# Temel Kavramlar

## Giriş

**Model,** gerçek dünyadaki bir olgunun bir anlatımıdır, bir tasviridir. Bu anlatma işine modelleme ve anlatımın kendisine de model denir.

**Simülasyon (Benzetim)**, model üzerinde olgunun irdelenmesidir, "model üzerinde deney yapmaktır". Simülasyon yoluyla elde edilen verilere sanal veri denir. Bu veriler, olgunun gerçeğinden elde edilen verilere (gözlemlere) benzemektedir.

## İstatistik Nedir?

İstatistik, rasgelelik içeren olaylar, süreçler, sistemler hakkında modeller kurmada, gözlemlere dayanarak bu modellerin geçerliliğini sınamada ve bu modellerden sonuç çıkarmada gerekli bazı bilgi ve yöntemleri sağlayan bir bilim dalıdır. Diğer bir ifade ile İstatistik, verilerin toplanması, düzenlenmesi, analiz edilmesi ve yorumlanmasını sağlayan bir disiplindir. Ayrıca istatistik, geleceğe yönelik tahminlerde bulunarak karanlığa ışık tutar ve araştırmacılara yol gösterir. Bu anlamda birçok bilim dalına yardımcı olur.

- ***Rasgelelik (kesin olmama);*** gelişigüzellik, tesadüfilik, rassallık, sözcüklerinin karşılığıdır. Olasılık bu rasgeleliğin ölçülmesi işine yarar. İstatistik, rasgelellik ortamında hesap yapabilmemizi sağlamaktadır (tavla zarının atılarak üste gelen nokta sayısının gözlenmesi, bir paranın atılması ve gelen yüzeyin gözlenmesi vb.).

  **Soru**: Sayıların üzerinde 1,2,3,4,5,6 sayılarından 25 tane ("kafadan") yazınız. Daha sonra k= round(1+5*rand(1,25)) komutu ile MATLAB programında 25 tane rasgele sayı üretiniz ve sonuçları karşılaştırınız.

İstatistiğin kullanım alanları;

- Tıp+Biyoloji=Biyoistatistik
- Ekonomi=Ekonometri
- Psikoloji=Psikometri
- Sosyoloji=Sosyometri
- Tarih=Kliometri

- Nüfus istatistikleri, çevre istatistikleri, spor istatistikleri, milli eğitim istatistikleri gibi daha birçok alanda kullanılır.

İstatistiğin amacı betimsel amaçlı istatistik ve çıkarımsal amaçlı istatistik olmak üzere ikiye ayrılır. Betimsel amaçlı istatistik, kitledeki tüm birimlerden ilgili değişken ya da değişkenler bakımından veri toplandığında bu verileri kullanarak kitleyi özetlemeyi (betimlemeyi) amaçlar. Bu amaçla çokluk dağılımları oluşturulur, grafikler çizilir ya da parametreler hesaplanır.

Çıkarımsal istatistikte ise, kitlelerden rasgele seçilen örneklerden toplanan verileri kullanarak kitle parametrelerini tahmin etmeyi ya da parametrelerle ilgili olan iddiaların doğru olup olmadığının araştırılmasını amaçlar. Tahmin ve hipotez testleri çıkarımsal istatistiğin temel konularıdır.

Verilerin toplanması bir istatistiksel araştırmanın en önemli aşamalarından birisidir.

**Temel Kavramlar**

**Kitle (Evren, Popülasyon):** Belirli bir özelliğe sahip bireylerin veya birimlerin tümünün oluşturduğu topluluğa **kitle** denir. Örneğin; fen fakültesinde okuyan öğrenciler, diş dolgusu yaptıranların oluşturduğu topluluk, eczacılık fakültesinden mezun olan öğrenciler vb. birer kitledir.

**Örneklem:** Örnekleme yöntemlerinden yararlanılarak bir kitleden seçilen, aynı özellikleri taşıyan ve kitleyi temsil edebilecek nitelikteki ve nicelikteki bireylerin oluşturduğu topluluğa **örneklem** denir.

**Parametre:** Kitleyi tanımlayan sayısal değerlere **parametre** denir. Kitle ortalaması, kitle varyansı, kitle korelasyon katsayısı birer parametredir.

**Veri:** Rasgelelik içeren olgulardan elde edilen ölçüm (gözlem) değerlerine **veri** denir.

**Değişken:** Birimlerin farklı değerler alabildikleri nitelik ya da niceliklerine **değişken** denir.

**Rasgele Değişken:** Değişkenin aldığı değer rasgele bir gözlemin veya bir deneyin sonucu ise rasgele değişken adını alır.

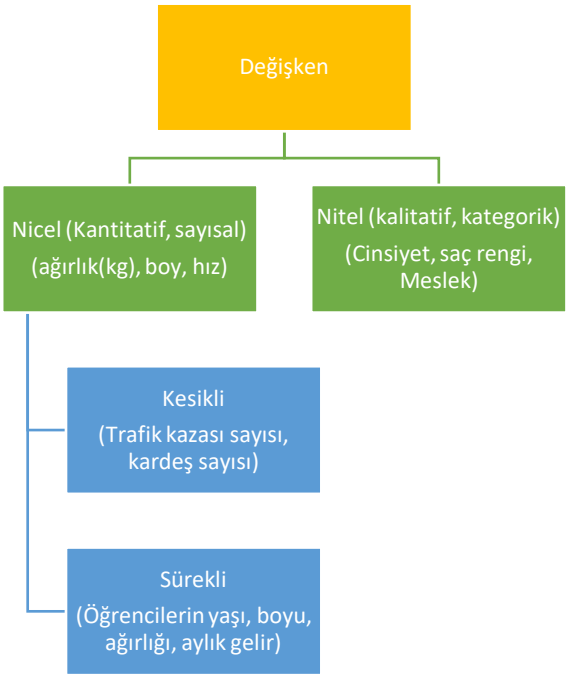
**Ölçme:** Araştırma konusu ile ilgili sayısal bilgiler elde etme işidir.

**Ölçek:** Sayısal verileri elde etmek için kullanılan araç ve gereçlerdir.

**Değişken Türleri ve Ölçme Düzeyleri**

Bir değişken sayısal değerlerle ölçülebiliyorsa, bu değişkene **nicel değişken** denir. Örneğin ağırlık, boy uzunluğu, bir hastalığın iyileşme süresi gibi. Nicel değişenler, kesikli ve sürekli değişkenler olarak ikiye ayrılır. Belli bir aralıkta her değeri alan değişkenler **sürekli değişkenler**, her değeri alamayan değişkenler **kesikli değişkenler** olarak adlandırılır. Örneğin hane halkı sayısı kesikli, hanedeki kişilerin ağırlıkları sürekli değişkenlerdir.

**Nitel değişken** ise, sayısal değerler ile ölçülemeyen değişkenlerdir. Bunlara **kategorik değişken** adı da verilir.




İstatistiksel tekniklerin kullanılabilmesi için ilgili birimlerden bilgi toplanmasında değişkenlerin nasıl ölçüldüğü çok önemlidir. Değişkenlere örnek olarak, zaman nedir? ve nasıl ölçülür? Kütle nedir ve nasıl ölçülür? Sıcaklı, ağırlık nedir ve nasıl ölçülür? Cinsiyet nedir ve nasıl ölçülür? Tansiyon nedir ve nasıl ölçülür? vb. örnekler verilebilir. Değişkenlerin ölçülmesi genel olarak dört başlık altında açıklanabilir.

**1) Sınıflama (Nominal/Sırasız) Ölçme Düzeyi**

Birimlere niteliklerine göre belirli isimler verilir. Bu ölçme düzeyinde değişkenler sınıflandırılarak ölçülür. Herhangi bir sıralama yapılmaz. Örneğin, cinsiyet; kadın, erkek gibi. Hastanın kan grubu, hastanın mesleği, göz rengi, doğum yeri, medeni durum, kan grubu, tarımda kullanılan gübre türü gibi değişkenlere ait ölçümler için sınıflama ölçme düzeyi kullanılır.

### 2) Sıralama (Ordinal) Ölçme Düzeyi

Sıralama ölçme düzeyinde değişkenlerin aldığı değerler önem derecesi ya da üstünlüklerine göre sıralanır. Örneğin, ağrı derecelerine göre hastalar, çok şiddetli ağrı, şiddetli ağrı, orta şiddetli ağrı, az hissedilen ağrı ve hissedilmeyen ağrı gibi değerler alabiliyorlar. Kategoriler arasında büyüklük, küçüklük ve eşitlik sıralamaları yapılabilir. Katılım düzeyi (Kesinlikle Katılıyorum, Katılıyorum, Kararsızım, Katılmıyorum, Kesinlikle Katılmıyorum), sıklık düzeyi (Hiç, Nadiren, Genellikle, Her Zaman), öğrenim durumu (İlköğretim, Lise, Lisans, Yüksek Lisans) , hastalık evresi (evre I, evre II, evre III), gelir düzeyi (düşük, orta, yüksek) vb. değişkenler için sıralama ölçme düzeyi kullanılır.

### 3) Eşit Aralıklı Ölçme Düzeyi

Sıcaklık, zeka düzeyi (IQ) gibi değişkenlerin ölçüldüğü ölçme düzeyidir. Başlangıç için anlamlı olabilecek doğal bir sıfır değeri yoktur. Örneğin, sıcaklık değerinin sıfır olması sıcaklığın olmadığı anlamına gelmez.

### 4) Oranlama Ölçme Düzeyi

Yaş, ağırlık, ücret, not, hız gibi anlamlı bir sıfır değerine sahip olan ölçme düzeyidir. Bu ölçme düzeyinde başlangıç "0" noktasıdır.

Sınıflama ve sıralama ölçme düzeyi nitel değişkenler için eşit aralıklı ve oranlama ölçme düzeyi ise nicel değişkenler için kullanılır.

## Verilerin Düzenlenmesi ve Grafikler

Toplanan veriler herhangi bir düzenlemeden geçirilmemiş ise bu tür verilere ham veri denir. Verileri özetlemek, özelliklerini ortaya koymak için yapılan çizelgelere frekans tabloları denir. Toplanan verilerin kullanıma sunulmasında yararlandığımız geometrik şekillere grafik adı verilir. Frekans tablolarının tamamlayıcısı olarak düşünülebilir. En yaygın kullanılanları çubuk grafiği, diyagram, histogram, dal-yaprak grafiği ve daire grafiğidir.

### 1) Nicel Veriler

Örneği çözmeye başlamadan önce, frekans tablosunu oluşturabilmek için bazı tanımlar verilecektir.

**Sınıf:** Değişkenin değer aralığı birbirinden kesin olarak ayrılmış gruplara bölünebilir, bu gruplara sınıf adı verilir. Sınıfların (grupların) sayısını belirlemek için en çok kullanılan yöntem Sturges Kuralı'dır. n gözlem sayısı, k sınıf sayısı olmak üzere aşağıda verilen formül kullanılır.

$k = 1+3.3 \log(n)$

Sınıf sayısının bulunması sınıf aralığının belirlenmesini kolaylaştırır.

**Dağılım Sınırları:** Örneklemdeki en küçük değer ile en büyük değere dağılım sınırları denir.

**Dağılım Genişliği:** Bir örneklemde en büyük değer ile en küçük değer arasındaki farka dağılım genişliği denir, *R* ile gösterilir.

**Sınıf Aralığı:** İki sınıf arasındaki farka sınıf aralığı denir, *c* ile gösterilir.

$$c = \frac{\text{En büyük değer - En küçük değer}}{\text{Sınıf sayısı}}$$

**Alt Sınır:** Bir sınıfın en küçük değeridir, As ile gösterilir.

**Üst Sınır:** Bir sınıfın en büyük değeridir, Üs ile gösterilir.

**Sınıf Değeri:** Bir sınıfın alt sınır ve üst sınır değerlerinin ortalaması sınıf değeridir, S ile gösterilir.

**Frekans (Sıklık):** Bir sınıfa düşen veri sayısına frekans denir, *f* ile gösterilir. Frekansların toplamı veri sayısına eşit olmalıdır. Yani, $\sum_{i=1}^{n} f_i = n$ dir.

**Göreli Frekans (Frekans Yüzdesi):** Her sınıfa düşen veri sayısının toplam veri sayısına oranına denir, *p* ile gösterilir. $p_i = \frac{f_i}{n}, i = 1,2,...,k \qquad \sum_{i=1}^{k} pi = 1$

**Örnek 1.** 50 araç sürçüşünün sabah işe giderken kaç km yol yaptıkları sorulmuş ve aşağıdaki cevaplar alınmıştır.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 7 | 4 | 13 | 6 | 3 | 7 | 11 | 12 | 6 |
| 12 | 8 | 3 | 11 | 5 | 5 | 8 | 17 | 13 | 4 |
| 20 | 11 | 3 | 8 | 4 | 2 | 9 | 16 | 11 | 14 |
| 9 | 10 | 2 | 7 | 3 | 2 | 9 | 5 | 10 | 16 |
| 6 | 16 | 9 | 7 | 2 | 6 | 4 | 9 | 5 | 17 |

Bu verileri kullanarak,

a) 5 sınıflı bir frekans dağılımı oluşturunuz.
b) Oransal frekans sütununu oluşturunuz.
c) Sınıf orta değerini bulunuz.
d) Histogram ve diyagram grafiğini çiziniz.

Sınıf sayısı $k=5$ olarak verilmiştir.

Sınıf aralığı ise,

$n = 50;\ k = 5\ olsun. Enb = 20;\quad Enk = 2 \qquad R = 20 - 2 = 18$

$$c = \frac{18}{5} = 3.6 \cong 4$$

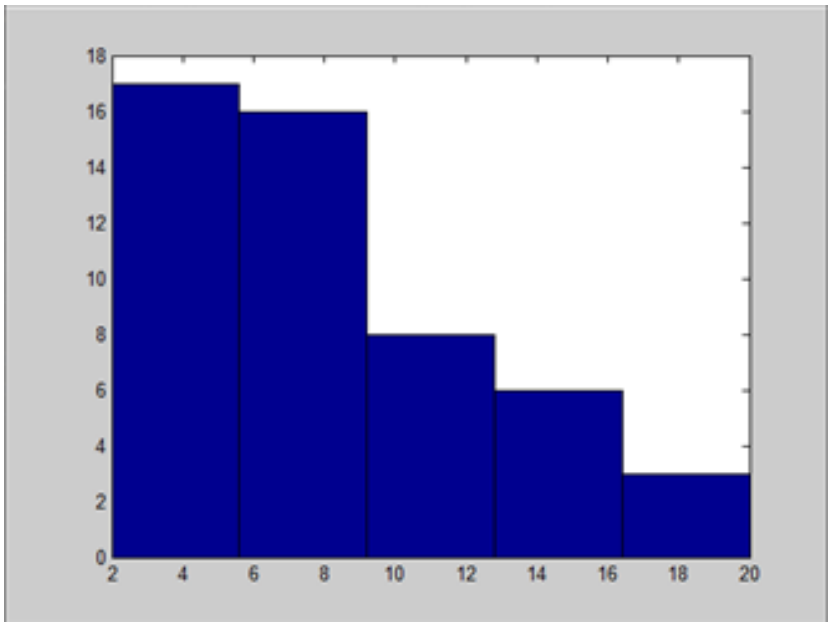
**Çözüm 1.:**

| **Sınıflar** | **Frekans ( $f_i$ )** | **Göreli Frekans ( $p_i$ )** | **Sınıf ( $S_i$ )** |
|---|---|---|---|
| $2 \le X < 6$ | 17 | 17/50 = 0.34 | (2+6)/2= 4 |
| $6 \le X < 10$ | 16 | 16/50 = 0.32 | 8 |
| $10 \le X < 14$ | 10 | 10/50 = 0.2 | 12 |
| $14 \le X < 18$ | 6 | 6/50 = 0.12 | 16 |
| $18 \le X < 22$ | 1 | 1/50 = 0.02 | 20 |
| **Toplam** | **50** | **1** | |

- Histogramda her bir sınıfın frekansı ayrı ayrı dikdörtgenler kullanılarak ifade edilir. Histogram, sürekli nicel veriler için kullanılır. Histogramlar, verinin şekli (simetrik ya da çarpık), dağılımı, merkezi ve sapan değerler hakkında bilgi sahibi olmamızı sağlar.
- Diyagramda ise her sınıfın sınıf orta değeri ( $S_i$ **)** ile sınıf frekansı kullanılır.

*Matlab Kodu*

```
Y=[5 7 4 13 6 3 7 11 12 6 12 8 3 11 5 5 8 17 13 4 20 11
38 4 2 9 16 11 14 9 10 2 7 3 2 9 5 10 16 6 16 9 7 2 6 4 9
5 17]
hist(Y,5)
```




**Birikimli (kümülatif) Frekans Tablosu**

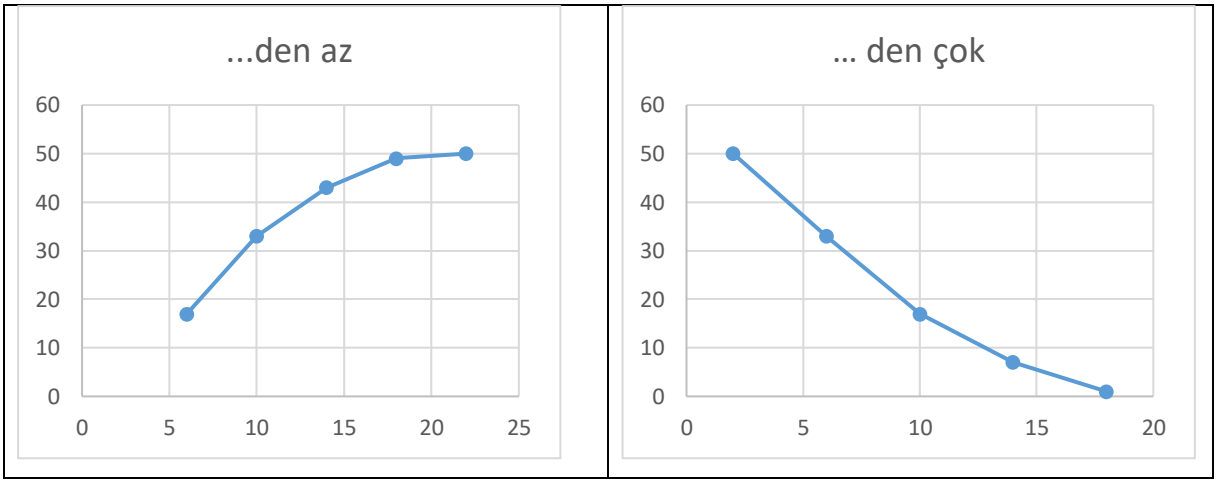
Her sınıfın üst sınırı ile bir sonraki sınıfın alt sınırı değerlerinin ortalaması sınıf ara değeri (*Sad*) olarak alınırsa, bu değerden daha az değer gösteren verilerin toplamı frekans sütununa yazıldığında den daha az frekans bulunmuş olur. Aynı şekilde, sınıf ara değerinden daha çok değer gösteren verilerin toplamı frekans sütununa yazıldığında den daha çok frekans bulunmuş olur. Sınıf ara değeri, den daha az ve den daha çok sütunları tamamlandıktan sonra oluşan tabloya birikimli frekans tablosu denir.

Aynı soru için birikimli frekanslar incelenmiştir. Birikimli frekanslar **"…den az" ve "….den çok"** grafikleri ile gösterilir.

| **…..den az** | **Sürücü Sayısı** |
|---|---|
| 6 km den daha az yol yapanların | 17 |
| 10 km den daha az yol yapanların | 17+16=33 |
| 14 km den daha az yol yapanların | 17+16+10=43 |
| 18 km den daha az yol yapanların | 17+16+10+6=49 |
| 22 km den daha az yol yapanların | 17+16+10+6+1=50 |

Veya

| **…..den çok** | **Sürücü Sayısı** |
|---|---|
| 2 km den daha çok yol yapanların | 17+16+10+6+1=50 |
| 6 km den daha çok yol yapanların | 16+10+6+1=33 |
| 10 km den daha çok yol yapanların | 10+6+1=17 |
| 14 km den daha çok yol yapanların | 6+1=7 |
| 18 km den daha çok yol yapanların | 1 |




**Grafik 1.3.** 50 araç sürücüsünün işe giderken yaptığı yol (km) için birikimli frekans dağılımlarının gösterimi.

### 2) Nitel Veriler

Nitel veriler için sınıflama ve sıralama ölçme düzeyinin kullanılabileceği daha önce ifade edilmişti. Sınıflanabilen ve sıralanabilen verilerde sınıflar birbirinden bağımsız olduğu için frekans tablosunda sadece sınıf, frekans ve göreli frekans sütunları yer alır.

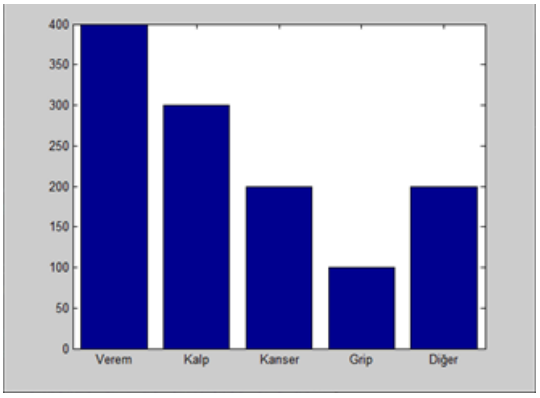
**Örnek 2.:** Bir hastaneye müracaat eden hastaların 400'ü verem, 300'ü kalp, 200'ü kanser, 100'ü grip ve 200'ü diğer hastalık türüne sahiptir. Bu verilere ilişkin frekans tablosu Tablo 1.4 de verilmiştir.

**Tablo 2.:** Hastalık türüne göre hasta sayısı

| Sınıf(Hastalık Türü) | Frekans(Hasta Sayısı) | Göreli Frekans |
|---|---|---|
| Verem | 400 | 400/1200 |
| Kalp | 300 | 300/1200 |
| Kanser | 200 | 200/1200 |
| Grip | 100 | 100/1200 |
| Diğer | 200 | 200/1200 |
| | 1200 | |

## Çubuk Grafiği

Çubukları grafiği, kesikli nicel verilerde ve nitel verilerde kullanılır. Çubuk grafiğinde sınıflar, tabanları eşit ve birbirine bitişik olmayan dikdörtgenlerle temsil edilir. Çubuk grafiği hem dikey hem yatay olarak çizilebilir. Tablo 1.4. de verilen frekans tablosu için çubuk grafiği aşağıda verilmiştir.



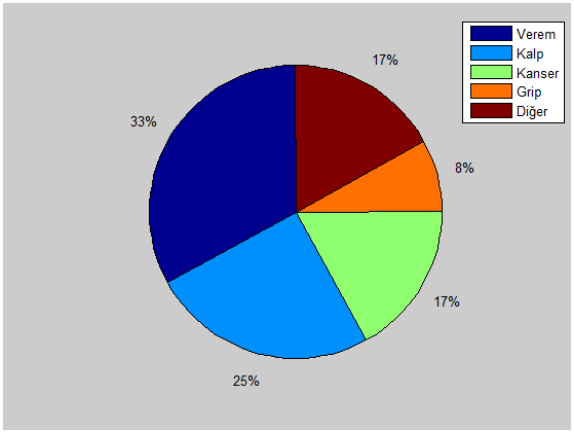

**Grafik1.1.** Hastaların hastalık türlerine göre grafiği

*Matlab Kodu*

```
hasta=[400 300 200 100 200];
bar(hasta)
set(gca,'xticklabel',{'Verem','Kalp','Kanser','Grip','Diğer'});
```

## Daire Grafiği

Her sınıfa düşen frekansın bir dairenin parçası ile gösterildiği grafik türüdür. Bu grafiği çizebilmek için göreli frekanslar hesaplanır. Her sınıfa ilişkin göreli frekans $360^o$ ile çarpılarak o sınıfa ilişkin daire dilimleri bulunur. Tüm sınıflar için yapıldığında daire tamamlanmış olur. Daha çok sınıflandırılabilen verilerde kullanılır. Tablo 1.4. de verilen frekans tablosu için daire grafiği aşağıda verilmiştir.




*Matlab Kodu*
*hasta=[ 0.33 0.25 0.17 0.08 0.17];*
*pie(hasta);*

**Dal-Yaprak Grafiği (Stemplot – Stem-and-leaf plot)**

Kesikli gözlem değerlerinin iki kısımda ifadesi ile oluşturulur. Değişkenin o gözlem değerini alma aralığı çok genişse uygun bir yöntemdir.

**Örnek 3.** Bir araç yıkama istasyonuna 30 gün içinde gelen araçların sayısı aşağıda verilmiştir. Dal-yaprak grafiğini oluşturunuz.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 22 | 75 | 63 | 24 | 8 | 17 | 19 | 26 | 35 | 53 |
| 58 | 57 | 69 | 70 | 16 | 9 | 5 | 28 | 37 | 53 |
| 34 | 29 | 7 | 44 | 46 | 19 | 35 | 18 | 60 | 13 |

**Çözüm 3.**

Veri seti büyükten küçüğe doğru sıralanır. Her gözlem değeri dal ve yaprak olarak ayrılır. Örneğin; iki basamaklı tamsayıların onlar basamağındaki rakam "dal", birler basamağındaki rakam ise "yaprak" olarak adlandırılır. Bu örnekte dal "onlar" , yaprak ise "birler" olarak tanımlanmıştır.

| **Dal** | **Yaprak** | |
|---|---|---|
| 0 | 5,8,9,7 | → Gözlem değerleri: 5,8,9,7 |
| 1 | 3,6,7,8,9,9 | → Gözlem değerleri: 13,16,17,18,19,19 |
| 2 | 2,4,6,8,9 | |
| 3 | 4,5,5,7,8 | |

| | |
|---|---|
| 4 | 4,6 |
| 5 | 3,8,7 |
| 6 | 0,3,9 |
| 7 | 0,5 → Gözlem değerleri: 0,5 |

*Matlab Kodu*

```
S=[22 75 63 24 8 17 19 26 35 53 58 57 69 70 16 9 5 28 37 53 34 29 7 44 46 19 35 18 60 13]
stemleafplot(S)
```

```
 0 | 5 7 8 9
 1 | 3 6 7 8 9 9
 2 | 2 4 6 8 9
 3 | 4 5 5 7
 4 | 4 6
 5 | 3 3 7 8
 6 | 0 3 9
 7 | 0 5
```