değişkenlik ölçüsü "%25 Kırpılmış Aralık Ölçüsü (25% Trimmed Range)" olarak da adlandırılır. Yani veriler küçükten büyüğe sıralandığında, verilerin başından ve sonundan %25'lik kısım atılarak kalan veriler ile Aralık ölçüsü hesaplanır. x verisinin KAD değeri R da şöyle hesaplanır:

```
x<-c(30,50,60,70,80,95)
```

```
IQR(x)
```

```
##[1] 25
```

KAD ölçüsü, özellikle aykırı değer etkisini ortadan kaldıran bir değişkenlik ölçüsüdür.

2.3. Asimetri ve Basıklık Ölçüleri

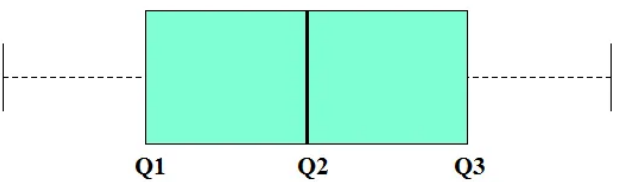
Verilerin normal dağılıma göre nasıl dağıldıklarını tanımlayan asimetri ve basıklık ölçüleri de Tanımsal İstatistik ölçüler olarak kabul edilmektedir. Asimetri ölçüleri verilerin simetrisi hakkında bilgi verir. Verinin simetrik dağılması verinin çan eğrisine uygun olduğunu gösterirken, sağa asimetrik olması; verilerin çoğunun aritmetik ortalamanın altında toplandığını, sola asimetrik olması ise; verilerin çoğunun aritmetik ortalamanın üzerinde toplandığını gösterir. Asimetri ölçüleri arasında en çok kullanılan ve en hassas (tüm değerler hesaba katıldığından) olan ölçü Momentlere Dayanan Asimetri Ölçüsü'dür. Momentlere dayanan asimetri ölçüsü dışında Pearson Asimetri Ölçüsü ve Bowley Asimetri Ölçüsü de ele alınacaktır.

2.3.1. Bowley Asimetri Ölçüsü

Yule'un Asimetri Katsayısı olarak da adlandırılan Bowley Asimetri Ölçüsü kartillere dayanır. Tüm değerleri hesaba katmaması ve veri setinin başından ve sonundan %25'lik kısmı gözardı etmesi nedeniyle zayıf bir asimetri ölçüsü olarak kabul edilmektedir.

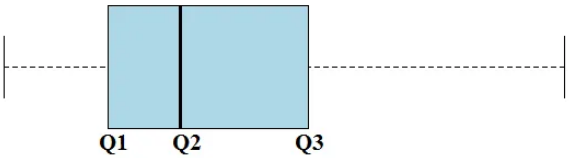
$$S_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)} \tag{26}$$

$S_B = 0$ olması verinin simetrik olduğunu gösterir. Aşağıda kutu diyagramları ile simetrik ve asimetrik durumlar tanımlanmıştır.



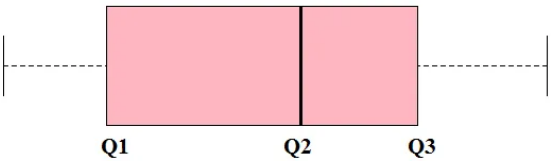
Şekil 1 Kutu Diyagramı (Simetrik Durum)

$S_B > 0$ olması sağa asimetrik durumu tanımlar. Yani verilerin çoğunun aritmetik ortalamanın altında toplandığını gösterir.



Şekil 2 Kutu Diyagramı (Sağa Asimetrik Durum)

$S_B < 0$ olması sola asimetrik durumu tanımlar. Yani verilerin çoğunun aritmetik ortalamanın üzerinde toplandığını gösterir.



Şekil 3 Kutu Diyagramı (Sola Asimetrik Durum)

R programında "KbMvtSkew" paketi ile Bowley Asimetri ölçüsü doğrudan hesaplanabilmektedir.

```
install.packages("KbMvtSkew")
```

```
library(KbMvtSkew)
```

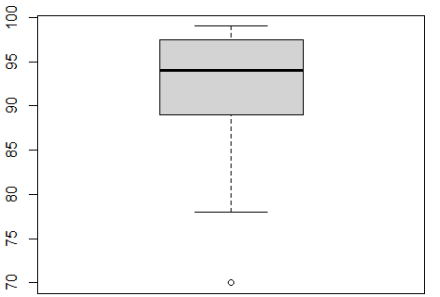
```
data=c(90,97,94,99,98,99,96,88,93,97,99,90,87,78,70)
```

```
BowleySkew(data)
```

```
## [1]  -0.1764706
```

Veriler sola asimetriktir. Kutu diyagramı ile de sola asimetri durumu görülmektedir.

```
boxplot(data)
```



Şekil 4 Sola Asimetrik Verilerin Kutu Diyagramı

2.3.2. Pearson Asimetri Ölçüsü

Pearson, mod ve medyanın yer aldığı iki asimetri ölçüsü geliştirmiştir. Ancak mod ve medyan verideki tüm değerleri hesaba katmadığından Pearson'un asimetri ölçüleri asimetriyi hassas bir şekilde ölçememektedir.

$$S_{P1} = \frac{\bar{X} - MOD}{\sigma} \tag{27}$$

$$S_{P2} = \frac{3(\bar{X} - MEDYAN)}{\sigma} \tag{28}$$

"KbMvtSkew" paketi ile Pearson Asimetri ölçüsü de hesaplanabilmektedir.

```
library(KbMvtSkew)
PearsonSkew(data)
##[1] -1.254945
```

Pearson Asimetri Ölçüsü'nde de Bowley'de olduğu gibi 0 simetrik durumu. 0'ın üzerinde bir katsayı sağa asimetriyi, 0'ın altında bir katsayı da sola asimetriyi temsil etmektedir.

2.3.3. Fisher'in Momentlere Dayanan Asimetri ve Basıklık Ölçüsü

Aritmetik ortalama etrafında momentlere dayanan bu asimetri ölçüsü uygulamada en sık kullanılan asimetri ölçüsü olarak bilinmektedir.

$$S_F = E\left[\left(\frac{(x-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} \quad (29)$$

Simetri durumu sıfır üzerinden değerlendirilmektedir.

$S_F = 0$ Simetrik dağılım

$S_F > 0$ Sağa asimetrik dağılım

$S_F < 0$ Sola asimetrik dağılım

Basıklık ölçüsü de aritmetik ortalama etrafında dördüncü momente dayanmaktadır.

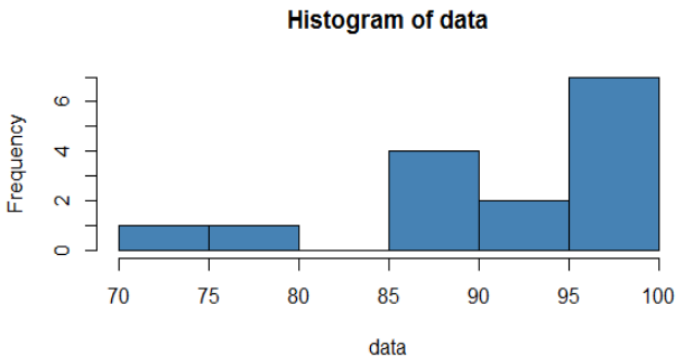
$$K_F = E\left[\left(\frac{(x-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} \quad (30)$$

Basıklık ölçüsü verilerin "normal" yükseklikte olup olmadığını gösterir. Basıklıkta temel alınan değer 3'tür. 3'ün üzerinde bir basıklık değeri verinin "sivri", 3'ün altında olan değerler ise verinin "basık" olduğunu gösterir. Sivri bir veri; verilerin belirli bir değer aralığında anlamlı bir şekilde toplandığını gösterirken, basık bir veri değerlerin anlamlı bir aralıkta toplanmadığı anlamına gelir.

R'da "moments" paketinde momentlere dayanan asimetri ve basıklık ölçüleri skewness( ) ve kurtosis( ) komutları ile hesaplanmaktadır.

```
install.packages("moments")

library(moments)

data=c(90,97,94,99,98,99,96,88,93,97,99,90,87,78,70)
```

```
hist(data,col='steelblue')
```



Şekil 5 Verilerin Histogramı

```
skewness(data)
```

```
[1] -1.391777
```

```
kurtosis(data)
```

```
[1] 4.177865
```

Verilerin "skewness" değerinin 0 ve "kurtosis" değerinin 3 olması verinin normal dağıldığını gösterir. Histogramda verilerin simetrisi hakkında bilgi verir. Ancak verilerin normal dağılıp dağılmadıkları test de edilebilmektedir. Burada hipotezler aşağıdaki gibidir:

H_0: *Veriler Normal Dağılmaktadır.*
H_1: *Veriler Normal Dağılmamaktadır.*

Tek değişkenli normallik testleri arasında en sık kullanılanlarından biri Jarque-Bera Testi'dir.

```
jarque.test(data)
```

```
##Jarque-Bera Normality Test
##data:  data
##JB = 5.7097, p-value = 0.05756
##alternative hypothesis: greater
```

p>0.05 olduğundan verilerin normal dağıldığı söylenebilir.

2.4. Grafikler-Veri Görselleştirme

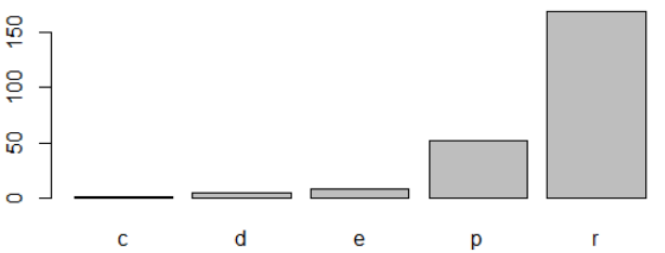
Verilerin özetlenmesinde grafikler verinin kolaylıkla anlaşılmasına sağlayan görsellerdir. Büyük verilerin analizi ile birlikte verinin görselleştirilmesinin önemi artmış ve günümüzde veri görselleştirme;veriden bilgi üretmeyi sağlayan bir araştırma alanı olarak nitelendirilmektedir.

Tanımsal istatistikte; pasta diyagramı, sütun diyagramı, histogram, kutu diyagramı ve serpilme diyagramı en sık kullanılan tanımlayıcı görsellerdir.

-Pasta ve Sütun Grafikleri

Sütun grafikleri kategorik değişkenlerin (nitel ve sınırlı sayıda şıkkı olan kesikli değişken) görsel olarak olarak özetlenmesinde kullanılır. "mpg" verisindeki "fl" yakıt tipi nitel değişkeninin R da sütun diyagramı aşağıdaki komutlar ile çizilir:

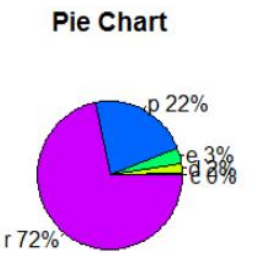
```
library(ggplot2)
barplot(table(mpg$fl))
```



Şekil 6 "fl" Değişkeninin Histogramı

Aynı verinin pasta diyagramının çizilmesi için R da kullanılan komutlar aşağıdaki gibidir:

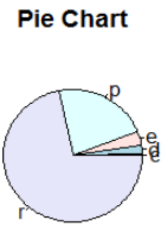
```
table(mpg$fl)
## c   d   e   p   r
## 1   5   8  52 168
slices <- c(1,5,8,52,168)
lbls <- c("c","d","e","p","r")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
+     main="Pie Chart")
```



Şekil 7 "fl" Değişkeninin Pasta Diyagramı

%'lerin görünmediği bir pasta diyagramı çiziminde kullanılan R kodları aşağıda verilmiştir.

```
slices <- c(1,5,8,52,168)
lbls <- c("c","d","e","p","r")
pie(slices, labels = lbls, main="Pie Chart")
```



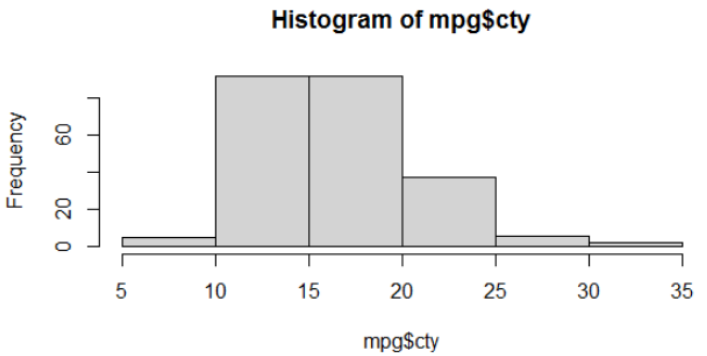
Şekil 8

"fl" Değişkeninin%'leri İçermeyen Pasta Diyagramı

-Histogramlar

Histogramlar nicel değişkenler için çizilen grafiklerdir. "ggplot2" paketinde "mpg" veri setinde yer alan "cty" (galon başına mil) nicel değişkenin histogramı hist( ) komutu ile çizilir.

```
library(ggplot2)
hist(mpg$cty)
```



Şekil 9 "cty" Değişkeninin Histogramı

Sturges kuralını uygulayarak grup sayıları belirlensin. Sturges kuralının formülü:

$$Grup\ sayısı = 1 + 3{,}22 \times logn \quad (31)$$

"mpg" veri setindeki cty nicel değişkeninin birim sayısı 234'tür.

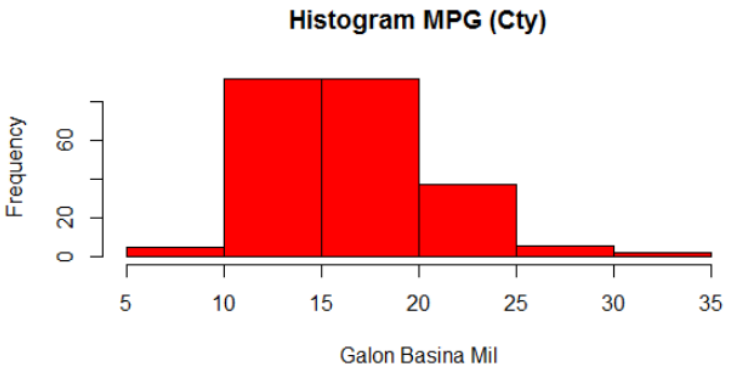
```
str(mpg$cty)
```

Buna göre grup sayısı;

$Grup\ sayısı = 1 + 3{,}22 \times log234$

$Grup\ sayısı = 8{,}62$ yani 9 kabul edilebilir. Burada 9 R'da "breaks" sayısıdır.

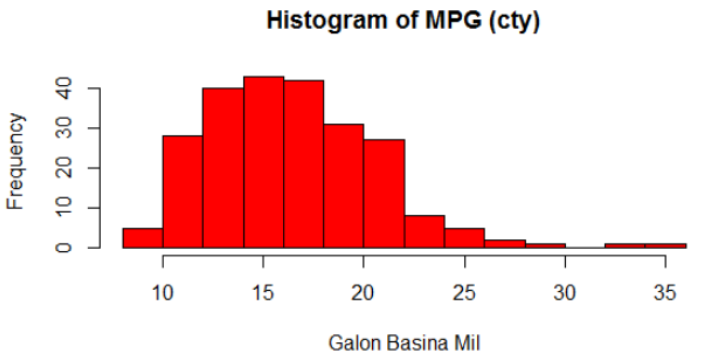
```
hist(mpg$cty,
+      xlab = "Galon Basina Mil",
+      main = "Histogram MPG (Cty)",
+      breaks = 9,
+      col = "red",
+      border = "black")
```



Şekil 10 Sturges Kuralına Göre "cty" Değişkeninin Histogramı

9 gruplu histogramda son gruplarda frekans sayıları çok düşük olduğundan belirgin bir şekilde görülememektedir. Aynı veride 12 gruplu bir histogram çizildiğinde görsel aşağıdaki gibidir:

```
hist(mpg$cty,
+      xlab = "Galon Basina Mil",
+      main = "Histogram of MPG (cty)",
+      breaks = 12,
+      col = "red",
+      border = "black")
```

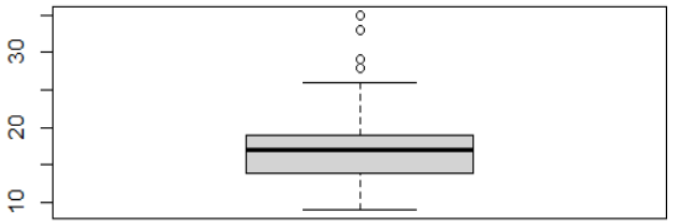


Şekil 11 "cty" Değişkeninin 12 gruplu Histogramı

-Kutu Diyagramı

Kutu diyagramları nicel değişkenler ile çizilmektedir. Kutu diyagramı, tek değişkenli aykırı gözlemlerin saptanmasında sıklıkla kullanılmakla birlikte verinin dağılımı hakkında fikir vermektedir. Verinin minimum, maksimum, Q1, Q2 ve Q3 değerleri kullanılarak çizilmektedir."ggplot2" paketindeki "mpg" verisinin "cty" değişkeni R da çizilsin.

```
library(ggplot2)
boxplot(mpg$cty)
```



Şekil 12 "cty" Değişkeninin Kutu Diyagramı

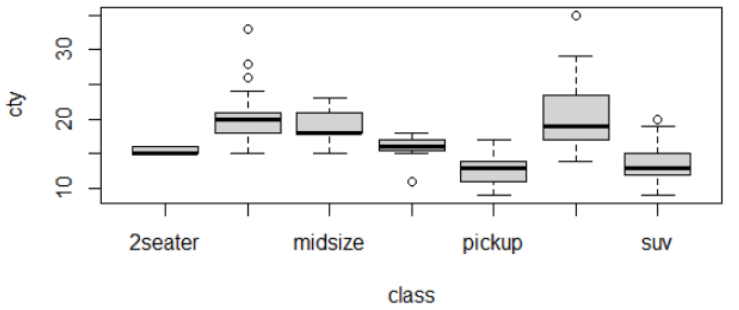
```
summary(mpg$cty)
```

```
##Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 9.00   14.00   17.00      16.86   19.00   35.00
```

Kutu diyagramında dikdörtgenin altındaki çizgi verinin minimum değerini, dikdörtgenin alt uzun kenarı verinin 1. Kartil değerini (Q1), dikdörtgenin içerisindeki çizgi 2. Kartili (medyan=Q2), dikdörtgenin üst uzun kenarı 3. Kartili (Q3) gösterir. Dikdörtgenin üzerindeki çizgi ise verinin maksimum değeridir. Maksimum değerin üzerinde 4 nokta (veri) verideki aykırı gözlemleri göstermektedir.

Kutu diyagramında aynı zamanda bir nicel değişken, başka bir nitel değişkenin kategorilerine göre çizilebilir ve bu kategorilere göre kutu diyagramları karşılaştırılabilir. "cty(galon başına mil)" nicel değişkeni aynı veri setinin "class(otomobil tipi)" nitel değişkeninin kategorilerine göre çizilen kutu diyagramları aşağıdaki gibidir:

```
boxplot(cty ~ class, data = mpg)
```

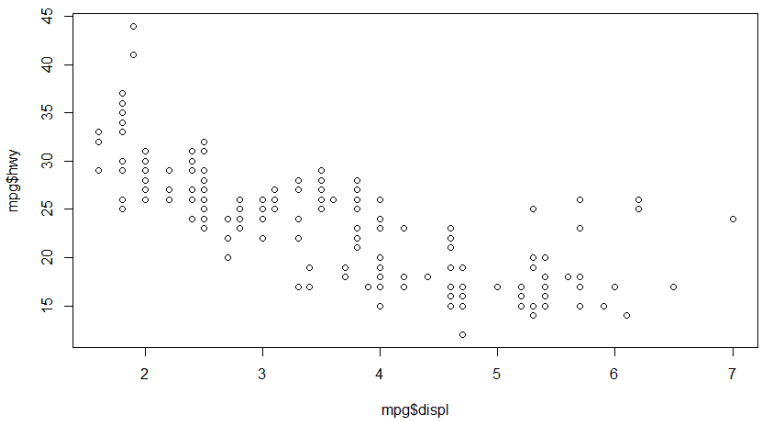


Şekil 13 "class" ve "cty" Değişkenlerinin Kutu Diyagramı

-Serpilme Diyagramı

Serpilme diyagramı iki nicel değişken arasındaki ilişkinin biçimini gösteren bir görseldir. Genellikle iki değişken arasında doğrusal bir ilişkinin olup olmadığının analaşılması için kullanılır. Aşağıdaki örnekte mpg verisinde yer alan motor silindir hacmi(displ) ve galon başına mil miktarı(hwy) değişkenleri arasında bir serpilme diyagramı çizilmektedir.

```
plot(mpg$hwy~mpg$displ)
```



Şekil 14 "displ" ve "hwy" Değişkenlerinin Serpilme Diyagramı

Bu bölümde ortalama, medyan, mod, varyans ve standart sapma gibi tanımlayıcı istatistik ölçüleri ile histogram, kutu grafiği ve çubuk grafik gibi görselleştirme yöntemleri ele alınmıştır. Örnek R kodlarıyla ölçüler ve grafikler uygulamalı olarak gösterilmiştir. Böylece uygulamacılar kendi verilerini özetleyip görselleştirebileceklerdir.

Kaynakça

Crawley, M. J. (2012). *The R book*. Wiley.

Dalgaard, P. (2008). *Introductory statistics with R*. Springer.

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using R*. SAGE.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2016). *Introduction to the practice of statistics*. W. H. Freeman.

Triola, M. F. (2018). *Elementary statistics*. Pearson.

Verzani, J. (2014). *Using R for introductory statistics*. CRC Press.

Zuur, A. F., Ieno, E. N., & Meesters, E. H. W. G. (2009). *A beginner's guide to R*. Springer.

OLASILIK KURALLARI VE OLASILIK DAĞILIMLARI

ÇİĞDEM ARICIGİL ÇİLAN[3]

Giriş

Olasılık bir olayın mümkün sonuçlarından birinin ortaya çıkma şansı olarak tanımlanabilir. Bir olay sonlu ya da sonsuz bir popülasyonda meydana gelebilir Sonlu popülasyonda olayların sayısı bilinmektedir. Bu nedenle kesin olasılıklar hesaplanabilmektedir. Örneğin 1000 işletme olan bir sektörde; küçük işletmelerin sayısı 500, orta ölçekli işletmelerin sayısı 300, büyük işletmelerin sayısı 200 olsun. Bu durumda tesadüfi olarak büyük işletme seçme olasılığı 200/1000, orta ölçekte işletme seçme olasılığı 300/1000 ve küçük işletme seçme olasılığı 500/1000'dir. Tüm bu oranların toplamı 1'e eşit olacaktır.

Sonsuz bir popülasyonda ise olay sayısı sonsuz olduğundan hesaplanan olasılıklar yaklaşık olasılıklar olacaktır. Burada olasılıklar kısmi frekans esas alınarak hesaplanır. Örneğin para atma olayında tura gelmesi olasılığı N'e yani deneme sayısına (yani paranın kaç kez atıldığına) bağlı olarak farklılaşacaktır. N deneme

[3] Prof.Dr., İstanbul Üniversitesi İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı, Orcid: 0000-0002-7862-7028

sayısı arttıkça sonucun sonlu olasılığa yaklaştığı yani formülünün aşağıdaki gibi olacağı kabul edilir.

$$P(T) = \frac{n(T)}{N} \quad (1)$$

Burada;

$P(T): Tura\ gelme\ olasiliği$

$n(T): N\ denemede\ Tura\ sayı$

$N: Toplam\ Deneme\ Sayısıdır.$

Hilesiz bir parada tura olasılığının 0.5 olması beklenir. Para 500 kez atılmış olsun ve 245 defa tura gelmiş olsun. Bu durumda tura gelme olasılığı $P(T) = \frac{245}{500}$ yani 0,49'dur. Sonsuz (kısmi frekanslara dayanan) olasılıklar N deneme sayısı arttıkça sonlu olasılıklara yaklaşmaktadır.

R da bir para atma deneyi tanımlanmak istendiğinde, öncelikle bir paranın karakterlerden oluşan bir vektör olarak tanımlanması gerekmektedir.

```
coin <- c("heads", "tails")
coin
```

```
## [1] "heads" "tails"
```

Para atma deneyinde sample() fonksiyonunu kullanmak da mümkündür.

```
sample(coin, size = 1)
```

```
## [1] "heads"
```

Burada size=1 "heads" ve "tails" karakterlerinden oluşan vektörden örnek büyüklüğü 1 olan bir örnek seçimini tanımlamaktadır.

```
sample(coin, size = 2)
## [1] "heads" "tails"
```

İadesiz çekimle örnek büyüklüğü 2 olan bir örnek çekilmesi durumu ise aşağıdaki gibi tanımlanmaktadır.

```
sample(coin, size = 2, replace = FALSE)
## [1] "tails" "heads"
```

İadeli çekim ancak örnek büyüklüğü 2'den fazla olması durumunda mümkündür. Aynı tesadüfi sonuçları elde edebilmek için set.seed( ) fonksiyonunun kullanılması gerekmektedir.

```
set.seed(1257)
sample(coin, size = 3, replace = TRUE)
```

```
##[1] "tails" "heads" "heads"
```

```
set.seed(1257)
sample(coin, size = 3, replace = TRUE)
```

```
##[1] "tails" "heads" "heads"
```

Hilesiz bir para atıldığında yazı ve turanın gelme olasılığı 0.5 olarak kabul edilmektedir. Ancak farklı olasılıklar da tanımlanabilmektedir. Örneğin tura gelme olasılığı 0.30 yazı gelme olasılığı 0.70 olsun.

```
sample(coin, size = 5, replace = TRUE, prob = c(0.3, 0
.7))
```

```
##[1] "tails" "heads" "tails" "tails" "tails"
```

Bir para 100 kez atılmış olsun.

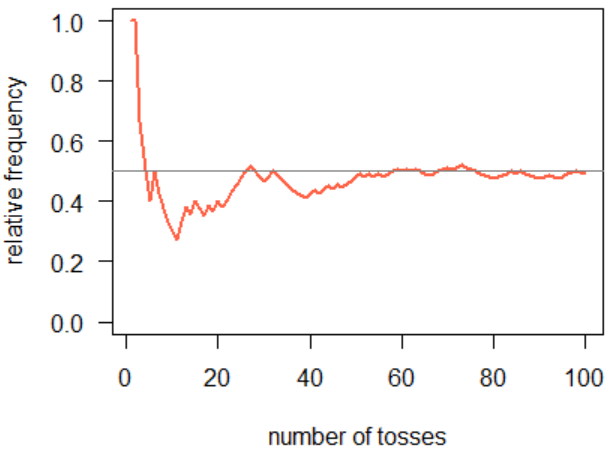
```
num_flips <- 100              #para atılma sayısı
coin <- c('heads', 'tails')
flips <- sample(coin, size = num_flips, replace = TRUE
)
freqs <- table(flips)
freqs
```

```
##flips
##heads tails
## 49    51
```

Burada sonuç 50 tura 50 yazı elde edilebildiği gibi bizim örneğimizde tura sayısı 49, yazı sayısı 51 elde edilmiştir.

Paranın atış sayısı arttıkça tura-yazı gelme olasılığının 0.5'e yaklaştığı bir grafik ile gösterilsin. Öncelikle para atış sayısı arttıkça tura sayısı kümülatif bir fonksiyon olarak tanımlansın.

```
heads_freq <- cumsum(flips == 'heads') / 1:num_flips
plot(heads_freq,      # vector
     type = 'l',      # line type
     lwd = 2,         # width of line
     col = 'tomato',  # color of line
     las = 1,         # orientation of tick-mark labels
     ylim = c(0, 1),  # range of y-axis
     xlab = "number of tosses",    # x-axis label
     ylab = "relative frequency")  # y-axis label
abline(h = 0.5, col = 'gray50')
```



Şekil 1 Tura Sayısı-Kısmi Frekanslar Grafiği

3.1. Olasılık Kuralları

Bu alt başlıkta olasılık kuralları marjinal, ortak ve koşullu olasılıklar temel alınarak ele alınacaktır. A ve B olayından oluşan bir örneklem uzayında; A'nın gerçekleşmesi olasılığı $P(A)$, B'nin gerçekleşmesi olasılığı $P(B)$ ile gösterilir. Bunlar marjinal olasılıklar olarak tanımlanır. A ve B olaylarının birlikte gerçekleşme olasılıkları olaylar bağımsız ise $P(A \cap B) = P(A) * P(B)$'ye eşit

olur. Bu olasılık ortak (joint) olasılıktır. A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı $P(B|A)$, B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı $P(A|B)$ koşullu olasılıklardır. Burada $P(A|B) = \frac{P(A \cap B)}{P(B)}$ eşitliği sağlanır. Bayes Teoremi ise bir olayın gerçekleşme olasılığını, bu olayla ilgili eldeki yeni bilgilere göre hesaplamaya yarar. Koşullu olasılıkları tersine çevirmemize olanak sağlayan bir teoremdir.

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_{j=1}^{n} P\left(B\middle|A_j\right)\,P(A_j)} \quad (2)$$

Bir örnek ile marjinal, ortak, koşullu ve Bayes olasılıklarının nasıl hesaplandığı gösterilebilir. Bu örnekte bir dijital pazarlama kampanyasında, hedef kitleye gösterilen bir reklama tıklayan kullanıcıların davranışları analiz edilmek istensin.

Ziyaretçilerin %20'si gençtir (18-25 yaş arası)

$P(Genç) = 0.20$ (Marjinal Olasılık)

Ziyaretçilerin %80'si genç değildir

$P(\overline{Genç}) = 0.80$ (Marjinal Olasılık)

Genç kullanıcıların reklamı tıklama olasılığı %15'dir.

$P(Tıklama|Genç) = 0{,}15$ (Koşullu Olasılık)

Genç olmayan kullanıcıların reklamı tıklama olasılığı %3'tür.

$P(\text{Tıklama} \mid \overline{Genç}) = 0{,}03$ (Koşullu Olasılık)

$$P(Genç|Tıklama) = \frac{P(Tıklama|Genç)\,P(Genç)}{P(Tıklama|Genç)\,P(Genç) + P\left(\text{Tıklama}\middle|\overline{Genç}\right)\,P(\overline{Genç})} = 0.555$$

(Bayes Koşullu Olasılık)

2x2 boyutlu bir çapraz tablo oluşturularak olasılıklar hesaplanacaktır:

```
tablo <- matrix(c(430,142, 370, 130), byrow =TRUE,
ncol=2)
colnames(tablo) <- c('Başarılı', 'Başarısız')
rownames(tablo) <- c('Kadın', 'Erkek')
```

```
##        Başarılı Başarısız  Sum
##Kadın      430       142    572
##Erkek      370       130    500
##Sum        800       272   1072
```

-Ortak Olasılıklar

```
prop.table(tablo)
```

```
##     Başarılı     Başarısız
##Kadın 0.4011194 0.1324627
##Erkek 0.3451493 0.1212687
```

Burada en yüksek ortak olasılık olan 0.4011, 1072 kişi içerisinde kadın ve başarılı olanların oranını göstermektedir.

$$P(Kadın \cap Başarılı) = \frac{430}{1072} = 0.4011$$

-Marjinal Olasılıklar

```
addmargins(prop.table(tablo))
```

```
##     Başarılı   Başarısız   Sum
##Kadın 0.4011194 0.1324627 0.5335821
##Erkek 0.3451493 0.1212687 0.4664179
##Sum   0.7462687 0.2537313 1.0000000
```

En yüksek marjinal olasılık olan 0.746, 1072 kişi içerisinde başarılı olanların oranıdır.

$$P(Başarılı) = \frac{800}{1072} = 0.746$$

-Koşullu Olasılıklar

```
prop.table(tablo,1) #Satır koşullu
```

```
##       Başarılı Başarısız

##Kadın 0.7517483 0.2482517

##Erkek 0.7400000 0.2600000
```

En yüksek koşullu olasılık (satır) olan 0.74, 500 erkek içerisindeki başarılı olma oranıdır.

$$P(Başarılı|Erkek) = \frac{370}{500} = 0.74$$

```
prop.table(tablo,2)  #sütun koşullu
```

```
##      Başarılı Başarısızolan 0.74

##Kadın   0.5375 0.5220588

##Erkek   0.4625 0.4779412
```

En yüksek koşullu olasılık (sütun) olan 0.5375, 800 başarılı içerisindeki kadın oranıdır.

$$P(Kadın|Başarılı) = \frac{430}{800} = 0.5375$$

-Bayes Olasılığı p(Kadın/Başarılı)

```
P_K<-0.533

BconK<-0.7517

P_E<-0.4664

BconE<-0.74

BconK<-0.7517

P_E<-0.4664

Bayes<- P_K*BconK/(P_K*BconK +P_E*BconE)

Bayes
```

```
[1] 0.5372222
```

3.2. Olasılık Dağılımları

Olasılık dağılımları bir rastgele değişkenin değerlerinin nasıl dağıldığını açıklar. Bu başlık altında uygulamada sık kullanılan olasılık dağılımları ele alınacaktır.

-Binom Dağılımı

İki mümkün sonucu (bu mümkün sonuçlar "başarılı", "başarısız" olarak tanımlanır) olan bir deneyde her denemenin "başarılı" olma olasılığının p olduğu durumda, n deneme sonucunda başarı sayısı olan x'in dağılımına Binom dağılımı denir. Dağılımın olasılık fonksiyonu aşağıdaki gibidir:

$$p(x) = \frac{n!}{x!(n-x)!}\ p^x(1-p)^{n-x} \qquad x = 0,1,2,\ldots,n \tag{3}$$

Burada her deneme birbirinden bağımsızdır yani bir denemenin sonucu diğer bir denemeyi etkilemez. Bu denemeler "Bernoulli Denemeleri" olarak bilinir. Dağılımın beklenen değeri ve varyansı aşağıdaki gibi hesaplanmaktadır:

$$\begin{aligned} E(x) &= np \\ VAR(x) &= np(1-p) \end{aligned} \tag{4}$$

Tesadüfen seçilen bir tüketiciye bir markanın bilinirliğinin araştırılması amacıyla marka ile ilgili "Doğru" ve "Yanlış" şıklarının yer aldığı 10 soruluk bir test uygulanmıştır. Daha önce tüketicilere uygulanan bu test sonucunda soruların doğru yanıtlanma oranı 0,70 olarak belirlenmiştir. Tüketicinin tüm soruları doğru yanıtlama olasılığı kaçtır?

$$p(x=10) = \frac{10!}{10!\,(10-10)!}\ 0{,}70^{10}\ \ (1-0{,}70)^{10-10}$$

$p(x=10) = 0{,}70^{10} = 0{,}028$

```
dbinom(10, size=10, prob=0.7)
```

```
##[1] 0.02824752
```

-Poisson Dağılımı

Bir olayın belirli bir zaman periyodunda (t), x kez meydana gelme olasılığının hesaplanmasını sağlayan bir olasılık dağılımıdır. Bu durumda olasılık fonksiyonu aşağıdaki gibidir:

$$p(x) = \frac{e^{-\lambda}\ \lambda^{x}}{x!} \qquad x = 0,1,2,..n \quad (e = 2,718) \qquad (5)$$

Burada λ , bir olayın t zaman periyodunda ortalama meydana gelme sayısıdır. Poisson dağılımına uygun bir değişkenin beklenen değeri $E(x)$ ve varyansı $V(x)$ λ 'ya eşittir.

$$E(x) = V(x) = \lambda \qquad (6)$$

Uygulamada bazen sayım değerlerinin varyansı ortalamayı aşmaktadır. Örneğin Poisson olasılık fonksiyonundan hareketle belirli sayıda kaza olması olasılığı hesaplanırken tüm haftalarda ortalama kaza sayısının sabit ve 2 olduğu varsayılsın. Uygulamada bu varsayım geçerliliğini yitirmektedir. Dolayısıyla varyans ortalama değerine eşit olmamaktadır. Bu durum "Aşırı Yayılım (Overdispersion)" olarak tanımlanmaktadır. Olasılık hesaplamalarında "Aşırı Yayılım" faktörü Poisson dağılımının yetersiz kalmasına neden olmakla birlikte Poisson dağılımı kategorik verilerin analizinde en çok temel alınan kesikli olasılık dağılımı olma özelliğini korumaktadır. Dağılımın şekli λ'ya bağlıdır. x, λ'nın üstünde değerler aldıkça olasılık değerinin azaldığı görülecektir.

Trafik kazalarının yoğun yaşandığı bir bölgede ayda ortalama 2 kazanın ölümle sonuçlandığı kabul edilsin. Belirli bir ayda ölümle

sonuçlanan kaza olmaması olasılığı hesaplansın. x kesikli tesadüfi değişkeni ölümle sonuçlanan kazaların sayısı olarak tanımlandığında bir ayda hiç kaza olmama durumu $(x = 2)$'dir. Ayda ortalama 2 kaza olduğundan $\lambda = 2$'dir. Buna göre hiç kaza meydana gelmemesi olasılığı;

$$p(x = 0) = \frac{e^{-2}\, 2^{0}}{0!} = 0{,}135\text{'dir.}$$

```
lambda <- 2
prob <- dpois(0, lambda)
prob
##[1] 0.1353353
```

-Hipergeometrik Dağılım

Hipergeometrik dağılım, sonlu bir anakütleden iadesiz yapılan örneklemde belirli sayıda başarı elde etme olasılığını verir.

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \qquad (7)$$

Burada:

N: *Anakütle birim sayı*

K: *Anakütlede aranan özellik sayısısı*

n: *Örnek birim sayısı*

k: *Örnekte aranan özellik sayısı*

Bir lojistik şirketinin deposunda 200 koli bulunsun ve bu kolilerin 50'si bozuk ürün içersin. Depodan rastgele 10 koli seçilsin. 3 tanesinin ve en fazla 3 tanesinin bozuk olma olasılığı R'da aşağıdaki gibi hesaplanır:

```
N <- 200
K <- 50
```

```
n <- 10
k <- 3


# Tam 3 bozuk koli olasılığı
dhyper(k, K, N-K, n)

##[1] 0.2567614
# En fazla 3 bozuk koli olasılığı
phyper(3, K, N-K, n)

##[1] 0.7804348
```

-Uniform Dağılımı (Kesikli)

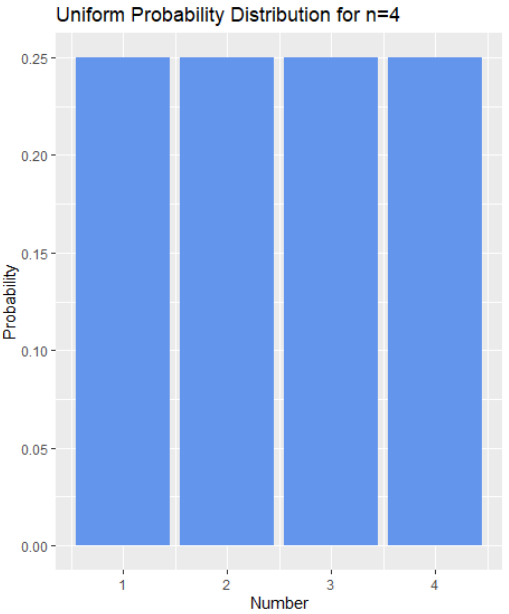
Uniform dağılımında, bir olayın 2'den fazla mümkün sonucunun her birinin gerçekleşme olasılığı eşit kabul edilmektedir. Gerçek yaşam uygulamalarına çok uygun olmayan bu dağılım ancak hilesiz para ve hilesiz zar atıldığında doğru çalışmaktadır. Sonsuz örnek büyüklünde de dağılım Uniform dağılıma yakınsar. Ancak küçük örneklerde mümkün sonuçların gerçekleşme olasılıklarının dağılımı farklılaşabilir.

$$P(X = x) = \frac{1}{n} \qquad (8)$$

4 mümkün sonucunun eşit olasılıkla gerçekleşeceği varsayılsın.

```
library(ggplot2)
x <- c(1:4)
d <- 1/4
p <- rep(d,4)
data <- data.frame(x,p)
plot <- ggplot(data=data, aes(x=x, y=p))+
  geom_bar(stat="identity", fill="cornflowerblue") +
```

```
  labs(title="Uniform Probability Distribution for
n=4", x="Number", y="Probability")
plot
```

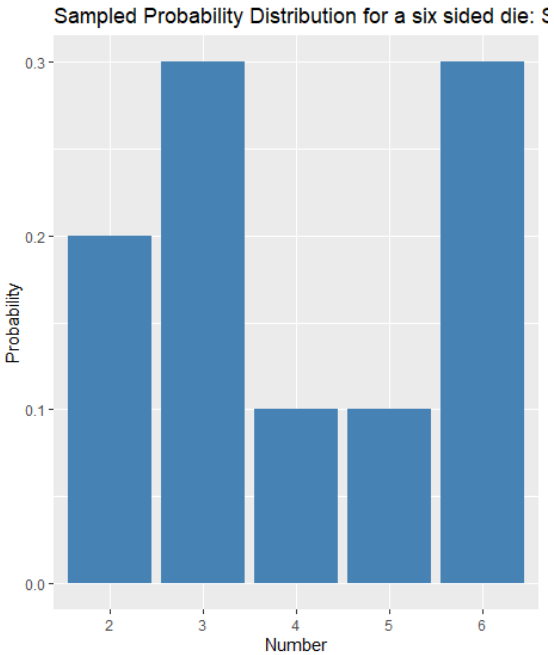


Şekil 1 Eşit Olasılıklı Dağılım

6 mümkün sonuçlu olayda olayların gerçekleşme olasılıkları eşit varsayılmasın.

```
library(ggplot2)


x <- c(1:6)
y <- sample(x,10,replace=TRUE)
t <- table(y)
data <- as.data.frame(t)
plot <- ggplot(data=data, aes(x=y, y=Freq/10))+
  geom_bar(stat="identity", fill="steelblue") +
```

```
  labs(title="Sampled Probability Distribution for
a six sided die: Sample Size = 10", x="Number",
y="Probability")

plot
```



Şekil 2 Eşit Olmayan Olasılıklı Dağılım

-Uniform Dağılımı (Sürekli)

Uniform Dağılım, belli bir aralıkta tüm değerlerin eşit sıklıkla gerçekleştiği dağılımdır.

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x < a \; veya \; x > b \end{cases} \quad (9)$$

Bir üretim hattında paketleme süresi 8 ile 12 dakika arasında değişiyor olsun ve tüm süreler eşit olasılıkla gerçekleşsin. Bir paketin 9 dakikadan kısa sürede tamamlanma olasılığı R'da aşağıdaki gibi hesaplanır:

```
min_sure <- 8
max_sure <- 12
```

9 dakikadan kısa olasılık

```
punif(9, min = min_sure, max = max_sure)
## [1] 0.25
```

-Normal Dağılım

Normal Dağılımda x değişkeninin ortalaması μ ve standart sapması σ ile tanımlanır.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad (10)$$

X değişkeni değerleri ortalaması 0 ve standart sapması 1 olacak şekilde dönüştürüldüğünde (standardize edildiğinde) standart normal değişkenin dağılım fonksiyonu;

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2} \qquad (11)$$

X değerleri Z skorlarına aşağıdaki formül ile dönüştürülür:

$$Z = \frac{X - \mu}{\sigma}$$

Standart Normal dağılım da orijinal x değerleri olan Normal dağılım gibi çan eğrisi şeklindedir ve simetriktir. Çan eğrisinin x ekseninde Z değerleri y ekseninde $f(Z)$ olasılık değerleri yer alır. Çan eğrisinin yüksekliği yani fonksiyonun tepe noktası ($x = 0$ olduğunda $1/\sqrt{(2 \times 3{,}1416)}$) 0.3989'a eşittir.

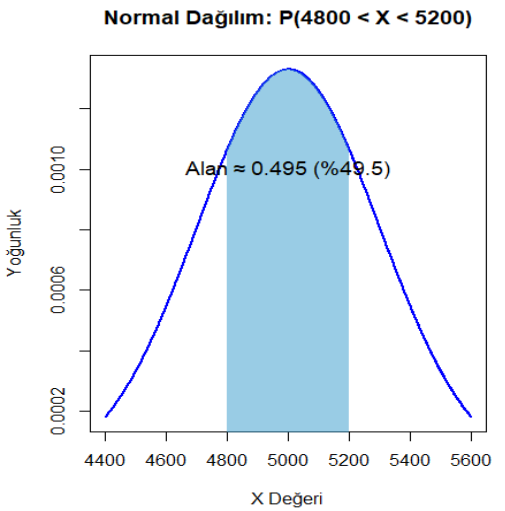
Bir firmanın aylık satış miktarı ortalama 5000 birim, standart sapması 300 birim olsun. Bir ayda satışların 4.800 ile 5.200 birim arasında olma olasılığı hesaplansın. Hesaplamalar kümülatif dağılım temel alınarak yapılacaktır.

```
mu <- 5000

sigma <- 300

pnorm(5200, mean = mu, sd = sigma) - pnorm(4800,
mean = mu, sd = sigma)
```

```
## [1] 0.4950149
```

Problemin grafik gösterimi R da aşağıdaki kodlar ile elde edilir.

```
library(ggplot2)
mu <- 5000
sigma <- 300
# Veri çerçevesi
df <- data.frame(x = seq(4400, 5600, length.out =
1000))
df$y <- dnorm(df$x, mean = mu, sd = sigma)
# Grafik
ggplot(df, aes(x, y)) +
  geom_line(color = "blue", size = 1.2) +
  geom_area(data = subset(df, x >= 4800 & x <= 5200),
            aes(x=x, y=y), fill = "skyblue", alpha =
0.5) +
  labs(title = "Normal Dağılım: P(4800 < X < 5200)",
       x = "X Değeri", y = "Yoğunluk") +
  theme_minimal()
```



Şekil. 3 Probleme İlişkin Çan Eğrisi

Bu bölümde olasılık kavramı ve temel ilkeleri açıklanmış, Normal, Binom ve Poisson gibi yaygın kullanılan dağılımlar ele alınmıştır. "R" programında bu dağılımlar için olasılık hesaplamaları yapılmış ve simülasyon örnekleri gösterilmiştir. Böylece

uygulayıcılar, dağılımların özelliklerini kavrayarak kendi veri setlerinde ilgili R kodlarını kullanarak uygulama yapabileceklerdir.

Kaynakça

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury Press.

Dalgaard, P. (2008). *Introductory statistics with R*. Springer.

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using R*. SAGE.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics*. McGraw-Hill.

Rice, J. A. (2006). *Mathematical statistics and data analysis*. Duxbury.

Ross, S. M. (2014). *Introduction to probability and statistics for engineers and scientists*. Academic Press.

Verzani, J. (2014). *Using R for introductory statistics*. CRC Press.

GÜVEN ARALIĞI VE HİPOTEZ TESTLERİ

ÇİĞDEM ARICIGİL ÇİLAN[4]

Giriş

Bir araştırma sonucu elde edilen bulgular kullanılarak anakütle ile ilgili bazı kararlar alınır. Bu kararları alabilmek için anakütlenin birer karakteristiği olan parametrelere gereksinim duyulur. Ancak, anakütlenin çok büyük olması, geniş bir coğrafi alana yayılması, zaman kısıtı, araştırma bütçesinin yetersiz olması ve birimlere uygulanan hasar verici testlerden dolayı anakütleyi gözlemlemek ve dolayısıyla parametrelere ulaşmak mümkün olamamaktadır. Bunun yerine anakütleden tesadüfi seçilen örnekleme ait istatistikleri kullanarak anakütle parametrelerini tahmin etmek veya bu parametreler hakkındaki iddiaları araştırmak (hipotez testleri) bir çıkar yol olarak görülmektedir. Anakütle parametreleri hakkındaki bu çıkarımları yapabilmek örnek istatistiklerinin dağılımlarına dayanmaktadır.

4.1. Örnekleme Dağılımı

Bir anakütleye ulaşmak her zaman mümkün değildir. Anakütle sonlu olsa bile maliyet, anakütlenin büyüklüğü, zaman

[4] Prof.Dr., İstanbul Üniversitesi İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı, Orcid: 0000-0002-7862-7028

kısıtları anakütleye erişimi engelleyebilir. Bu durumda anakütleyi temsil eden bir örnek ile ölçümler yapılabilir. Ancak örnekten hesaplanan değerler örnekten örneğe değişkenlik gösterir. Ortaya çıkan bu değişkenlik Örnekleme Dağılımı kavramını ortaya çıkarmıştır. Örnekleme dağılımı, aynı büyüklükteki çok sayıda örnekten hesaplanan bir istatistiğin (ortalama, oran, varyans) olası tüm değerlerinin dağılımıdır.

Örnekleme dağılımında tüm mümkün örnek ortalamalarının beklenen değeri anakütlenin ortalamasına eşittir.

$$E(\bar{X}) = \mu \quad (1)$$

Tüm örnek ortalamalarının standart sapması yani standart hata aşağıdaki formülle hesaplanır:

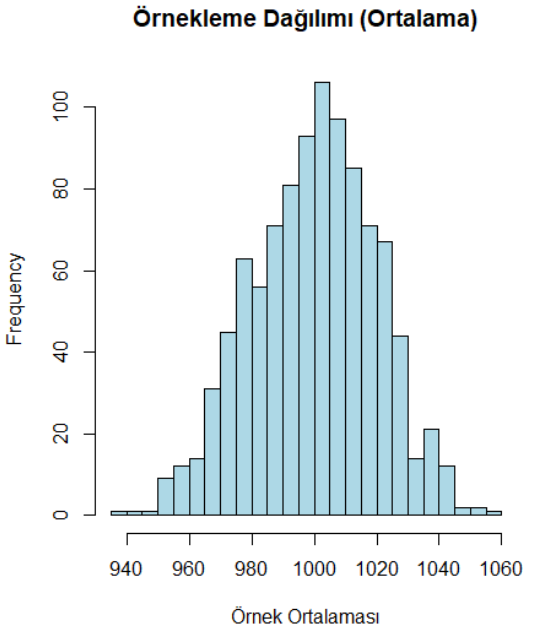
$$\sigma_{\bar{x}=\frac{\sigma}{\sqrt{n}}} \quad (2)$$

Örnek büyüklükleri 30 ve üzerinde olduğunda örnek ortalamalarının dağılımı yaklaşık olarak normal dağılır. Bu kural Merkezi Limit Teoremi olarak adlandırılır.

Bir işletme, ürettiği ampullerin ömrünü incelemek istemektedir. Anakütle ortalaması μ=1000 saat, standart sapması σ=100 saattir. 25'lik örnekler alındığında örnek ortalamalarının standart sapması yani ortalamaların standart hatası $\sigma_{\bar{x}} = 100 / \sqrt{25} = 20$ saattir. Standart hata burada 25 birimden oluşan örneklerin ortalama ömürlerinin anakütle ortalama ömrü olan 1000 saatten ortalama 20 saat saptığını göstermektedir.

```
set.seed(123)

# Anakütele verisinin oluşturulması

population <- rnorm(100000, mean = 1000, sd = 100)
```

```
n <- 25          # örneklem büyüklüğü
num_samples <- 1000  # örnek sayısı
# Örnek ortalamalarını depolamak için vektör
sample_means <- numeric(num_samples)
# Tek tek örnek alıp ortalamaları hesaplama
for (i in 1:num_samples) {
+     sample_means[i] <- mean(sample(population, n))
+ }
# Örnekleme dağılımı istatistikleri
cat("Örnekleme ortalaması:", mean(sample_means), "\n")
#Örnekleme ortalaması: 999.9258
cat("Standart hata:", sd(sample_means), "\n")
#Standart hata: 20.00926
# Histogram ile görselleştirme
hist(sample_means, breaks = 30, col = "lightblue",
+      main = "Örnekleme Dağılımı (Ortalama)",
+      xlab = "Örnek Ortalaması")
```



Şekil 1 Örnek Ortalamalarının Dağılımı

4.2. Bilinmeyen Anakütle Ortalamasının Güven Aralığı

Güven Aralığı, anakütleyi temsil ettiği varsayılan bir örneğin ortalamasından belirli bir güven düzeyinde anakütle parametrelerinin tahmin edilmesidir. Anakütle ortalaması, varyansı ve oranının tahmin edilmesinde kullanılır.

Anakütle ortalamasının tahmininde $(\hat{\mu})$ anakütle standart sapmasının (σ) bilindiği durumlarda Standart Normal Dağılım(Z) temel alınır ve $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ formülü kullanılır. Anakütle standart sapmasının bilinmediği durumlarda ise t dağılımı kullanılır. $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$

Bir e-ticaret şirketi, müşterilerin ortalama sipariş tutarını tahmin etmek istesin. 30 müşteriden oluşan örnekte günlük ortalama

sipariş tutarı 2500 TL, standart sapma 400 TL olsun. %95 güven düzeyinde anakütle ortalama sipariş tutarı R ile aşağıdaki gibi tahmin edilir:

```
xbar <- 2500

s <- 400

n <- 30

alpha <- 0.05

tcrit <- qt(1 - alpha/2, df = n-1)

se <- s / sqrt(n)

lower <- xbar - tcrit * se

upper <- xbar + tcrit * se

c(lower=lower, upper=upper)

##lower    upper

##2350.638 2649.362
```

%95 güvenle günlük ortalama sipraiş tutarının 2350.638 TL ile 2649.362 TL arasında olduğu söylenebilir.

4.3. Bilinmeyen Anakütle Oranının Güven Aralığı

Anakütledeki nitel bir değişkenin belirli bir şıkkının gerçekleşme oranını ($\widehat{P}$)tahmin etmek için kullanılan güven aralığıdır. Oran tahmininde nitel değişkenin belirli bir özelliğinin Normal Dağılım yaklaşımıyla tahmin edilebilmesi için örnek büyüklüğünün en az 100 ($n \geq 100$) olması önerilir. 100'ün altındaki

örnek büyüklüklerinde np≥5 ve n(1-p)≥5 olması gerekir. Burada p ilgilenilen özelliğin örnek oranıdır.

Bir otel müşterilerinin memnuniyet oranını tahmin etmek istesin. Anket yapılan 200 müşteriden 160'ı memnun olduğunu belirtmiş olsun.%95 güvenle tüm müşterilerin memnuniyet oranı R'da tahmin edilsin.

```
n <- 200

x<-160

phat <- x / n

alpha <- 0.05

zcrit <- qnorm(1 - alpha/2)

se <- sqrt(phat * (1 - phat) / n)

lower <- phat - zcrit * se

upper <- phat + zcrit * se

c(lower=lower, upper=upper)

##  lower     upper

   0.7445638 0.8554362
```

4.4. Hipotez Testleri

Geçerliliği bilimsel olarak araştırılabilen iddialar hipotez olarak tanımlanabilir. Bu iddialar anakütleye ilişkindir. Her iddia sıfır hipotezi ve alternatif hipotez olmak üzere iki hipotez ile yazılır. Örneğin anakütle ortalamasına ilişkin bir iddia çift taraflı ve tek taraflı hipotezler olarak aşağıdaki gibi yazılabilir.

$H_0: \mu = \mu_0$ $H_0: \mu = \mu_0$ $H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$ $H_1: \mu > \mu_0$ $H_1: \mu < \mu_0$

Diğer parametreler için de benzer hipotezler oluşturulabilir. Testin α hatası belirlenir. Burada α oranı doğru bir H_0 hipotezinin red edilmesi olasılığıdır. Her testin bir test istatistiği formülü vardır. Hesaplanan test istatistiği ile ilgili tablo değeri karşılaştırılır. Genellikle hesaplanan değerin mutlak değeri tablo değerinden büyükse H_0 hipotezi red edilir. Ya da karar için yazılım çıktılarında yer alan p değeri kullanılır.$p < \alpha$ ise H_0 hipotezi red edilir. Eşitlik içeren iddialar H_0'a, eşitlik içermeyen iddialar H_1'e yazılır. Yani araştırmacının iddiası H_0 da olabilir H_1 de olabilir. Ancak test edilen hipotez daima H_0'dır. Test sonunda ya H_0 red edilir ya da red edilemez.

Hipotez testleri başlığı altında uygulamada en sık kullanılan testlere yer verilecektir. Testlerin varsayımları dikkate alınarak hangi testin (parametrik / parametrik olmayan) uygulanacağına karar verilecektir. Testlerin uygulamasında "wooldridge" paketindeki, ceosal2 ve wage2 verileri kullanılacaktır.

```
install.packages("wooldridge")
library(wooldridge)
data("ceosal2")
dim(ceosal2)

##[1] 177  15
str(ceosal2)

##'data.frame':  177 obs. of  15 variables:

## $ salary  : int  1161 600 379 651 497 1067 945 1261 503 1
094 ...
```

```
## $ age     : int  49 43 51 55 44 64 59 63 47 64 ...

## $ college : int  1 1 1 1 1 1 1 1 1 1 ...

## $ grad    : int  1 1 1 0 1 1 0 1 1 1 ...

## $ comten  : int  9 10 9 22 8 7 35 32 4## 39 ...

## $ ceoten  : int  2 10 3 22 6 7 10 8 4 5 ...

## $ sales   : num  6200 283 169 1100 351 19000 536 4800 610
2900 ...

## $ profits : int  966 48 40 -54 28 614 24 191 7 230 ...

##$ mktval  : num  23200 1100 1100 1000 387 3900 623 2100 45
4 3900 ...

## $ lsalary : num  7.06 6.4 5.94 6.48 6.21 ...

## $ lsales  : num  8.73 5.65 5.13 7 5.86 ...

## $ lmktval : num  10.05 7 7 6.91 5.96 ...

## $ comtensq: int  81 100 81 484 64 49 1225 1024 16 1521 ..
.

## $ ceotensq: int  4 100 9 484 36 49 100 64 16 25 ...

## $ profmarg: num  15.58 16.96 23.67 -4.91 7.98 ...

##- attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

4.4.1. Tek Örneklem t Testi

177 CEO nun bilgilerinden hareketle CEO ortalama yaşının 50 olduğu iddiası test edilsin.

$H_0: \mu = 50$
$H_1: \mu \neq 50$

α=0.05

Test istatistiği:

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \quad (3)$$

Burada, $\bar{x} = 56{,}42$ (örnekteki 177 CEO'nun ortalama yaşı, μ_0 iddia edilen değer $s_{\bar{x}}$ ise ortalamanın standart hatasıdır. $s_{\bar{x}}$ örnek ortalamasının anakütle ortalamasından ortalama olarak net kadar saptığını gösterir ve $\frac{s}{\sqrt{n}}$ formülü ile hesaplanır. Anakütle standart sapması bilinmediğinden burada s örneğin sapması, n örnek büyüklüğüdür. t istatistiğini R da manuel olarak hesaplayalım.

```
m=mean(ceosal2$age)
s=sd(ceosal2$age)
pay=(m-50)
payda=s/sqrt(177)
pay/payda
```

```
##[1] 10.15655
```

Doğrudan test kodu ile çözüm şöyledir:

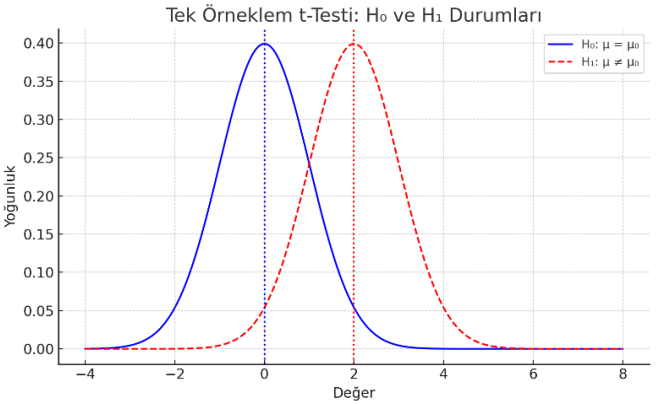
```
t_test_age <- t.test(ceosal2$age, mu=50)
t_test_age
```

```
## t = 10.157, df = 176, p-value < 2.2e-16

##alternative hypothesis: true mean is not equal to 50

##95 percent confidence interval:

## 55.18008 57.67868
```

```
##sample estimates:

##mean of x

##56.42938
```

Test sonucunda p değerinin sıfıra çok yakın bir değer (0.000) olduğu görülmektedir. Buna göre tüm CEOların ortalama yaşının 50 olduğu iddiası red edilmektedir.

Ortalama testlerinde normal dağılım varsayımı bulunmaktadır.



Şekil 2 Tek Örneklem t-Testi: H_0 ve H_1 Dağılımları

Buna göre CEO yaşlarının normal dağılıp dağılmadığı test edilecektir.

```
shapiro.test(ceosal2$age)

## Shapiro-Wilk normality test

##data:  ceosal2$age

##W = 0.98753, p-value = 0.1195  #Normal dağılım p>0.0
5 durumda sağlanır.
```

"age" değişkeni normal dağılmaktadır. Normal dağılımın sağlanamaması durumda (p<0.05) Wilcoxon Testi ((*Wilcoxon Signed-Rank Test*) uygulanır.

```
wilcox.test(ceosal2$age, mu=50)
```

Ortalama testi üst kuyruk veya alt kuyruk testi olarak da uygulanabilir. CEO'ların ortalama yaşının 50'nin üzerinde oldu iddiası üst kuyruk testi ile test edilir.

$H_0: \mu = 50$
$H_1: \mu > 50$

```
t.test(ceosal2$age, mu = 50, alternative = "greater")
```

```
##One Sample t-test

##data:  ceosal2$age

##t = 10.157, df = 176, p-value < 2.2e-16

##alternative hypothesis: true mean is greater than 50

##95 percent confidence interval:

## 55.38263      Inf

##sample estimates:

##mean of x

## 56.42938
```

H_0 red edilir. İddianın geçerli olduğu söylenebilir. Alt kuyruk testlerinde de;

$H_0: \mu = 50$
$H_1: \mu < 50$

```
t.test(ceosal2$age, mu = 50, alternative = "less")
```

```
##One Sample t-test

##data:  ceosal2$age

##t = 10.157, df = 176, p-value = 1

##alternative hypothesis: true mean is less than 50

##95 percent confidence interval:

##-Inf 57.47613

##sample estimates:

##mean of x

## 56.42938
```

Testin sonucunda H_0 red edilemez.

4.4.2. İki Bağımsız Örnek t Testi

Kolej mezunu olan ve olmayan CEO'ların maaş ortalamalarının farklı olup olmadığı test edilsin.

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

α=0.05

Burada test istatistiği varyansaların eşit olup olmamasına göre değişmektedir. Öncelikle iki anakütlenin varyanslarının eşitliği test edilsin.

$H_0: \sigma_1^2 = \sigma_2^2$
$H_0: \sigma_1^2 \neq \sigma_2^2$

```
var.test(salary ~ college, data=ceosal2)
```

```
## F test to compare two variances

##data:  salary by college

##F = 1.1652, num df = 4, denom df = 171, p-value = 0.6559

##alternative hypothesis: true ratio of variances is not equ
al to 1

##95 percent confidence interval:

##0.407209 9.663981

##sample estimates:

##ratio of variances

##            1.16521
```

Sonuca göre H_0 hiptotezi red edilemez. Varyanslar eşit kabul edilebilir. Bu durumda ortak varyans (pooled variance) formülü kullanılır.

$$S_P = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \tag{4}$$

t istatistiği;

$$t = \frac{\overline{x_1}-\overline{x_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{5}$$

Varyansların bağımsız olması durumunda standart hata;

$$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{6}$$

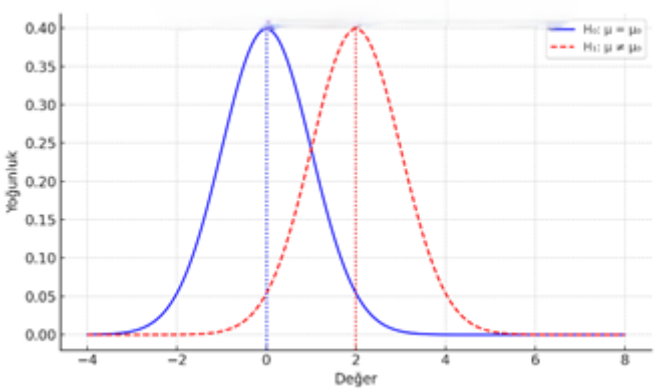
formülleri ile hesaplanır.

```
t.test(salary ~ college, data=ceosal2, var.equal=TRUE)
```

```
##Two Sample t-test

##data:  salary by college

##t = 0.88866, df = 175, p-value = 0.3754

##alternative hypothesis: true difference in means between g
roup 0 and group 1 is not equal to 0

##95 percent confidence interval:

## -289.3891  763.4518

##sample estimates:

##mean in group 0 mean in group 1

##       1096.2000         859.1686
```

p>0.05 olduğundan $H_0: \mu_1 = \mu_2$ hipotezi red edilemez. Yani CEO maaşları CEO'nun kolej mezunu olup olmamasına bağlı değildir.

İki ortalama testinde her bir grubun normal dağıldığı varsayılmaktadır.



Şekil 3 İki Bağımsız Örneklem t-Testi: H_0 ve H_1 Dağılımları

Bu varsayım aşağıdaki kodlar ile test edilebilir.

```
group1 <- ceosal2$salary[ceosal2$college == "1"]
group2 <- ceosal2$salary[ceosal2$college == "0"]
shapiro.test(group1)
##Shapiro-Wilk normality test
##data:  group1
##W = 0.77529, p-value = 5.79e-15
shapiro.test(group2)
##Shapiro-Wilk normality test

##data:  group2

##W = 0.91093, p-value = 0.4732
```

Sonuçlardan grup1'in normal dağılmadığı grup2'nin ise normal dağıldığı görülmektedir. Bu durum her iki grubun da normal dağılması gerektiği varsayımını sağlamamaktadır. Bu nedenle iki bağımsız örneklem t testi yerine parametrik olmayan alternatif bir test uygulanır.

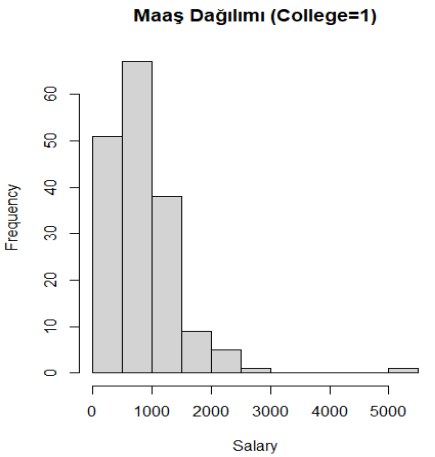
$H_0: Medyan_1 = Medyan_2$
$H_1: Medyan_1 \neq Medyan_2$

```
wilcox.test(salary ~ college, data=ceosal2)

##Wilcoxon rank sum test with continuity correction

##data:  salary by college

##W = 538, p-value = 0.3412
```

```
##alternative hypothesis: true location shift is not e
qual to 0
```

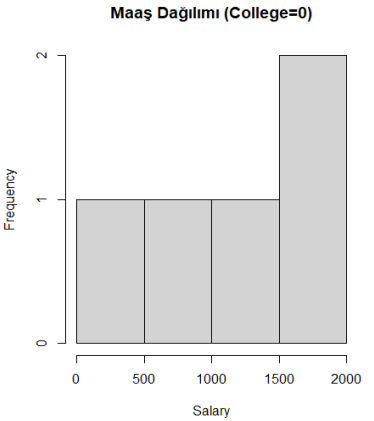
Test sonucuna göre koleji bitirme durumuna göre ortalama maaşlarda fark olmadığı söylenebilir. Aşağıda CEO'ların kolejli olma durumlarına göre histogramlar çizilmiştir.

```
hist(ceosal2$salary[ceosal2$college==1],     main='Maaş
Dağılımı (College=1)', xlab='Salary')
```



Şekil 4 Kolejli Olanların Maaş Dağılımı

```
hist(ceosal2$salary[ceosal2$college==0],     main='Maaş
Dağılımı (College=0)', xlab='Salary')
```



Şekil 5 Kolejli Olmayanların Maaş Dağılımı

İki Bağımsız Örnek t –Testi üst kuruk veya alt kuyruk hipotezlerinin test edilmesinde de kullanılmaktadır. Üst kuyruk testlerinde;

```
t.test(group1, group2, alternative = "greater", var.equal = TRUE)
```

alt kuyruk testlerinde;

```
t.test(group1, group2, alternative = "less", var.equal = TRUE)
```

kodları kullanılır. Örneğin kolej mezunu olanların ortalama maaşlarının olmayanlara göre daha fazla olduğu iddiası test edilsin.

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 > \mu_2$

```
t.test(group1, group2, alternative = "greater", var.equal = TRUE)
```

```
## Two Sample t-test

##data:  group1 and group2

##t = -0.88866, df = 175, p-value = 0.8123
```

```
##alternative hypothesis: true difference in means is
greater than 0

##95 percent confidence interval:

## -678.0971       Inf

##sample estimates:

##mean of x mean of y

## 859.1686 1096.2000
```

p>0.05 olduğundan H_0 red edilemez. Yani kolej mezunu olanların ortalama maaşlarının olmayanlardan daha fazla olduğu iddiası geçerli değildir.

4.4.3. Eşleştirilmiş Örneklerde t Testi

Tek bir örneğe müdahale yapıldığında örneğin müdahale öncesi ve müdahale sonrası arasında anlamlı bir fark olup olmadığını araştıran testtir. Bu müdahale hasta bir örnek grubuna uygulanan bir tedavi ya da bir ameliyat olabildiği gibi bir çalışan grubuna uygulanan bir eğitim de olabilir. Yapılan müdahalenin anlamlı bir fark yaratıp yaratmadığı araştırılır.

$d_i = X_{1i} - X_{2i}$ i.gözlemin müdahale öncesi ve sonrası arasındaki fark

$\bar{d}$: Farkların ortalaması

s_d: Farkların standart sapması

n: Gözlem sayısı

$H_0: \mu_d = 0$
$H_1: \mu_d \neq 0$

α=0.05

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{7}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} \tag{8}$$

R da "öncesi" ve "sonrası" verileri oluşturulsun.

```
set.seed(42)

n <- 20

before <- round(rnorm(n, mean = 50, sd = 8), 1) # eğitim öncesi

before
```

```
##[1] 61.0 45.5 52.9 55.1 53.2 49.2 62.1 49.2 66.1 49.
5 60.4 68.3 38.9 47.8## 48.9

##[16] 55.1 47.7 28.7 30.5 60.6
```

```
after  <- round(before + rnorm(n, mean = 5, sd = 5), 1)
# eğitim sonrası artış bekleniyor

after
```

```
##[1] 64.5 41.6 57.0 66.2 67.7 52.0 65.8 45.4 73.4 51.
3 67.7 76.8 49.1 49.##8 56.4

##[16] 51.5 48.8 29.4 23.4 65.8
```

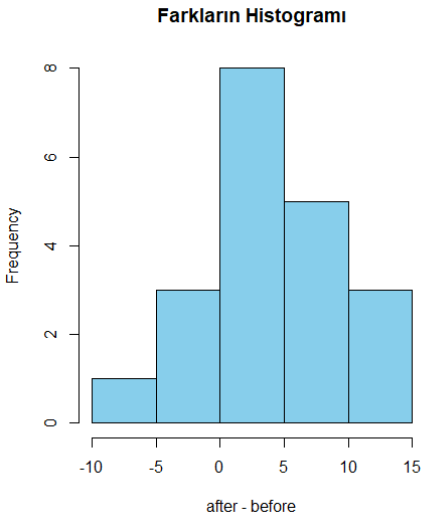
Veriler bir çalışan grubunun eğitim öncesi ve eğitim sonrası skorları olsun. Eşleştirilmiş örneklerin t testinde, "farkların (d) normal dağılması" varsayımı bulunmaktadır.

```
d <- after - before

shapiro.test(d)
```

```
##Shapiro-Wilk normality test

##data:  d

##W = 0.9797, p-value = 0.9301
```

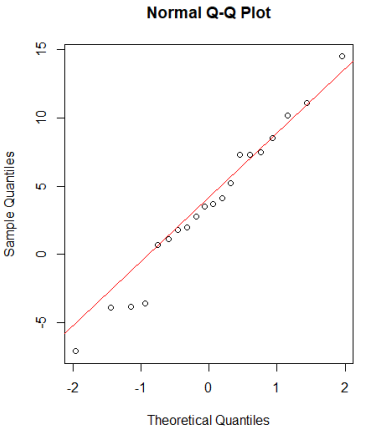
p>0.05 olduğundan farkların normal dağıldığı söylenebilir. Grafikler ile de bu sonucu desteklemek mümkündür.

```
hist(d, main="Farkların Histogramı", xlab="after - before", col="skyblue")
```



Şekil 6 Farkların (d) Dağılımı

```
qqnorm(d)

qqline(d, col="red")
```



Şekil 7 Farkların Q-Q Grafiği

Farkların normal dağılmaması durumunda Wilcoxon İşaret Sıra Testi uygulanır.

```
# Wilcoxon Signed-Rank Testi (çift kuruklu)

wilcox.test(after, before, paired = TRUE, alternative
= "two.sided")

#  after > before (tek kuyruklu) hipotezinin testi

wilcox.test(after, before, paired = TRUE, alternative
= "greater")
```

kodları kullanılır.

Eğitim öncesi ve eğitim sonrası arasında anlamlı fark olup olmadığı test edilsin.

$H_0: \mu_d = 0$
$H_1: \mu_d \neq 0$

```
t.test(after, before, paired = TRUE, alternative =
"two.sided")


##Paired t-test
```

```
##data:  after and before

##t = 2.9341, df = 19, p-value = 0.008515

##alternative hypothesis: true mean difference is not equal
to 0

##95 percent confidence interval:

## 1.044854 6.245146

##sample estimates:mean difference 3.645
```

Eğitim sonrası ile öncesi arasında anlamlı fark vardır.

Eğitim sonrası skorların eğitim öncesine göre anlamlı olarak büyük olup olmadığı (yani eğitimin etkin olup olmadığı) test edilsin.

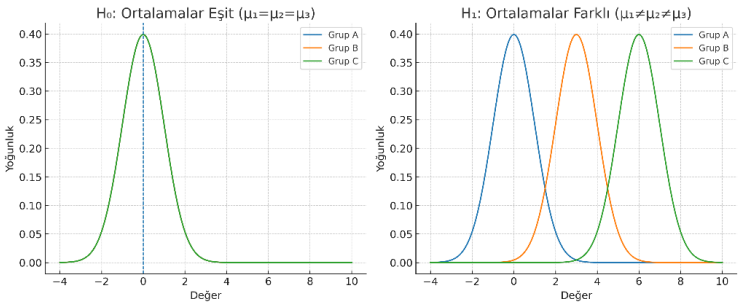
$H_0: \mu_d = 0$
$H_1: \mu_d > 0$

```
t.test(after, before, paired = TRUE, alternative =
"greater")

##Paired t-test

##data:  after and before

##t = 2.9341, df = 19, p-value = 0.004258

##alternative hypothesis: true mean difference is grea
ter than 0

##95 percent confidence interval:

## 1.496913      Inf

##sample estimates:mean difference  3.645
```

p değeri<0.05 eğitim etkindir. Eğitim sonrası skorlar anlamlı bir şekilde artmıştır.

4.4.4. Varyans Analizi(ANOVA)

Ücretlerin eğitim gruplarına göre farklılık gösterip göstermediği test edilsin. ANOVA ikiden fazla anakütlenin (grubun) aritmetik ortalamasının eşitliğinin test edilmesinde kullanılmaktadır. Testin uygulanabilmesi için "normallik" ve varyansların eşitliği" varsayımlarının sağlanması gerekmektedir.



Şekil 8 Ortalamaların Eşit Ve Eşit Olmadığı Durumlar

$H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

$H_1: \mu_i \neq \mu_j \;\; i \neq j \;\; i\, ve\, j = 1,2, \ldots k$

α=0.05

$$F = \frac{MSB}{MSW} \tag{9}$$

Toplam Kareler (SST)

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (X_{ij} - \bar{X})^2 \tag{10}$$

$SST: Kareler\ Toplamı$

$k: Grup\ Sayısı$

$X_{ij}: i.grubun\ j.gözlemi$

n_i: $i.\ grubun$ örnek büyüklüğü

$\bar{X}$: $Grup\ ayrımı\ yapmaksızın\ tüm\ değerlerin\ ortalaması$

Gruplararası Kareler (SSB)

$$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 \quad (11)$$

$\bar{X}_i$: $i.\ grubun\ ortalaması$

Gruplariçi Kareler (SSW)

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (X_{ij} - \bar{X}_i)^2 \quad (12)$$

$$MSB = \frac{SSB}{k-1} \quad (13)$$

$$MSW = \frac{SSW}{N-1} \quad (14)$$

Hesaplanan F istatistiği $F_{\alpha,k-1,n-k}$ tablo değerinden büyükse H_0 hipotezi red edilir. R'da karar verilirken p değeri temel alınır. p<α ise H_0 hipotezi red edilir.

Uygulamada "wooldridge paketindeki "wage2" verisi kullanılacaktır. CEO ücretlerinin eğitim gruplarına göre farklılaşıp farklılaşmadığı test edilsin. Öncelikle veri setinde nicel olan eğitim yılı verisi 3 gruplu nitel bir değişkene dönüştürülecektir.

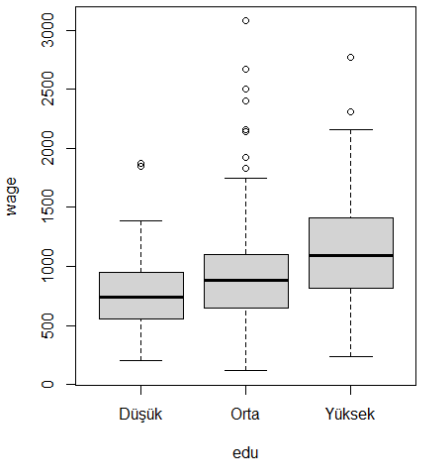
- Grup 1: Düşük eğitim (<12 yıl)
- Grup 2: Orta eğitim (12–15 yıl)
- Grup 3: Yüksek eğitim (16+ yıl)

```
library(wooldridge)
library(car)    # Levene testi için
library(carData)
```

```
# Veri setinin yüklenmesi
data("wage2")
# Eğitim düzeyini 3 kategoriye ayrılması
wage2$educ_group <- cut(wage2$educ,
                        breaks = c(0, 11, 15, Inf),   # 0-11, 1-15, 16+
                        labels  =  c("Düşük",  "Orta", "Yüksek"),
                        right = TRUE)
wage2$educ_group <- as.factor(wage2$educ_group)
edu<-wage2$educ_group
table(edu)
##Düşük   Orta Yüksek
##  88    600    247

# Tek yönlü ANOVA
anova_model <- aov(wage ~ edu, data = wage2)
summary(anova_model)
##              Df    Sum Sq   Mean Sq F value     Pr(>F)
##educ_group    2    12925398     6462699   43.09    <2e-16 ***
##Residuals     932  139790770    149990
##Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# Ortalama ücretler (gruplara göre)
aggregate(wage ~ edu, data = wage2, mean)
```

```
##educ_ group     wage

##1      Düşük  774.250

##2       Orta  908.555

##3     Yüksek 1143.368
```

```
boxplot(wage~ edu, data = wage2)
```



Şekil 9 Eğitim Durumunda Göre Kutu Diyagramları

F testi sonucuna göre ücretler (wage) eğitim gruplarına göre farklılaşmaktadır (p=0.000). $H_0: \mu_1 = \mu_2 = \mu_3$ hipotezi red edilir. Farklılığın hangi gruptan kaynaklandığını görebilmek için Tukey HSD prosedürü uygulanır.

```
TukeyHSD(anova_model)
```

```
##Tukey multiple comparisons of means

##    95% family-wise confidence level
```

```
##Fit: aov(formula = wage ~ educ_group, data = wage2)

##$educ_group

##                diff         lwr         upr           p
adj

##Orta-Düşük    134.3050    30.52675    238.0833     0.0
069089

##Yüksek-Düşük 369.1184    256.25280    481.9840     0.0
000000

##Yüksek-Orta    234.8134  166.08337    3 03.5435     0.0
000000
```

Tüm ikili grupların ortalamaları birbirinden anlamlı olarak farklıdır.

Yukarıdaki ANOVA uygulamasında varsayımlar dikkate alınmamıştır. ANOVA sonuçları varsayımların sağlanması durumunda geçerlidir. Varsayımların sağlanamaması durumunda ANOVA'nın parametrik olmayan alternatifi olan Kruskal Wallis Testi uygulanır.

-Grupların Varyanslarının Eşitliği Varsayımı

Bu varsayım aşağıda hipotez ile test edilir.

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$
$H_1: \sigma_1^2 \neq \sigma_2^2 = \sigma_3^2$

```
leveneTest(wage ~ edu, data = wage2)

##Levene's Test for Homogeneity of Variance (center =
median)

##Df F value    Pr(>F)
```

```
##group   2  11.465 1.205e-05 ***
##      932
```

```
##Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

Grupların varyanslarının eşitliği p<0.05 olduğundan sağlanmamıştır.

-Grupların Normallliği Varsayımı

```
by(wage2$wage, edu, shapiro.test)
```

##edu: Düşük

##Shapiro-Wilk normality test

##data: dd[x,]

##W = 0.91675, **p-value = 3.005e-05**

##edu: Orta

##Shapiro-Wilk normality test

##data: dd[x,]

##W = 0.91288**, p-value < 2.2e-16**

##edu: Yüksek

##Shapiro-Wilk normality test

##data: dd[x,]

##W = 0.97047, **p-value = 5.277e-05**

p değerleri 0.05'den küçük olduklarından gruplar normal dağılmamaktadır. Bu sonuçlara göre Krukal-Wallis testinin uygulanması gerekmektedir.

$H_0: Medyan_1 = Medyan_2 = Medyan_3$

$H_1: Medyan_1 \neq Medyan_2 \neq Medyan_3$

```
kruskal.test(wage ~ edu, data = wage2)
```

```
##Kruskal-Wallis rank sum test

##data:  wage by edu

##Kruskal-Wallis chi-squared = 78.565, df = 2, p-value
< 2.2e-16
```

Grupların medyanları eşit kabul edilemez. H_0 red edilir. Hangi grupların arasında fark olduğu Çoklu Karşılaştırma (Dunn Testi) ile araştırılır

```
install.packages("FSA")

library(FSA)

dunnTest(wage2$wage ~edu, method = "bonferroni")
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison

##p-values adjusted with the Bonferroni method.

##    Comparison         Z      P.unadj          P.adj

##1   Düşük - Orta -3.431489 6.002780e-04   1.800834e-03

##2 Düşük - Yüksek -7.634632 2.264664e-14   6.793993e-14

##3  Orta - Yüksek -7.355932 1.896001e-13   5.688002e-13
```

Tüm grupların medyanlarının anlamlı olarak birbirinden farklı olduğu söylenebilir.

4.4.5. Pearson Ki-Kare Testi (X^2)

Ki-Kare Testi iki nitel değişken arasında ilişki olup olmadığının araştırılmasında kullanılır.

H_0: İki nitel değişken arasında ilişki yok

H_1: İki nitel değişken arasında ilişki var

$$X^2 = \sum \frac{(f-f')^2}{f'} \quad (15)$$

f: Gözlenen Frekans (Kontenjans tablosundaki hücrelerin frekansı)

f': Beklenen Frekans

$$f' = \frac{Satır\ Toplamı \times Sütun\ Toplamı}{Genel\ Toplam} \quad (16)$$

Hesaplanan X^2 istatistiği $\chi^2_{\alpha,(r-1)(c-1)}$ tablo değerinden büyükse H_0 hipotezi red edilir. R da karar verilirken p değeri temel alınır. p<α ise H_0 hipotezi red edilir.

r: Kontenjans tablosunun satır sayısı

c: Kontenjans tablosunun sütun sayısı

Burada f kontenjans tablosunda gözlenen frekans f' beklenen frekansdır.

Uygulamada ceosal2 veri setinde CEO'ların yaşı (age) ile kolej mezunu olup olmadıkları arasındaki ilişki test edilecektir. Öncelikle nicel değişken olan "age" değişkeni "age_group" nitel değişkenine dönüştürülür.

```
library(wooldridge)
```

```
data("ceosal2", package="wooldridge")
#Kolej Mezun Durumu
ceosal2$college_fac <- factor(ceosal2$college,
                              levels = c(0, 1),
                              labels  =  c("NoCollege",
"College"))
# Yaşı Gruplama
ceosal2$age_group <- cut(
  ceosal2$age,
  breaks = c(0, 50, 60, 100),
  labels = c("<50", "50-60", "60+"),
  right = TRUE
)
tab <- table(ceosal2$age_group,ceosal2$college_fac)
tab
 ##           NoCollege     College
 ## <50           0            38
 ## 50-60         2            84
 ## 60+           3            50
```

```
#Satır ve sütun toplamlarını ekleme
addmargins(tab)
 ##          NoCollege  College Sum
 ## <50           0       38      38
 ## 50-60         2       84      86
 ## 60+           3       50      53
 ## Sum           5       172     177


# Ki-kare bağımsızlık testi
chisq.test(tab)
##Pearson's Chi-squared test
##data:  tab
##X-squared = 2.7351, df = 2, p-value = 0.2547
```

p>0.05 olduğundan CEO yaşı ile kolej mezunu olmak arasında anlamlı bir ilişki yoktur.

```
chi_res <- chisq.test(tab)
# Beklenen frekanslar
chi_res$expected
  ##      NoCollege   College
  ##<50    1.073446    36.92655
```

```
  ##50-60   2.429379    83.57062

  ##60+     1.497175     51.50282


# Gözlenen frekanslar

chi_res$observed
  ##            NoCollege   College

  ##<50            0            38

  ##50-60          2            84

  ##60+            3            50
```

"Yaş-Kolejli olma durumu" çapraz tablosunda 1. satır ve 1. sütun hücresi (1,1) sıfırdır. Yani veri setinde 50 yaşın altında kolej mezunu olmayan yoktur. Tabloda 5'den küçük gözlenen frekans olması durumunda Fisher'in Kesin Testi (Fisher's Exact Test) kullanılabilir.

```
fisher.test(tab)
##Fisher's Exact Test for Count Data

##data:  tab

##p-value = 0.4243

##alternative hypothesis: two.sided
```

Bu sonuca göre de CEO'ların yaşı ile kolejli olma durumu arasında ilişki yoktur. Yani yaş arttıkça kolejli olma durumunda anlamlı bir değişiklik olmamaktadır. H_0 red edilemez.

Bu bölümde örneklem dağılımları, güven aralıkları ve hipotez testleri ele alınmıştır. R programında tek ve iki örneklem t-testleri, Varyans Analizi (ANOVA) ve Ki-Kare testleri örneklerle uygulanmıştır. Böylece uygulamacılar, R kodları ile kendi verileri ile istatistiki çıkarımlar yapabilecek ve hipotezleri testlerini uygulayabileceklerdir.

Kaynakça

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences*. Pearson.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury Press.

Dalgaard, P. (2008). *Introductory statistics with R*. Springer.

Howell, D. C. (2013). *Statistical methods for psychology*. Cengage Learning.

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using R*. SAGE.

Montgomery, D. C., & Runger, G. C. (2018). *Applied statistics and probability for engineers*. Wiley.

Verzani, J. (2014). *Using R for introductory statistics*. CRC Press.

REGRESYON VE KORELASYON ANALİZİ

MUSTAFA CAN[5]

Giriş

Regresyon ve Korelasyon Analizleri genellikle birlikte uygulanmaktadır. Korelasyon Analizi'nde nicel değişkenler arasındaki ilişkinin doğrusallık düzeyi ölçülürken, Regresyon Analizi'nde değişkenler arasındaki neden-sonuç ilişkileri araştırılmaktadır.

5.1. Korelasyon Analizi

Doğrusal ilişkinin ölçülmesinde genellikle Pearson koreslasyon katsayısı kullanılmaktadır.

X ve Y değişkenleri için;

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2(Y_i-\bar{Y})^2}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

- X_i, Y_i: i. gözlem değerleri
- $\bar{X}$, $\bar{Y}$: değişkenlerin ortalamalarıdır.

[5] Dr.Öğretim Üyesi, İstanbul Üniversitesi İşletme Fakültesi Sayısal Yöntemler Anabilim Dalı, Orcid: 0000-0002-7786-5198

r_{XY} -1 ile +1 arasında değer alır. r_{XY}'in 0 veya 0'a yakın olması değişkenler arasında doğrusal ilişki olmadığını gösterir. r_{XY}'in +1'e yakın olması güçlü pozitif ilişkiyi, -1'e yakın olması ise negatif güçlü bir ilişkiyi temsil eder. Korelasyon Analizi sadece iki nicel değişkenin % kaç birlikte hareket ettiğini gösterir. Ancak değişkenler arasındaki nedensellik ölçülemez. R uygulanmasında, "datarium" paketinde yer alan "marketing" verisi temel alınacaktır. "marketing" datası 200 birim ve 4 değişkenden oluşmaktadır. Veri setinde "sales" bağımlı değişken, "youtube", "facebook" ve "newspaper" bağımsız değişkenlerdir. "sales" satışları, "facebook", "youtube" ve "newspaper" bu platformalara ödenen reklam bütçelerini (000 $) göstermektedir. "marketing" verisi ile; Korelasyon Analizi, Basit Regresyon Analizi ve Çoklu Regresyon Analizi uygulanacak ve bu yöntemlerin varsayımlarının sağlanıp sağlanmadığı test edilecektir. Analizlerin yapılabilmesi için gerekli olan paketler şöyledir:

```
install.packages("tidyverse")
install.packages("ggpubr")
library(tidyverse)
library(ggpubr)
theme_set(theme_pubr())
install.packages("datarium")
library(datarium)
data("marketing", package = "datarium")
data("marketing")
dim(marketing)

## [1] 200   4
str(marketing)
##data.frame':      200 obs. of  4 variables:
##$ youtube  : num  276.1 53.4 20.6 181.8 217 ...
```

```
##$ facebook : num  45.4 47.2 55.1 49.6 13 ...
##$ newspaper: num  83 54.1 83.2 70.2 70.1 ...
##$ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
data("marketing") # Veri setinin yüklenmesi
names(marketing)  # Değişken isimleri

##[1] "youtube"   "facebook"  "newspaper" "sales"

vars <- marketing[,
c("youtube","facebook","newspaper","sales")]
cor_mat <- cor(vars, use = "complete.obs")
#Korelasyon matrisi,
cor_mat <- cor(vars, use = "complete.obs")
cor_mat

##            youtube    facebook   newspaper   sales
##youtube   1.00000000 0.05480866 0.05664787 0.7822244
##facebook  0.05480866 1.00000000 0.35410375 0.5762226
##newspaper 0.05664787 0.35410375 1.00000000 0.2282990
##sales     0.78222442 0.57622257 0.22829903 1.0000000

cor(marketing$sales, marketing$youtube, use = "complet
e.obs")
#İki değişken arasındaki  korelasyon katsayısı
## [1] 0.7822244
```

Satışlar ile youtube bütçesi arasında güçlü bir pozitif doğrusal ilişkivardır.

5.2. Basit Regresyon Analizi

Bağımlı değişken olan "sales" ile "youtube bütçesi" arasında pozitif bir doğrusal ilişki (0.7822) saptanmıştı. Basit Regresyon Analizi ile bu ilişkinin nedensel bir ilişki olup olmadığı araştırılacaktır.

$$Y = a + bX + \varepsilon \quad (2)$$

Burada:

Y: Bağımlı değişken (tahmin edilmek istenen değişken)

X: Bağımsız değişken (açıklayıcı değişken)

a: Sabit terim (X=0 iken Y'nin aldığı değer, kesişim noktası)

b: Regresyon katsayısı (X'teki bir birimlik değişimin Y üzerindeki etkisi)

ε: Hata terimi (modelin açıklayamadığı kısım)

R da uygulamanın yapılabilmesi için gerekli olan paketler aşağıdakigibidir:

```
# Gerekli paketler
install.packages("datarium")
install.packages("car")
install.packages("lmtest")
library(datarium)
library(car)
library(lmtest)
# marketing datası
data("marketing")
```

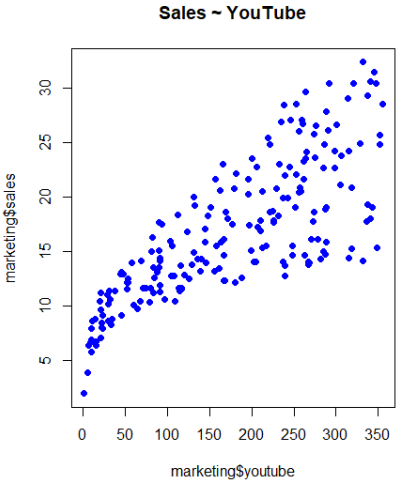
```
head(marketing)

##youtube    facebook newspaper sales

##1  276.12    45.36   83.04  26.52

##2   53.40    47.16   54.12  12.48

##3   20.64    55.08   83.16  11.16

##4  181.80    49.56   70.20  22.20

##5  216.96    12.96   70.08  15.48

##6   10.44    58.68    90.00  8.64

#Basit Regresyon Modeli

model <- lm(sales ~ youtube, data = marketing)

summary(model)

##Call:

##lm(formula = sales ~ youtube, data = marketing)

##Residuals:

##    Min       1Q   Median       3Q      Max

##-10.0632  -2.3454  -0.2295   2.4805   8.6548

##Coefficients:

##             Estimate  Std. Error   t value    Pr(>|t|)

##(Intercept) 8.439112   0.549412   15.36   <2e-16 ***

##youtube     0.047537   0.002691   17.67   <2e-16 ***
```

```
##Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

##Residual standard error: 3.91 on 198 degrees of freedom

##Multiple R-squared:  0.6119,    Adjusted R-squared:  0.609
9

##F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Modelde bağımlı değişken “sales” bağımsız değişken “youtube” dur. Modelin regresyon katsayısı olan b katsayısı (0.047537) p değerine göre anlamlıdır. b katsayısı; “youtube” harcamaları bir birim arttığında “sales” yani satışların 0.0475 birim arttığını göstermektedir. R^2 değeri olan 0.6119, “youtube” daki değişkenliğin “sales” deki değişkenliğin %61.79 oranında açıkladığını göstermektedir. F istatistiğine göre bu oran anlamlıdır. Modelin varsayımları sağlayıp sağlamadığı araştırılacaktır.

5.2.1. Doğrusallık Varsayımı

```
plot(marketing$youtube,    marketing$sales,    pch=19,
col="blue",

     main="Sales ~ YouTube")

abline(model, col="red", lwd=2)
```



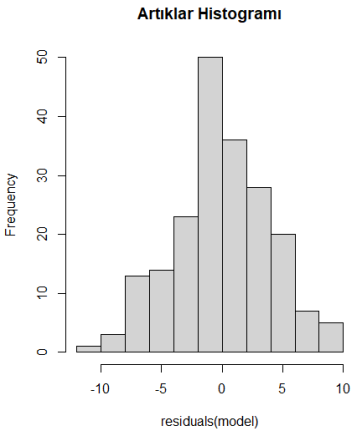
Şekil 1 Youtube Bütçesi ile Satışlar Arasındaki Serpilme Diyagramı

"sales" ile "youtube" değişkenleri arasında güçlü bir doğrusal ilişki olduğu korelasyon katsayısı (0.7822244) ile saptanmıştı. Yukarıdaki Serpilme diyagramı ile de doğrusallık görülmektedir.

5.2.2. Hataların Normal Dağılması

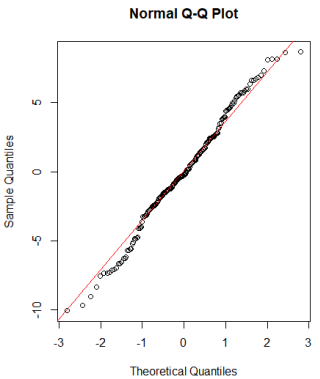
H_0: *Normal dağılıma uygun*
H_1: *Normal dağılıma uygun değil*

```
hist(residuals(model), main="Artıklar Histogramı")
```



Şekil 2 Model Artıklarının Dağılımı

```
qqnorm(residuals(model));      qqline(residuals(model),
col="red")
```



Şekil 3 Q-Q Grafiği

Bazen grafik ile değerlendirme yanıltıcı olabileceğinden normallik testlerine göre de karar verilebilmektedir.

```
shapiro.test(residuals(model))
```

```
## Shapiro-Wilk normality test
```

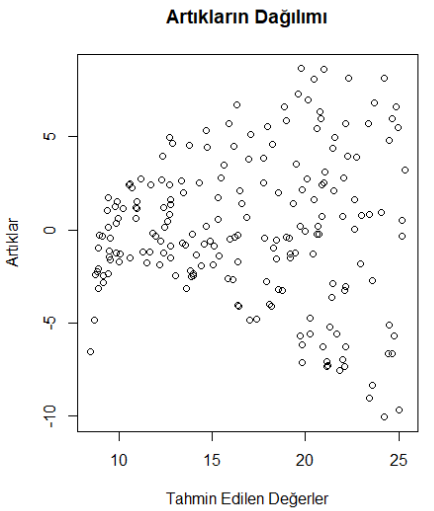
```
##data:  residuals(model)

##W = 0.99053, p-value = 0.2133
```

Teste göre modelin hataları normal dağılmaktadır. H_0 red edilememektedir.

5.2.3. Hataların Sabit Varyanslılık Varsayımı (Homoscedastisity)

H_0: *Sabit Varyanslılık sağlanır*
H_1: *Sabit Varyanslılık sağlanmaz*

```
plot(fitted(model), residuals(model),
     xlab="Tahmin Edilen Değerler", ylab="Artıklar",
     main="Artıkların Dağılımı")
abline(h=0, col="red")
```



Şekil 4 Artıkların Dağılımı

Burada artıklar sıfır etrafında tesadüfi olarak dağılmamaktadır. Sabit varyanslılık varsayımının sağlanmadığı görülmektedir. Değişen varyanslılık (Heteroskedastisite) sorunu vardır. Varsayım Breusch-Pagan Testi ile test edilecektir.

```
bptest(model)

##studentized Breusch-Pagan test

##data:  model

##BP = 48.038, df = 1, p-value = 4.18e-12
```

Hem grafikten hem de test sonucundan sabit varyanslılığın sağlanmadığı (H_0 red edilir) değişen varyanslılığın olduğu görülmektedir.

5.2.4. Hataların Bağımsızlığı

H_0: *Hatalar Bağımsızdır.*
H_1: *Hatalar Bağımsız Değildir.*

```
install.packages("lawstat")

library(lawstat)

# Artıklar(residuals)

res <- residuals(model)

runs.test(res)

##Runs Test - Two sided

##data:  res

##Standardized Runs Statistic = -0.99245, p-value = 0.321
```

```
##p>0.05 olduğundan hataların normalliği varsayımı
sağlanmaktadır.
```

Test sonucuna göre H_0hipotezi red edilemez. Artıkların tesadüfi olduğu söylenebilir.

5.3. Çoklu Regresyon Analizi

Marketing verisinde "sales" (satışlar) bağımlı değişkeni ile en düşük ilişkiye sahip değişken "newspaper" (gazete reklam bütçesi) arasındaki ilişki (0.2282), "sales" değişkenin "youtube" (youtube reklam bütçesi) ve "facebook"(facebook bütçesi) ile ilişkisine göre oldukça düşüktür. Bu nedenle çoklu regresyon modelinde "sales" bağımlı değişken "youtube" ve "facebook" değişkenleri bağımsız değişkenleri olarak yer alacaktır.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (3)$$

Burada:

$Y \rightarrow$ Bağımlı değişken (tahmin edilmek istenen)

$X_1, X_2, \ldots, X_k \rightarrow$ Bağımsız değişkenler

$\beta_0 \rightarrow$ Sabit terim (intercept)

$\beta_1, \beta_2, \ldots, \beta_k \rightarrow$ Regresyon katsayıları (her bir X'in Y üzerindeki marjinal etkisi)

$\varepsilon \rightarrow$ Hata terimi (modelin açıklayamadığı kısım)

```
#Çoklu Regresyon Modeli

model_multi <- lm(sales ~ youtube + facebook, data =
marketing)

summary(model_multi)

##Call:

##lm(formula = sales ~ youtube + facebook, data = marketing)
```

```
##Residuals:
##Min       1Q      Median       3Q      Max
-10.5572  -1.0502   0.2906   1.4049   3.3994
##Coefficients:
##            Estimate    Std. Error  t value  Pr(>|t|)
##(Intercept) 3.50532     0.35339     9.919    <2e-16 ***
##youtube      0.04575     0.00139    32.909   <2e-16 ***
##facebook     0.18799     0.00804    23.382   <2e-16 ***
##Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
##Residual standard error: 2.018 on 197 degrees of freedom
##Multiple R-squared:  0.8972,    Adjusted R-squared:  0.896
2
##F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Modelin tüm kısmi regresyon katsayılarının anlamlı olduğu görülmektedir. Modelin bağımsız değişkenler tarafından açıklama oranı R^2 oldukça yüksek (0.8972) ve anlamlıdır. Hangi bağımsız değişkenin bağımlı değişkendeki değişkenliği daha fazla açıkladığı "standardize beta" katsayıları ile değerlendirilecektir.

```
install.packages("lm.beta")
library("lm.beta")
beta<-lm.beta(model_multi)
beta
```

##Call:

##lm(formula = sales ~ youtube + facebook, data = marketing)

Standardized Coefficients::

##(Intercept) youtube facebook

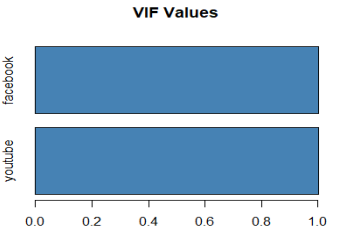
NA **0.7529042** 0.5349569

"youtube" değişkeni "facebook" değişkenine göre "sales" değişkenindeki değişkenliği daha fazla açıklamaktadır. Modelin varsayımlarını sağlayıp sağlamadığı incelenecektir.

5.3.1. Bağımsız Değişkenler Arasında İlişki Olmaması Varsayımı

Bu varsayımın sağlanmaması durumunda Çoklu Doğrusal Bağlantı(ÇBS) sorunu ortaya çıkmaktadır. Modelde ÇBS sorunun olup olmadığının ortaya çıkarılmasında VIF(Variance Inflation Factor) sıklıkla kullanılır. VIF=1 olması durumu bağımsız değişkenler arasında ilişki olmadığı anlamına gelir. VIF değerinin 1 ile 5 arasında olması "kabul edilebilir düzeyde" bir ilişkinin olduğunu,VIF değerinin 5'den büyük olması (bazı kaynaklara göre 10'dan büyük olması) ise güçlü bir ÇBS sorunu olduğunu gösterir.

```
library(carData)
library(car)
vif_values<-vif(model_multi)
vif_values
## youtube facebook
1.003013 1.003013
barplot(vif_values, main = "VIF Values", horiz = TRUE,
col = "steelblue")
```



Şekil 5 Modelin VIF Değerleri

VIF değerlerinin 1' eşit ve 5'ten (bazı kaynaklara göre 10'dan) küçük olduğu görülmektedir. Bu durumda çoklu doğrusal bağlantı sorunun olmadığı yani varsayımın sağlandığı söylenebilir.

5.3.2. Artıkların (Hataların) Normalliği Varsayımı

```
res<-resid(model)
shapiro.test(res)
##Shapiro-Wilk normality test
##data:  res
##W = 0.91804, p-value = 4.19e-09
```

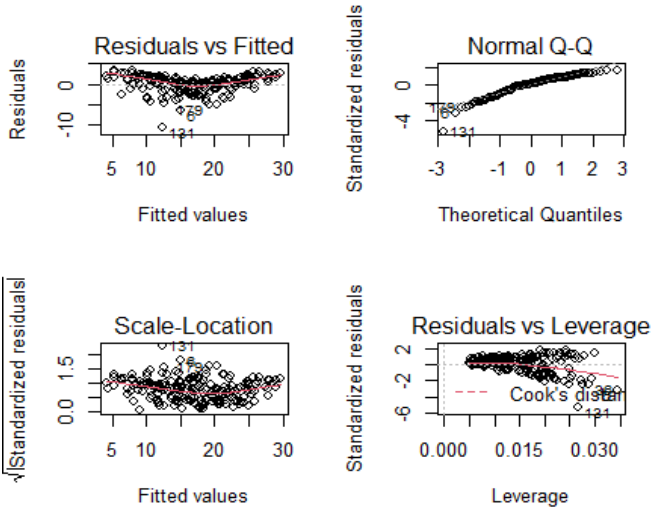
Modelin hataları normal dağılmamaktadır (p=0,000<0,05).

5.3.3. Sabit Varyanslılık Varsayımı (Homoscedastisity)

Varsayımın sağlanıp sağlanmadığı hem görseller ile hem de test ile araştırılacaktır.

```
par(mfrow=c(2,2))
plot(model)
```

,



Şekil 6 Modelin Varsayımlarına İlişkin Grafikler

```
install.packages("lmtest")
lmtest::bptest(model)
##studentized Breusch-Pagan test
##data:  model
##BP = 4.8093, df = 2, p-value = 0.0903
```

Testin anlamlılık düzeyi p=0.0903 (>0.05) olduğundan Heteroskedastisite sorununun olmadığı yani varsayımın sağlandığı söylenebilir.

Bu bölümde Korelasyon Analizi, Basit Regresyon Analizi ve Çoklu Regresyon Analizi açık kaynak kodlu bir program olan R'daki "datarium" paketindeki "marketing" verisi kullanılarak ele alınmıştır. Böylece uygulamacılar bölümde yer alan kodları doğrudan R da uygulayabilecek ve kendi verilerine kodları uyarlayabileceklerdir.

Kaynakça

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. Wiley.

Faraway, J. J. (2014). *Linear models with R*. CRC Press.

Fox, J. (2016). *Applied regression analysis and generalized linear models*. SAGE.

Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill.

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using R*. SAGE.

Torgo, L. (2017). *Data mining with R: Learning with case studies* (2nd ed.). Chapman & Hall/CRC.

Weisberg, S. (2014). *Applied linear regression*. Wiley.

FİNANSAL GÖSTERGELER İLE ÇEVRENİN SÜRDÜRÜLEBİLİRLİĞİ ARASINDAKİ İLİŞKİNİN KANONİK KORELASYON ANALİZİ İLE İNCELENMESİ :R UYGULAMASI

GÜLGÜN ERKAN[6]

Giriş

Günümüzde ülkelerin ekonomik ve finansal performansları yalnızca büyüme oranlarıyla değil, aynı zamanda çevresel sürdürülebilirlik göstergeleriyle de değerlendirilmektedir. Finansal sistemin sağlamlığı ve derinliği, yeşil yatırımların finansmanında ve sürdürülebilir kalkınma politikalarının desteklenmesinde kritik bir rol oynamaktadır. Özkaynak yapısı güçlü, takipteki kredi oranı düşük ve sermaye piyasaları gelişmiş olan ülkeler, yenilenebilir enerji yatırımlarını daha kolay fonlayabilmekte ve emisyon azaltımı hedeflerini daha etkin biçimde gerçekleştirebilmektedir (World Bank, 2021).

Nitekim Almanya gibi sermaye piyasaları derin ve finansal sistemi istikrarlı ülkelerde, yenilenebilir enerji tüketim oranı yükselmiş, kişi başına düşen emisyonlar ise göreli olarak daha düşük

[6] Dr.Öğretim Üyesi, Çanakkale Onsekiz Mart Üniversitesi Biga İktisadi ve İdari Bilimler Fakültesi, İşletme Bölümü, Orcid: 0000-0001-7108-131X

seyretmiştir. Buna karşılık finansal sistemi daha kırılgan olan ülkelerde çevresel göstergeler olumsuz eğilimler gösterebilmektedir. Bu nedenle, ülkelerin finansal göstergeleri ile çevresel sürdürülebilirlik arasındaki ilişkiyi ortaya koymak, sadece ekonomik büyüme açısından değil, aynı zamanda yeşil dönüşümün başarısı için de stratejik öneme sahiptir (IPCC, 2022).

Bu bölümde, ülkelerin finansal göstergeleri (sermaye/aktif oranı, takipteki krediler, özel sektör kredileri/GSYH, piyasa değeri/GSYH) ile çevresel sürdürülebilirlik göstergeleri (CO_2 emisyonu, yenilenebilir enerji payı, korunan alan, metan emisyonu) arasındaki ilişki Kanonik Korelasyon Analizi ile incelenecektir. Çalışmada kullanılan veri, R'daki WDI paketi üzerinden Dünya Bankası'nın World Development Indicators (WDI) veri tabanından elde edilmiştir. Böylece hem teorik çerçeve hem de uygulamalı istatistiki bulgular birlikte sunulacaktır.

6.1. Kanonik Korelasyon Analizi

Kanonik Korelasyon Analizi, iki değişkenler kümesi arasındaki çok boyutlu ilişkiyi ölçmeyi amaçlar. Her bir değişkenler kümesinin doğrusal bileşimleri gizli değişkenleri oluşturur. Amaç iki gizli değişken arasındaki en yüksek korelasyonu sağlayacak doğrusal bileşenleri elde etmektir.

X ve Y değişken setlerinden oluşsun.

$$X = (X1, X2, \dots, Xp)' \text{ve } Y = (Y1, Y2, \dots, Yq)' \quad (1)$$

Değişkenlerin doğrusal bileşimleri şöyledir;

$$U = a'X \text{ ve } V = b'Y \quad (2)$$

Burada $\boldsymbol{a}$, X değişkenleri kümesinin katsayı vektörü $\boldsymbol{b}$, Y değişkenlerinin katsayı vektörüdür.

$$U = a'X = a1X1 + a2X2 + \cdots + apXp \quad (3)$$

$$V = b'Y = b1Y1 + b2Y2 + \cdots + bqYq \quad (4)$$

$$\rho = Corr(U,V) = \frac{a\prime \sum XYb}{\sqrt{a\prime \sum XXa\ b\prime \sum YY\ b}} \quad (5)$$

Burada ρ U ile V arasındaki en büyük kanonik korelasyon katsayılarıdır.

ΣXX : X değişkenlerinin kovaryans matrisi (p×p)

ΣYY : Y değişkenlerinin kovaryans matrisi (q×q)

ΣXY : X ile Y arasındaki kovaryans matrisi (p×q)

p: X kümesindeki değişken sayısı

q: Y kümesindeki değişken sayısı

6.2. R ile Uygulama

2021 yılı pandemi etkisiyle ekonomik ve sosyal göstergelerde ciddi değişimlerin yaşandığı yıldır. Bu nedenle uygulamada R'ın "WDI" paketinde yer alan 2021 yılı için 12 ülkenin 4 finansal göstergesi (X seti) ve 4 sürdürülebilirlik göstergesi ($Y\ seti$) kullanılacaktır. Covid döneminin hemen sonrası dönemdir. Analize dahil edilecek ülkeler: 1-Türkiye (TUR),2-Polonya (POL),3- Almanya (DEU),4- Fransa (FRA),5-İtalya (ITA),6-İspanya (ESP),7-Birleşik Krallık (GBR),8- Hollanda (NLD), 9-İsveç (SWE) ,10-ABD (USA),11-Çekya (CZE),12-Avusturya (AUT)

Verideki kısaltmaların temsil ettiği değişkenler aşağıdaki gibidir:

Tablo 1 Değişken Listesi

WDI Kodu	Verideki İsim	Açıklama
FB.BNK.CAPA.ZS	bank_capital_assets	Banka Sermayesi / Aktifler Oranı
FB.AST.NPER.ZS	npl_ratio	Takipteki Krediler Oranı
FS.AST.PRVT.GD.ZS	credit_private_gdp	Özel Sektör Kredileri / GSYH
CM.MKT.LCAP.GD.ZS	mktcap_gdp	Halka Açık Şirketlerin Piyasa Değeri / GSYH
EN.ATM.CO2E.PC	co2_pc	Kişi Başına CO_2 Emisyonu
EG.FEC.RNEW.ZS	renew_share	Yenilenebilir Enerji Tüketim Payı
ER.PTD.TOTL.ZS	protected_land	Korunan Kara Alanı Oranı
EN.ATM.METH.KT.CE	methane_kt	Metan Emisyonu (bin ton CO_2 eşdeğeri)

R uygulaması adımları ve çıktıları aşağıdaki gibidir:

```
install.packages("CCA")

install.packages("CCP")

install.packages(WDI)

library(CCA)

library(CCP)

library(WDI)

#Verinin tanımlanması

veri<-WDI(country =
c("TUR","POL","DEU","FRA","ITA","ESP",

                "GBR","NLD","SWE","USA","CZE","AUT"),

    indicator = c("FB.BNK.CAPA.ZS",   #
bank_capital_assets

                  "FB.AST.NPER.ZS",   # npl_ratio

                  "FS.AST.PRVT.GD.ZS",#
credit_private_gdp

                  "CM.MKT.LCAP.GD.ZS",# mktcap_gdp
```

```
                    "EN.ATM.CO2E.PC",   # co2_pc
                    "EG.FEC.RNEW.ZS",   # renew_share
                    "ER.PTD.TOTL.ZS",   # protected_land
                    "EN.ATM.METH.KT.CE" # methane_kt ),
    start = 2021, end = 2021)
veri
df <- data.frame(
  country =
c("Turkey","Poland","Germany","France","Italy","Spain"
,
              "United
Kingdom","Netherlands","Sweden","United States","Czech
Republic","Austria"),
  bank_capital_assets =
c(12.50,10.80,7.50,7.90,8.60,9.10,6.80,7.20,8.90,7.00,
9.80,8.10),
  npl_ratio           =
c(3.10,3.60,1.20,1.80,3.30,2.90,1.50,1.30,0.80,1.00,2.
20,1.70),
  credit_private_gdp  =
c(76,62,108,103,85,94,148,132,139,216,86,102),
  mktcap_gdp          =
c(38,34,65,92,46,59,128,120,112,183,43,54),
  co2_pc              =
c(4.50,8.10,8.10,4.60,5.70,5.10,4.70,8.30,3.40,14.90,9
.70,7.40),
  renew_share         =
c(16,17,17,14,18,21,18,14,37,12,17,33),
  protected_land      =
c(5.30,39.80,37.80,32.00,21.60,27.20,28.70,26.50,14.50
,14.00,22.00,28.30),
  methane_kt          =
c(430,420,820,540,470,460,370,220,160,720,190,180))
```

```
 print(df)   #Bu komut çalıştırıldığında veri setine
ulaşılacaktır

# X ve Y kümeleri

Xvars <-
c("bank_capital_assets","npl_ratio","credit_private_gd
p","mktcap_gdp")

Yvars <-
c("co2_pc","renew_share","protected_land","methane_kt"
)

X <- scale(as.matrix(df[, Xvars]))   # Finansal set

Y <- scale(as.matrix(df[, Yvars]))   #
Sürdürülebilirlik set

# Kanonik Korelasyon Modeli

library(CCA); library(CCP)

fit <- cc(X, Y)

rho  <- fit$cor              # kanonik korelasyonlar

rho

## [1] 0.95537106 0.65891065 0.50068107
0.07367527
```

İlk kanonik korelasyon katsayısı (0.9553) yorumlandığında ülkelerin finansal göstergeleri ile çevresel sürdürülebilirlik göstergeleri arasında çok güçlü bir pozitif doğrusal ilişki olduğu söylenebilir.

```
#Kanonik Korelasyon Katsayılarının Anlamlılığı (Wilks
Lambda, Bartlett yaklaşımı)

n <- nrow(X); p <- ncol(X); q <- ncol(Y)

wilks <- p.asym(rho, n, p, q, tstat = "Wilks")

##Wilks' Lambda, using F-approximation (Rao's F):

##         stat       approx    df1    df2    p.value
```

```
##1to 4:  0.0367993 1.56495908  16 12.85783 0.2113694
##2to 4:  0.4216905 0.58296500   9 12.31929 0.7884518
##3to 4:  0.7452511 0.47512102   4 12.00000 0.7534539
##4to 4:  0.9945720 0.03820369   1  7.00000 0.8505872
```

"1 to 4" ifadesi şunu gösteriyor: 1'den 4'e kadar olan bütün kanonik korelasyon katsayıları birlikte test ediliyor. "2 to 4" sadece ρ_2, ρ_3, ρ_4 katsayılarının test edildiğini, "3 to 4" sadece ρ_3 ve ρ_4 test edildiğini "4 to 4" ise sadece ρ_4 katsayısının test edildiğini gösteriyor. Sonuçlara göre hiçbir test sonucu anlamlı değil. Anlamsız sonuçların nedeni sadece 12 ülke ve tek yıl (2021) ile çalışılmış olması olabilir. Daha uzun dönem ortalamaları veya daha geniş ülke seti kullanıldığında yüksek kanonik korelasyon katsayılarının anlamlı çıkacağı düşünülmektedir.

```
#Değişkenlerin Özvektör katsayıları
xcoef <- fit$xcoef   # X kümesi için
ycoef <- fit$ycoef   # Y kümesi için
round(xcoef, 3)
```

```
##                      [,1]   [,2]   [,3]   [,4]
##bank_capital_assets  0.592  1.018  0.957  0.097
##npl_ratio           -0.069  0.469 -1.455 -0.725
##credit_private_gdp   3.210 -1.329 -0.718 -1.973
##mktcap_gdp          -2.316  2.401 -0.133  2.239
```

Bu katsayılara göre 1. kanonik vektörde mutlak değeri en yüksek olan değişken (3.210) "credit_private_gdp" (Özel sektör kredilerinin GSYH'ye oranı) değişkenidir. Bu değişken finansal sistemin derinliğini ve kredi bağımlılığını en iyi temsil eden değişkendir.

```
round(ycoef, 3)
```

```
##                [,1]   [,2]   [,3]   [,4]
##co2_pc          0.684 -0.442 -0.352  0.674
##renew_share     0.613 -0.911  0.566 -0.293
##protected_land -0.668 -0.702 -0.247  0.208
##methane_kt      0.438 -0.347 -0.244 -1.078
```

Çevre değişkenleri arasında 1. Kanonik vektörde mutlak olarak en yüksek değişkenler sırasıyla "co2_pc" (0.684) ve protected_land (-0.668)'dir. Yani çevresel boyutta hem karbondioksit emisyonu (çevresel baskı) hem de korunan alan oranı (çevresel koruma politikaları) öne çıkmaktadır.

```
# Ülkelerin Kanonik Katsayıları
U <- X %*% fit$xcoef   # X tarafı kanonik skorlar
V <- Y %*% fit$ycoef   # Y tarafı kanonik skorlar
U
```

```
##        [,1]        [,2]        [,3]         [,4]
##[1,]  0.5983343  1.74126798  1.30944112 -0.939704702
##[2,] -0.9137118  1.20052396 -0.15259625 -0.941127476
##[3,]  0.1023144 -1.81162776  0.71191280 -0.008554034
##[4,] -1.5423516  0.28556124  0.04287457  1.111923817
##[5,] -0.4863847 -0.38184147 -1.37789107 -1.352216553
##[6,] -0.2397665   0.11328492 -0.68513555 -0.819920461
```

```
##[7,] -0.2339442  -0.09002699 -1.00936846  0.882987877
 ##[8,] -0.9152300   0.14972475 -0.18185559  1.428259658
 ##[9,]  0.6580114   0.29018585  1.43576038  1.179897891
 ##[10,] 2.3680610  0.47229125  -1.47686923  0.707617833
 ##[11,] 0.2395931  -0.38151155  0.95034806 -0.650531873
 ##[12,] 0.3650746  -1.58783218  0.43337922 -0.598631975
```

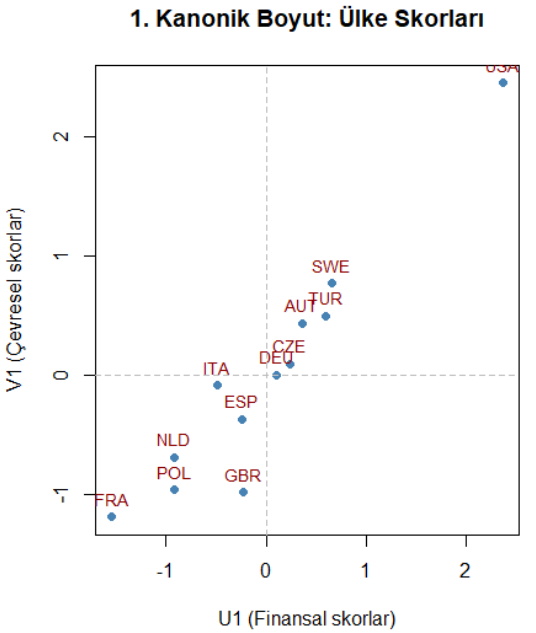
v

```
##         [,1]         [,2]         [,3]       [,4]
## [1,]  0.501906041  2.11711727  0.48846177 -0.8918498
## [2,] -0.961626267 -0.91021427 -0.67862033  0.6082059
## [3,]  0.002657934 -1.42802695 -1.09209475 -1.4791102
## [4,] -1.190678356  0.28756732 -0.45635484 -0.8010664
## [5,] -0.082618419  0.50137119  0.05427246 -0.5773943
## [6,] -0.366477290 -0.14845349  0.21674593 -0.6528419
## [7,] -0.980885900  0.30800515  0.10654309 -0.1320217
## [8,] -0.686603251  0.68130626 -0.36359837  1.5119727
## [9,]  0.773614502 -0.43266245  2.25081783 -0.3585672
##[10,]  2.455624074  0.04888453 -1.51873635  0.1821231
##[11,]  0.095464048  0.49257030 -0.15223079  1.7562109
```

```
##[12,]  0.439622884 -1.51746487  1.14479436  0.8343390
```

1.kanonik vektörlerin (U,V) aynı yönde olan ((+ +) veya (- -)) katsayıları temel alındığında **USA(10), SWE(9), TUR(1)** → yüksek finansal gelişmişlik ve yüksek çevresel skorlarla öne çıkmaktadır. Burada USA(10), (2.368, 2.45) değerleri ile finansal gelişmişlik ile çevresel sürdürülebilirlik ilişkisi en yüksek ülkedir.

```
#Grafikle Gösterim
U1 <- U[,1]
V1 <- V[,1]
countries <- c("TUR","POL","DEU","FRA","ITA","ESP",
               "GBR","NLD","SWE","USA","CZE","AUT")


plot(U1, V1,
     xlab = "U1 (Finansal skorlar)",
     ylab = "V1 (Çevresel skorlar)",
     main = "1. Kanonik Boyut: Ülke Skorları",
     pch = 19, col = "steelblue")
text(U1, V1, labels = countries, pos = 3, cex = 0.9,
col = "darkred")
abline(h = 0, v = 0, lty = 2, col = "gray")
```



Şekil 1 Ülkelerin Finansal ve Çevresel Sürdürülebilirlik Skorları

```
#Kanonik yükler (Orijinal değişkenlerin kendi kümeleri
ile ilişkisi)

loadings_X <- cor(X, U)   # p x r

loadings_Y <- cor(Y, V)   # q x r

loadings_X

##                        [   ,1]      [,2]        [,3]
[,4]

##bank_capital_assets -0.03298317  0.58605194   0.5055
630  -0.6323459
```

```
##npl_ratio            -0.36611231  0.41008448  -0.1265
708  -0.8256951

##credit_private_gdp   0.61861548 -0.05969366  -0.3877
065   0.6807608

##mktcap_gdp           0.42800581  0.05487317  -0.4044
219   0.8063764
```

Finansal tarafta özel sektör kredilerinin GSYH'ye oranı (credit_private_gdp) 1. kanonik boyutu en iyi temsil eden değişkendir. Yani finansal derinlik arttıkça U_1 de artıyor.

```
loadings_Y
```

```
##                  [,1]         [,2]       [,3]       [
,4]

##co2_pc           0.5694234 -0.1752000 -0.6721299  0.
439662803

##renew_share      0.1525385 -0.4516620  0.8790125  0.
008392406

##protected_land  -0.6497966 -0.6699384 -0.3576150  0.
032532523

##methane_kt       0.1892416 -0.1164300 -0.7281109 -0.
648449067
```

Kişi başına CO_2 emisyonları pozitif, korunan alan oranı negatif yüke sahiptir. Yani çevresel skor yükseldikçe emisyonlar artmakta, ama korunan alanlar azalmaktadır.

Ekonomilerde finansal derinlik (kredi hacmi) arttıkça çevre üzerinde baskının arttığı (CO_2 yükseliyor), buna karşılık çevre koruma kapasitesinin (korunan alan) azaldığı görülmektedir.

```
#Çapraz yükler

cross_X_on_V <- cor(X, V)

##                                  [,1]          [,2]
[,3]          [,4]

##bank_capital_assets -0.03151117   0.38615587    0.25
312583  -0.0465882
##npl_ratio           -0.34977310   0.27020903   -0.06
337161  -0.06083331
##credit_private_gdp   0.59100733  -0.03933279   -0.19
411731   0.05015523
##mktcap_gdp           0.40890437   0.03615652   -0.20
248640   0.05941000
```

Özel sektör kredilerinin GSYH'ye oranı (credit_private_gdp) değişkeni 0.591 ağırlık katsayısı ile "çevrenin sürdürülebilirliği" kümesi ile en yüksek ilişkiye sahiptir.

```
cross_Y_on_U <- cor(Y, U)

##                      [,1]          [,2]          [,3]
[,4]

##co2_pc             0.5440106  -0.11544112 -0.3365227  0
.0323922764
```

```
##renew_share      0.1457309  -0.29760491  0.4401049  0
.0006183128

##protected_land -0.6207969  -0.44142956 -0.1790511  0
.0023968424

##methane_kt       0.1807960  -0.07671695 -0.3645513 -0
.0477746611
```

Korunan alan (protected_land **)** ve karbon emisyonları (co2_pc) değişkenleri sırasıyla -0.6207 ve 0.544 ağırlıkları ile "finansal göstergeler" kümesi ile en yüksek ilişkiye sahiptirler.

Sonuç ve Öneriler

Bu çalışmada 2021 yılına ait veriler kullanılarak NATO ülkelerinin (Avusturya hariç) finansal ve çevresel göstergeleri arasındaki ilişkiler R programlama dili aracılığıyla, CCA paketi kullanılarak kanonik korelasyon analizi (CCA) yöntemiyle incelenmiştir. Bulgular, ilk kanonik boyutun en güçlü ve anlamlı ilişkiyi yansıttığını ortaya koymuştur. Finansal tarafta en baskın değişken özel sektör kredilerinin GSYH'ye oranı olurken, çevresel tarafta bu boyut CO_2 emisyonları (pozitif yük) ve korunan alanların oranı (negatif yük) ile karakterize edilmiştir. Bu durum, 2021 itibarıyla finansal derinliğin artmasının çevresel baskıları artırdığını, yani emisyonların yükseldiğini ve korunan alanların görece azaldığını göstermektedir. Ülke skorları incelendiğinde, ABD, İsveç ve Türkiye'nin finansal–çevresel boyutta yüksek değerler sergilediği, Fransa, Polonya ve Hollanda'nın düşük düzeyde kaldığı, Almanya, İtalya, İspanya, Çekya ve Avusturya'nın ise orta seviyede dengeli bir görünüm sunduğu görülmektedir.

2021 sonrası gelişmeler dikkate alındığında bu bulgular, daha güncel politika çerçeveleriyle birlikte değerlendirilmelidir. Özellikle

Avrupa Birliği'nin Sürdürülebilir Finans Taksonomisi ve Yeşil Mutabakat kapsamında 2022'den itibaren yürürlüğe giren düzenlemeler, finansal yatırımların çevresel etkilerinin daha sıkı denetlenmesine yol açmıştır. Bu bağlamda kredi/GSYH oranının çevresel baskılarla pozitif ilişkisinin azaltılması için, yeşil finansal araçların (yeşil tahviller, sürdürülebilir krediler) payının artırılması ve bankacılık sermaye yeterliliği düzenlemelerinde çevresel risklerin entegre edilm**esi** kritik hale gelmiştir.

Sonuç olarak, 2021 verilerine göre finansal gelişme ve çevresel baskı arasında güçlü bir pozitif bağ bulunmuştur. Ancak 2021 sonrası dönemde politika ortamının hızla evrildiği göz önüne alındığında, finansal büyümenin sürdürülebilir kalkınma hedefleriyle uyumlu hale getirilmesi teknik düzeyde mümkündür.

Bu nedenle ülkeler arasında farklı finansal ve çevresel profiller göz önüne alınarak, yüksek emisyon profiline sahip finansal açıdan gelişmiş ülkelerde çevresel riskleri fiyatlayan düzenlemeler, daha düşük finansal kapasiteye sahip ülkelerde ise yeşil finansın erişilebilirliğini artıran mekanizmalar ön plana çıkmalıdır. Böylece finansal derinleşme, çevresel sürdürülebilirlik ile çelişmek yerine onu destekleyen bir araç haline getirilebilir.

Kaynakça

Arel-Bundock, V., Enevoldsen, N., & Yetman, C. (2022). *WDI: World Development Indicators (World Bank)*. R package version 2.7.6

Härdle, W. K., & Simar, L. (2015). *Applied Multivariate Statistical Analysis*. Springer.

IPCC. (2022). *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. Wiley.

Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, 84(1), 37–48.

World Bank. (2021). *World Development Indicators (WDI)*. Washington, DC: The World Bank.


R ile
Uygulamalı
İstatistik
BİDGE