Tanımlayıcı İstatistik

IST 108 Olasılık ve İstatistik
Bahar 2016

Yrd. Doç. Dr. Ferhat Dikbıyık

Veri kümelerini tanımlamak

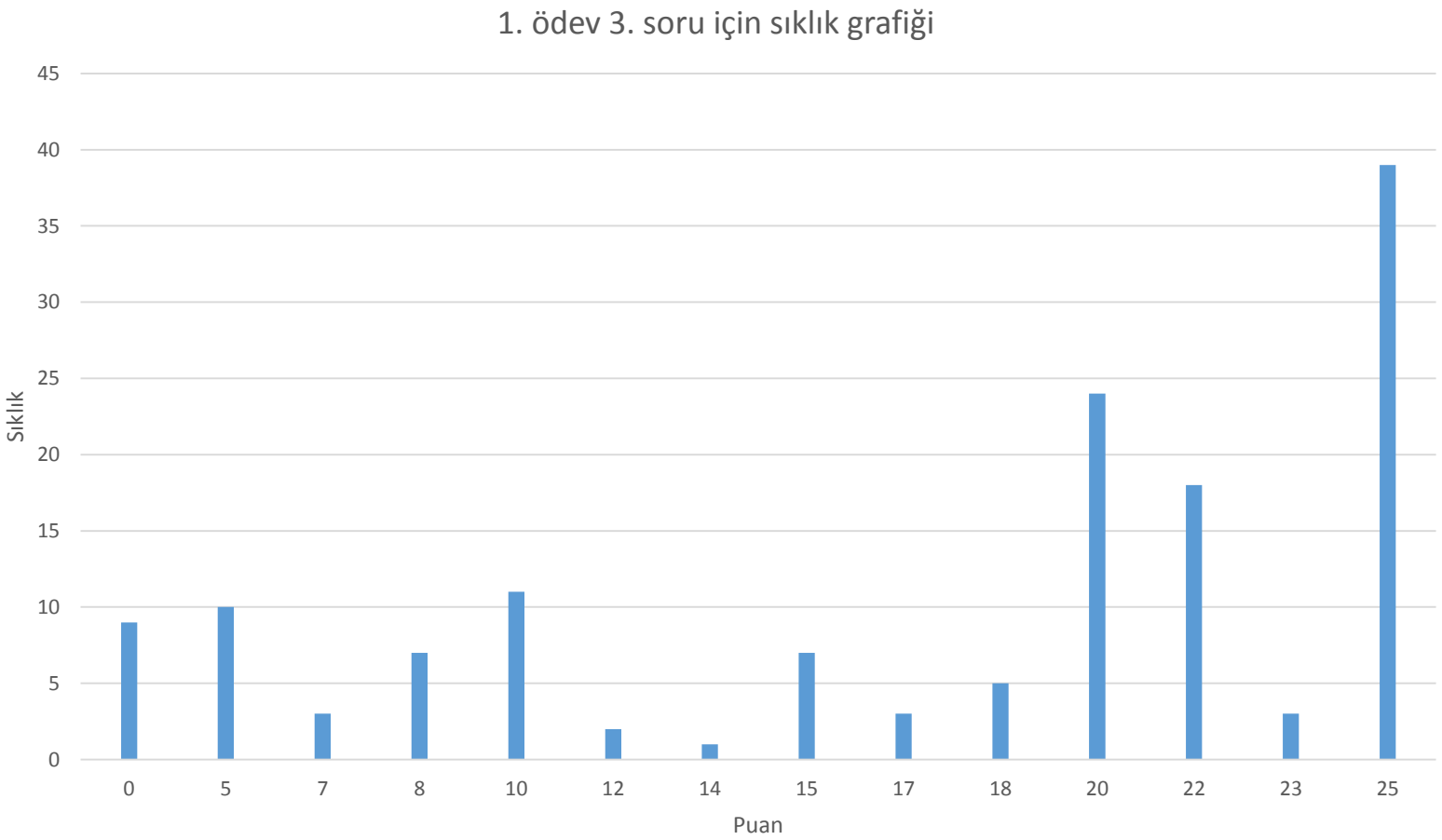
- Bir çalışmanın sonucunda elde edilen veriler açık, öz, ve gözlemcinin verinin temel karakteristiği ile ilgili bilgileri hızlıca elde edebileceği şekilde sunulmalıdır.
- Yıllar boyunca tablolar ve grafikler verilerinin sunulmasında çok faydalı olmuşlardır. Tablolar ve grafikler verinin aralığı, yoğunluk derecesi, ve simetrisi gibi bir çok bilgiyi bize vermektedirler.

Sıklık (frekans) tablosu

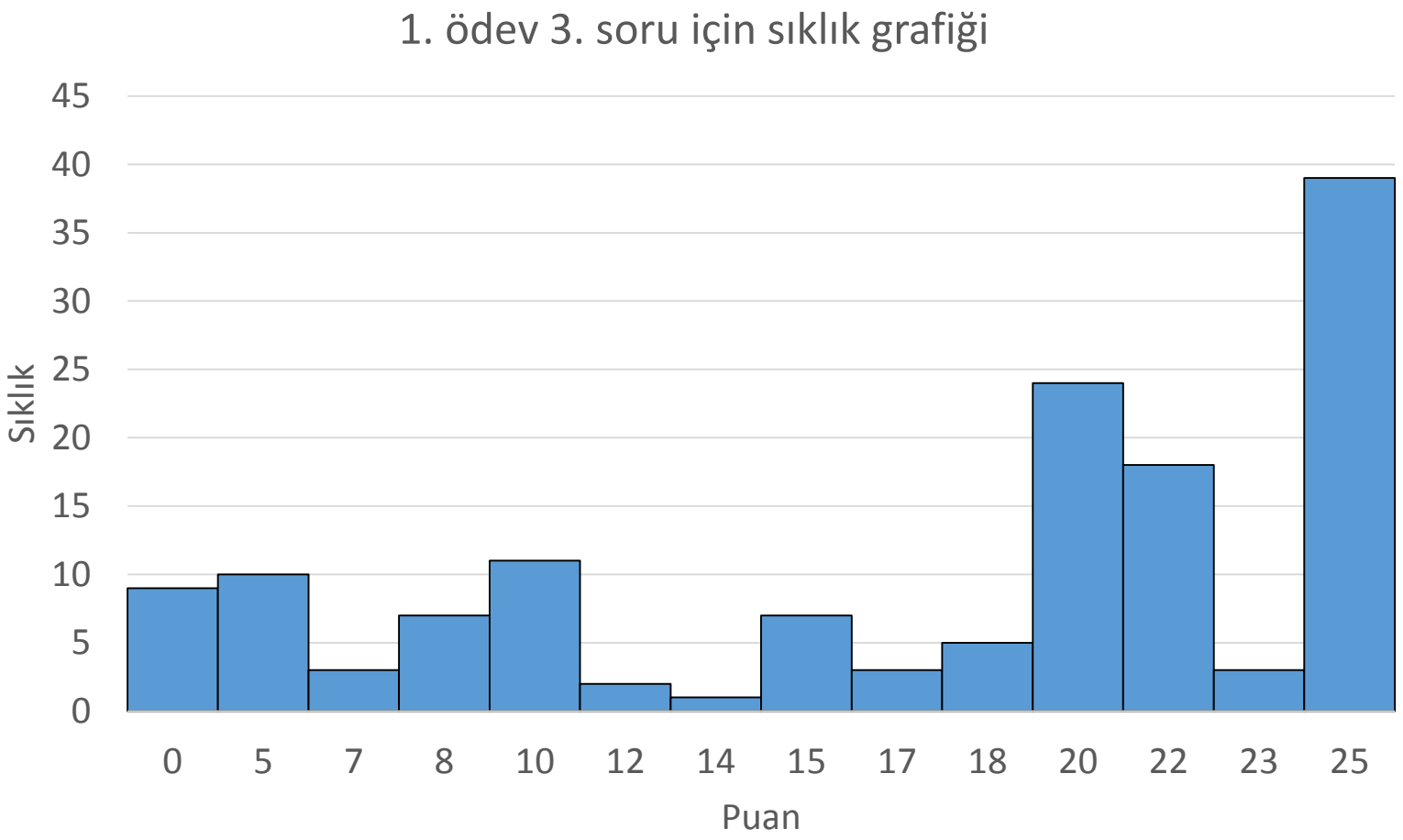
- Az sayıda farklı değerlere sahip bir veri kümesi bir sıklık (frekans) tablosu ile gösterilebilir. Aşağıda ödevin 3. sorusu için verilen 14 farklı puana ait sıklık tablosu var.

Puan	Sıklık
0	9
5	10
7	3
8	7
10	11
12	2
14	1
15	7
17	3
18	5
20	24
22	18
23	3
25	39

Çizgi Grafiği



Sütun Grafiği



Sıklık Poligonu

Göreceli Sıklık Tabloları ve Grafikleri

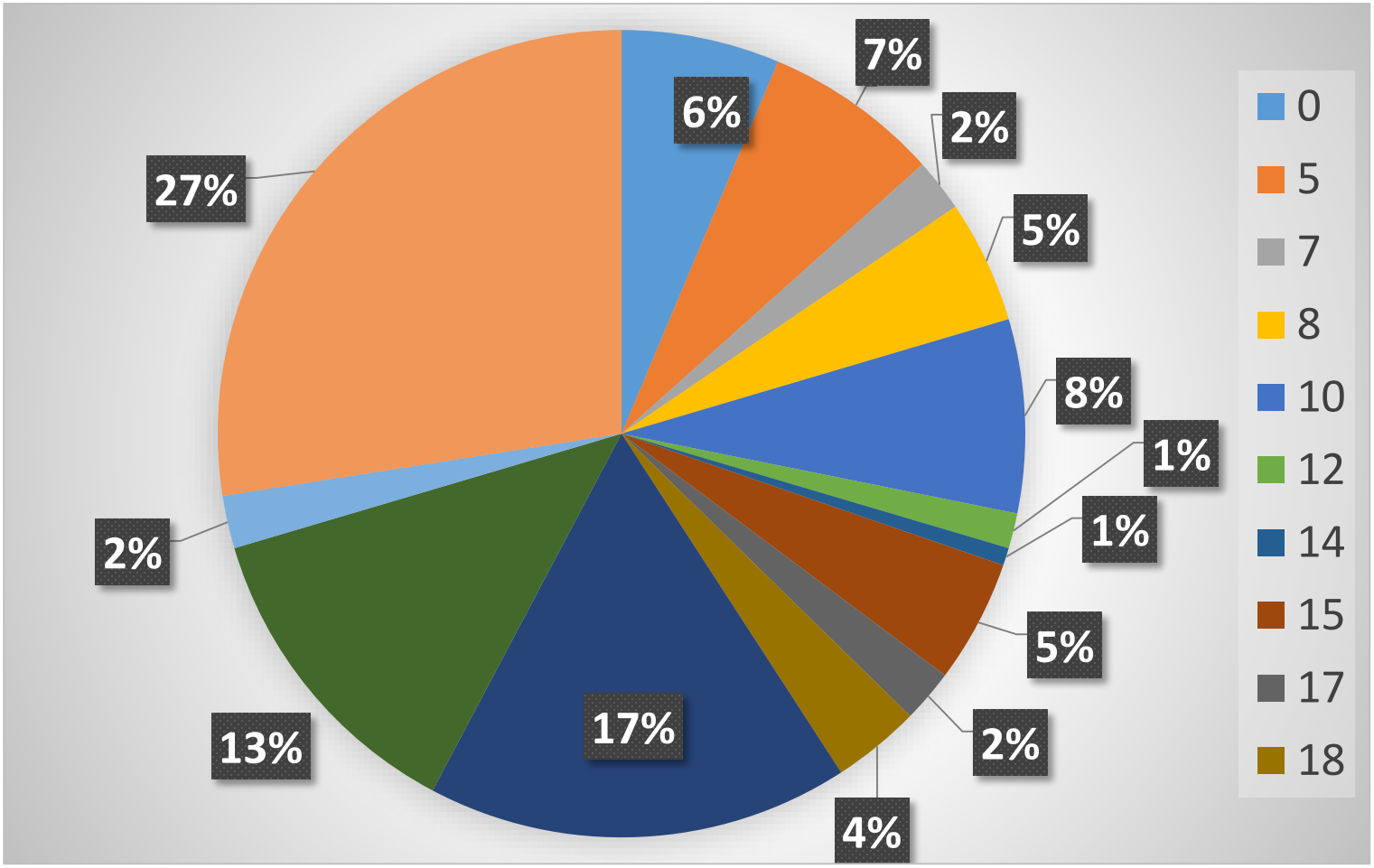
- n adet veri içeren bir veri kümesi düşünelim. Eğer f belirli bir değere ait bir frekans (sıklık) bilgisi ise f/n göreceli frekans olarak adlandırılır.

- Yani bir veri değerine ait göreceli frekans o değerin o verinin kaçta kaçında göründüğünü verecektir.

Göreceli Frekans Tablosu

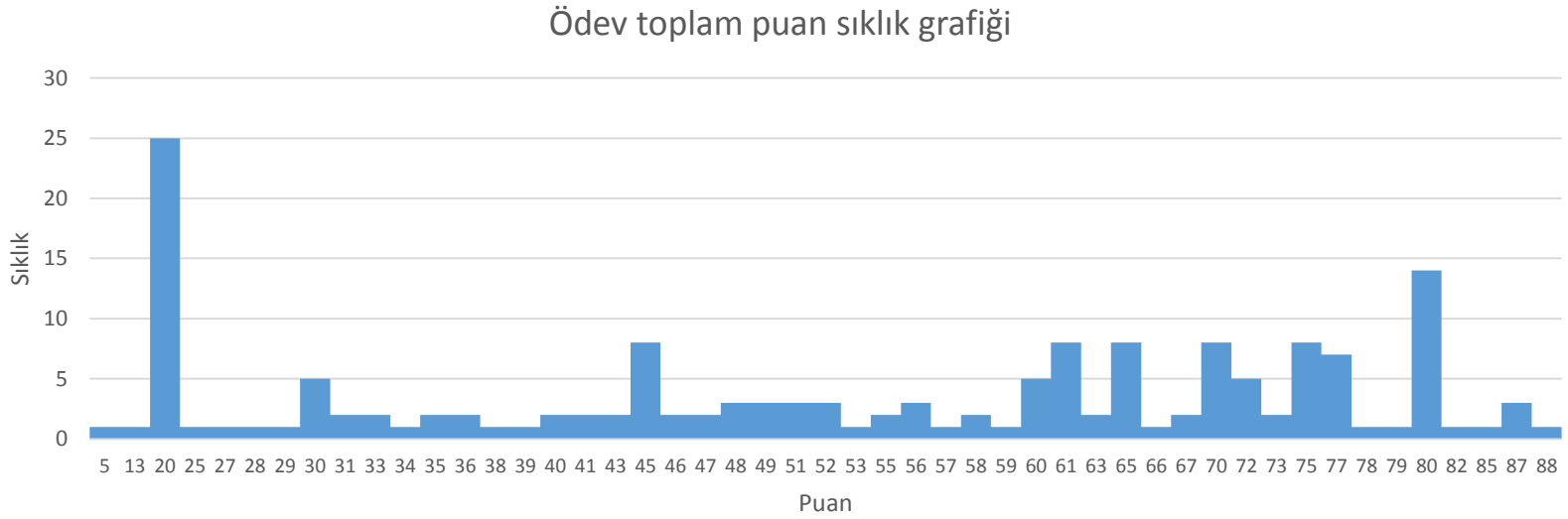
Puan	Göreceli Frekans
0	9/142
5	5/71
7	3/142
8	7/142
10	11/142
12	1/71
14	1/142
15	7/142
17	3/142
18	5/142
20	12/71
22	9/71
23	3/142
25	39/142

Göreceli frekans grafiği



Veri gruplama, Histogram, Kök ve Yaprak Gösterimi

- Eğer verilerin sahip olduğu farklı değerler çok fazla ise yukarıdaki gösterimler çok pratik olmayabilir.
- Örneğin toplam vize notlarınıza bakıldığında 49 farklı değer var ve aşağıda sıklık sütun grafiği görünüyor.

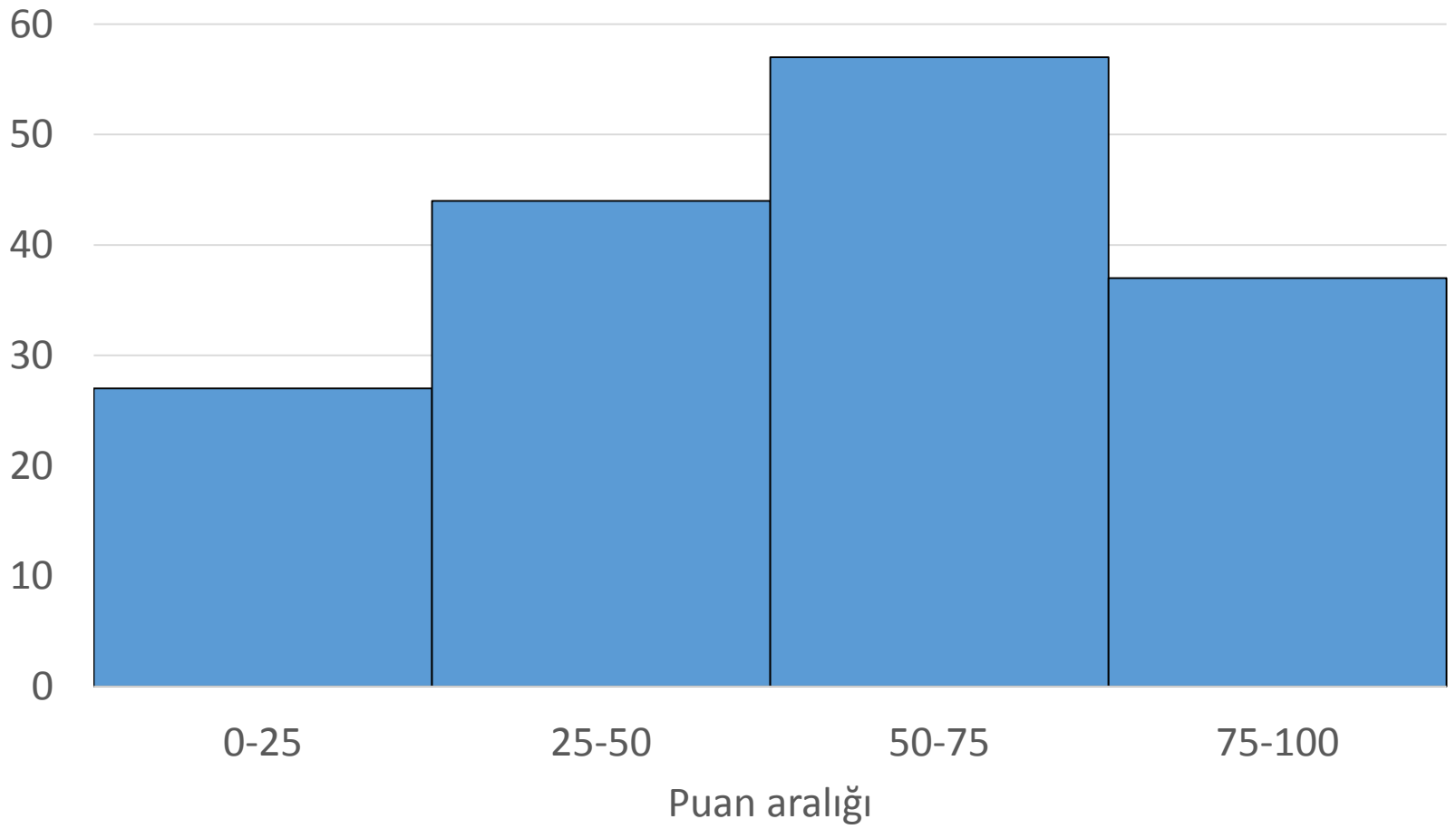


Veri gruplama

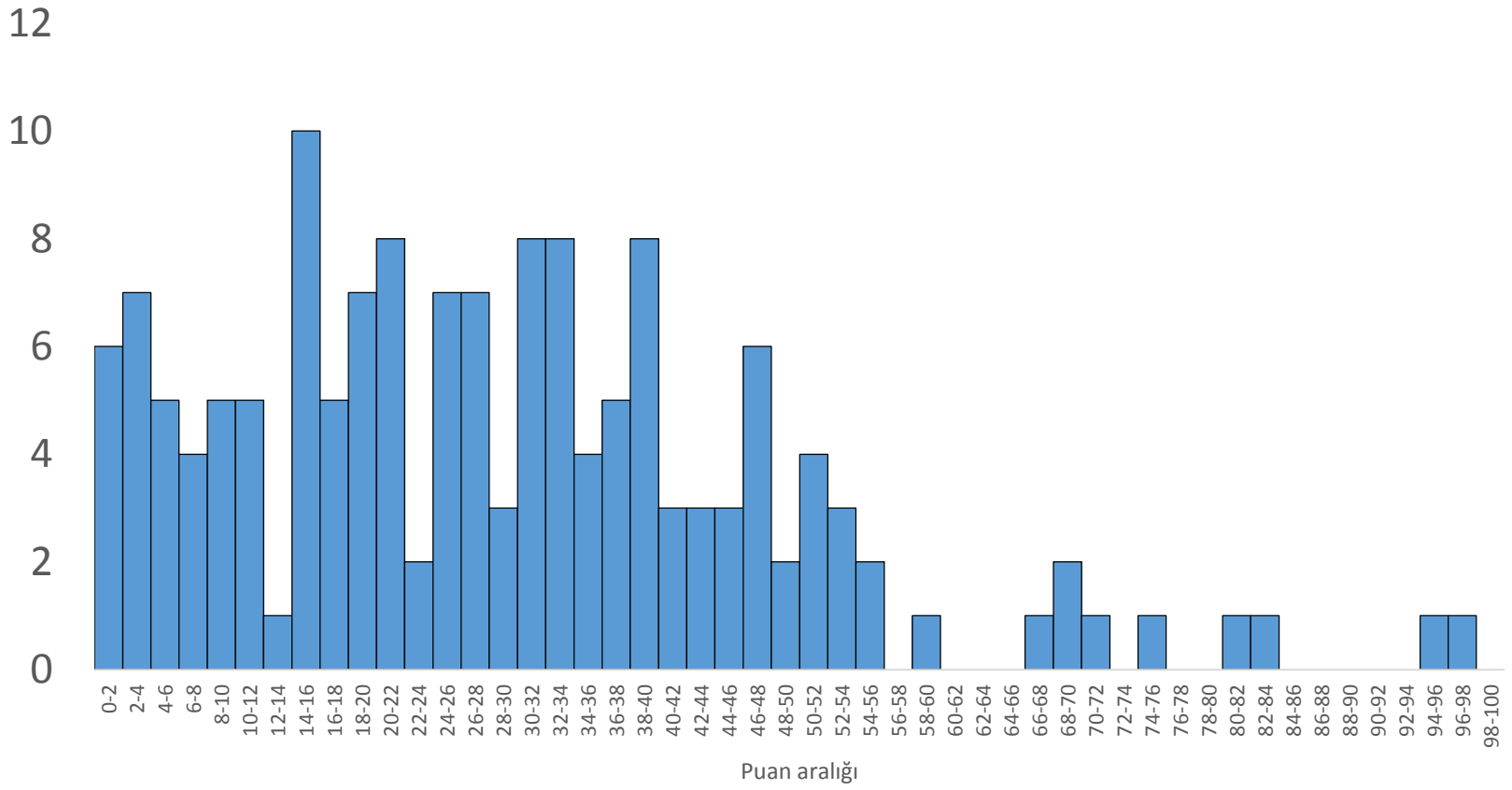
- Bu gibi durumlarda veri gruplayarak *aralıklara* bölmek daha faydalı olacaktır. Seçilen aralık sayısı çok önemlidir. Çünkü;
  - Eğer çok az sayıda aralık seçilirse çok fazla bilgi kaybı olacaktır.
  - Eğer çok fazla sayıda aralık seçilirse, her aralığın frekansı anlamlı bir örüntü çıkarmak için çok küçük olacaktır.

- Aralıkların başlangıç ve bitiş noktaları sınır olarak ifade edilir. Biz dersimizde aralık gösterimleri için sol-uç dahil gösterimini kullanacağız.
  - Örneğin 20-30 aralığı 20 ve 20'den büyük ve 30'dan küçük değerler aralığını gösterir.

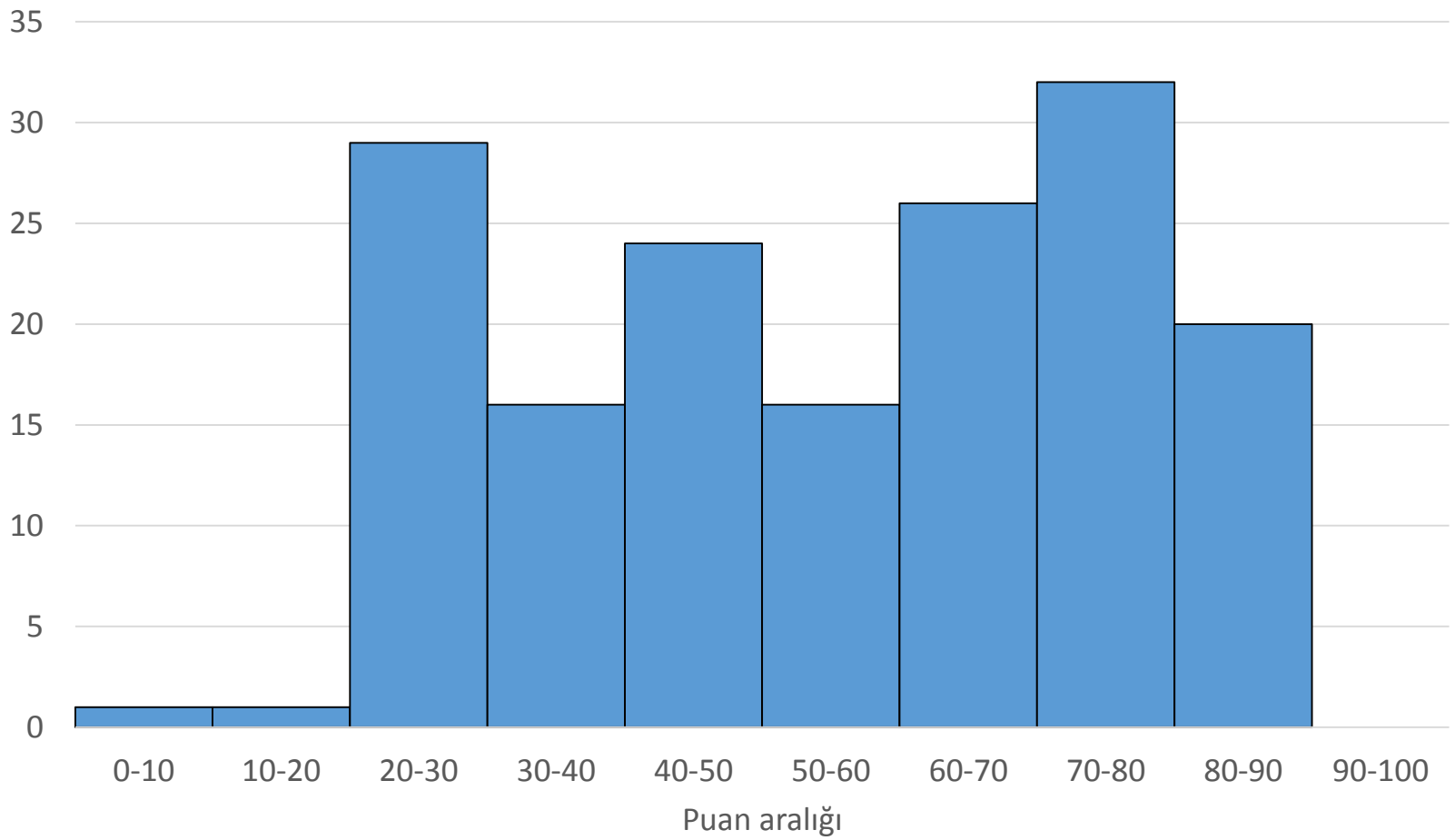
Veri gruplama
Örnek: Az sayıda aralık seçimi



Veri gruplama
Örnek: Çok sayıda aralık seçimi

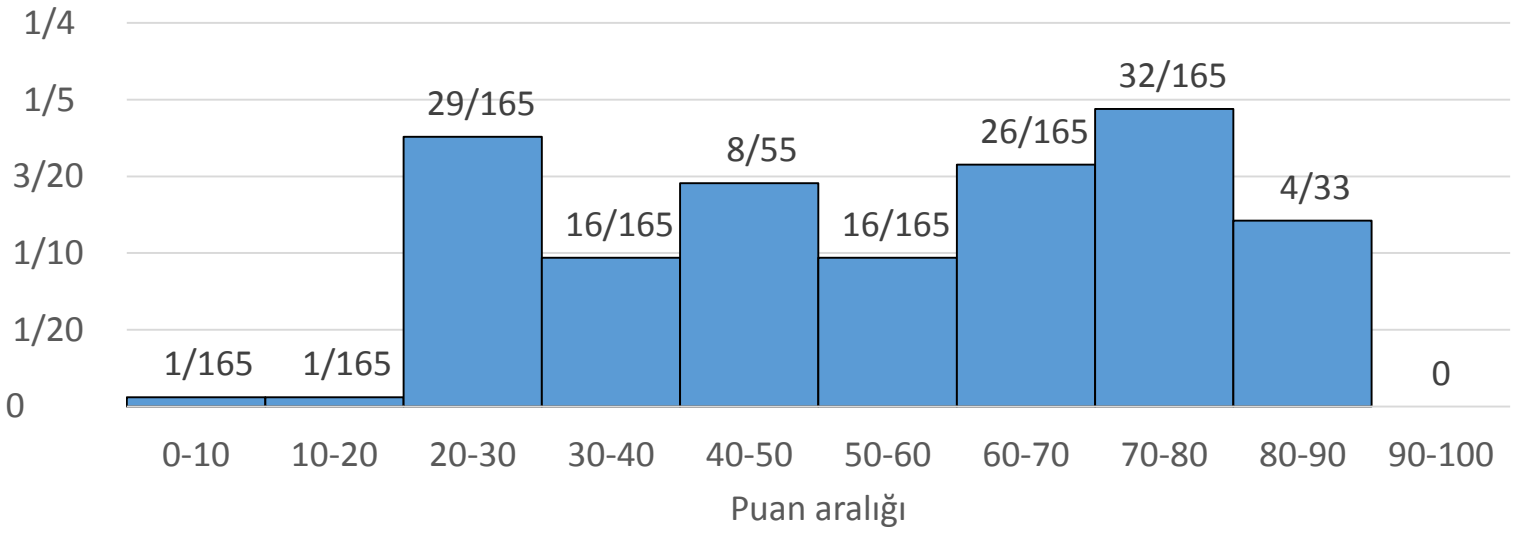


Veri gruplama Örnek

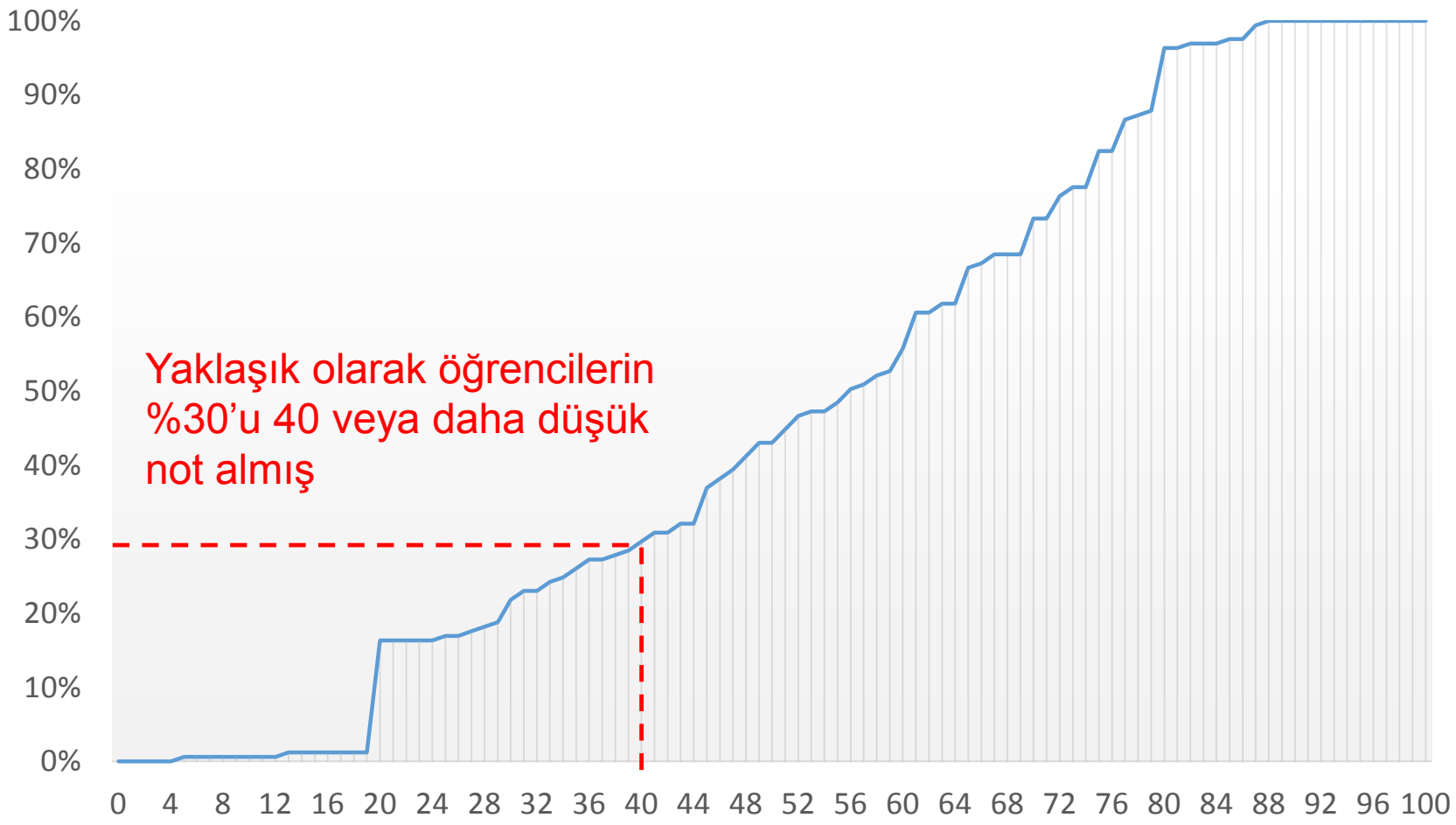


Histogram

- Veri aralıklarının sıklıklarını (frekanslarını) yada göreceli frekanslarını veren birbirine bitişik sütunlardan oluşmuş sütün grafiğine histogram denir. Bir önceki slaytta frekans histogram, aşağıda ise göreceli frekans histogram bulunmakta.



Kümülatif (Göreceli) Frekans Grafiği



Kök-yaprak gösterimi

- Küçük ve orta ölçekli verileri göstermede kök ve yaprak gösterimi iyi bir yoldur.
- Bu gösterim her bir değeri kök ve yaprak olmak üzere iki parçaya ayırmak ile mümkün olabilir.
- Örneğin tüm veri değerleri iki basamaklı ise, onlar basamağı kök ve birler basamağı yaprak olabilir.
- Örneğin 62 sayısı için kök 6 ve yaprak 2 olacaktır. Eğer 62 ve 67 sayıları varsa gösterim;

Kök	Yaprak
6	2;7

Kök yaprak gösterimi
Örnek: Ödev 3. soru puanları

Kök	Yaprak
0	0; 5; 7; 8
1	0; 2; 4; 5; 7; 8
2	0; 2; 3; 5

Veri kümelerini özetlemek

- Günümüzde yapılan bir çok deney büyük verilerle başa çıkmak zorundadır.
- Örneğin 1951 yılında tıbbi istatistikçiler R. Doll ve A. B. Hill İngiltere'deki tüm doktorlara bir anket gönderdiler ve yaklaşık olarak 40.000 cevap aldılar. Sorular yaş, yeme alışkanlıkları ve sigara içme sıklıkları ile ilgiliydi. Cevap verenler 10 yıl boyunca gözlemlenerek öldükleri zaman ölüm sebepleri araştırıldı.
- Bu kadar büyük bir veriden mantıklı veriler çıkartılabilmesi için verilerin önemli ölçüde özetlenmesi gerekmekteydi.

Örnekleme Ortalaması

- n adet sayısal değerden oluşan bir veri kümesi için, örnekleme ortalaması bu sayıların aritmetik ortalamasıdır.

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

- Örnekleme ortalamasının hesaplanması genellikle şu şekilde basitleştirilebilir.

$$y_i = ax_i + b, \quad i = 1,2,\ldots,n$$

$$\bar{y} = \sum_{i=1}^{n} (ax_i + b) / n = \sum_{i=1}^{n} (ax_i) / n + \sum_{i=1}^{n} b / n = a\bar{x} + b$$

Örnek 1

- Aşağıdaki veri setinin ortalaması nedir?

280, 278, 272, 276, 281, 279, 276, 281, 289, 280

Örnek 1

- Aşağıdaki veri setinin ortalaması nedir?

  280, 278, 272, 276, 281, 279, 276, 281, 289, 280

- Bu veri setine ait değerleri direk olarak toplamak yerine bu değerlerden 280'i çıkarıp ortalama almak daha kolaydır. $y_i = x_i - 280$

$$0, -2, -8, -4, 1, -1, -4, 1, 9, 0$$

$$\bar{y} = -0{,}8$$

$$\bar{x} = \bar{y} + 280 = 279{,}2$$

Örnekleme Ortalaması

- Bazen frekans tablosunda listelenen $f_1, f_2, \ldots, f_k$ frekanslarına sahip k adet farklı $v_1, v_2, \ldots, v_k$ değerinin örnekleme ortalaması ile ilgileniriz. $n = \sum_{i=1}^{k} f_i$ adet veriden oluşan böyle bir veri kümesi için örnekleme ortalaması

$$\bar{x} = \sum_{i=1}^{n} f_i v_i / n$$

- Yani örnekleme ortalaması veri kümesindeki farklı değerlerin ağırlıklı ortalamasıdır.

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \ldots \frac{f_k}{n} v_k$$

Örnek 2

- Aşağıdaki tabloda belirli bir gruptaki kişilere ait yaş sıklık grafiği verilmiştir. Bu grubun yaş ortalaması nedir?

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

Örnek 2

- Aşağıdaki tabloda belirli bir gruptaki kişilere ait yaş sıklık grafiği verilmiştir. Bu grubun yaş ortalaması nedir?

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

$$\bar{x} = \frac{15 \times 2 + 16 \times 5 + 17 \times 11 + 18 \times 9 + 19 \times 14 + 20 \times 13}{54} = 18{,}24$$

Örnekleme medyanı

- Bir veri kümesinin merkezini gösteren bir diğer istatistik ise örnekleme medyanıdır. Kabaca, veri kümesi artan sıra ile dizildiğinde tam ortada kalan değerdir diyebiliriz.
- Veri kümesindeki değerleri en küçükten en büyüğe sırala. Eğer n tek ise, medyan $(n+1)/2$'deki değerdir, eğer n çift ise, medyan $n/2$ ve $n/2 + 1$'deki değerlerin ortalamasıdır.

Ortalama ve medyan

- Hem örnekleme ortalaması hem de medyan faydalı istatistiklerdir.
- Örnekleme ortalaması, diğer verilere göre çok büyük veya çok küçük olan uç değerlerden etkilenirken, medyan bu değerlerden etkilenmez.
- Sabit bir vergi oranı kullanılan bir yerde vergiden gelecek gelir hesaplanırken, vergi verenlerin gelirlerinin ortalamasını almak daha mantıklıdır.
- Fakat, eğer orta sınıf için apartmanlar yapılacaksa ve bu apartmanların fiyatını ödeyebilecek nüfusun oranı tespit edilmeye çalışılıyor ise bu durumda medyan daha kullanışlıdır.

Örnek 3

- Aşağıdaki tabloda belirli bir gruptaki kişilere ait yaş sıklık grafiği verilmiştir. Bu veriye ait medyan nedir?

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

Örnek 3

- Aşağıdaki tabloda belirli bir gruptaki kişilere ait yaş sıklık grafiği verilmiştir. Bu veriye ait medyan nedir?

Yaş	Sıklık
15	2
16	5
17	11
18	9
19	14
20	13

54 veri olduğu için medyan 27. ve 28. verinin ortalamasıdır. Bu nedenle 18,5 olur.

Örnekleme Modu

- Merkezi eğilimi gösteren bir başka istatistik ise örnekleme modudur. Mod, en yüksek sıklıkla veri kümesinde bulunan değerdir.
- Eğer tek bir değer en yüksek sıklıkla görünmüyorsa, bu durumda en yüksek frekansa sahip değerler modal değerler olarak adlandırılır.

Örnekleme Varyansı

- Bir veri kümesinin merkezi eğiliminin yanı sıra, veri kümesindeki verilerin yayılımı veya değişkenliği hakkında da bilgi edinmek isteriz.
- Bu anlamda değerlerin ortalamadan uzaklıklarının karesinin ortalamasını veren değer, örnekleme varyansı, bize böyle bir bilgi sunar ve s^2 ile gösterilir.

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$$

Örnekleme Varyansı

- Aşağıdaki eşitlikler varyansın hesaplanmasında faydalı olabilir.

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$y_i = ax_i + b, \quad i = 1,2,\ldots,n \quad \Rightarrow \quad \bar{y} = a\bar{x} + b$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \Rightarrow \quad s_y^2 = a^2 s_x^2$$

Örnekleme Standart Sapması

- Standart sapma, varyansın pozitif kareköküdür.

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)}$$

Örnek 4

- Aşağıdaki veri setlerinin varyansını bulunuz.

- A = 3, 4, 6, 7, 10
- B = -20, 5, 15, 24

Örnek 4

- Aşağıdaki veri setlerinin varyansını bulunuz.

- A = 3, 4, 6, 7, 10
  - Ortalama = 6
  - $s^2 = \frac{-3^2+-3^2+0^2+1^2+4^2}{4} = 7{,}5$
- B = -20, 5, 15, 24
  - Ortalama = 6
  - $s^2 = \frac{-26^2+-1^2+9^2+18^2}{3} = 360{,}67$

Örnek 5

- Aşağıdaki veri setine varyans nedir?
  - 25, 20, 21, 18, 13, 13, 7, 9, 18

Örnek 5

- Aşağıdaki veri setine varyans nedir?
  - 25, 20, 21, 18, 13, 13, 7, 9, 18
  - Her bir değerden 18'i çıkarırsak daha rahat hesaplama yaparız.
  - 7, 2, 3, 0, -5, -5, -11, -9, 0
  - Bu veri setinin varyansı, original veri setinin varyansına eşittir. (Neden?)
  - Bu yeni veri setini y_i değerleri ile gösterelim.
  - $\bar{y} = -18 \qquad \sum_{i=1}^{9} y_i^2 = 314$
  - $s^2 = \frac{314-9(4)}{8} = 34{,}75$

Chebyshev Eşitsizliği

- Ortalaması ve standart sapması verilmiş bir veri setinde, Chebyshev eşitsizliği, herhangi bir $k \geq 1$ değeri için, verinin en az $\%100(1 - 1/k^2)$'si $\bar{x} - ks$ ile $\bar{x} + ks$ arasında olduğunu söyler. Örneğin;
  - $k = 3/2$ için, Chebyshev eşitsizliğine göre, verinin en az $\%100(1 - 4/9) = \%55{,}56$'sı örnekleme ortalamasından en fazla $1{,}5s$ uzaklıktadır.
- Formal tanım:

$$S_k = \{i, 1 \leq i \leq n : | x_i - \bar{x} | < ks\}$$

$$N(S_k) = | S_k |$$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Eşlenmiş veri kümeleri ve örnekleme korelasyon katsayısı

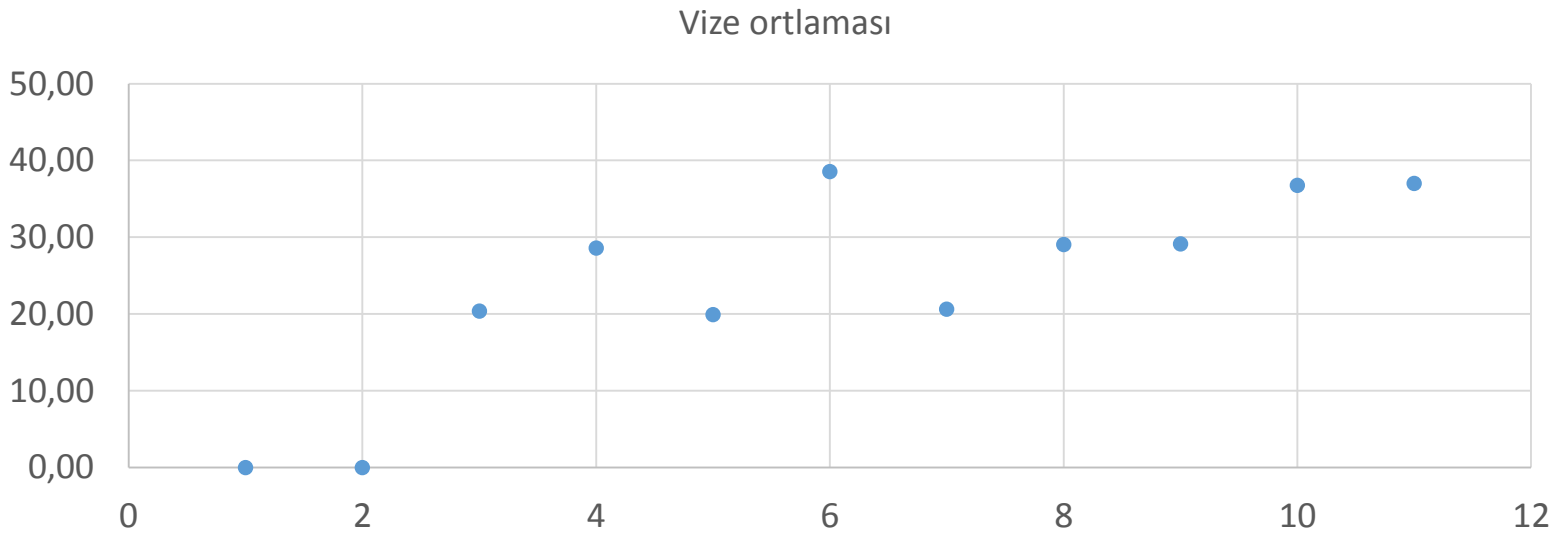
- Sıklıkla aralarında bir ilişki bulunan veri çiftlerine sahip veri kümeleri ile ilgileniriz. Böyle bir veri kümesinde, eğer her eleman bir x ve bir de y değerine sahip ise, i. veri noktasını (x_i, y_i) veri çifti ile ifade ederiz.

1. Ödev puan aralığı	Vize ortlaması
0-10	0,00
10-20	0,00
20-30	20,38
30-40	28,58
40-50	19,90
50-60	38,56
60-70	20,63
70-80	29,03
80-90	29,14
90-100	36,74
100-110	37,02

Dağıtılmış Diagram Gösterimi
Örnek: 1. ödev ve vize ortalaması

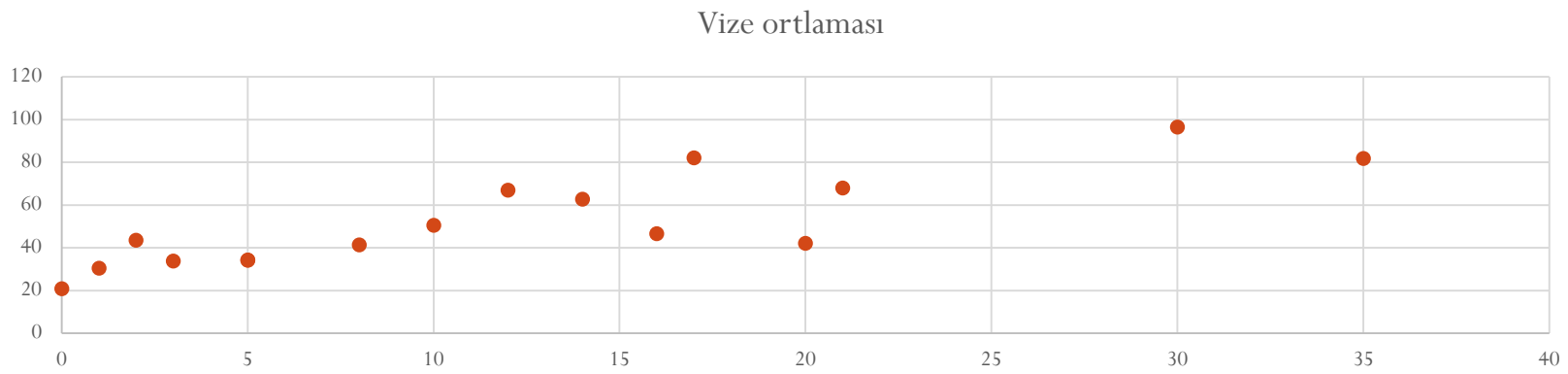
Eşlenmiş verilerde sormamız gereken sorular:

- Büyük x değerleri büyük y değerleri ile ve küçük x değerleri küçük y değerleri ile eşleşiyor mu
- yada büyük x değerleri küçük y değerleri ile ve küçük x değerleri büyük y değerleri ile eşleşiyor mu
- Dağıtımlı diyagram bu sorulara cevaplamada bize kabaca bilgi verir.



Dağıtılmış Diagram Gösterimi
Örnek: 4. soru ve vize ortalaması

Örneğin bu diyagramda 4. soru ve vize notu arasında ciddi bir ilişki olduğundan söz edilebilir.



Örnekleme Korelasyon Katsayısı

- Eşleşmiş bu veriler arasındaki ilişkiyi sayısal olarak ölçebilmek için bir istatistik bilgi gerekir.
- Bu bilgi örnekleme korelasyon katsayısı, *r* ile ifade edilebilir.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Örnekleme Korelasyon Katsayısı Özellikleri

- $-1 \leq r \leq 1$
- $y_i = ax_i + b$ ve $b > 0$ ise $r = 1$.
- $y_i = ax_i + b$ ve $b < 0$ ise $r = -1$.
- Eğer r, x_i ve y_i $(i = 1,2, \ldots, n)$ arasında korelasyon katsayısı ise r aynı zamanda $a + bx_i$ ve $c + dyi$ $(i = 1,2, \ldots, n)$ değerlerine sahip veri kümesi için de korelasyon katsayısıdır.

Örnek