

Olasılık ve İstatistik

HAFTA 10

Tanımlayıcı İstatistik

Dr. Öğretim Üyesi Burcu ÇARKLI YAVUZ

bcarkli@sakarya.edu.tr

Tanımlayıcı İstatistik

- Tanımlayıcı istatistik ***verilerin organize edilmesi, özetlenmesi ve sunulması*** için uygun ve bilgilendirici yöntemleri kapsar.
- Temelde iki tür tanımlayıcı istatistik yaklaşımı vardır:
 - **Grafik ve Tablo Teknikleri**
 - **Sayısal teknikler (nicel yöntemler)**
- Bazı durumlarda **ortalama, mod, medyan, varyans** ve **standart sapma** gibi sayısal tanımlayıcı istatistikler, bazen de **histogram, ogive ve çizgi grafiği** gibi grafik teknikler karar verme sürecine destek olurlar.
 - Türkiye İstatistik Kurumu verilerine göre Ağustos 2019'da işsizlik oranı %14'tür. (Yani Türkiye genelinde 15 yaş ve üzeri kişilerin **ortalama** %14'ü işsizdir)

Çıkarımsal İstatistik

- Çıkarımsal İstatistikler bize; ***ana kütle ile ilgili yorum yapma ve sonuç çıkarma*** süreçlerini gerçekleyebilmek adına uygulanabilecek yöntemleri sunar.
 - SiGames FM 2013 oyununda Türkiye satışlarını artıracığını **öngörerek** «Türkçe» dilini de eklemiştir.
- Çıkarım yapmak çoğu zaman bizi büyük maliyet ve zaman israfından kurtarmakla birlikte, sonuçta yapılan işlem bir tahmin sürecidir ve sürecin en temel ögesi **tahmin hatalarıdır**.
- Bu amaçla Çıkarımsal istatistiksel yöntemler, **Güvenirlilik ölçütleri, Güven Seviyesi ve Anlamlılık Seviyesi** değerleri ile birlikte sunulurlar.

Grafik ve Tablo Teknikleri

- Veri serilerinin grafik teknikler ve tablolar ile sunulması karar vericilere önemli avantajlar sağlamaktadır.
- Her teknik, her tür veri ile kullanılmaz. Bu bağlamda veri türünün kategorik veya nümerik olması durumunda hangi yöntemleri kullanacağımızı inceleyelim.

Kategorik veriler için kullanabileceğimiz sunum teknikleri:

- Tablo teknikleri
 - Frekans dağılımı
 - Göreceli frekans dağılımı
- Grafik teknikler
 - Pasta grafiği
 - Sütun grafiği

Kategorik Veriler İçin Grafik ve Tablo Teknikleri

- Kategorik veriler üzerinde yapabileceğimiz tek tablo gösterimi **frekansların** sayılmasıdır. Bu tür verileri kategoriler ve frekansların sunulduğu **Frekans Dağılımı** tabloları ile özetleyebiliriz.
- **Frekans dağılımları**; kategorik verilerin birbiri ile örtüşmeyecek gruplar içerisindeki sayısının gösterildiği tablolardır.
- **Göreceli Frekans dağılımı**; frekans tablolarında verilerin sayım değerlerinin yanında toplam veri içerisindeki yüzdelerinin sunulmasıdır.
- Göreceli frekans değerleri gruptaki sayımın toplam örnek sayısına bölünmesi ile elde edilir.

Örnek

- Bir öğrenci işleri bürosu son yıllarda mezun olan öğrenciler üzerinde bir araştırma yaparak, hangi tür işlere yerleştirildiklerini belirlemeye çalışmıştır.
- Bu verileri yeni dönemde hangi tür firmaların iş görüşmeleri için kampüs organizasyonlarına davet edileceğine karar verme için kullanacaktır.
- Çalışma alanları aşağıda sunulmuştur.
 1. Muhasebe
 2. Finans
 3. Yönetim
 4. Pazarlama ve Satış
 5. Diğer

Öğrencilerin yerleştiği iş alanları

1	1	2	4	1	4	2	4	5	2	5	4	1	1	4	2	3
4	5	1	4	1	3	2	4	3	1	2	5	4	2	3	3	2
5	4	1	4	1	4	5	5	1	4	2	4	2	2	5	2	5
1	5	3	4	1	4	1	2	1	3	4	2	4	5	5	1	2
2	1	4	3	3	1	4	1	1	1	1	2	4	1	4	3	2
2	4	1	1	2	4	4	4	5	4	5	1	1	3	2	1	3
3	1	5	3	1	3	2	1	1	1	5	3	2	3	4	2	5
1	3	1	1	1	4	2	4	4	2	1	4	4	5	5	2	1
4	4	2	5	3	2	4	1	1	4	3	2	4	2	3	1	1
1	2	1	1	4	1	4	3	4	4	2	3	1	4	5	3	3
1	4	1	2	4	1	4	5	2	2	2	5	4	4	4	1	4
4	1	4	4	1	2	4	2	2	3	2	1	4	4	3	4	1
3	4	5	3	3	1	5	1	4	2	2	1	5	5	4	1	1
1	4	3	2	2	1	1	4	2	3	1	3	3	2	2	3	
4	2	2	1	4	2	3	1	1	5	1	1	2	1	1	1	

Çözüm

- Çözüm aşamasında ilk yapılacak olan verilen değerlerin tek tek sayılarak her bir kategorideki **frekansların** hesaplanması olacaktır. Daha sonra bu frekans değerleri toplam örnek sayısına bölünerek **göreceli frekans** değerleri de hesaplanıp tabloya eklenmelidir.

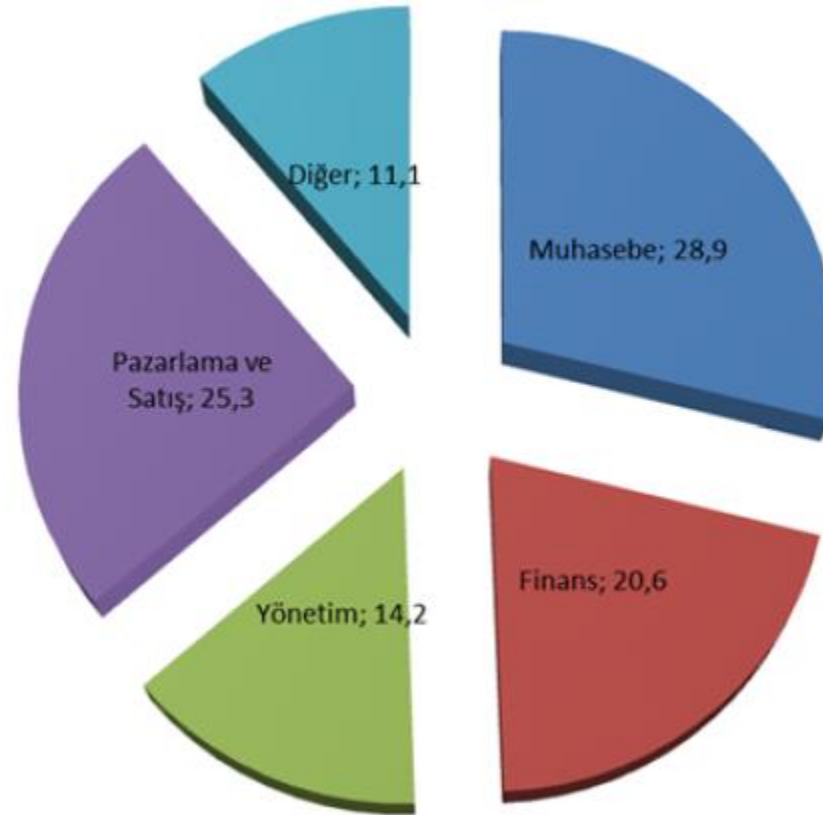
Çalışma Alanı	Frekans	Göreceli Frekans
Muhasebe	73	28,9
Finans	52	20,6
Yönetim	36	14,2
Pazarlama ve Satış	64	25,3
Diğer	28	11,1
Toplam	253	100

- Toplam satırındaki değerler kontrol amaçlıdır. Yani bu örnek için toplam örnek sayınız 253 olmalı ve göreceli frekans değerlerinin kümülatifi her zaman 100 olmalıdır. Aksi halde sayma işlemini tekrar kontrol etmelisiniz.

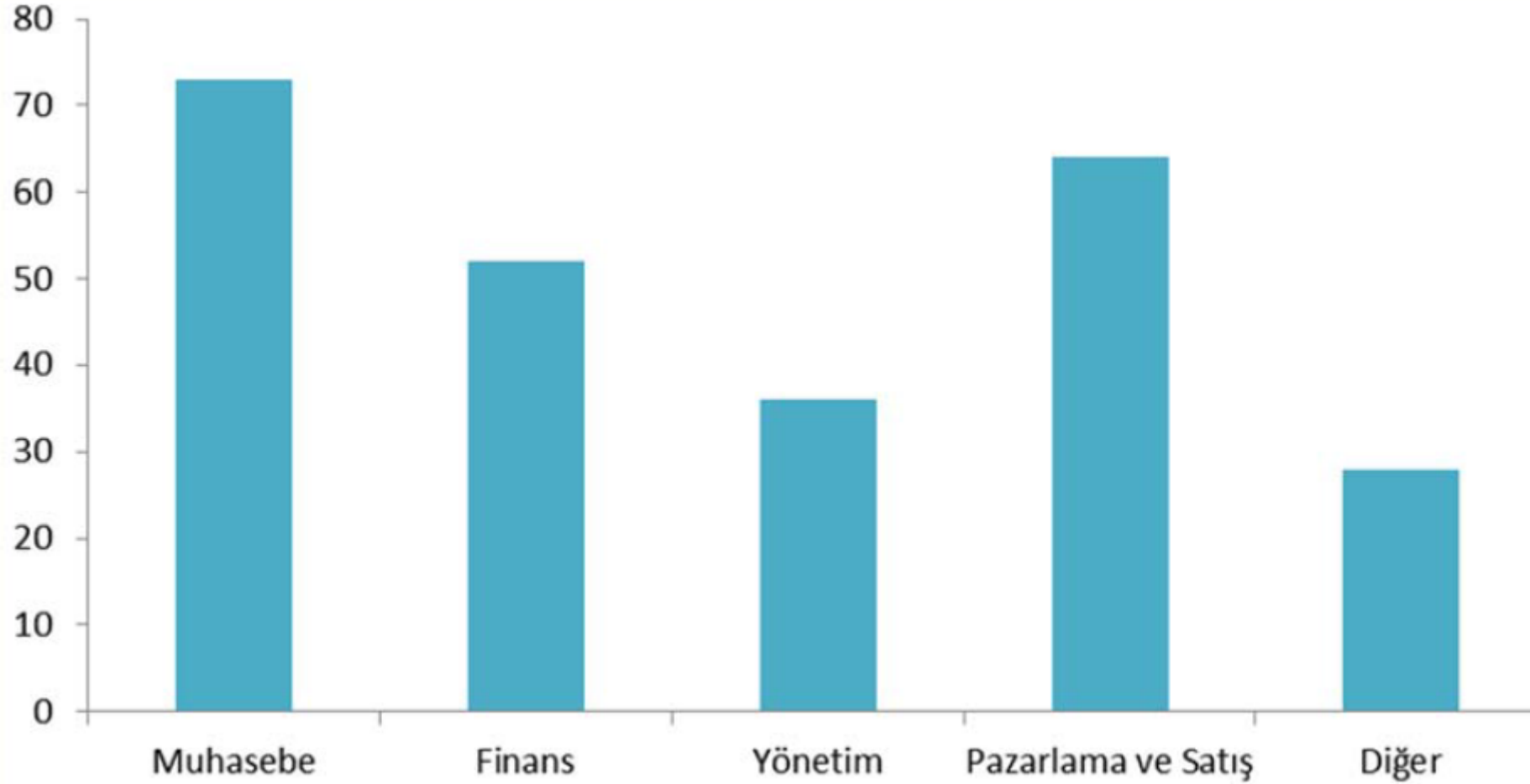
Sütun ve Pasta Grafikleri

- Sütun ve pasta grafikleri frekans değerleri üzerinden ve EXCEL yardımıyla çizilebilirler.
- Excelde Frekans değerleri seçilerek Ekle sekmesinde Grafikler kısmında istenilen grafik iki veya üç boyutlu olarak, istenilen renk düzeninde eklenebilir.

Önceki örnekteki frekans tabloları için pasta grafiği



Önceki örnekteki frekans tabloları için sütun grafiği



Nümerik veriler için kullanabileceğimiz sunum teknikleri:

- Tablo teknikleri
 - Frekans dağılımı
 - Göreceli frekans dağılımı
- Grafik teknikler
 - Histogram
 - Stemplot
 - Ogive

Nümerik Veriler İçin Grafik ve Tablo Teknikleri

Histogram

- Histogram nümerik verileri özetleyen güçlü bir görsel araç olmakla beraber, olasılıkları açıklayabildiği için de ayrıca dikkate değerdir.
- Ayrıca verinin dağılımı hakkında bilgiler sunar.
- Histogram çizmek için frekans dağılımı tablosu çizilmesi şarttır.
- Fakat grafik nümerik veriler ile çizildiğinden gruplar belli değildir. Veri adedine bağlı olarak bütün veri genelde 8-12 eşit sınıfa bölünür. Daha sonra frekans sayımında bu sınıflara düşen değerler belirlenir.
- Son olarak ise sınıf frekansları sütun grafiği ile gösterilir.
- Dikkat!!! Sınıf sayısını belirlemeden önce serinin en büyük ve en küçük değerleri belirlenmelidir.

Örnek

- Aşağıda veri seti verilmiş olan şehirlerarası arama tutarlarına ait histogram grafiğini çiziniz.
(Verilerin tamamı gösterilmemiştir)

42.19	39.21	75.71	8.37	...	114.67	15.30
38.45	48.54	88.62	7.18	...	27.57	75.49
29.23	93.31	99.50	11.07	...	64.78	68.69
89.35	104.88	85.00	1.47	...	45.81	35.00
...
74.01	93.57	23.31	9.01	...	3.03	41.38
56.01	0	11.05	84.77	...	9.16	45.77

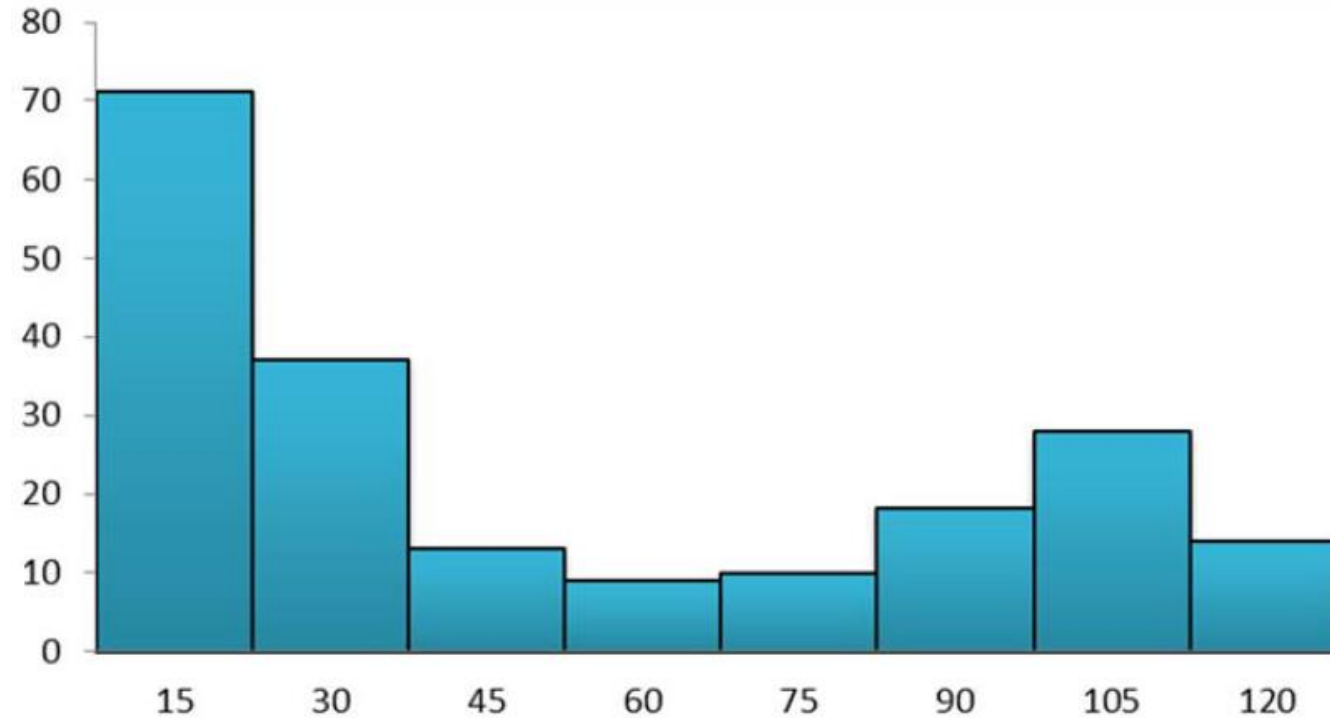
Çözüm

- Verideki en büyük değer 114,67 ve en küçük değer 0 olduğundan 0-120 arasında 15'er birimlik 8 adet sınıf belirlemek histogram çizmek için yeterli olacaktır. (10'arlık 12 sınıfa bölebileceğimize de dikkat ediniz. Bu karar tamamen sezgiseldir.)

Sınıflar	Frekanslar
0-15	71
16-30	37
31-45	13
46-60	9
61-75	10
76-90	18
91-105	28
106-120	14
Toplam	200

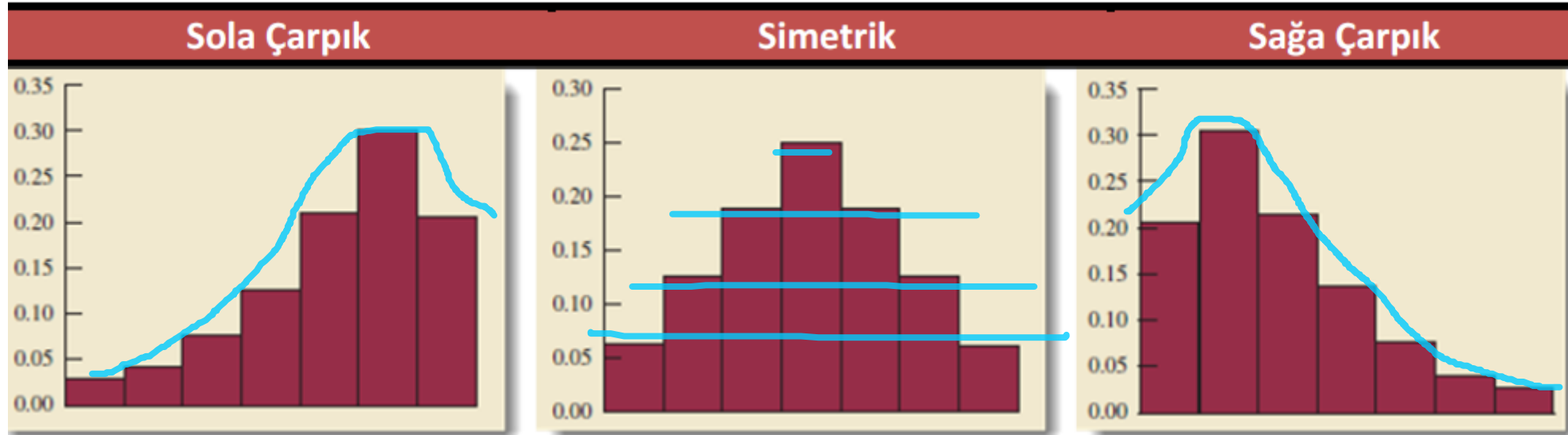
Frekans tablosu çizildikten sonra sütun grafiği hazırlanır.

- Hazırlanan grafik ile verilerin nerede en çok toplandığı açıkça görülebilmektedir. Histogramlar'da dağılımı net bir şekilde anlayabilmek için her bir sınıfın tepe noktalarını birleştiren bir çizgi çekmek de tercih edilmektedir.



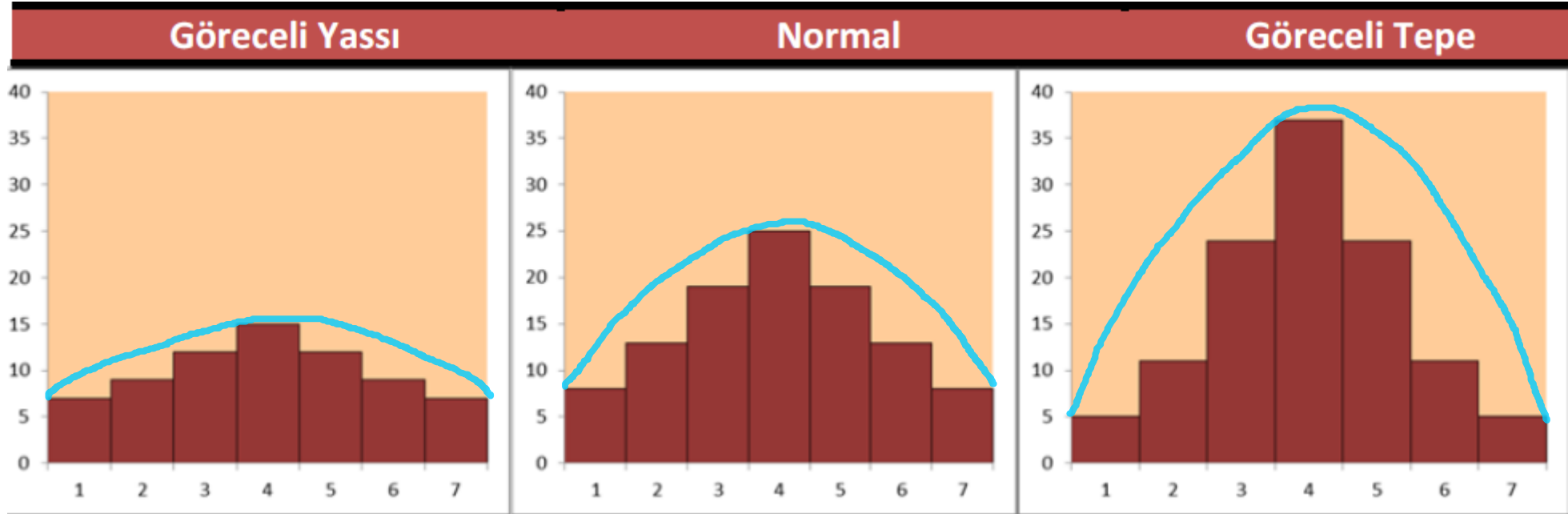
Çarpıklık

- Histogramlar her zaman simetrik değildirler. Verilerin görsel incelemesinde bazı durumlarda sağa veya sola doğru bir toplanma olabilir. İstatistik biliminde bu durum Çarpıklık olarak adlandırılır.



Basıklık

- Histogramlar ortalamaya bağılı olarak (orta sınıftaki veri sayısına) bağılı olarak dik ve yassı olabilmektedir. Bu duruma ise frekans dağılımlarının basıklığı adı verilir.



Stem-plot diyagramı

- Bir dağılım şekli üzerinde yorum yapmamızı sağlayacak histogram benzeri bir görsel gösterimdir.
- Nümerik değişkenlerle çizim yapılabilir.
- Literatürde Dal-Yaprak diyagramı olarak da geçmektedir.
- Stem-plot diyagramını çizmede ilk adım; hangi basamakların dal, hangi basamakların yaprak olarak ayrılacağıdır.
- İkinci aşamada ise belirlenen dal basamakları alt alta gelecek şekilde yaprak frekansları sağa doğru sıralanır. Burada önemli nokta frekansların sola hizalı olarak yerleştirilmesidir.
- Frekanslar sayı olarak ifade edilebileceği gibi işaretlerle (x, *. o) ifade edilebilir.

Örnek

- Bir grup çalışanın libre (lb) cinsinden ağırlıkları aşağıdaki tabloda sunulmuştur. Stemplot diyagramını çiziniz.

173	165	171	175	188
183	177	160	151	169
162	179	145	171	175
168	158	186	182	162
154	180	164	166	157

Çözüm

- Diyagramı çizmede ilk aşama dal ve yaprakların belirlenmesidir.
- Sorumuzda sayıların 140 ile 190 arasında değiştiği görülmektedir.
- Dal olarak ilk iki hane seçilirse, 14, 15, 16, 17, 18 gibi beş farklı grup oluşur.
- Yapraklar ise son haneler olur.
- 173 ve 183 değerleri için dal ve yaprak aşağıdaki gibidir.

Ağırlık	Dal	Yaprak
173	17	3
183	18	3

Çözüm

➤ İkinci aşamada her bir rakam teker teker diyagram içerisine frekansları sayılarak atanır.

DAL	YAPRAK							
14	5							
15	4	8	1	7				
16	2	8	5	0	4	6	9	2
17	3	7	9	1	5	1	5	
18	3	0	6	2	8			

Ogive

- Ogive Kümülatif (Toplamalı) Göreceli Frekansların grafiksel temsilidir. Bu görsel metot sayesinde verinin hangi değerinin hangi orana denk geldiği grafik olarak yorumlanabilir.
- Literatürde s-Eğrisi olarak da adlandırıldığı görülmektedir.
- Bu eğriyi çizmek için öncelikle göreceli frekanslar (frekansların yüzdesel değerleri diğer bir deyişle frekansların 0-1 arasına normalleştirilmiş halleri) hesaplanır.
- Daha sonra başka bir sütunda göreceli frekansların kümülatifleri bulunur.
- Kümülatif göreceli değerler ise iki boyutlu bir grafikte noktasal olarak yerleştirilip, bu noktalar birleştirilir.

Örnek

➤ Şehirlerarası telefon görüşmeleri ile ilgili soru için ogive çizelim.

Sınıflar	Frekanslar
0-15	71
16-30	37
31-45	13
46-60	9
61-75	10
76-90	18
91-105	28
106-120	14
Toplam	200

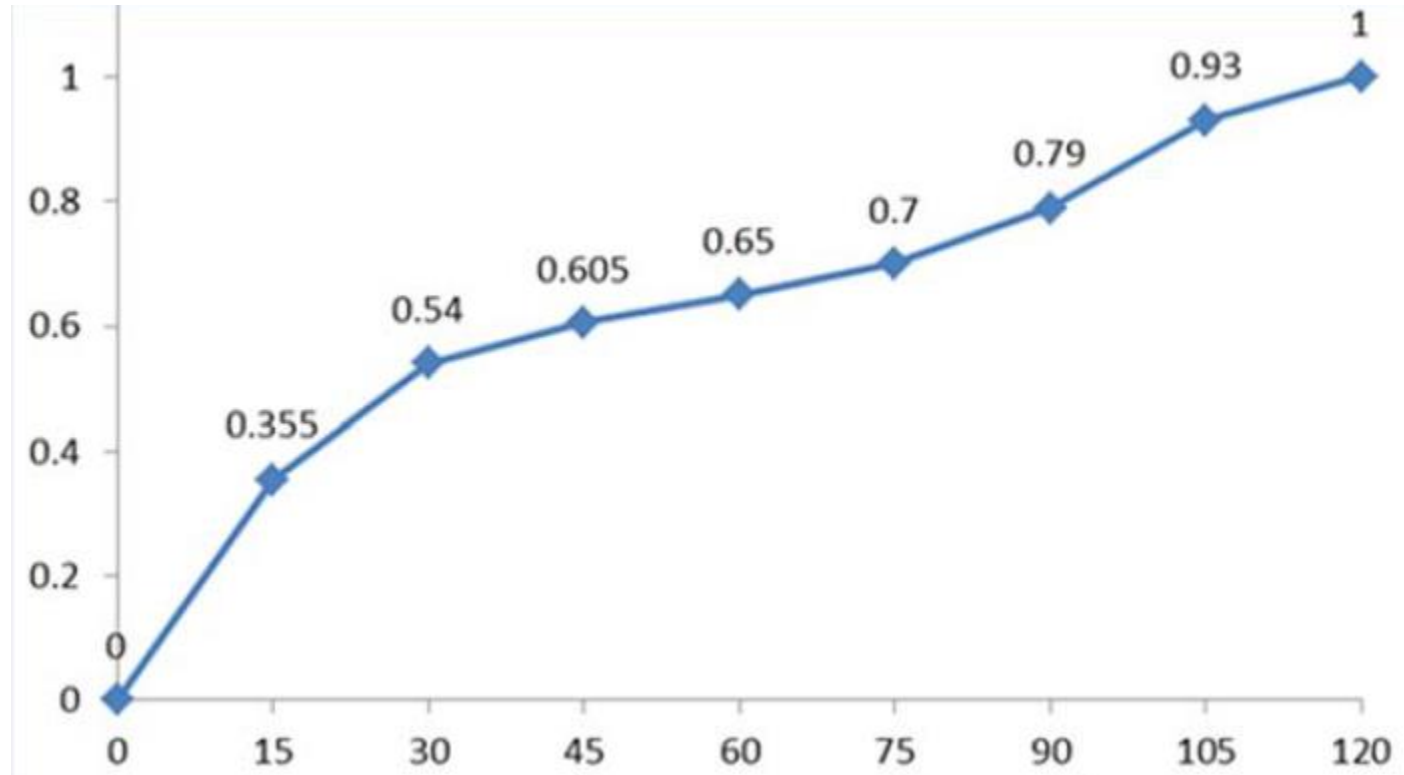
Çözüm

- Önceki soruda 8 sınıfa bölünüp frekanslar hesaplanmıştı. Şimdi aynı frekansları için göreceli frekansları ve Kümülatif frekansları hesaplayalım.

Limitler	Göreceli Frekanslar	Kümülatif Frekanslar
0 to 15	$71/200=0.355$	0.355
15 to 30	$37/200 = 0.185$	0.540
30 to 45	$13/200 = 0.065$	0.605
45 to 60	$9/200 = 0.045$	0.650
60 to 75	$10/200 = 0.050$	0.700
75 to 90	$18/200 = 0.090$	0.790
90 to 105	$28/200 = 0.140$	0.930
105 to 120	$14/200 = 0.070$	1.00
Total	200	

Çözüm

- Daha sonra kümülatif değerleri, x eksenini görüşme süreleri ve y eksenini yüzdeler olacak şekilde iki boyutlu grafiğe noktalar olarak yerleştirelim. Son olarak ise noktaları birleştirip ilgili s-Eğrisini çizelim.



Sayısal Teknikler

➤ Tek bir değişken için tanımlayıcı teknikler;

- Merkezi eğilim ölçütleri (ortalama, mod, medyan),
- Değişkenlik ölçütleri (Değişim aralığı, varyans ve standart sapma) ve
- Göreceli durum ölçütleri (yüzdebirlik ve çeyreklik) şeklindedir.

➤ İki veya daha fazla değişken durumlarında ise bu teknikler her bir değişken için ayrı ayrı hesaplanarak kullanılmakla beraber birden fazla değişkenin ilişkilerinin özetlenmesi Doğrusal İlişki Ölçütleri (Korelasyon, Kovaryans) ile gerçekleştirilebilir

Merkezi Eğilim Ölçütleri

Aritmetik Ortalama

- En önemli merkezi eğilim ölçütü Aritmetik ortalamadır. Basit şekli ile bütün gözlem değerlerinin toplam gözlem adedine bölünmesi ile hesaplanır.
- Bu değer bize veri setinin merkezi konumunu gösteren önemli bir istatistiktir. Aritmetik ortalama anakütle ve örneklem için aşağıdaki formüllerle hesaplanır:

$$\text{Ana Kütle Ortalaması: } \mu = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{Örneklem Ortalaması: } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Merkezi Eğilim Ölçütleri

Aritmetik Ortalama

- Aykırı (uç) değerlerden önemli ölçüde etkilenmektedir.
- Bu bağlamda küçük veri setlerinde ortalamanın çok uzağında bir veri, veri setine dâhil olduğunda ortalama birden bire yükselebilir. Bu yükselme verinin yorumlanmasında sıkıntılara yol açabilir.
- Örneğin bir milyoner gelir seviyesi ortalama olan bir mahalleye taşındığında hane halkı gelirlerinin ortalamasının bir anda çok fazla yükselmesi sonucunu doğurur ki, bu durumda hane halkları ile ilgili bir genelleme yapma imkânı da ortadan kalkmış olur.

Örnek

- Bir sınıftaki öğrencilerin ağırlıkları aşağıda verilmiştir. Aritmetik ortalamayı hesaplayınız.
- Eğer 140 kg ağırlığında yeni bir öğrenci sınıfa kayıt yaptırırsa, yeni ortalama ne olur?

89	77	90	101	66	66	76	59	64	75
88	65	75	72	70	64	66	68	82	80

Çözüm

- Birinci durumda ortalama aşağıdaki formülle hesaplanır:

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{89 + 77 + 90 + \dots + 80}{20} = 74,65$$

- 140 kg ağırlığında yeni bir öğrencinin sınıfa katılması durumunda ise yeni ortalama aşağıdaki gibidir:

$$= \frac{89 + 77 + 90 + \dots + 80 + \mathbf{140}}{\mathbf{21}} = \mathbf{77,77}$$

Çözüm

- Ortalama değerinin 3 birimden daha fazla arttığına dikkat ediniz.
- Ayrıca ilk durumda ortalama değeri verileri %50 - %50 şeklinde bölmektedir. (10 kişi 74,65 kg den daha hafif, 10 kişi ise daha ağırdır)
- İkinci durumda ise verilerin $\frac{2}{3}$ ü ortalamanın altında kalmaktadır. (21 kişinin 14 tanesi 77,77 kg den daha hafiftir)

Aritmetik Ortalama

- Aritmetik ortalama bir merkezi konum (central location) ölçütüdür. Fakat uç değerler olduğunda ve örnek sayısı az olduğunda bu özelliğini tam olarak yansıtmamaktadır.
- Her ne kadar uç değerler aritmetik ortalamanın tanımlayıcı özelliğini azaltsa da, veri seti büyüdükçe uç değerlerin etkisinin keskin düşüslere veya yükselmelere neden olmayacağı da aşikardır.
- **Dikkat!!!** Aritmetik ortalama aykırı değerlerden en çok etkilenen ölçüttür. Bu yüzden ortalama hesaplanmadan verinin büyüklüğünü inceleyip, aykırı değer olup olmadığı sorgulanmalı gerekirse aritmetik ortalama yerine diğer merkezi eğilim ölçütleri tercih edilmelidir.
- Her veri setinde sadece 1 adet aritmetik ortalama vardır.
- Sadece nicel veriler için aritmetik ortalama hesaplanabilir.

Frekans serisi için aritmetik ortalama

Örnek Bir işletmede aynı parçayı üreten işçilerin bu parçayı üretim sürelerinin dağılımı aşağıdaki gibi gözlenmiştir. Parça üretim süresinin aritmetik ortalamasını bulunuz.

Parça üretim süresi (dk) (X_i)	İşçi sayısı (f_i)	$f_i \cdot X_i$
12	2	24
13	5	65
14	10	140
15	7	105
16	4	64
Toplam	28	398

$$\bar{X} = \frac{\sum_{i=1}^s f_i X_i}{\sum_{i=1}^s f_i} = \frac{398}{28}$$
$$\bar{X} = 14,21 \text{ dk}.$$

Gruplanmış seriler için aritmetik ortalama

$$\bar{x} = \frac{\sum_{i=1}^k m_i f_i}{\sum_{i=1}^k f_i}$$

f : frekans

k : sınıf sayısı

$i = 1, 2, 3, \dots, k$

m : sınıf orta noktası

$$\sum_{i=1}^k f_i = n$$

- Sınıflanmış serilerde her bir sınıf içindeki değerlerin neler olduğu bilinmediğinden ve yalnızca her bir sınıfın frekans değerleri bilindiğinden dolayı sınıfı temsil etmek üzere sınıf orta noktaları hesaplamada kullanılır.

Gruplanmış seriler için aritmetik ortalama

Örnek Bir işyerinde yapılan telefon görüşmelerinin süresinin dağılımı için aşağıdaki gruplanmış seri verilmiştir. Buna göre görüşme süresinin aritmetik ortalamasını bulunuz.

Görüşme süresi	Görüşme sayısı (f_i)	m_i	$f_i m_i$
0 - 2	5	1	5
2 - 4	10	3	30
4 - 6	40	5	200
6 - 8	30	7	210
8 - 10	25	9	225
Toplam	110		670

$$\bar{X} = \frac{\sum_{i=1}^k f_i m_i}{\sum_{i=1}^k f_i} = \frac{670}{110}$$
$$\bar{X} = 6,09 \text{ dakika}$$

Ağırlıklı (tartılı) aritmetik ortalama

Örnek Aşağıda bir öğrencinin almış olduğu dersler, notları ve kredileri verilmiştir. Not ortalamasını tartılı aritmetik ortalama cinsinden hesaplayınız.

Dersler	Notlar (X_i)	Kredi (t_i)	$t_i X_i$
İstatistik	70	3	210
Matematik	60	4	240
İktisat	50	3	150
İşletme	80	2	160
Toplam	260	$\Sigma t_i = 12$	$\Sigma t_i X_i = 760$

$$\bar{X}_T = \frac{\sum t_i X_i}{\sum t_i} = \frac{760}{12}$$
$$\bar{X}_T = 63,33 \text{ puan}$$

Merkezi Eğilim Ölçütleri

Medyan (Ortanca)

- Önceden sıralanmış olan veri seti içerisindeki tam orta değere Medyan veya Ortanca denir.
- Gözlem sayısı tek olduğu durumlarda orta değer direkt alınır. (19 gözlem varsa 10. değer)
- Eğer gözlem sayısı çift ise ortadaki iki değer aritmetik ortalaması medyan olarak değerlendirilir. (20 gözlem varsa 10. ve 11. değerlerin aritmetik ortalaması)

Örnek

- Önceki örnekteki aşağıdaki veri setinde 20 ve 21 kişilik durumlar için medyan değerini bulunuz.

59	64	64	65	66	66	66	68	70	72	
75	75	76	77	80	82	88	89	90	101	140

Çözüm

- 20 veri olduğunda 10 ve 11. değerlerin aritmetik ortalamasının alındığı ve 21 veri olduğunda ise direkt olarak 11. değer medyan olarak seçildiğine dikkat ediniz.

$$\text{Medyan}_{20} = \frac{72 + 75}{2} = 73,5$$

$$\text{Medyan}_{21} = 75$$

- Ayrıca sonuçlar karşılaştırıldığında medyanın, aritmetik ortalamaya göre daha az yükseldiği görülebilir. (Medyan +1,5 – Ortalama +3,12)

Merkezi Eğilim Ölçütleri

Mod

- Bir veri seti içerisinde en yüksek frekansa sahip değere, bir başka deyişle veri seti içerisinde en çok rastlanan değere Mod adı verilir.
- Mod değeri örneklem sayısının az olmasından en çok etkilenen ölçüttür.
- Örnek sayısının az olduğu durumlarda mod değerinin güvenilirliği sorgulanmalıdır.
- Veri setinde birden fazla Mod değeri ile karşı karşıya kalınabilir.
- Yada hiç tekrar eden değer yoksa mod yoktur.
- Mod, nitel ve nicel veriler için kullanılabilir.

Örnek

➤ Önceki örnekteki aşağıdaki veri setinde 20 ve 21 kişilik durumlar için mod değerini bulunuz.

59	64	64	65	66	66	66	68	70	72	
75	75	76	77	80	82	88	89	90	101	140

Çözüm

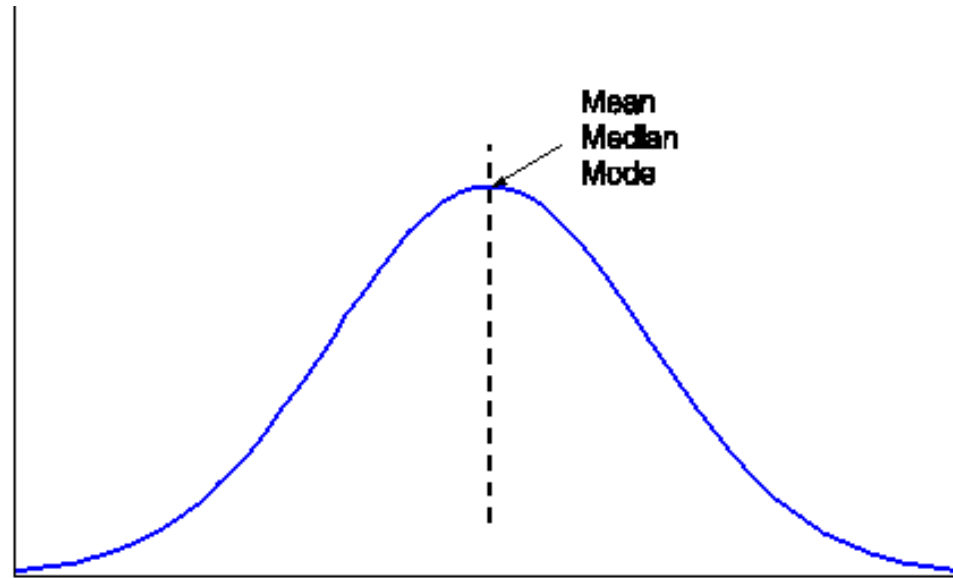
- 20 örnek için de 21 örnek için de mod 66 dır.

59	64	64	65	66	66	66	68	70	72	
75	75	76	77	80	82	88	89	90	101	140

- 21. örnek mod değerini değiştirmemiştir. Mod uç değerlerden en az etkilenen merkezi eğilim ölçütüdür.

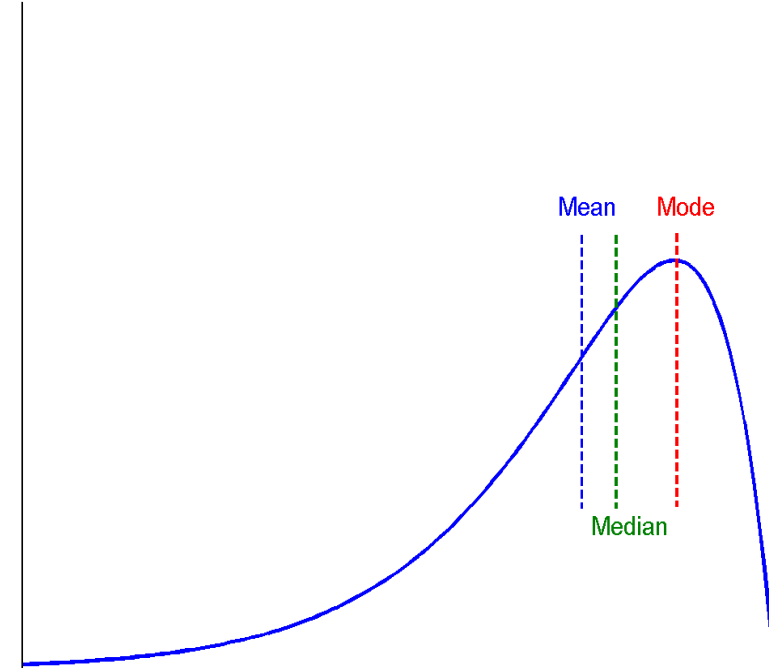
Ortalama, Medyan ve Mod ilişkisi

- Aritmetik ortalama, mod ve medyan arasındaki ilişki veri dağılımının çarpıklığı hakkında bilgi verir.
- **Simetrik dağılımda ortalama, mod, medyan değerleri birbirine eşittir.**



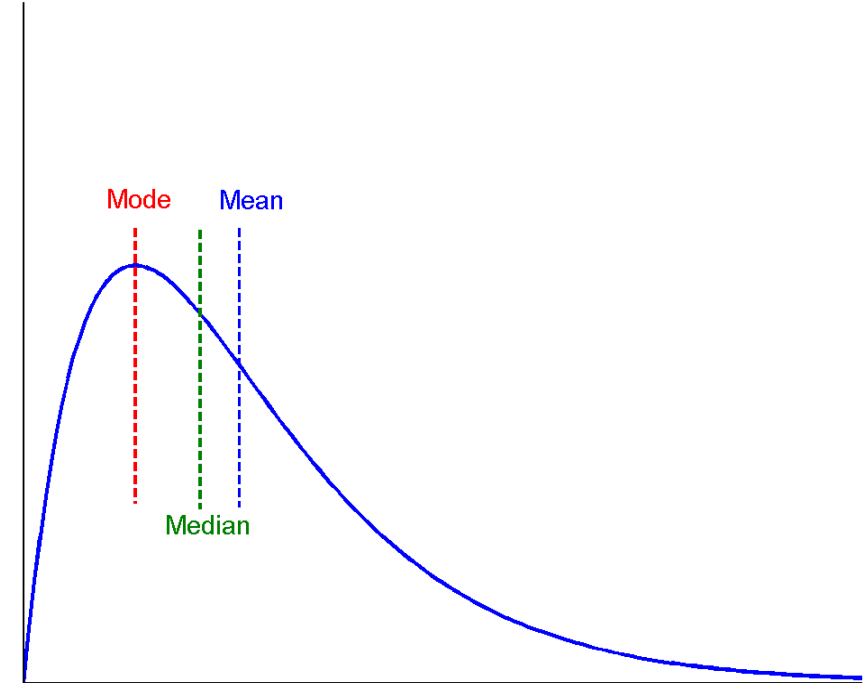
Ortalama, Medyan ve Mod ilişkisi

- Ortalama < medyan < mod ise **sola çarpık dağılım** gözlemlenir.
- Ortalama negatif tarafa doğru kayıyor. Puanların yarısından fazlası ortalamanın üzerinde kalır. Dağılım yüksek puanlarda yığılma gösterir.
- Grafik öğrenci başarısını gösteriyor ise, öğrencilerin çoğunun başarısı yüksektir, sınav kolaydır, öğrenme düzeyi yüksektir gibi yorumlar yapılabilir.



Ortalama, Medyan ve Mod ilişkisi

- Ortalama>medyan>mod ise **sağa çarpık dağılım** gözlemlenir.
- Ortalama pozitif tarafa doğru kayıyor. Puanların yarısından fazlası ortalamamanın altında kalır. Dağılım düşük puanlarda yığılma gösterir.
- Grafik öğrenci başarısını gösteriyor ise, öğrencilerin çoğunun başarısı düşüktür, sınav zordur, öğrenme düzeyi düşüktür gibi yorumlar yapılabilir.



Kaynaklar

- Bilişim Teknolojileri için İşletme İstatistiği ders notları, Dr. Öğr. Üyesi Halil İbrahim Cebeci
- İstatistik ders notları, Prof. Dr. Mehmet Aksaraylı